Stephen M. Robinson

Convexity in Finite-Dimensional Spaces

September 12, 2019

Springer

Preface

Sections with an asterisk before their titles contain material that is not required for the understanding of later sections, though there may be references to it and it is likely to be useful in providing perspective on the subject.

(Remainder of preface to be written)

Contents

Part I
Convex Sets

Chapter 1
Convex sets

This chapter introduces convex sets and explains some basic concepts, such as the convex hull. With only these basic concepts we will be able to prove two important structural theorems, due to Carathéodory and to Shapley and Folkman respectively. By adding some information about affine sets we can construct the *simplex,* a kind of elementary convex set that serves as a building block for more complicated convex sets. Then by combining properties of the simplex with some elementary properties of interiors we will define the *relative interior* of a convex set. In contrast to the situation with interiors, every nonempty convex set has a nonempty relative interior, and we will use that fact in nearly every topice area in the rest of the book.

To read this chapter you will need to know several properties of affine sets (translated linear spaces). These are often included in linear algebra courses, but all of the necessary material is in Section A.6 in Appendix A.

1.1 Convex sets

Almost everything we do in this book will involve sets having a property called *convexity.* These sets turn out to be considerably more interesting and useful than the simple affine sets appearing in Appendix A. This section contains basic definitions and results about convex sets that we use throughout the rest of the book.

1.1.1 Basic ideas

Definition 1.1.1. A subset C of $\mathbb{R}^n$ is *convex* if whenever c_1 and c_2 belong to C and $\lambda \in (0,1)$, the point $(1-\lambda)c_1 + \lambda c_2$ belongs to C. □

This says that C contains the *line segment* between each pair of its points, where we define the line segment $[c_1, c_2]$ between c_1 and c_2 to be the set $\{(1-\lambda)c_1 + \lambda c_2 \mid$

$0 \le \lambda \le 1\}$. We also use the notation $(c_1, c_2]$, $[c_1, c_2)$, and (c_1, c_2) for segments lacking one or both endpoints.

We call a collection $\lambda_1, \ldots, \lambda_k$ of nonnegative numbers that sum to 1 *convex coefficients*, and we say a linear combination $\sum_{i=1}^{I} \lambda_i x_i$ of points $x_i \in \mathbb{R}^n$ is a *convex combination* if the λ_i are convex coefficients.

Proposition 1.1.2. *A subset C of $\mathbb{R}^n$ is convex if and only if it contains every convex combination of finitely many of its points.*

The proof parallels that given for affine sets in Proposition A.6.2.

Proof. (if) This is immediate, since the hypothesis implies that C satisfies Definition 1.1.1.

(only if) Suppose that C is convex. If it is empty or a singleton, then the conclusion holds. Therefore suppose C contains at least two points. We know that it contains all convex combinations of two-point subsets, because it is convex. Therefore let $k > 2$ and suppose C contains all convex combinations of $k-1$ or fewer points. Let $x_1, \ldots, x_k$ be points of C and let $x := \sum_{i=1}^{k} \lambda_i x_i$ be a convex combination of these. As $k \neq 1$ not all of the λ_i can be 1, and by renumbering if necessary we can assume that $\lambda_k < 1$. Then

$$x = (1-\lambda_k)[\sum_{i=1}^{k-1} (1-\lambda_k)^{-1} \lambda_i x_i] + \lambda_k x_k.$$

The summation in the first term is a convex combination of $k-1$ points of C; hence by the induction hypothesis it is in C. The definition of convexity then tells us that $x \in C$. □

Any convex set has an affine hull, and we define the *dimension* $\dim C$ and the *parallel subspace* $\operatorname{par} C$ of C to be those of $\operatorname{aff} C$. By using the definition of convexity one can show that the closure of a convex set is convex.

The definition of convexity shows that the intersection of an arbitrary collection of convex sets is convex. If we specify a subset of $\mathbb{R}^n$ and take the collection to consist of all convex sets containing that subset, we obtain the *convex hull* of the subset.

Definition 1.1.3. S be a subset of $\mathbb{R}^n$. The *convex hull* of S, written $\operatorname{conv} S$, is the intersection of all convex subsets of $\mathbb{R}^n$ that contain S. □

As the intersection of convex sets is convex, the convex hull of a set S is itself a convex set. As it contains S, it is the smallest convex set containing S. It need not be closed, even if S is closed (Exercise 1.1.24), so the operators conv and cl need not commute.

An alternate way of obtaining $\operatorname{conv} S$ is to build it up from S using the construction of Proposition 1.1.2, as we now show.

Proposition 1.1.4. *If S is a subset of $\mathbb{R}^n$, then* $\operatorname{conv} S$ *is the set of all convex combinations of finitely many points of S.*

Proof. Write C for the set of all such convex combinations. If T is a convex set containing S, then a finite collection of points from S is a subset of T, and by Proposition 1.1.2 any convex combination of such a collection must belong to T; therefore $C \subset T$. As T was arbitrary, we have proved that $C \subset \operatorname{conv} S$. On the other hand, a convex combination of two points of C is, by the definition of C, also a convex combination of some finite subset of S, and therefore it lies in C. This means that C is convex, and it certainly contains S; therefore $C \supset \operatorname{conv} S$, so that $C = \operatorname{conv} S$. □

Using Proposition 1.1.4, one can show easily that the convex-hull operator commutes with an affine transformation: that is, if S is a subset of $\mathbb{R}^n$ and T is an affine map from $\mathbb{R}^n$ to $\mathbb{R}^m$, then $T(\operatorname{conv} S) = \operatorname{conv} T(S)$. The situation with inverse images is not quite so nice (Exercise 1.1.28).

We can use the associativity of Minkowski addition to show that it commutes with the convex hull operation.

Proposition 1.1.5. *If $S_1, \ldots, S_N$ is a finite collection of nonempty subsets of $\mathbb{R}^n$, then*

$$\operatorname{conv} \Sigma_{i=1}^N S_i = \Sigma_{i=1}^N \operatorname{conv} S_i. \tag{1.1}$$

Proof. It suffices to prove (1.1) for $N = 2$ because the associativity of Minkowski addition allows us to extend the result for $N = 2$ to any finite number of sets.

($\subset$). Let A and B be nonempty subsets of $\mathbb{R}^n$ and suppose $z \in \operatorname{conv}(A+B)$. We will show that $z \in \operatorname{conv} A + \operatorname{conv} B$. Proposition 1.1.4 says that z is a convex combination of a finite set of points of $A+B$. Each of these points is the sum of a point of A and a point of B. Therefore there are a positive integer I, points $a_i + b_i \in A+B$ for $i = 1, \ldots, N$, and convex coefficients $\gamma_1 \ldots, \gamma_I$, such that

$$z = \Sigma_{i=1}^I \gamma_i (a_i + b_i) = \Sigma_{i=1}^I \gamma_i a_i + \Sigma_{i=1}^I \gamma_i b_i \in \operatorname{conv} A + \operatorname{conv} B,$$

so that $\operatorname{conv}(A+B) \subset \operatorname{conv} A + \operatorname{conv} B$.

($\supset$). Now suppose that $y \in \operatorname{conv} A + \operatorname{conv} B$.There are then positive integers I and J, points $a_1, \ldots a_I$ in A and $b_1, \ldots b_J$ in B, and convex coefficients $\alpha_1, \ldots, \alpha_I$ and $\beta_1, \ldots, \beta_J$ such that

$$y = \Sigma_{i=1}^I \alpha_i a_i + \Sigma_{j=1}^J \beta_j b_j = \Sigma_{i=1}^I \Sigma_{j=1}^J \alpha_i \beta_j (a_i + b_j) \in \operatorname{conv}(A+B),$$

So $\operatorname{conv} A + \operatorname{conv} \subset \operatorname{conv}(A+B)$, and then (1.1) holds for $N = 2$ and therefore for any integer $N \geq 2$. □

If S is a subset of $\mathbb{R}^n$, then the *diameter* of S is $\operatorname{diam} S := \sup\{\|s - s'\| \mid s, s' \in S\}$; it may be $+\infty$, and it is $-\infty$ by convention if S is empty. Although taking the convex hull may enlarge a set, it does not increase its diameter.

Proposition 1.1.6. *If S is a subset of $\mathbb{R}^n$, then* $\operatorname{diam} \operatorname{conv} S = \operatorname{diam} S$.

Proof. If S is empty then so is $\operatorname{conv} S$ and there is nothing to prove. If S is nonempty, then as it is a subset of $\operatorname{conv} S$ we have $\operatorname{diam} S \leq \operatorname{diam} \operatorname{conv} S$. On the other hand, if x and x' are two points of $\operatorname{conv} S$ then by Proposition 1.1.4 we can write

$$x = \sum_{i=1}^{I} \lambda_i s_i, \qquad x' = \sum_{j=1}^{J} \mu_j s'_j,$$

where the λ_i and μ_j are convex coefficients and the points s_i and s'_j belong to S. Then the products $\mu_i \lambda_j$ are also convex coefficients, so

$$\|x - x'\| = \|\sum_{i=1}^{I} \lambda_i s_i - \sum_{j=1}^{J} \mu_j s'_j\| = \|\sum_{i=1}^{I} \sum_{j=1}^{J} \lambda_i \mu_j (s_i - s'_j)\| \leq \sum_{i=1}^{I} \sum_{j=1}^{J} \lambda_i \mu_j \|s_i - s'_j\|.$$

As $\|s_i - s'_j\| \leq \operatorname{diam} S$ for each i and j, we have $\operatorname{diam} \operatorname{conv} S \leq \operatorname{diam} S$. □

1.1.2 Dimensionality results

In the proof of Proposition 1.1.6 we used I points to form x and J points to form x', because at this point we do not know how many points of S we might need to represent a given point of $\operatorname{conv} S$. For example, if the set S consists of three points in $\mathbb{R}^2$ forming the vertices of a triangle, and if x belongs to the interior of that triangle, then no two points of S suffice to represent x: we need to use three.

In this example we had a set S whose affine hull had dimension $d = 2$, and we needed $1 + d = 3$ points to represent x. The following theorem will show that for any set $S \subset \mathbb{R}^n$ one never needs more than $1 + \dim \operatorname{aff} S$ points to represent an element of $\operatorname{conv} S$.

Theorem 1.1.7. *Let S be a subset of $\mathbb{R}^n$. If $x \in \operatorname{conv} S$ then there is a finite set $T \subset S$ having the following properties:*

- *The points of T are in general position, and T contains no more than $1 + \dim \operatorname{aff} S$ points;*
- *x is a convex combination, with strictly positive coefficients, of the points of T.*

The set T will depend on the chosen element x, and in general it will not be unique.

Proof. If S is empty there is nothing to prove. Therefore suppose $x \in \operatorname{conv} S$, and use Proposition 1.1.4 to find an integer $k \geq 0$, a set $s_0, \ldots, s_k$ of elements of S, and a set of convex coefficients $\lambda_0, \ldots, \lambda_k$ with $x = \sum_{i=0}^{k} \lambda_i s_i$. We may suppose that k is the smallest integer for which x can be expressed as such a convex combination, and therefore in particular that all of the λ_i are positive. Define $T := \{s_0, \ldots, s_k\}$.

First suppose that the s_i are not in general position. The λ_i satisfy the linear system

$$\sum_{i=0}^{k} \begin{pmatrix} s_i \\ 1 \end{pmatrix} \lambda_i = \begin{pmatrix} x \\ 1 \end{pmatrix}, \tag{1.2}$$

and by Theorem A.6.12 the vectors in parentheses on the left side of (1.2) are linearly dependent. Let $\mu_0, \ldots, \mu_k$ be numbers, not all zero, satisfying

$$\sum_{i=0}^{k} \begin{pmatrix} s_i \\ 1 \end{pmatrix} \mu_i = 0,$$

where the element on the right-hand side is the origin of $\mathbb{R}^{n+1}$. As the μ_i sum to zero but are not all zero, at least one is positive. Let ρ be the minimum of the ratios λ_i/μ_i for those i such that $\mu_i > 0$, and for each i let $\tau_i = \lambda_i - \rho\mu_i$. The τ_i are then convex coefficients, and $x = \sum_{i=0}^{k} \tau_i s_i$. But at least one of the τ_i is zero (choose any i for which the minimum in the definition of ρ was attained), and this contradicts our minimality assumption about k. Therefore in fact the $1+k$ points s_i of T are in general position, so $\dim \operatorname{aff} T = k$. But $T \subset S$, so $\operatorname{aff} T \subset \operatorname{aff} S$ and $k = \dim \operatorname{aff} T \leq \dim \operatorname{aff} S$. The number of points is then $1+k = 1+\dim \operatorname{aff} T \leq 1+\dim \operatorname{aff} S$. □

Theorem 1.1.7 includes the very well known *Carathéodory's theorem* [6].

Corollary 1.1.8 (Carathéodory's theorem, 1911). *If S is a subset of $\mathbb{R}^n$ then each point of* $\operatorname{conv} S$ *is a convex combination of no more than $n+1$ points of S.*

Proof. Let $x \in \operatorname{conv} S$. Theorem 1.1.7 says that x is a convex combination of a set of no more than $1+\dim \operatorname{aff} S$ points of S. As $S \subset \mathbb{R}^n$, $\dim \operatorname{aff} S \leq n$ so the number of points is not more than $n+1$.

We noted earlier that the convex hull of a closed set need not be closed. As an example of the application of Theorem 1.1.7, we show that if the set is taken to be compact instead of just closed, then the convex hull is also compact.

Proposition 1.1.9. *If S is a compact subset of $\mathbb{R}^n$, then* $\operatorname{conv} S$ *is compact.*

Proof. If S is empty there is nothing to prove, so assume $S \neq \emptyset$ and let

$$\Lambda = \{(\lambda_1, \ldots, \lambda_{n+1}) \mid \lambda_i \geq 0 \text{ for each } i, \sum_{i=1}^{n+1} \lambda_i = 1\}, \qquad K = \prod_{i=1}^{n+1} S.$$

Then Λ and K are compact subsets of $\mathbb{R}^{n+1}$ and $\mathbb{R}^{(n+1)n}$ respectively; the points of K are just the $(n+1)$-tuples $(s_1, \ldots, s_{n+1})$ with each $s_i \in S$. Let f be the function defined on $\Lambda \times K$ by

$$f(\lambda_1, \ldots, \lambda_{n+1}, s_1, \ldots, s_{n+1}) = \sum_{i=1}^{n+1} \lambda_i s_i.$$

Then any point of the set $f(\Lambda \times K)$ is a convex combination of elements of S, hence an element of $\operatorname{conv} S$. But Carathéodory's theorem tells us that in fact $f(\Lambda \times K)$ includes *every* point of $\operatorname{conv} S$. Therefore $\operatorname{conv} S$ is the image of the compact set $\Lambda \times K$ under the continuous function f, so by Proposition C.1.2 it is compact. □

For any subset S of $\mathbb{R}^n$ we have $\operatorname{conv} S \supset S$, so $\operatorname{cl}\operatorname{conv} S \supset \operatorname{cl} S$. We already noted that the closure of a convex set is convex, so $\operatorname{cl}\operatorname{conv} S$ is a convex set containing $\operatorname{cl} S$, and therefore it contains $\operatorname{conv}\operatorname{cl} S$. As we noted above, the two sets may not be

equal. However, suppose S is bounded. Then $\operatorname{cl} S$ is compact, so $\operatorname{conv}\operatorname{cl} S$ is compact by Proposition 1.1.9. As $\operatorname{cl} S \supset S$ we have $\operatorname{conv}\operatorname{cl} S \supset \operatorname{conv} S$. The set on the left is compact, hence *a fortiori* closed; by applying the closure operator we have $\operatorname{conv}\operatorname{cl} S \supset \operatorname{cl}\operatorname{conv} S$. Therefore in this case the operators conv and cl actually commute. The situation is much simpler for the convex-hull and affine-hull operators; see Exercise 1.1.26.

The following theorem, attributed to Lloyd Shapley and Jon Folkman, has a proof similar to that of Theorem 1.1.7.

Theorem 1.1.10 (Shapley-Folkman, 1966). *Let $\{S_\alpha \mid \alpha \in A\}$ be a finite family of subsets of $\mathbb{R}^n$. For each $x \in \operatorname{conv}\sum_{\alpha\in A} S_\alpha$ there is a subset B of A, containing no more than n elements, such that*

$$x \in \sum_{\alpha\in B} \operatorname{conv} S_\alpha + \sum_{\alpha\in A\setminus B} S_\alpha.$$

Proof. Fix $x \in \operatorname{conv}\sum_{\alpha\in A} S_\alpha$. Proposition 1.1.5 shows that $\operatorname{conv}\sum_{\alpha\in A} S_\alpha = \sum_{\alpha\in A}\operatorname{conv} S_\alpha$, so for each $\alpha \in A$ we can find a point $x_\alpha \in \operatorname{conv} S_\alpha$ and a positive integer N_α so that

$$x = \sum_{\alpha\in A} x_\alpha, \qquad x_\alpha = \sum_{i=1}^{N_\alpha} \lambda_i^\alpha x_i^\alpha,$$

with $\{\lambda_i^\alpha \mid i = 1,\dots,N_\alpha\}$ being convex coefficients for each α and with $x_i^\alpha \in S_\alpha$ for each α and i.

For each choice of the points x_α and of their representations as convex combinations of points x_i^α, the quantity $\sum_{\alpha\in A} N_\alpha$ is a positive integer. We select the points x_α and their representations so that the sum attains its minimum value over all such choices; in particular, this will force each coefficient λ_i^α to be positive.

Now let $B = \{\alpha \in A \mid N_\alpha > 1\}$. If $\alpha \in B$ then x_α does not equal any of the x_i^α, and if $\alpha \notin B$ then $x_\alpha \in S_\alpha$. Thus if B has no more than n elements we are finished.

Suppose then that B has more than n elements; we will produce a contradiction. In this case the set $\{x_\alpha - x_1^\alpha \mid \alpha \in B\}$ is linearly dependent in $\mathbb{R}^n$, so that there are scalars $\{\mu_\alpha \mid \alpha \in B\}$, not all zero, with $\sum_{\alpha\in B} \mu_\alpha(x_1^\alpha - x_\alpha) = 0$. We have

$$x - \sum_{\alpha\in A\setminus B} x_\alpha = \sum_{\alpha\in B} x_\alpha. \tag{1.3}$$

We will introduce a parameter $\theta \in \mathbb{R}^n$ by modifying the sum on the right in (1.3). For each $\alpha \in B$ and for $1 \le i \le N_\alpha$ let

$$\lambda_i^\alpha(\theta) := [1 - \theta\mu_\alpha]\lambda_i^\alpha + \delta(i)\theta\mu_\alpha,$$

where the parameter θ is a real number and $\delta(i)$ is 1 if $i = 1$ and zero otherwise. We have $\lambda_i^\alpha(0) = \lambda_i^\alpha$, and for fixed $\alpha \in B$ and any θ,

$$\sum_{i=1}^{N_\alpha} \lambda_i^\alpha(\theta) = \sum_{i=1}^{N_\alpha} \lambda_i^\alpha - \theta\mu_\alpha \sum_{i=1}^{N_\alpha} \lambda_i^\alpha + \theta\mu_\alpha = 1 - \theta\mu_\alpha + \theta\mu_\alpha = 1. \tag{1.4}$$

Also,

$$\sum_{\alpha\in B}\sum_{i=1}^{N_\alpha}\lambda_i^\alpha(\theta)x_i^\alpha=\sum_{\alpha\in B}\sum_{i=1}^{N_\alpha}\{[1-\theta\mu_\alpha]\lambda_i^\alpha+\delta(i)\theta\mu_\alpha\}x_i^\alpha. \tag{1.5}$$

The inner sum on the right in (1.5) is

$$\sum_{i=1}^{N_\alpha}\{[1-\theta\mu_\alpha]\lambda_i^\alpha x_i^\alpha+\delta(i)\theta\mu_\alpha x_i^\alpha=x_\alpha-\theta\mu_\alpha x_\alpha+\theta\mu_\alpha x_1^\alpha=x_\alpha-\theta\mu_\alpha[x_1^\alpha-x_\alpha],$$

so that

$$\sum_{\alpha\in B}\sum_{i=1}^{N_\alpha}\lambda_i^\alpha(\theta)x_i^\alpha=\sum_{\alpha\in B}\{x_\alpha-\theta\mu_\alpha[x_1^\alpha-x_\alpha]\}=\sum_{\alpha\in B}x_\alpha. \tag{1.6}$$

For each $\alpha\in B$ and for $i=1,\ldots,N_\alpha$ we have $\lambda_i^\alpha(0)=\lambda_i^\alpha>0$, and for each θ and each $\alpha\in B$, $\sum_{i=1}^{N_\alpha}\lambda_i^\alpha(\theta)=1$. In addition, for each $\alpha\in B$, $x_\alpha\neq x_1^\alpha$ so that no λ_1^α can be 1. Therefore, for those $\alpha\in B$ for which $\mu_\alpha\neq 0$, the quantities $\lambda_i^\alpha(\theta)$ will change as θ increases from 0. As their sum remains 1, some will increase and some will decrease. Let θ_* be the smallest value of θ for which one of the $\lambda_i^\alpha(\theta)$ takes the value zero. Then (1.3) still holds, but the total number of positive coefficients $\lambda_i^\alpha(\theta_*)$ has decreased by at least one, which is impossible because of our earlier choice to minimize the total number of positive coefficients. This contradiction shows that the number of elements in B does not exceed n. □

The Shapley-Folkman theorem has an interesting consequence, described in Example 1.1.12 below. To state the example we need the following definition.

Definition 1.1.11. Let P and Q be subsets of $\mathbb{R}^n$.

a. The *distance* from a point $x\in\mathbb{R}^n$ to Q is $d_Q(x)=\inf_{q\in Q}\|x-q\|$ ($+\infty$ by convention if Q is empty). We will sometimes write $d(x,Q)$ instead of $d_Q(x)$.
b. If Q is nonempty, the *excess* $e(P,Q)$ of P over Q is $\sup_{p\in P}d_Q(p)$ ($-\infty$ by convention if P is empty).
c. If P and Q are nonempty, the *Pompeiu-Hausdorff distance* $\rho(P,Q)$ between P and Q is $\max\{e(P,Q),e(Q,P)\}$.

□

In the literature, the Pompeiu-Hausdorff distance is often called the *Hausdorff metric*. An equivalent way of writing it is

$$\rho(P,Q)=\inf\{\mu\geq 0\mid P\subset Q+\mu B^n,\ Q\subset P+\mu B^n\},$$

where B^n is the Euclidean unit ball in $\mathbb{R}^n$: that is, the set of all $x\in\mathbb{R}^n$ with $\|x\|\leq 1$.

Example 1.1.12. Suppose that we have a sequence of nonempty subsets $\{S_i\mid i=1,2,\ldots\}$ of some compact convex set K in $\mathbb{R}^n$. For any positive integer k define

$$A_k=k^{-1}\sum_{i=1}^{k}S_i,\qquad C_k=\operatorname{conv}A_k.$$

Thus the A_k are averages of the sets $S_1,\dots,S_k$, and $A_k \subset C_k$. The quantity $e(C_k,A_k)$ is a measure of the nonconvexity of A_k.

Let k be large compared to n, and let x be any point of C_k, so that $kx \in \operatorname{conv}\sum_{i=1}^k S_i$. According to the Shapley-Folkman theorem, there are points $x_1 \in \operatorname{conv} S_1,\dots,x_k \in \operatorname{conv} S_k$, of which all but n can actually be chosen with $x_i \in S_i$, such that $s := \sum_{i=1}^k x_i = kx$. We now replace those points x_i in the sum s that are not in S_i by points $x_i' \in S_i$, and call the resulting sum s'. Then $s' \in kA_k$. Further, we have $s = s' + r$, where r is a sum of no more than n terms of the form $x_i - x_i'$. As both x_i and x_i' are in K, we have $\|r\| \le n \operatorname{diam} K$. Then

$$x = k^{-1}s = k^{-1}s' + k^{-1}r \in A_k + k^{-1}r.$$

As x was any point of C_k, we have

$$e(C_k,A_k) \le \|k^{-1}r\| \le (n/k)\operatorname{diam} K. \tag{1.7}$$

The right-hand side of (1.7) becomes as small as we please as k increases, so that the set averages A_k become almost convex as the number of sets being averaged increases. □

1.1.3 Simplices

This section introduces a class of convex sets, with very simple structure, that serve as building blocks for other convex sets. A set of this class is called a *simplex* (plural: *simplices*). Later in the section we define a class of *generalized simplices* useful for understanding the properties of unbounded convex sets.

Simplices

Definition 1.1.13. A *k-simplex* in $\mathbb{R}^n$ is the convex hull σ of a set $\{x_0,\dots,x_k\}$ of $k+1$ points of $\mathbb{R}^n$ that are in general position. The points x_i are the *vertices* of σ. The *barycenter* of σ is the point in σ whose barycentric coordinates with respect to the vertices are $(k+1)^{-1},\dots,(k+1)^{-1}$: that is, the average of its vertices. □

A point in the affine hull of a simplex actually belongs to the simplex if and only if its barycentric coordinates with respect to the vertices are convex coefficients. Thus another way to express the content of Theorem 1.1.7 is to say that $\operatorname{conv} S$ is a union of simplices whose vertices are points of S.

The simplex illustrates why the number of points required by Theorem 1.1.7 is the best possible if we specify only the dimensionality k of the set S appearing in the theorem. For any $k \ge 0$, take S to be the set of vertices of a k-simplex. As the barycentric coordinates are unique, no point whose barycentric coordinates are all positive can be written as a convex combination of k or fewer points of S.

Lemma 1.1.14. *Let A be a k-dimensional affine subset of $\mathbb{R}^n$, with $k \geq 0$; let c be a point of A and N be a neighborhood of c in A. Then N contains a k-simplex σ having barycenter c.*

Proof. As $c \in N$ there is some $\rho > 0$ such that $\{x \in A \mid \|x-c\| \leq \rho\} \subset N$. We will construct a simplex in $\mathrm{par}A$, then move it to A.

Define $y_0 = 0$ and choose k linearly independent vectors $y_1, \ldots, y_k$ in $\mathrm{par}A$, small enough so that $\mathrm{diam}\{y_0, \ldots, y_k\} \leq \rho$. The points $y_0, \ldots, y_k$ are in general position by Theorem A.6.12. The set $\tau := \mathrm{conv}\{y_0, \ldots, y_k\}$ is then a k-simplex in $\mathrm{par}A$, and by Proposition 1.1.6 $\mathrm{diam}\,\tau = \mathrm{diam}\{y_0, \ldots, y_k\} \leq \rho$. Write b for the barycenter of τ.

Now define $\sigma := \tau - b + c$. This is a k-simplex in A: the subtraction of b produces a k-simplex $\tau - b \subset \mathrm{par}A$ whose barycenter is 0, and because $c \in A$, addition of c moves that simplex to A and its barycenter to c.

If x is a point of σ, then

$$\|x-c\| \leq \mathrm{diam}\,\sigma = \mathrm{diam}\,\tau \leq \rho,$$

so that $x \in N$ and therefore $\sigma \subset N$. □

The next proposition tells us that any k-dimensional convex set has to contain a k-simplex. That fact will be important when we look at relative interiors of convex sets.

Proposition 1.1.15. *Let C be a convex subset of $\mathbb{R}^n$ having dimension $k \geq 0$, and let x_0 be a point of C. Then C contains a k-simplex, one vertex of which is x_0. Moreover, for any k-simplex σ that is contained in C,* $\mathrm{aff}\,\sigma = \mathrm{aff}\,C$.

Proof. As $x_0 \in C$, C contains a 0-simplex with vertex x_0. If $k = 0$ we are finished. If $k > 0$ then suppose that for some j with $0 \leq j < k$ we have found a j-simplex S_j with vertices $x_0, \ldots x_j$ belonging to C. We show that the construction can be extended from the dimension j to $j+1$ and thereby establish the result by induction.

The set C cannot be contained in the affine set $A_j := \mathrm{aff}\{x_0, \ldots, x_j\}$ because the dimension of A_j is j (the points $x_0, \ldots, x_j$ are in general position), whereas the dimension of C is $k > j$. Let x_{j+1} be a point of C not in A_j, and note that $x_1 - x_0, \ldots, x_{j+1} - x_0$ must be linearly independent because $x_{j+1} \notin A_j$; to see this assume dependence, use the fact that $x_1 - x_0, \ldots, x_j - x_0$ are linearly independent, and rearrange the terms to exhibit x_{j+1} as an affine combination of $x_0, \ldots, x_j$. The set $S_{j+1} = \mathrm{conv}\{x_0, \ldots, x_{j+1}\}$ is a $j+1$-simplex in C with vertices $x_0, \ldots, x_{j+1}$. We can now continue the induction to find a k-simplex $S_k \subset C$ as required.

For the statement about affine hulls, recall that $\mathrm{aff}\,C$ has dimension k, and as it contains C it certainly contains S_k. Then Corollary A.6.14 says that $\mathrm{aff}\,C = \mathrm{aff}\,S_k$.

□

Corollary 1.1.16. *Let C be a convex subset of $\mathbb{R}^n$ containing a point x_0. For each convex neighborhood Q of x_0,* $\dim(C \cap Q) = \dim C$.

Proof. Let $k = \dim C$ and suppose Q is a convex neighborhood of x_0. As $C \cap Q$ is a convex set contained in C, we have $\dim(C \cap Q) \leq \dim C$.

By Proposition 1.1.15, C contains a k-simplex σ whose vertices are $x_0, x_1, \ldots, x_k$. For $i = 0, \ldots, k$ and $\mu \in [0,1]$ let $v_i^\mu = x_0 + \mu(x_i - x_0)$ and define $\sigma_\mu = \operatorname{conv}\{v_0^\mu, \ldots, v_k^\mu\}$. If $\mu \in (0,1]$ then σ_μ is a k-simplex contained in C. If we take μ to be small enough then $\sigma_\mu \subset C \cap Q$. Then

$$\dim C = k = \dim \sigma_\mu \leq \dim(C \cap Q) \leq \dim C,$$

so C and $C \cap Q$ have the same dimension. □

Generalized simplices and convex hulls

It will be useful to extend the definition of convex hull to include what one might think of as points at infinity. For example, let x and v be elements of $\mathbb{R}^n$, with $v \neq 0$. Define the *halfline* from x in the direction of v to be the set $x + v\mathbb{R}_+$. We call v a *generator* of the halfline, and x its *endpoint.* Of course, any positive scalar multiple of v is also a generator.

This halfline is certainly a convex set, but it is not the convex hull of any finite set of points. However, even for very large μ the line segment $[x, x + \mu v]$ is the convex hull of x and $x + \mu v$. The trouble is that to get the entire halfline we have to let μ become infinitely large, and then the right-hand endpoint of the segment escapes to infinity.

To extend the definition of convex hull we need a new operation. For a finite set $T = \{t_1, \ldots, t_k\}$ we say that a point x is a *nonnegative linear combination* of points of T if there are nonnegative scalars $\mu_1, \ldots, \mu_k$ such that $x = \sum_{i=1}^k \mu_i t_i$.

Definition 1.1.17. Let S be a subset of $\mathbb{R}^n$. The *positive hull* of S is the set $\operatorname{pos} S$ consisting of the origin together with the set of all nonnegative linear combinations of finite subsets of S. □

There is an asymmetry in the treatment of the convex hull and the positive hull, as $\operatorname{conv}\emptyset = \emptyset$ but $\operatorname{pos}\emptyset$ is the origin. The purpose of including the origin is to ensure that $\operatorname{pos} S$ is never empty, even if S is.

Definition 1.1.18. Let S and T be subsets of $\mathbb{R}^n$, with S nonempty. The *generalized convex hull of the pair* (S,T) is the set

$$\operatorname{conv}(S,T) := \operatorname{conv} S + \operatorname{pos} T.$$

□

It is easy to show using this definition that $\operatorname{conv}(S,T)$ is convex.

By using this new construction we can extend the notion of general position and the applicability of Theorem 1.1.7.

Definition 1.1.19. Given a nonempty finite set $X = \{x_1, \ldots, x_r\}$ of points of $\mathbb{R}^n$ and a finite set $Y = \{y_1, \ldots, y_s\}$ of points of $\mathbb{R}^n$, we say that the pair (X,Y) is in *general position* if the vectors

$$\begin{pmatrix} x_1 \\ 1 \end{pmatrix}, \ldots, \begin{pmatrix} x_r \\ 1 \end{pmatrix}, \begin{pmatrix} y_1 \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} y_s \\ 0 \end{pmatrix}$$

are linearly independent. If (X, Y) is in general position then we call $\operatorname{conv}(X, Y)$ a *generalized simplex.* □

When the set Y is empty, this just reduces to the ordinary notions of general position and of simplex. If (X,Y) is in general position then the coefficients in the representation of a point of the generalized simplex $\operatorname{conv} X + \operatorname{pos} Y$ must be unique. The next theorem extends the representation of Theorem 1.1.7 to this more general situation.

Theorem 1.1.20. *Let X and Y be subsets of $\mathbb{R}^n$, with X nonempty. For each point z of $\operatorname{conv}(X,Y)$ there are finite subsets $\tilde{X}$ of X and $\tilde{Y}$ of Y such that $(\tilde{X}, \tilde{Y})$ is in general position and z belongs to the generalized simplex $\operatorname{conv}(\tilde{X}, \tilde{Y})$, with the coefficients in the representation of z being strictly positive.*

Proof. If Y is empty then Theorem 1.1.7 yields the required conclusion. Therefore assume that each of X and Y is nonempty.

By hypothesis $z \in \operatorname{conv} X + \operatorname{pos} Y$. Using Proposition 1.1.4 and the definition of $\operatorname{pos} Y$ we can find some finite subsets $X' = \{x_1 \ldots, x_r\}$ of X and $Y' = \{y_1, \ldots, y_s\}$ of Y such that with

$$u_i = \begin{pmatrix} x_i \\ 1 \end{pmatrix},\ i = 1, \ldots, r, \qquad v_j = \begin{pmatrix} y_j \\ 0 \end{pmatrix},\ j = 1, \ldots, s, \qquad w = \begin{pmatrix} z \\ 1 \end{pmatrix},$$

we can find nonnegative coefficients $\rho_1, \ldots, \rho_r$ and $\sigma_1, \ldots, \sigma_s$ such that

$$w = \sum_{i=1}^{r} \rho_i u_i + \sum_{j=1}^{s} \sigma_j v_j.$$

We can also suppose that we have chosen the coefficients so as to make the integer $r + s$ as small as possible, which in particular implies that the coefficients ρ_i and σ_j are all positive. If (X', Y') were not in general position then the vectors $u_1, \ldots, u_r, v_1, \ldots, v_s$ would be linearly dependent. We could then argue as we did in Theorem 1.1.7 to justify deleting one of the points. This would contradict the minimality of $r+s$; therefore (X', Y') must be in general position. □

Theorem 1.1.20 has an important consequence for sets that are generalized convex hulls of pairs (X,Y) in which both X and Y are finite. There is a special name for such sets.

Definition 1.1.21. A *finitely generated convex set* in $\mathbb{R}^n$ is the generalized convex hull $C = \operatorname{conv} X + \operatorname{pos} Y$ of finite subsets X and Y of $\mathbb{R}^n$. □

As $\operatorname{pos} Y$ is always nonempty, the sum C is empty if and only if X is empty.

Proposition 1.1.22. *Each finitely generated convex subset C of $\mathbb{R}^n$ is closed.*

Proof. If C is empty then it is certainly closed, so we may assume that X is nonempty. The first step is to show that C is closed in the special case in which it is a generalized simplex: that is, in which the pair (X,Y) is in general position. For this case suppose that $\{c_k\}$, $k = 1,2,\ldots$, is a sequence in C that converges to c_0. Let $X = \{x_1,\ldots,x_I\}$ and, if Y is nonempty, $Y = \{y_1,\ldots,y_J\}$. Then for each k there are convex coefficients ρ_i^k and nonnegative coefficients σ_j^k with

$$c_k = \sum_{i=1}^{I} \rho_i^k x_i + \sum_{j=1}^{J} \sigma_j^k y_j.$$

For the case $Y \neq \emptyset$ and for each k, define

$$A = \begin{bmatrix} x_1 & \ldots & x_I & y_1 & \ldots & y_J \\ 1 & \ldots & 1 & 0 & \ldots & 0 \end{bmatrix}, \qquad a_k = \begin{bmatrix} c_k \\ 1 \end{bmatrix}.$$

Here and in what follows, we proceed as if Y were nonempty; if it is empty then the procedure is the same except that the elements pertaining to Y do not appear. We next define

$$z_k = \begin{bmatrix} \rho_1^k & \ldots & \rho_I^k & \sigma_1^k & \ldots & \sigma_J^k \end{bmatrix}^*.$$

Then for each k we have $Az_k = a_k$ and $z_k \geq 0$, where we use the symbol $\geq$ to denote the partial order on $\mathbb{R}^{I+J}$ that makes $p \geq q$ if $p_r \geq q_r$ for $r = 1,\ldots,I+J$.

By assumption C is a generalized simplex, which means that the columns of A are linearly independent and therefore the kernel $\ker A$ is $\{0\}$. It follows that A is an invertible linear map of $\mathbb{R}^{I+J}$ onto its image $\mathrm{im}A \subset \mathbb{R}^{n+1}$. Write $\tau : \mathrm{im}A \to \mathbb{R}^{I+J}$ for the inverse linear map. The a_k belong to $\mathrm{im}A$, so that $z_k = \tau(a_k)$. As τ is linear and therefore continuous, and $\mathrm{im}A$ is a subspace and therefore closed, the fact that the a_k converge to the point $(c_0,1)^*$ implies that $(c_0,1)^*$ belongs to $\mathrm{im}A$ and that the z_k converge to the point

$$z_0 := \tau \begin{bmatrix} c_0 \\ 1 \end{bmatrix}.$$

As the limit of $\{z_k\}$, z_0 is nonnegative with $\sum_{i=1}^{I} (z_0)_i = 1$, and therefore $c_0 \in \mathrm{conv}(X,Y)$.

Now suppose that $C = \mathrm{conv}(X,Y)$ is finitely generated but not necessarily a generalized simplex. Each generalized simplex $S := \mathrm{conv}(\tilde{X},\tilde{Y})$ formed from subsets $\tilde{X} \subset X$ and $\tilde{Y} \subset Y$ is contained in C, so the union of all such S is a subset of C. But Theorem 1.1.20 says that each point of C is contained in such an S, so that in fact C is the union of all such generalized simplices. As C is finitely generated, there are only finitely many distinct generalized simplices in this union, so that C is a finite union of closed sets and is therefore closed. □

1.1.4 Exercises for Section 1.1

Exercise 1.1.23. Show that the inverse image of a convex set under an affine transformation is convex.

Exercise 1.1.24. Construct an example of a closed subset S of $\mathbb{R}^2$ whose convex hull is not closed. (You must prove that $\operatorname{conv} S$ is not closed.)

Exercise 1.1.25. Let C and D be convex sets in $\mathbb{R}^n$, and let μ and ν be real numbers. Show that $\mu C + \nu D$ is convex. Further, show that if μ and ν are nonnegative, then $\mu C + \nu C = (\mu + \nu)C$. Give an example to show that this equation is generally false for nonconvex sets.

Exercise 1.1.26. Show that for any subset S of $\mathbb{R}^n$, $\operatorname{aff} S = \operatorname{aff}\operatorname{conv} S$. (In particular, the affine-hull and convex-hull operators commute.)

Exercise 1.1.27. Show that if $E : \mathbb{R}^k \to \mathbb{R}^n$ is an affine transformation and U is a subset of $\mathbb{R}^k$, then $\operatorname{conv} E(U) = E(\operatorname{conv} U)$. (In particular, if U is convex then so is $E(U)$.) Give an example to show that $E(U)$ need not be closed even if U is closed and convex.

Exercise 1.1.28. Show that if $E : \mathbb{R}^k \to \mathbb{R}^n$ is an affine transformation and S is a subset of $\mathbb{R}^n$, then $\operatorname{conv} E^{-1}(S) \subset E^{-1}(\operatorname{conv} S)$. Show by example that the reverse inclusion need not hold in general, but prove that it holds if either S is convex or $S \subset \operatorname{im} E$.

Exercise 1.1.29. For $i = 1, \ldots, I$ let S_i be a subset of $\mathbb{R}^{n_i}$. Show that

$$\operatorname{conv} \prod_{i=1}^{I} S_i = \prod_{i=1}^{I} \operatorname{conv} S_i.$$

1.2 Relative interiors of convex sets

The definition of a convex set may seem almost like that of an affine set, and of course any affine set is convex. However, convex sets exhibit much more complex behavior than do affine sets. This section introduces some of that complexity.

For convex sets in $\mathbb{R}^n$ the affine hull is an extremely useful technical device. To see why, think about the usefulness of the interior of a set, convex or not. Obviously many convex sets will not have interiors (think of a line segment in $\mathbb{R}^2$), but we can define a substitute, called the *relative interior,* which is the interior of the set relative to its affine hull. This is the set's interior in the *relative topology* of the affine hull: that is, the topology induced on the affine hull by the topology of $\mathbb{R}^n$. The open sets of the relative topology are just the intersections of the open sets of $\mathbb{R}^n$ with the affine hull.

For example, the relative interior of a line segment between two distinct points of $\mathbb{R}^2$ consists of the line segment with its endpoints deleted. It's possible that every point in the set may be in the relative interior (for example if the set is affine, or in the case of the line segment if it contains neither of its endpoints), and in that case we call the set *relatively open.*

In the remainder of this section we first show that any nonempty convex subset of $\mathbb{R}^n$ has a nonempty relative interior. This fact highlights the usefulness of the relative interior, since a nonconvex set S in $\mathbb{R}^n$ certainly need not have an interior, either relative to $\mathbb{R}^n$ itself or relative to $\operatorname{aff} S$. Then we introduce two constructions using line segments that will help us to compute relative interiors of sets. Finally, we use those results to develop basic properties of the relative interior, including its usefulness in developing regularity conditions for set operations.

1.2.1 Nonemptiness of the relative interior

Remark A.6.19 showed that the barycentric coordinates are Lipschitz continuous. That fact has a useful consequence.

Lemma 1.2.1. *Let σ be a simplex of dimension $j \geq 0$ in $\mathbb{R}^n$ having vertices $\{v_0, \ldots, v_j\}$. The relative interior of σ is the set of points whose barycentric coordinates with respect to these vertices are all positive. Moreover,* $\operatorname{aff} \sigma = \operatorname{aff} \operatorname{ri} \sigma$.

Proof. If $j = 0$ then $\operatorname{ri} \sigma$, σ, and $\operatorname{aff} \sigma$ are all the same point, whose barycentric coordinate is 1, and the assertions hold. If $j \geq 1$ then σ has at least two vertices. Let P be the set of points of σ having all barycentric coordinates positive.

If $x \in \operatorname{ri} \sigma$ then $x \in \sigma$, so its (unique) barycentric coordinates are convex coefficients. If $x \notin P$ then at least one of these coefficients, say β_i, is zero. Let β_j be another coefficient. For $\varepsilon > 0$ the point $x(\varepsilon)$ whose barycentric coordinates $\beta(\varepsilon)$ are

$$\beta_k(\varepsilon) = \begin{cases} \beta_k - \varepsilon & \text{if } k = i, \\ \beta_k + \varepsilon & \text{if } k = j, \\ \beta_k & \text{otherwise,} \end{cases}$$

belongs to $\operatorname{aff} \sigma$ but does not belong to σ because $\beta(\varepsilon)$ is not a convex combination of the vertices ($\beta_i(\varepsilon) < 0$) and the barycentric coordinates are unique. But as $x(\varepsilon)$ is a continuous function of its barycentric coordinates, by choosing ε to be small we can bring $x(\varepsilon)$ as close to x as we wish. Therefore $x \notin \operatorname{ri} \sigma$, so $\operatorname{ri} \sigma \subset P$.

If $x \in P$ then, as the barycentric coordinates are continuous functions of x, there is a neighborhood N of x in $\operatorname{aff} \sigma$ such that whenever $x' \in N$ the barycentric coordinates of x' are all positive. Then $N \subset P \subset \sigma$, so that $x \in \operatorname{ri} \sigma$. Then $P \subset \operatorname{ri} \sigma$, so $P = \operatorname{ri} \sigma$.

As $\operatorname{ri} \sigma \subset \sigma$, we have $\operatorname{aff} \operatorname{ri} \sigma \subset \operatorname{aff} \sigma$. Thus, for the final claim it is enough to show the reverse inclusion, and we can do that by exhibiting each vertex of σ as an affine combination of points of P.

Construct the barycenter $b = \sum_{i=0}^{k}(k+1)^{-1}v_i$ of σ and, for $\varepsilon \in (0,1]$ and $i = 0,\ldots,j$, let w_i be the convex combination $(1-\varepsilon)v_i + \varepsilon b$. The barycentric coordinates of b are all positive, and those of the v_i are all nonnegative. Therefore each w_i belongs to P. Then $v_i = (1-\varepsilon)^{-1}(w_i - \varepsilon b)$ is the required affine combination. □

Proposition 1.2.2. *If C is a nonempty convex subset of $\mathbb{R}^n$, then* $\operatorname{aff}\operatorname{ri} C = \operatorname{aff} C$. *In particular,* $\operatorname{ri} C$ *is nonempty.*

Proof. Let the dimension of C be $k \geq 0$. Use Proposition 1.1.15 to find a k-simplex σ in C. That proposition also tells us that the affine hull of σ coincides with $\operatorname{aff} C$. Lemma 1.2.1 shows that $\operatorname{ri}\sigma$ is nonempty; as this is also the interior of σ relative to $\operatorname{aff} C$ and as $\sigma \subset C$, we have $\emptyset \neq \operatorname{ri}\sigma \subset \operatorname{ri} C$. Using Lemma 1.2.1 again, we have

$$\operatorname{aff}\sigma = \operatorname{aff}\operatorname{ri}\sigma \subset \operatorname{aff}\operatorname{ri} C \subset \operatorname{aff} C = \operatorname{aff}\sigma,$$

so that all of these sets are the same. □

One of our arguments in the above proof was that if D and E are subsets of $\mathbb{R}^n$ with $D \subset E$ and $\operatorname{aff} D = \operatorname{aff} E$, then $\operatorname{ri} D \subset \operatorname{ri} E$. This follows from elementary properties of interiors, because we are taking the interior relative to the common affine hull of the two sets. However, we can only make this claim if the sets have the same affine hull. To see this, let E be a closed line segment and D be one of its endpoints; then $D \subset E$, but the relative interiors of D and E are disjoint. This problem does not occur with interiors because the interior is taken with respect to an ambient space that contains both sets.

Any affine set is relatively open (it is simultaneously its own affine hull and relative interior), so Proposition 1.2.2 shows that for any convex set C we have $\operatorname{aff}\operatorname{ri} C = \operatorname{aff} C = \operatorname{ri}\operatorname{aff} C$. Therefore the affine-hull and relative-interior operators commute. If S is a nonconvex set then $\operatorname{ri} S$ may be empty even if S is not, and in that case $\operatorname{aff}\operatorname{ri} S$ obviously cannot equal $\operatorname{aff} S$. However, as $\operatorname{ri} S \subset S$ we always have $\operatorname{aff}\operatorname{ri} S \subset \operatorname{aff} S = \operatorname{ri}\operatorname{aff} S$.

Here is another consequence of nonemptiness of the relative interior.

Corollary 1.2.3. *Let C be a nonempty convex subset of $\mathbb{R}^n$. If H is any hyperplane containing C in one of its closed halfspaces, then H meets* $\operatorname{ri} C$ *if and only if H contains* $\operatorname{cl} C$.

Proof. (if) As C is nonempty, $\operatorname{ri} C$ is nonempty by Proposition 1.2.2. Therefore if H contains $\operatorname{cl} C$ it must meet $\operatorname{ri} C$.

(only if) For some $z^* \neq 0$ let $H = \{x \mid \langle z^*, x\rangle = \zeta\}$, and suppose that $c_0 \in H \cap \operatorname{ri} C$. With no loss of generality assume that C is in the lower closed halfspace of H, and let c be any point of C. For small positive μ, $c_0 + \mu(c_0 - c) \in C$ by the definition of relative interior. Then

$$\zeta \geq \langle z^*, (1+\mu)c_0 - \mu c\rangle = (1+\mu)\zeta - \mu\langle z^*, c\rangle,$$

where we used the fact that $c_0 \in H$. It follows that $\langle z^*, c\rangle \geq \zeta$. But $\langle z^*, c\rangle \leq \zeta$ because c is in the lower closed halfspace of H, so in fact $c \in H$ and therefore $C \subset H$. As H is closed, we also have $\mathrm{cl}\,C \subset H$. □

1.2.2 Line segments and the relative interior

Proofs in convexity often proceed by using simple geometric devices to develop information about a set or to manipulate it in some way. Among the most useful of these devices are some facts about the relationships between line segments and the relative interior. The following theorem and its corollaries develop these. It says that if x belongs to the closure of a convex set C and y belongs to the relative interior of C, then the entire line segment between x and y, except for the single point x, lies in $\mathrm{ri}\,C$.

In the following theorem, if $\rho \geq 0$ and if A is an affine set, we define $B_A(0,\rho)$ to be $(\mathrm{par}A) \cap B(0,\rho)$. This is a subset of $\mathrm{par}A$, and if $a \in A$ then $a + B_A(0,\rho) = A \cap B(a,\rho)$, a subset of A.

Theorem 1.2.4. *Let C be a convex subset of $\mathbb{R}^n$; let $r \in \mathrm{ri}\,C$ and $x \in \mathrm{cl}\,C$. Then $[r,x) \subset \mathrm{ri}\,C$.*

Proof. If $r = x$ then $[r,x) = \{(1-\lambda)r + \lambda r \mid \lambda \in [0,1)\} = \{r\} \subset \mathrm{ri}\,C$. Suppose then that $r \neq x$ and let $y \in [r,s)$. Let $A := \mathrm{aff}\,C$; we show that $y \in \mathrm{ri}\,C$ by showing that it has a neighborhood in A that is contained in C. If $y = r$ this is obvious, so let $\lambda \in (0,1)$ and $y = (1-\lambda)r + \lambda s$.

As $r \in \mathrm{ri}\,C$ there is a positive ρ such that

$$r + B_A(0,\rho) \subset C.$$

Let σ and ε be positive numbers with

$$(1-\lambda)\rho > \lambda\sigma + \varepsilon.$$

As $x \in \mathrm{cl}\,C$ there is some $c \in C$ with $\|s - c\| < \sigma$. Define $g := (1-\lambda)r + \lambda c$. Then

$$\|y - g\| = \lambda\|c - s\| < \lambda\sigma.$$

Using these relationships we conclude that

$$\begin{aligned} y + B_A(0,\varepsilon) &\subset [g + B_A(0,\lambda\sigma)] + B_A(0,\varepsilon) = g + B_A(0,\lambda\sigma + \varepsilon) \\ &\subset [(1-\lambda)r + \lambda c] + B_A(0,(1-\lambda)\rho) = (1-\lambda)[r + B_A(0,\rho)] + \lambda c \subset C, \end{aligned}$$

so that $y \in \mathrm{ri}\,C$. □

Theorem 1.2.4 shows, for example, that the relative interior of a convex set is convex (take $x \in \mathrm{ri}\,C$). Here is another useful technical result.

Corollary 1.2.5. *Let C be a nonempty convex subset of $\mathbb{R}^n$ and let c be a point of $\mathbb{R}^n$. The following are equivalent:*

a. $c \in \mathrm{ri}\, C$.
b. For each $y \in \mathrm{aff}\, C$ *there is a positive* ε *such that* $c + \varepsilon(c-y) \in C$.
c. For each $y \in C$ *there is a positive* ε *such that* $c + \varepsilon(c-y) \in C$.
d. There exist $r \in \mathrm{ri}\, C$ *and* $\varepsilon > 0$ *such that* $c + \varepsilon(c-r) \in \mathrm{cl}\, C$.

Proof. (a) implies (b). As $c \in \mathrm{ri}\, C$ there is a neighborhood N of c in $\mathrm{aff}\, C$ such that $N \subset C$. Let $y \in \mathrm{aff}\, C$ and choose $\varepsilon > 0$ small enough so that $c + \varepsilon(c-y) \in N \subset C$; then (b) holds.

(b) implies (c). $\mathrm{aff}\, C \supset C$.

(c) implies (d). As C is nonempty, so is $\mathrm{ri}\, C$ by Proposition 1.2.2. Take y to be $r \in \mathrm{ri}\, C$ and apply (c) to get (d).

(d) implies (a). If r and $\varepsilon > 0$ exist so that $c + \varepsilon(c-r) =: x \in \mathrm{cl}\, C$, then $c = (1+\varepsilon)^{-1}\varepsilon r + (1+\varepsilon)^{-1}x$ with $r \in \mathrm{ri}\, C$, $x \in \mathrm{cl}\, C$, and $\varepsilon > 0$. Theorem 1.2.4 says that then $c \in \mathrm{ri}\, C$. □

We define the *relative boundary* of a convex set C to consist of those points of $\mathrm{cl}\, C$ that do not belong to $\mathrm{ri}\, C$. Returning again to our line segment in $\mathbb{R}^2$, we see that the relative boundary consists of the two endpoints of the segment, whereas the boundary would be the entire segment. This relative boundary relates to the affine hull and the relative interior as the boundary does to the ambient space and the interior.

We commented earlier that Theorem 1.2.4 is a very useful device in proofs. Here is a proposition showing the particularly nice behavior of the relative interior and closure operators when applied to convex sets. Although one could not expect them to commute, they do almost the next best thing.

Proposition 1.2.6. *If C is a convex subset of $\mathbb{R}^n$, then*

$$\mathrm{cl}\,\mathrm{ri}\, C = \mathrm{cl}\, C, \qquad \mathrm{ri}\,\mathrm{cl}\, C = \mathrm{ri}\, C.$$

Proof. If C is empty there is nothing to prove, so assume C is nonempty. For the first formula, the inclusion $\subset$ follows from elementary properties of the closure operation and the fact that $\mathrm{ri}\, C \subset C$. For the other inclusion, let x be any point of $\mathrm{cl}\, C$ and use Proposition 1.2.2 to choose $c \in \mathrm{ri}\, C$. According to Theorem 1.2.4, the line segment $(x, c]$ lies in $\mathrm{ri}\, C$. As x belongs to the closure of this segment it belongs to $\mathrm{cl}\,\mathrm{ri}\, C$.

For the second formula, recall that any set and its closure have the same affine hull; therefore as $C \subset \mathrm{cl}\, C$ we have $\mathrm{ri}\, C \subset \mathrm{ri}\,\mathrm{cl}\, C$ by elementary properties of the interior operation. For the opposite inclusion let $z \in \mathrm{ri}\,\mathrm{cl}\, C$ and choose $x \in \mathrm{ri}\, C$. This x also belongs to the convex set $\mathrm{cl}\, C$, so by Corollary 1.2.5 we can prolong the line segment from x to z slightly beyond z to obtain $\mu > 0$ and a point $w := z + \mu(z-x)$ that still lies in $\mathrm{cl}\, C$. However, z lies in the relatively open line segment between $x \in \mathrm{ri}\, C$ and $w \in \mathrm{cl}\, C$, so by Theorem 1.2.4 z belongs to $\mathrm{ri}\, C$. □

Note that in the last part of the above proof we used Theorem 1.2.4 twice in succession, applied to two different sets.

Here is a result connecting forward affine images and relative interiors. Such a result certainly does not hold for interiors; what makes it work for convex sets is the fact that affine transformations commute with the affine-hull operator (Exercise A.6.32).

Proposition 1.2.7. *Let C be a convex subset of $\mathbb{R}^n$ and T an affine transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then $T(\operatorname{ri} C) = \operatorname{ri} T(C)$.*

Proof. If C is empty there is nothing to prove, so we may assume C is nonempty.

($\subset$). Let a_0 be a point of $T(\operatorname{ri} C)$; then $a_0 = T(c_0)$, where $c_0 \in \operatorname{ri} C$. The set $T(C)$ is nonempty and convex, so let $a_1 \in \operatorname{ri} T(C)$; then $a_1 = T(c_1)$ for some $c_1 \in C$. By part (c) of Corollary 1.2.5 there is some positive μ such that $c_2 := c_0 + \mu(c_0 - c_1)$ belongs to C. Then $T(c_2) = a_0 + \mu(a_0 - a_1) =: a_2$ belongs to $T(C)$, so that $a_0 \in (a_1, a_2)$. Use Theorem 1.2.4 to conclude that a_0 lies in $\operatorname{ri} T(C)$, so that $T(\operatorname{ri} C) \subset \operatorname{ri} T(C)$.

($\supset$). For a set S and a continuous function f whose domain includes $\operatorname{cl} S$ we always have $f(\operatorname{cl} S) \subset \operatorname{cl} f(S)$ (Exercise A.6.30). As $\operatorname{cl} C = \operatorname{cl}\operatorname{ri} C$ by Proposition 1.2.6, we find that

$$T(C) \subset T(\operatorname{cl} C) = T(\operatorname{cl}\operatorname{ri} C) \subset \operatorname{cl} T(\operatorname{ri} C). \tag{1.8}$$

As $\operatorname{ri} C$ and C have the same affine hull, $T(C)$ and $T(\operatorname{ri} C)$ have the same affine hull; this is also the affine hull of $\operatorname{cl} T(\operatorname{ri} C)$. Then we can take relative interiors in (1.8) to obtain

$$\operatorname{ri} T(C) \subset \operatorname{ri}\operatorname{cl} T(\operatorname{ri} C) = \operatorname{ri} T(\operatorname{ri} C) \subset T(\operatorname{ri} C),$$

so that $\operatorname{ri} T(C) \subset T(\operatorname{ri} C)$. □

We can now use the same device to deal with sums of relative interiors that we used previously with affine and convex hulls.

Corollary 1.2.8. *If C and D are two convex subsets of $\mathbb{R}^n$, then* $\operatorname{ri}(C+D) = \operatorname{ri} C + \operatorname{ri} D$.

Proof. We have $\operatorname{ri}(C \times D) = \operatorname{ri} C \times \operatorname{ri} D$ (Exercise 1.2.15). Now apply Proposition 1.2.7, using the linear operator L from $\mathbb{R}^{2n}$ to $\mathbb{R}^n$ given by $L(x,y) = x + y$, to obtain

$$\operatorname{ri}(C+D) = \operatorname{ri} L(C \times D) = L[\operatorname{ri}(C \times D)] = L(\operatorname{ri} C \times \operatorname{ri} D) = \operatorname{ri} C + \operatorname{ri} D.$$

□

1.2.3 Regularity conditions for set operations

As an illustration of the usefulness of the relative interior, we use it to augment the formula of Exercise A.6.32. That formula related forward affine images to affine hulls; this one deals with inverse images. It is not hard to show that the inverse image of an affine set or a convex set under an affine transformation is affine or convex, respectively; see Exercises A.6.27 and 1.1.23.

Proposition 1.2.9. *Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be an affine transformation and let S be a nonempty subset of $\mathbb{R}^m$. Then* $\operatorname{aff} T^{-1}(S) \subset T^{-1}(\operatorname{aff} S)$*, and if* $\operatorname{im} T$ *meets* $\operatorname{ri} S$ *then equality holds.*

Proof. $T^{-1}(\operatorname{aff} S)$ is affine and it contains $T^{-1}(S)$, so it also contains $\operatorname{aff} T^{-1}(S)$ by definition of the affine hull. To prove equality, assume that $\operatorname{im} T$ meets $\operatorname{ri} S$, so that there is some $x_0 \in \mathbb{R}^n$ with $T(x_0) =: s_0 \in \operatorname{ri} S$. Suppose that $x \in T^{-1}(\operatorname{aff} S)$, and let $T(x) =: f \in \operatorname{aff} S$. Then $f - s_0 \in \operatorname{par} S$, so for sufficiently small positive ε we have $s_\varepsilon := s_0 + \varepsilon(f - s_0) \in S$ by the definition of relative interior. However, if we define $x_\varepsilon := x_0 + \varepsilon(x - x_0)$ then $s_\varepsilon = T(x_\varepsilon)$ and therefore $x_\varepsilon \in T^{-1}(S)$. Then we have

$$x = (1 - \varepsilon^{-1})x_0 + \varepsilon^{-1}x_\varepsilon \in \operatorname{aff} T^{-1}(S),$$

which shows that in this case $\operatorname{aff} T^{-1}(S) = T^{-1}(\operatorname{aff} S)$. □

The condition that the image of T should meet the relative interior of S is a fundamental regularity condition that one sees throughout the study of convexity. In this case it's really needed, as equality may not hold without it (Exercise 1.2.14).

Proposition 1.2.9 has a useful corollary dealing with intersections of sets.

Corollary 1.2.10. *Let C_1 and C_2 be convex subsets of $\mathbb{R}^n$. Then* $\operatorname{aff}(C_1 \cap C_2) \subset (\operatorname{aff} C_1) \cap (\operatorname{aff} C_2)$*, and if* $(\operatorname{ri} C_1) \cap (\operatorname{ri} C_2) \neq \emptyset$ *then equality holds.*

Proof. Apply Proposition 1.2.9 with $C = C_1 \times C_2$ taking T to be the linear transformation $L : \mathbb{R}^n \to \mathbb{R}^{2n}$ given by $L(x) = (x, x)$ and using the result of Exercise A.6.26. □

Although Proposition 1.2.7 showed that forward images behave very well under the relative interior operation, with inverse images one has to be more careful. The next proposition shows that the same regularity condition helps to deal with this case.

Proposition 1.2.11. *Suppose D is a convex set in $\mathbb{R}^m$, and let T be an affine transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then*

$$T^{-1}(\operatorname{ri} D) \subset \operatorname{ri} T^{-1}(D), \qquad T^{-1}(\operatorname{cl} D) \supset \operatorname{cl} T^{-1}(D),$$

with equality in each case if $\operatorname{im} T$ *meets* $\operatorname{ri} D$*.*

Note that the inclusions go in opposite directions.

Proof. If D is empty there is nothing to prove, so we may assume D to be nonempty. We first show that $T^{-1}(\operatorname{ri} D) \subset \operatorname{ri} T^{-1}(D)$. If x belongs to $T^{-1}(\operatorname{ri} D)$ and y is any point of $\operatorname{ri} T^{-1}(D)$ then $T(x) \in \operatorname{ri} D$ and $T(y) \in D$. As $T(x) - T(y) \in \operatorname{par} D$, by the definition of relative interior there is some positive ε with $t := T(x) + \varepsilon(T(x) - T(y)) \in D$. Writing $z = x + \varepsilon(x - y)$, we have $T(z) = t$ because T is affine; as $t \in D$ we have $z \in T^{-1}(D)$. Applying Corollary 1.2.5, we find that $x \in \operatorname{ri} T^{-1}(D)$ as required.

To prove the inclusion involving closures, note that since T is continuous, $T^{-1}(\operatorname{cl} D)$ must be a closed set. This set contains $T^{-1}(D)$, so it must also contain $\operatorname{cl} T^{-1}(D)$ by elementary properties of closure.

Now assume that $\operatorname{im} T$ meets $\operatorname{ri} D$. We show that in this case the inclusions become equalities.

Let $w \in \operatorname{ri} T^{-1}(D)$; we prove $w \in T^{-1}(\operatorname{ri} D)$. Suppose that $v \in T^{-1}(\operatorname{ri} D)$, and use the definition of relative interior to find $\varepsilon > 0$ such that $u := w + \varepsilon(w - v) \in T^{-1}(D)$. Then $T(u) \in D$, and

$$T(w) = (1+\varepsilon)^{-1} T(u) + \varepsilon(1+\varepsilon)^{-1} T(v) \in \operatorname{ri} D,$$

where we used Theorem 1.2.4. Therefore $w \in T^{-1}(\operatorname{ri} D)$, as claimed.

Finally, let $x_0 \in T^{-1}(\operatorname{cl} D)$ and $x_1 \in T^{-1}(\operatorname{ri} D)$, and write $d_i = T(x_i)$ for $i = 0, 1$. Theorem 1.2.4 says that $(d_0, d_1] \subset \operatorname{ri} D$, so $(x_0, x_1] \subset T^{-1}(\operatorname{ri} D) \subset T^{-1}(D)$. Then $x_0 \in \operatorname{cl}(x_0, x_1] \subset \operatorname{cl} T^{-1}(D)$. □

There are cases in which this regularity condition doesn't suffice. One of these is Exercise 1.1.28, in which the condition $V \subset \operatorname{im} E$ cannot be replaced by, for example, the requirement that $\operatorname{im} E$ meet $\operatorname{ri} \operatorname{conv} V$ (Exercise 1.2.17).

We can use Proposition 1.2.11 together with some calculation to prove very useful formulas about commutativity of closure with intersection.

Proposition 1.2.12. *Let $\{C_\alpha \mid \alpha \in A\}$ be a collection of convex subsets of $\mathbb{R}^n$. Assume that $\bigcap_{\alpha \in A} \operatorname{ri} C_\alpha \neq \emptyset$. Then*

$$\operatorname{cl} \bigcap_{\alpha \in A} C_\alpha = \bigcap_{\alpha \in A} \operatorname{cl} C_\alpha,$$

and

$$\operatorname{ri} \bigcap_{\alpha \in A} C_\alpha \subset \bigcap_{\alpha \in A} \operatorname{ri} C_\alpha, \tag{1.9}$$

with equality when A is a finite set. In particular, the intersection of any finite collection of relatively open convex sets is relatively open.

Proof. Let $y \in \bigcap_{\alpha \in A} \operatorname{ri} C_\alpha$ and $x \in \bigcap_{\alpha \in A} \operatorname{cl} C_\alpha$. For each $\alpha \in A$ the segment $[y, x)$ lies in $\operatorname{ri} C_\alpha$ by Theorem 1.2.4; therefore $[y, x) \subset \bigcap_{\alpha \in A} \operatorname{ri} C_\alpha$. As $x \in \operatorname{cl}[y, x)$ it also belongs to $\operatorname{cl} \bigcap_{\alpha \in A} \operatorname{ri} C_\alpha$.

We then have

$$\bigcap_{\alpha \in A} \operatorname{cl} C_\alpha \subset \operatorname{cl} \bigcap_{\alpha \in A} \operatorname{ri} C_\alpha \subset \operatorname{cl} \bigcap_{\alpha \in A} C_\alpha \subset \bigcap_{\alpha \in A} \operatorname{cl} C_\alpha,$$

where the first inclusion came from the argument we just made and the last came from elementary properties of the closure operation. Since the same set appears at each end of the chain of inclusions, we have proved the first claim. We have also shown that

$$\operatorname{cl} \bigcap_{\alpha \in A} C_\alpha = \operatorname{cl} \bigcap_{\alpha \in A} \operatorname{ri} C_\alpha.$$

From this, by recalling that $\operatorname{ri} S = \operatorname{ri} \operatorname{cl} S$ for convex sets S, we obtain

$$\operatorname{ri}\cap_{\alpha\in A}C_\alpha = \operatorname{ri}\operatorname{cl}\cap_{\alpha\in A}C_\alpha = \operatorname{ri}\operatorname{cl}\cap_{\alpha\in A}\operatorname{ri}C_\alpha = \operatorname{ri}\cap_{\alpha\in A}\operatorname{ri}C_\alpha \subset \cap_{\alpha\in A}\operatorname{ri}C_\alpha,$$

and this proves the main part of the second claim.

For the rest of the second claim, suppose that A is finite, say $\{1,\dots,k\}$. Let $D = C_1 \times \dots \times C_k$, a subset of $\mathbb{R}^{nk}$. If we define a linear transformation L from $\mathbb{R}^n$ to $\mathbb{R}^{nk}$ by $Lx = (x,\dots,x)$, then $L^{-1}(\operatorname{ri}D) = \cap_{i=1}^k \operatorname{ri}C_i$ (assumed nonempty), and $\operatorname{ri}L^{-1}(D) = \operatorname{ri}\cap_{i=1}^k C_i$. By Proposition 1.2.11 these sets are the same.

The last claim is certainly true if the intersection is empty. If it is nonempty then the relative interiors of the sets being intersected have a point in common; then the second claim shows that the intersection is its own relative interior, so it is relatively open. □

The restriction to finite sets in the last claim is necessary (Exercise 1.2.18). Further, without the intersection condition the inclusion opposite to (1.9) may hold *strictly* (Exercise 1.2.19).

By combining some of the preceding results we can produce various useful formulas for dealing with sets encountered in practice. Here is an example.

Proposition 1.2.13. *Let $x_0,\dots,x_k$ be points in $\mathbb{R}^n$. Then*

$$\operatorname{ri}\operatorname{conv}\{x_0,\dots,x_k\} = \{\sum_{i=0}^k \lambda_i x_i \mid \lambda_i > 0,\ \sum_{i=0}^k \lambda_i = 1\},$$

and

$$\operatorname{ri}\operatorname{pos}\{x_0,\dots,x_k\} = \{\sum_{i=0}^k \lambda_i x_i \mid \lambda_i > 0\}.$$

Proof. For the first formula, let X be an $n\times(k+1)$ matrix whose columns are $x_0,\dots,x_k$ and let $H = \{x\in\mathbb{R}^{k+1} \mid \sum_{i=0}^k x_i = 1\}$. Then $\operatorname{conv}\{x_0,\dots,x_k\} = X(H\cap\mathbb{R}^{k+1}_+)$. Further, H (which is its own relative interior) meets $\operatorname{int}\mathbb{R}^{k+1}_+$. Applying Propositions 1.2.7 and 1.2.12, we find that

$$\operatorname{ri}\operatorname{conv}\{x_0,\dots,x_k\} = \operatorname{ri}X[(H\cap\mathbb{R}^{k+1}_+)] = X[\operatorname{ri}(H\cap\mathbb{R}^{k+1}_+)]$$
$$= X[H\cap(\operatorname{int}\mathbb{R}^{k+1}_+)] = \{\sum_{i=0}^k \lambda_i x_i \mid \lambda_i > 0,\ \sum_{i=0}^k \lambda_i = 1\}.$$

The proof for $\operatorname{pos}\{x_0,\dots,x_k\}$ is similar except that we only have to deal with $\mathbb{R}^{k+1}_+$ and not with H. □

1.2.4 Exercises for Section 1.2

Exercise 1.2.14. Show by example that in the situation of Proposition 1.2.9, if $\operatorname{im}T$ does not meet $\operatorname{ri}C$ then $T^{-1}(\operatorname{aff}C)$ may properly contain $\operatorname{aff}T^{-1}(C)$.

Exercise 1.2.15. Show that for two convex sets $C \subset \mathbb{R}^n$ and $D \subset \mathbb{R}^m$ one has $\operatorname{ri}(C \times D) = (\operatorname{ri} C) \times (\operatorname{ri} D)$. You may find Corollary 1.2.5 helpful.

Exercise 1.2.16. Exhibit a subset S of $\mathbb{R}^2$ and a point $x \in S$ having the properties that for each $y \in S$ and all sufficiently small positive μ we have $x + \mu(x-y) \in S$, but $x \notin \operatorname{ri} S$.

Exercise 1.2.17. Exhibit a linear transformation E from $\mathbb{R}$ to $\mathbb{R}^2$ and a subset V of $\mathbb{R}^2$ having no more than three points, with $\operatorname{im} E \cap \operatorname{ri} \operatorname{conv} V \neq \emptyset$, such that $E^{-1}(\operatorname{conv} V)$ and $\operatorname{conv} E^{-1}(V)$ are both nonempty but are not equal.

Exercise 1.2.18. Exhibit a family of intervals $\{C_i\}_{i=1}^{\infty}$ in $\mathbb{R}$ such that $\operatorname{ri} \cap_{n=1}^{\infty} C_n$ and $\cap_{i=1}^{\infty} \operatorname{ri} C_n$ are unequal.

Exercise 1.2.19. Exhibit two convex sets C_1 and C_2 such that $\operatorname{ri}(C_1 \cap C_2)$ properly contains $\operatorname{ri} C_1 \cap \operatorname{ri} C_2$.

Exercise 1.2.20. If C is a convex subset of $\mathbb{R}^n$ and $\operatorname{cl} C$ is affine, then C is affine.

1.3 Notes and references

The first published version of the Shapley-Folkman theorem seems to have appeared in [42, Appendix 2].

Chapter 2
Separation

A rough description of separation would say that if two convex sets don't overlap too much then one can put a hyperplane between them, thereby *separating* them. Different definitions of "too much overlap" and "between" lead to different kinds of separation, and an important question is what properties a problem needs to have in order to accomplish these different sorts of separation.

Many important constructions in convexity rely on the information that a separating hyperplane can provide. Moreover, the process of separation is closely connected with notions of distance between points and sets or between sets and sets, and therefore also with methods for approximating sets by other sets.

2.1 Definitions of separation

Suppose C and C' are two nonempty subsets of $\mathbb{R}^n$ and H is a hyperplane in $\mathbb{R}^n$. Here are four different ways in which H can separate C and C'.

Definition 2.1.1. *H separates C* and C' if each of C and C' lies in a different closed halfspace of H. □

Intuitively this seems to mean that C and C' should somehow be "apart," but this is not necessarily the case: for example, the hyperplane H separates itself from itself. We can eliminate this annoyance by introducing the stronger requirement of *proper separation.*

Definition 2.1.2. *H properly separates C* and C' if H separates C and C' with $C \cup C' \not\subset H$. □

Now H no longer separates itself from itself. However, one can properly separate an endpoint of the interval $[0,1]$ in $\mathbb{R}$ from the interval itself (take H to be the endpoint: it's a hyperplane in $\mathbb{R}$), so this kind of separation might seem too weak for some purposes. One can then introduce a third notion, that of *strict separation.*

Definition 2.1.3. *H strictly separates C* and C' if each of these two sets lies in a different open halfspace of H. □

This seems stronger, because the hyperplane lies between the sets and cannot meet either set. But even this does not suffice for some purposes: for example it is possible for two convex sets C and C' to be strictly separated, yet for points of C and of C' to be as close together as we please. An example in $\mathbb{R}^2$ is $C = \{(x_1,x_2) \mid x_2 \geq \exp x_1\}$ and $C' = \{(x_1,x_2) \mid x_2 \leq -\exp x_1\}$, with H being the x_1-axis. For that reason we define a fourth type of separation, called *strong separation:*

Definition 2.1.4. *H strongly separates C* and C' if there is a positive scalar ε such that H separates $C + \varepsilon B^n$ and $C' + \varepsilon B^n$, where B^n is the unit ball. □

Each of these four kinds of separation is stronger than those preceding it, but not all are equally useful. For convex analysis proper separation seems to be the most useful, with strong separation in second place.

The concept of *supporting hyperplane* is also often useful.

Definition 2.1.5. Let S be a subset of $\mathbb{R}^n$ and H be a hyperplane in $\mathbb{R}^n$. We say *H supports S* if S lies in one of the closed halfspaces of H and at least one point of S belongs to H. If in addition $S \not\subset H$ then we say that the support of S by H is *proper*, or that *H properly supports S*. □

The following section characterizes strong separation and then proper separation.

2.2 Conditions for separation

For any subset S of $\mathbb{R}^n$ and any $x \in \mathbb{R}^n$ we say that a point s is *closest in S to x* if $\|x - s\| = d[x,S]$. Such a point need not be unique: for example, if x is the origin of $\mathbb{R}^n$ and S is the sphere S^{n-1}, then every point of S is at the same distance from x and hence is a closest point in S to x.

The idea of separation is related to that of closeness, and one of the most useful facts about separation is the following characterization of the point in a convex set that is closest to a given point.

Lemma 2.2.1. *Let S be a subset of $\mathbb{R}^n$ containing a point s_0, and let $x \in \mathbb{R}^n$. Then for s_0 to be the closest point in S to x it suffices that*

$$\text{For each } s \in S,\ \langle x - s_0, s - s_0 \rangle \leq 0, \tag{2.1}$$

and if (2.1) *holds then s_0 is the unique closest point in S to x.*

If S is also convex, then (2.1) *is necessary and sufficient for s_0 to be the unique closest point in S to x.*

Proof. (Sufficiency). Suppose that (2.1) holds, and choose any $s \in \mathbb{R}^n$. Then

$$x - s = (x - s_0) - (s - s_0),$$

so that

$$\|x-s\|^2 = \|x-s_0\|^2 - 2\langle x-s_0, s-s_0\rangle + \|s-s_0\|^2. \tag{2.2}$$

The second and third terms on the right side of (2.2) are both nonnegative, so that $\|x-s_0\| \leq \|x-s\|$ and therefore s_0 is a closest point in S to x. Further, if $s \neq s_0$ then the third term is positive, so that $\|x-s_0\| < \|x-s\|$, and therefore s_0 is unique.

Necessity. Now suppose that S is convex and that s_0 minimizes $\|x-(\cdot)\|$ on S. Let s be any point of S, and for each $\mu \in (0,1]$ define $s_\mu := s_0 + \mu(s-s_0)$; then $s_\mu \in S$ by convexity. We have

$$x - s_\mu = (x-s_0) - \mu(s-s_0),$$

so that

$$\|x-s_\mu\|^2 = \|x-s_0\|^2 - 2\mu\langle x-s_0, s-s_0\rangle + \mu^2\|s-s_0\|^2.$$

Then

$$0 \geq \|x-s_0\|^2 - \|x-s_\mu\|^2 = 2\mu\langle x-s_0, s-s_0\rangle - \mu^2\|s-s_0\|^2. \tag{2.3}$$

Dividing the right side of (2.3) by 2μ and letting $\mu \downarrow 0$ shows that for each $s \in S$, $0 \geq \langle x-s_0, s-s_0\rangle$. Therefore when S is convex (2.1) is necessary; we have already shown that it is sufficient and that then the closest point is unique. □

2.2.1 Strong separation

Lemma 2.2.1 allows us to characterize those points that we can strongly separate from a convex set.

Proposition 2.2.2. *Let C be a nonempty convex subset of $\mathbb{R}^n$, and let x be a point of $\mathbb{R}^n$. A hyperplane strongly separating $\{x\}$ and C exists if and only if $x \notin \mathrm{cl}C$.*

Proof. (only if) If such a hyperplane exists, there is a nonzero element y^* of $\mathbb{R}^n$, together with real numbers η and ε with ε positive, such that for each $c \in C$ and any elements v and v' of B^n,

$$\langle y^*, x+\varepsilon v\rangle \leq \eta \leq \langle y^*, c+\varepsilon v'\rangle. \tag{2.4}$$

There is no loss in assuming that we have scaled y^* and η so that y^* has norm 1. By subtracting the left side of (2.4) from the right side and rearranging terms we find that

$$\varepsilon\langle y^*, v-v'\rangle \leq \langle y^*, c-x\rangle.$$

By choosing $v = y^* = -v'$ and then using the Schwarz inequality, we obtain

$$2\varepsilon \leq \langle y^*, c-x\rangle \leq \|c-x\|.$$

This means that each point of C lies at a distance at least 2ε from x, so that x cannot belong to $\mathrm{cl}C$.

(if) If $x \notin \mathrm{cl}C$ then let c' be any point of C and consider the intersection D of $\mathrm{cl}C$ with the closed ball $x + \|c' - x\|B^n$. This D is compact and nonempty, so the continuous function $\|x - (\cdot)\|$ attains its minimum on D at some point c_0, which then actually minimizes $\|x - (\cdot)\|$ on all of $\mathrm{cl}C$. We have $\|x - c_0\| > 0$ because $x \notin \mathrm{cl}C$. As $\mathrm{cl}C$ is convex, Lemma 2.2.1 says that for each $c \in \mathrm{cl}C$,

$$0 \geq \langle x - c_0, c - c_0 \rangle = \|x - c_0\|^2 - \langle x - c_0, x - c \rangle. \tag{2.5}$$

Define $y^* = (x - c_0)/\|x - c_0\|$ and $\varepsilon = (1/2)\|x - c_0\|$; then divide (2.5) by $\|x - c_0\|$ and rewrite the result as $\langle y^*, x - c \rangle \geq 2\varepsilon$, which implies that for any $c \in \mathrm{cl}C$,

$$\langle y^*, x \rangle - \varepsilon \geq \langle y^*, c \rangle + \varepsilon. \tag{2.6}$$

Let $\eta = \varepsilon + \sup_{c \in \mathrm{cl}C} \langle y^*, c \rangle$; by (2.6) the supremum is not greater than $\langle y^*, x \rangle - 2\varepsilon$, so η is finite. We have then for any $c \in \mathrm{cl}C$ and any v and v' in B^n

$$\langle y^*, c + \varepsilon v \rangle \leq \langle y^*, c \rangle + \varepsilon \leq \eta \leq \langle y^*, x \rangle - \varepsilon \leq \langle y^*, x + \varepsilon v' \rangle,$$

so that the hyperplane $H_0(y^*, \eta)$ separates $\{x\} + \varepsilon B^n$ and $(\mathrm{cl}C) + \varepsilon B^n$. This means it strongly separates $\{x\}$ and $\mathrm{cl}C$, and therefore also $\{x\}$ and C. □

Proposition 2.2.2 lets us prove an external representation for the closure of the convex hull, complementing the internal representation in Proposition 1.1.4.

Theorem 2.2.3. *Let S be a subset of $\mathbb{R}^n$. Then $\mathrm{cl}\,\mathrm{conv}\,S$ is the intersection of all closed halfspaces of $\mathbb{R}^n$ that contain S.*

Proof. If S is empty the claim is evidently true, so we can suppose that S is nonempty. Write C for the intersection of all closed halfspaces containing S. Any closed halfspace containing S is closed and convex; therefore it contains $\mathrm{cl}\,\mathrm{conv}\,S$ as well, so C must also contain $\mathrm{cl}\,\mathrm{conv}\,S$. For the opposite inclusion, suppose x does not belong to $\mathrm{cl}\,\mathrm{conv}\,S$. Use Proposition 2.2.2 to find a nonzero point $y^* \in \mathbb{R}^n$ and a real number η such that for each $z \in \mathrm{cl}\,\mathrm{conv}\,S$, $\langle y^*, x \rangle > \eta \geq \langle y^*, z \rangle$. The closed halfspace consisting of all v with $\langle y^*, v \rangle \leq \eta$ contains S but not x; accordingly, x does not belong to C, so $\mathrm{cl}\,\mathrm{conv}\,S$ contains C. □

For another application of this representation, we introduce the polarity operation, which plays an important part in many aspects of convexity.

Definition 2.2.4. C be a nonempty subset of $\mathbb{R}^n$. The *polar* of C is the set C° of all x^* in $\mathbb{R}^n$ such that for each $x \in C$, $\langle x^*, x \rangle \leq 1$. □

As C° is the intersection over $x \in C$ of a collection of closed convex sets of the form $\{\, x^* \mid \langle x^*, x \rangle \leq 1 \,\}$, it is closed, convex, and nonempty (because it always contains the origin). Also, the definition shows that if $C_1 \supset C_2$ then $C_1^\circ \subset C_2^\circ$. One can check that if C happens to be a subspace, then $C^\circ = C^\perp$.

Here is a fundamental property of this polarity operation.

Theorem 2.2.5. *Let S be a nonempty subset of $\mathbb{R}^n$. Then $S^{\circ\circ} = \mathrm{cl}\,\mathrm{conv}(S \cup \{0\})$.*

Proof. Write T for $\mathrm{cl}\,\mathrm{conv}(S\cup\{0\})$.

($\supset$). Let $x\in T$. If a point x^* belongs to S° then by definition $H_-(x^*,1)$ contains S. It contains the origin too, so it contains $S\cup\{0\}$. Then Theorem 2.2.3 says that x belongs to $H_-(x^*,1)$. Therefore

$$x\in\cap_{x^*\in S^\circ}H_-(x^*,1)=S^{\circ\circ},$$

so that $S^{\circ\circ}\supset T$.

($\subset$). Suppose that $x\notin T$. By Proposition 2.2.2 there is a closed halfspace $H_-(x^*,\xi)$ containing $S\cup\{0\}$ that does not contain x. As this halfspace contains the origin, $\langle x^*,x\rangle>\xi\geq 0$. By scaling x^* and ξ appropriately we can arrange that $\xi\leq 1$ and that $\langle x^*,x\rangle>1$. The first of these inequalities shows that $x^*\in(S\cup\{0\})^\circ\subset S^\circ$; then the second shows that $x\notin S^{\circ\circ}$. Therefore $S^{\circ\circ}\subset T$. □

Corollary 2.2.6. *If C is a convex subset of $\mathbb{R}^n$ whose closure contains the origin, then $C^{\circ\circ}=\mathrm{cl}\,C$.*

Proof. The hypothesis shows that $\mathrm{cl}\,C\supset C\cup\{0\}$, and as $\mathrm{cl}\,C$ is convex and closed we have the first inclusion in

$$\mathrm{cl}\,C\supset\mathrm{cl}\,\mathrm{conv}(C\cup\{0\})\supset\mathrm{cl}\,C;$$

the second follows from $\mathrm{conv}(C\cup\{0\})\supset C$. Then Theorem 2.2.5 shows that

$$C^{\circ\circ}=\mathrm{cl}\,\mathrm{conv}(C\cup\{0\})=\mathrm{cl}\,C.$$

□

Corollary 2.2.6 contains as a very special case the familiar equality of L and $L^{\perp\perp}$ for a subspace L of $\mathbb{R}^n$.

Our results so far have concerned strong separation of a point and a set. With very little additional effort we can characterize conditions under which we can strongly separate two nonempty convex sets.

Theorem 2.2.7. *Let C and D be nonempty convex subsets of $\mathbb{R}^n$. Then C and D can be strongly separated if and only if $0\notin\mathrm{cl}(C-D)$.*

Proof. (only if) Strong separation implies that for some nonzero $y^*\in\mathbb{R}^n$ and some real η and positive ε, one has for each $c\in C$ and $d\in D$,

$$\langle y^*,c\rangle\geq\eta+\varepsilon,\qquad\langle y^*,d\rangle\leq\eta-\varepsilon.$$

Subtracting and using the Schwarz inequality, we find that

$$\|y^*\|\,\|c-d\|\geq\langle y^*,c-d\rangle\geq 2\varepsilon,$$

so that $\|c-d\|\geq 2\varepsilon/\|y^*\|$. Therefore the origin cannot belong to the closure of $C-D$.

(if) Suppose that $0 \notin \mathrm{cl}(C-D)$. Proposition 2.2.2 then tells us that we can strongly separate the origin from $\mathrm{cl}(C-D)$. In particular, this means that there is a nonzero y^* such that for some positive ε and all $c \in C$ and $d \in D$,

$$\langle y^*, c-d\rangle - \varepsilon \geq \langle y^*, 0\rangle + \varepsilon;$$

that is,

$$\langle y^*, c\rangle - \varepsilon \geq \langle y^*, d\rangle + \varepsilon.$$

It follows that

$$\inf_{c\in C}\langle y^*, c\rangle - \varepsilon \geq \sup_{d\in D}\langle y^*, d\rangle + \varepsilon. \tag{2.7}$$

Let η be any real number between the left and right sides of (2.7); then if $c \in C$ and $d \in D$ and if u and v are any elements of B^n,

$$\langle y^*, d+\varepsilon u\rangle \leq \eta \leq \langle y^*, c+\varepsilon v\rangle,$$

so that the hyperplane $H_0(y^*, \eta)$ strongly separates C and D. □

A possibly surprising feature of Theorem 2.2.7 is that it does not say that we can strongly separate two closed convex sets if and only if they are disjoint. The reason is that such a statement is not generally true (Exercise 2.2.14). However, it becomes true if we add the requirement that one of the sets be compact. The next proposition shows why.

Proposition 2.2.8. *If S and T are subsets of $\mathbb{R}^n$, with S closed and T compact, then $S+T$ is closed.*

Proof. If either set is empty there is nothing to prove, so assume each is nonempty. If $x \in \mathrm{cl}(S+T)$ then there are sequences $\{s_k\} \subset S$ and $\{t_k\} \subset T$ such that $\{s_k+t_k\}$ converges to x. As T is compact, by thinning the sequence if necessary we can assume that the t_k converge to $t \in T$; then the s_k must also converge to some s, and as S is closed we have $s \in S$. Therefore $x = s+t \in S+T$, so $S+T$ is closed. □

Corollary 2.2.9. *Let C and D be nonempty convex sets in $\mathbb{R}^n$, with C closed and D compact. Then C and D can be strongly separated if and only if they are disjoint.*

Proof. Saying that C and D are disjoint is equivalent to saying that the origin does not belong to $C-D$. Under our hypotheses Proposition 2.2.8 guarantees that the set $C-D$ is closed, so the assertion follows from Theorem 2.2.7. □

2.2.2 Proper separation

Having characterized strong separation, we now do the same for proper separation. The procedure is similar to that for strong separation: first we deal with a point and a set, then extend the result to two sets. However, the arguments are more delicate.

Proposition 2.2.10. *Let C be a nonempty convex subset of $\mathbb{R}^n$, and let x be a point of $\mathbb{R}^n$. A hyperplane properly separating $\{x\}$ and C exists if and only if $x \notin \mathrm{ri}\, C$.*

Proof. (only if) Suppose that $H_0(y^*, \eta)$ is a hyperplane properly separating $\{x\}$ and C. If $x \notin H_0(y^*, \eta)$ then x is not even contained in $\mathrm{cl}\, C$, much less in $\mathrm{ri}\, C$. Therefore suppose that $x \in H_0(y^*, \eta)$ and suppose that we have chosen y^* so that $C \subset H_-(y^*, \eta)$. The hypothesis of proper separation says that $C \not\subset H_0(y^*, \eta)$, so by Corollary 1.2.3 $\mathrm{ri}\, C$ does not meet this hyperplane and therefore cannot contain x.

(if) Suppose $x \notin \mathrm{ri}\, C$. If $x \notin \mathrm{cl}\, C$, then by Proposition 2.2.2 we can strongly separate $\{x\}$ and $\mathrm{cl}\, C$, which certainly suffices. Therefore suppose $x \in \mathrm{cl}\, C$. We have $\mathrm{ri}\,\mathrm{cl}\, C = \mathrm{ri}\, C$, so $x \notin \mathrm{ri}\,\mathrm{cl}\, C$ and so it is in the relative boundary of $\mathrm{cl}\, C$. This means that there are points $x_n \in \mathrm{aff}\, C$ converging to x, none of which belongs to $\mathrm{cl}\, C$. Project each x_n onto $\mathrm{cl}\, C$ to obtain points c_n; as $x \in \mathrm{cl}\, C$ we have

$$\|c_n - x\| \leq \|c_n - x_n\| + \|x_n - x\| \leq 2\|x_n - x\|,$$

so the c_n also converge to x. Apply Lemma 2.2.1 to show that for each n and each $c \in \mathrm{cl}\, C$,

$$\langle x_n - c_n, c - c_n \rangle \leq 0. \tag{2.8}$$

Let $z_n := (x_n - c_n)/\|x_n - c_n\|$; these points all belong to $\mathrm{par}\, C$ and have norm 1, so by thinning the sequence we may suppose they converge to $z_0 \in \mathrm{par}\, C$ with norm 1. If we divide (2.8) by $\|x_n - c_n\|$ and take the limit, we obtain for each $c \in C$ the inequality $\langle z_0, c - x \rangle \leq 0$. If we let $\zeta = \langle z_0, x \rangle$ then $C \subset H_-(z_0, \zeta)$ and $x \in H(z_0, \zeta)$.

To show that this separation is proper, let $c \in \mathrm{ri}\, C$; then as $z_0 \in \mathrm{par}\, C$ we have for sufficiently small positive ε the inclusion $c + \varepsilon z_0 \in C$, so that

$$\zeta \geq \langle z_0, (c + \varepsilon z_0) \rangle = \langle z_0, c \rangle + \varepsilon.$$

Therefore $\langle z_0, c \rangle < \zeta$, showing that $\mathrm{ri}\, C$ is actually disjoint from $H_0(z_0, \zeta)$ and, in particular that the separation is proper. □

Here is the characterization of proper separation.

Theorem 2.2.11. *Let C and D be nonempty convex subsets of $\mathbb{R}^n$. Then C and D can be properly separated if and only if $\mathrm{ri}\, C \cap \mathrm{ri}\, D = \emptyset$.*

Proof. As $\mathrm{ri}(C - D) = \mathrm{ri}\, C - \mathrm{ri}\, D$ (Corollary 1.2.8), $\mathrm{ri}\, C$ fails to meet $\mathrm{ri}\, D$ exactly when the origin does not belong to $\mathrm{ri}(C - D)$. By Proposition 2.2.10 this is equivalent to saying that the origin can be properly separated from $C - D$ by a hyperplane. Therefore we only need to show that such separation occurs if and only if C and D can be properly separated.

(Only if). If $H_0(y^*, \eta)$ properly separates the origin from $C - D$, we can suppose that

$$C - D \subset H_-(y^*, \eta), \qquad 0 \in H_+(y^*, \eta), \qquad (C - D) \cup \{0\} \not\subset H_0(y^*, \eta). \tag{2.9}$$

The first statement in (2.9) implies that for each $c \in C$ and $d \in D$, $\langle y^*, c \rangle \leq \langle y^*, d \rangle + \eta$, and the second statement shows that $\eta \leq 0$. Taking the supremum on the left over

$c \in C$ and then the infimum on the right over $d \in D$, we obtain

$$\sup_{c\in C}\langle y^*, c\rangle \le \inf_{d\in D}\langle y^*, d\rangle + \eta,$$

and if we take η_0 to be any real number between the left and right sides of this inequality we find that

$$C \subset H_-(y^*, \eta_0), \qquad D \subset H_+(y^*, \eta_0 - \eta) \subset H_+(y^*, \eta_0). \tag{2.10}$$

Accordingly, $H_0(y^*, \eta_0)$ separates C and D.

If $\eta = 0$, then if C and D were both contained in $H_0(y^*, \eta_0)$ we would have $C - D \subset H_0(y^*, 0)$. As the origin also belongs to that hyperplane, we would then have a contradiction to the third statement in (2.9). Therefore in this case $H_0(y^*, \eta_0)$ properly separates C and D.

If $\eta < 0$, then (2.10) shows that for each $c \in C$ and $d \in D$,

$$\langle y^*, c\rangle \le \eta_0 < \eta_0 - \eta \le \langle y^*, d\rangle,$$

so that no point of D can belong to $H_0(y^*, \eta_0)$ and therefore in this case also the separation is proper.

(If). Suppose that $H_0(y^*, \eta)$ properly separates C and D. Then for each $c \in C$ and $d \in D$, $\langle y^*, c - d\rangle$ will remain on the same side of the origin and, for at least one pair $\hat{c} \in C$ and $\hat{d} \in D$, we have $\langle y^*, \hat{c}\rangle \ne \langle y^*, \hat{d}\rangle$. The first statement shows that $H_0(y^*, 0)$ separates $C - D$ from the origin. The separation must also be proper because $\langle y^*, \hat{c} - \hat{d}\rangle \ne 0$ so that $\hat{c} - \hat{d} \notin H_0(y^*, 0)$. □

Theorem 2.2.5 applied strong separation to show that the polar of S° was the closure of the convex hull of $S \cup \{0\}$. We can now use proper separation to learn something about the structure of S°.

Theorem 2.2.12. *Let S be a nonempty subset of $\mathbb{R}^n$. Then*

a. If $0 \notin \operatorname{ri}\operatorname{conv} S$ then S° contains a halfline from the origin.
b. If $0 \in \operatorname{ri}\operatorname{conv} S$, let $\delta = d[0, \operatorname{rb}\operatorname{conv} S]$ and let L be the subspace parallel to $\operatorname{conv} S$. Then there is a subset D of $B(0, \delta^{-1}) \cap L$ such that $S^\circ = D + L^\perp$.

Proof. If $0 \notin \operatorname{ri}\operatorname{conv} S$, separate $\{0\}$ properly from $\operatorname{conv} S$ to produce a nonzero s^* and $\sigma \in \mathbb{R}$ such that for each $s \in S$,

$$0 = \langle s^*, 0\rangle \ge \sigma \ge \langle s^*, s\rangle,$$

and $\{0\} \cup \operatorname{conv} S \not\subset H_0(s^*, \sigma)$. Then S° contains the halfline $s^*\mathbb{R}_+$, which proves (a).

For (b), suppose $0 \in \operatorname{ri}\operatorname{conv} S$; then $\operatorname{aff}\operatorname{conv} S$ is the parallel subspace L. For each $s^* \in S^\circ$ let $s^* = s_L^* + s_N^*$ be the unique decomposition of s^* into $s_L^* \in L$ and $s_N^* \in L^\perp$, and define $D = \{s_L^* \mid s^* \in S^\circ\}$.

If $s^* \in S^\circ$ then $s^* = s_L^* + s_N^*$ with $s_L^* \in D$ and $s_N^* \in L^\perp$, so $s^* \in D + L^\perp$ and therefore $S^\circ \subset D + L^\perp$. Now suppose $t^* \in D + L^\perp$. Then there are $s^* \in S^\circ$ and $n^* \in L^\perp$ such that $t^* = s_L^* + n^*$. Let $s \in S$. Then

$$\langle t^*, s\rangle = \langle s_L^* + n^*, s\rangle = \langle s^*, s\rangle + \langle n^* - s_N^*, s\rangle \le 1 + 0 = 1,$$

because $s \in L$ and therefore $\langle n^* - s_N^*, s\rangle = 0$. so $t^* \in S^\circ$ and therefore $D + L^\perp \subset S^\circ$. We then have $S^* = D + L^\perp$.

To establish the bound on D, let $\varepsilon \in (0, \delta)$; then $\varepsilon B_L \subset \operatorname{conv} S$, where $B_L = B^n \cap L$. The definition of D above shows that each element of D has the form s_L^* where for some $s^* \in S^\circ$ we have $s^* = s_L^* + s_N^*$ with $s_L^* \in L$ and $s_N^* \in L^\perp$. If $s_L^* = 0$ then certainly $\|s_L^*\| \le \varepsilon^{-1}$. Otherwise, by choice of ε we have $\varepsilon s_L^*/\|s_L^*\| \in \operatorname{conv} S$, so there exist points $s_0, \ldots, s_k$ of S and convex coefficients $\mu_0, \ldots, \mu_k$ with

$$\varepsilon s_L^*/\|s_L^*\| = \sum_{i=0}^{k} \mu_i s_i.$$

As $\langle s^*, s_i\rangle \le 1$ and $\langle s_N^*, s_i\rangle = 0$ for each i, we have

$$1 \ge \sum_{i=0}^{k} \mu_i \langle s^*, s_i\rangle = \sum_{i=0}^{k} \mu_i \langle s_L^*, s_i\rangle = \langle s_L^*, \sum_{i=0}^{k} \mu_i s_i\rangle = \langle s_L^*, \varepsilon s_L^*/\|s_L^*\|\rangle = \varepsilon \|s_L^*\|,$$

and therefore $\|s_L^*\| \le \varepsilon^{-1}$. It follows that no element of D has norm greater than ε^{-1}. As ε can be arbitrarily close to δ, we have $D \subset B(0, \delta^{-1}) \cap L$. □

2.2.3 Exercises for Section 2.2

Exercise 2.2.13. Let C be a nonempty closed convex subset of $\mathbb{R}^n$. Define the *Euclidean projector* π_C on $\mathbb{R}^n$ to be, for each $x \in \mathbb{R}^n$, the point $c(x) \in C$ closest to x. Now define the *normal component* $\nu_C(x)$ to be $x - \pi_C(x)$.

1. Show that for each x and x' in $\mathbb{R}^n$ the following inequalities hold.

$$\langle \nu_C(x) - \nu_C(x'), \pi_C(x) - \pi_C(x')\rangle \ge 0; \tag{2.11}$$

$$\|\pi_C(x) - \pi_C(x')\|^2 + \|\nu_C(x) - \nu_C(x')\|^2 \le \|x - x'\|^2; \tag{2.12}$$

$$\langle \pi_C(x) - \pi_C(x'), x - x'\rangle \ge \|\pi_C(x) - \pi_C(x')\|^2; \tag{2.13}$$

$$\langle \nu_C(x) - \nu_C(x'), x - x'\rangle \ge \|\nu_C(x) - \nu_C(x')\|^2. \tag{2.14}$$

2. Show that each of π_C and ν_C is Lipschitz continuous on $\mathbb{R}^n$ with Lipschitz modulus 1. For each of these functions show that this Lipschitz constant cannot be improved, by exhibiting a nonempty closed convex subset C of $\mathbb{R}^2$ and points x and x' in $\mathbb{R}^2$ such that the norm of the difference of the function values at x and x' is $\|x - x'\|$.

Exercise 2.2.14. Give an example in $\mathbb{R}^2$ to show that there exist nonempty closed convex sets C and D that are disjoint but cannot be strongly separated.

Exercise 2.2.15. Let C be a convex subset of $\mathbb{R}^n$ and let $x \in C$. Show that C has a supporting hyperplane at x if and only if x is in the boundary of C, and that if the support is proper then x is in the relative boundary of C.

Chapter 3
Cones

3.1 Basic properties and some applications

This section defines cones and develops some elementary properties. It also illustrates some ways in which they occur in applications by providing two prominent members of the class of *theorems of the alternative* from linear algebra.

3.1.1 Definitions

Definition 3.1.1. A subset K of $\mathbb{R}^n$ is a *cone* if whenever $x \in K$ and μ is a positive real number, $\mu x \in K$. □

A cone is thus a set that is closed under *positive* scalar multiplication. Some authors use a different definition, requiring a cone to be closed under nonnegative scalar multiplication, so it is a good idea to check what definition an author is using. One important reason for using positive scalar multiplication is Theorem 3.1.7 below.

A cone may, but need not, contain the origin; it may be open or closed or neither. It may or may not be convex; if it is, we call it a *convex cone*. The definition also implies that the forward and inverse images of a cone under a linear transformation are cones, and the Cartesian product of a finite collection of cones is a cone.

Any subspace is a cone, but many interesting cones are not subspaces. Three familiar convex cones are

- $\mathbb{R}^n_+$, the nonnegative orthant of the space $\mathbb{R}^n$;
- S^n_+, the *semidefinite cone,* consisting of the positive semidefinite symmetric matrices in $\mathbb{R}^{n\times n}$;
- $N^{m\times n}$, the cone of matrices in $\mathbb{R}^{m\times n}$ whose entries are all nonnegative.

We have already seen the polarity operation for sets. For cones it takes a special form.

Proposition 3.1.2. *Let K be a nonempty cone in $\mathbb{R}^n$. The polar K° is a nonempty closed convex cone, and we have*

$$K^\circ = \{x^* \in \mathbb{R}^n \mid \textit{for each } x \in K,\ \langle x^*, x\rangle \le 0\}. \tag{3.1}$$

When K is convex we have

$$(\operatorname{par} K)^\perp = \operatorname{lin}(K^\circ). \tag{3.2}$$

Proof. We already know that K° will be a closed convex set, and it must contain the set shown on the right in (3.1). To see that they are the same set, suppose that $x^* \in K^\circ$, so that $\langle x^*, x\rangle \le 1$ for each $x \in K$. Then there cannot be any $x \in K$ with $\langle x^*, x\rangle > 0$, as otherwise we could multiply x by a sufficiently large positive scalar so that $\langle x^*, \alpha x\rangle > 1$. But $\alpha x \in K$, so this contradicts $x^* \in K^\circ$. Therefore (3.1) holds, and it shows that K° is a cone.

Suppose that K is convex and that $v^* \in (\operatorname{par} K)^\perp$ and $u^* \in K^\circ$. Then for any $x \in K$ we have $2x \in K$ so $x = 2x - x \in \operatorname{par} K$. Then for each $\alpha \in \mathbb{R}$,

$$\langle u^* + \alpha v^*, x\rangle = \langle u^*, x\rangle \le 0,$$

which implies that $v^* \in \operatorname{lin}(K^\circ)$. Therefore $(\operatorname{par} K)^\perp \subset \operatorname{lin}(K^\circ)$.

To show the other inclusion, let v^* be any element of $\operatorname{lin} K^\circ$ and u^* be any element of K°. For each $x \in K$ and each $\alpha \in \mathbb{R}$ we have $\langle u^* + \alpha v^*, x\rangle \le 0$, which implies that $\langle v^*, x\rangle = 0$. Let L be the linear space spanned by v^*; then $K \subset L^\perp$ and therefore

$$\operatorname{par} K = \operatorname{aff} K \subset L^\perp.$$

But then $L \subset (\operatorname{par} K)^\perp$, and as L was the span of an arbitrary element of $\operatorname{lin}(K^\circ)$, we have $\operatorname{lin}(K^\circ) \subset (\operatorname{par} K)^\perp$. □

Sometimes we have to add cones to other sets. If the other set is bounded and the result is a cone, the following proposition can be useful.

Proposition 3.1.3. *Let P and K be nonempty subsets of $\mathbb{R}^n$, with P bounded and K a closed cone. If $P + K$ is a cone then $P \subset K$.*

Proof. Suppose $P + K$ is a cone. As $0 \in K$ because K is closed, we have $P = P + \{0\} \subset P + K$ so that for any positive μ, $\mu P \subset P + K$. Choose $p \in P$: then for some $p' \in P$ and $k' \in K$ we have $\mu p = p' + k'$ and so $p - \mu^{-1} p' = \mu^{-1} k' \in K$. Therefore $d_K(p) \le \mu^{-1}\|p'\| \le \mu^{-1}\pi$, where the ball $B(0,\pi)$ contains P. As μ can be arbitrarily large and K is closed, this implies that $p \in K$ and therefore $P \subset K$. □

Certain operations make sets into cones, such as the conical hull defined next.

Definition 3.1.4. If S is a subset of $\mathbb{R}^n$, the *conical hull* of S, written $\operatorname{cone} S$, is the intersection of all cones containing S. □

At least one cone ($\mathbb{R}^n$) contains S, so the intersection in Definition 3.1.4 is not over an empty collection, though the intersection itself can be empty: the empty set is

also a cone. As the intersection of any collection of cones is a cone, this definition says that cone S is the smallest cone containing S. As the next proposition shows, we can also describe cone S as the *cone generated by* S: that is, the set $\cup_{\alpha>0}\alpha S$, where for a subset S of $\mathbb{R}^n$ and a real number α,

$$\alpha S := \{\alpha s \mid s \in S\}.$$

Under this definition, cone S contains the origin if and only if S does. Some definitions in the literature differ from this one.

Proposition 3.1.5. *For any subset S of $\mathbb{R}^n$,* cone $S = \cup_{\alpha>0}\alpha S$.

Proof. Write G for $\cup_{\alpha>0}\alpha S$. G is a cone (use the definition), and it contains S because α assumes the value 1 along with the other positive reals, so cone $S \subset G$. On the other hand, if K is any cone containing S then K must contain every positive multiple of any element of S, so $K \supset G$. Definition 3.1.4 then shows that cone $S \supset G$, so cone $S = G$. □

If C is a convex set then so is cone C, though the converse is generally false. To see this, suppose that x and x' belong to cone C and that $\lambda \in [0,1]$. By Proposition 3.1.5, there are points c and c' of C and positive real numbers μ and μ' such that $x = \mu c$ and $x' = \mu' c'$. Let $\gamma = (1-\lambda)\mu + \lambda\mu'$; then

$$(1-\lambda)x + \lambda x' = \gamma[(1-\lambda)\mu\gamma^{-1}c + \lambda\mu'\gamma^{-1}c'].$$

The quantity in square brackets on the right is a convex combination of c and c'. It is therefore in C, so the left-hand side belongs to cone C.

3.1.2 Relative interiors of cones

The next theorem shows that in the convex case the conical hull and relative interior operators commute, and it also provides a good example of the proper separation theorem in action. Its proof uses the following lemma about separation of two sets, one of which is a cone.

Lemma 3.1.6. *Let K and S be nonempty subsets of $\mathbb{R}^n$, and suppose that K is a cone. If a hyperplane $H_0(y^*,\eta)$ separates K and S, then we may choose the orientation of (y^*,η) so that $H_0(y^*,0)$ separates K and S, $y^* \in K^\circ$, and $\eta \geq 0$. If the separation by $H_0(y^*,\eta)$ is proper, then so is the separation by $H_0(y^*,0)$.*

Proof. By multiplying (y^*,η) by -1 if needed, we can arrange that $K \subset H_-(y^*,\eta)$ and $S \subset H_+(y^*,\eta)$. If any $k \in K$ had $\langle y^*,k\rangle > 0$ then by taking a large positive γ we could obtain $\langle y^*,\gamma k\rangle > \eta$, a contradiction; therefore $K \subset H_-(y^*,0)$, so that $y^* \in K^\circ$. Moreover, as the origin belongs to cl K we must have $\eta \geq 0$; as $S \subset H_+(y^*,\eta)$ we also have $S \subset H_+(y^*,0)$ so that $H_0(y^*,0)$ separates K and S.

Now suppose that the separation by $H_0(y^*,\eta)$ is proper. If $\eta > 0$ then S is disjoint from $H_0(y^*,0)$ so the separation is proper, while if $\eta = 0$ then the separation is proper by hypothesis. □

Theorem 3.1.7. *If C is a convex subset of $\mathbb{R}^n$, then* $\operatorname{ri}\operatorname{cone} C = \operatorname{cone}\operatorname{ri} C$.

Proof. If C is empty there is nothing to prove, so assume C is nonempty.

($\subset$) If $x \notin \operatorname{cone}\operatorname{ri} C$, define $X = \operatorname{cone}\{x\}$. Then X and $\operatorname{ri} C$ are disjoint, so the proper separation theorem tells us there is a nonzero $y^* \in \mathbb{R}^n$ and a real η such that $H_0(y^*,\eta)$ properly separates X and C. As X is a cone we can use Lemma 3.1.6 to choose the orientation of (y^*,η) so that $H_0(y^*,0)$ properly separates X and C, $y^* \in X^\circ$, and $\eta \geq 0$. Therefore in fact we have for each $c \in C$,

$$\langle y^*, c\rangle \geq \eta \geq 0 \geq \langle y^*, x\rangle.$$

Moreover, proper separation guarantees that there is some $\hat{c} \in C$ with $\langle y^*, x\rangle < \langle y^*, \hat{c}\rangle$.

Let ε be any positive number and define $p = x + \varepsilon(x - \hat{c})$. Then

$$\langle y^*, p\rangle = \langle y^*, x\rangle + \varepsilon(\langle y^*, x\rangle - \langle y^*, \hat{c}\rangle) < \langle y^*, x\rangle \leq 0.$$

This implies that for each positive ξ we have $\langle y^*, \xi p\rangle < 0 \leq \eta$, so $\xi p \notin C$, and therefore $p \notin \operatorname{cone} C$. Now Corollary 1.2.5 shows that $x \notin \operatorname{ri}\operatorname{cone} C$, so that $\operatorname{ri}\operatorname{cone} C \subset \operatorname{cone}\operatorname{ri} C$.

($\supset$) Let $x \in \operatorname{cone}\operatorname{ri} C$, and let z be an arbitrary point of $\operatorname{cone} C$. We will find $\varepsilon > 0$ with $x + \varepsilon(x - z) \in \operatorname{cone} C$, which by Corollary 1.2.5 will establish that $x \in \operatorname{ri}\operatorname{cone} C$.

The assumptions about x and z mean that there are positive numbers σ and τ such that $\tau x \in \operatorname{ri} C$ and $\sigma z \in C$. As $\tau x \in \operatorname{ri} C$ there is some positive δ_0 such that $\tau x + \delta_0(\tau x - \sigma z) \in C$, and since C is convex it is easy to see that the inclusion holds also if δ_0 is replaced by any number $\delta \in (0, \delta_0)$.

Choose $\delta \in (0, \delta_0)$ so small that $1 + \delta^{-1} > \tau^{-1}\sigma$, and define positive numbers ε and α by

$$1 + \delta^{-1} = (1 + \varepsilon^{-1})(\tau^{-1}\sigma), \qquad \alpha = \varepsilon^{-1}\delta\sigma.$$

Then $\delta\sigma = \varepsilon\alpha$, and

$$\begin{aligned}(1+\delta)\tau &= \delta\tau(1+\delta^{-1}) = \delta\tau(1+\varepsilon^{-1})(\tau^{-1}\sigma)\\ &= \varepsilon\alpha(1+\varepsilon^{-1}) = (1+\varepsilon)\alpha.\end{aligned}$$

We know that C contains the point $y := \tau x + \delta(\tau x - \sigma z)$. However,

$$\begin{aligned}\alpha^{-1}y &= \alpha^{-1}[\tau x + \delta(\tau x - \sigma z)] = \alpha^{-1}[(1+\delta)\tau x - \delta\sigma z]\\ &= \alpha^{-1}[(1+\varepsilon)\alpha x - \varepsilon\alpha z] = x + \varepsilon(x - z).\end{aligned}$$

Therefore $x + \varepsilon(x - z)$ belongs to $\operatorname{cone} C$, so that $x \in \operatorname{ri}\operatorname{cone} C$ and therefore $\operatorname{ri}\operatorname{cone} C \supset \operatorname{cone}\operatorname{ri} C$. □

Corollary 3.1.8. *Let C be a nonempty convex cone in $\mathbb{R}^n$. Then* $\mathrm{ri}\,C$ *and* $\mathrm{cl}\,C$ *are convex cones, and the three cones* $\mathrm{ri}\,C$*, C, and* $\mathrm{cl}\,C$ *all have the same polar.*

Proof. As C is convex, both $\mathrm{ri}\,C$ and $\mathrm{cl}\,C$ are convex. That $\mathrm{cl}\,C$ is a cone is easy to show using the definition of a cone, and Theorem 3.1.7 shows that $\mathrm{ri}\,C = \mathrm{ri}\,\mathrm{cone}\,C = \mathrm{cone}\,\mathrm{ri}\,C$, so $\mathrm{ri}\,C$ is a cone.

We have $\mathrm{ri}\,C \subset C \subset \mathrm{cl}\,C$, and by taking polars we find that

$$(\mathrm{cl}\,C)^\circ \subset C^\circ \subset (\mathrm{ri}\,C)^\circ. \tag{3.3}$$

Let $x^* \in (\mathrm{ri}\,C)^\circ$; then as $\mathrm{cl}\,C = \mathrm{cl}\,\mathrm{ri}\,C$ by Proposition 1.2.6, for any point $x \in \mathrm{cl}\,C$ there is a sequence $\{c_k\} \subset \mathrm{ri}\,C$ converging to x. As $\langle x^*, c_k\rangle \le 0$ for each k, we have $\langle x^*, x\rangle \le 0$ also, and therefore $x^* \in (\mathrm{cl}\,C)^\circ$. Then $(\mathrm{ri}\,C)^\circ \subset (\mathrm{cl}\,C)^\circ$, so the three sets in (3.3) are the same. □

Corollary 3.1.9. *If C is a nonempty convex cone in $\mathbb{R}^n$, then $C^{\circ\circ} = \mathrm{cl}\,C$.*

Proof. As $C^\circ = (\mathrm{cl}\,C)^\circ$ by Corollary 3.1.8, it suffices to show that for a nonempty closed convex cone K one has $K^{\circ\circ} = K$. Such a K must contain the origin, so the result follows from Corollary 2.2.6. □

3.1.3 Two theorems of the alternative

Theorems of the alternative are results about the solvability of systems of linear equations and/or inequalities, in which cones play important parts. They are often useful in the analysis of problems from various application areas. This section presents two of the best known theorems of the alternative, due to Farkas and to Gordan. After developing some additional tools we will show in Section 3.2 how to generalize each of these.

Proposition 3.1.10 (Farkas lemma, 1898). *Let A be an $m \times n$ matrix and let $a \in \mathbb{R}^m$. Exactly one of the following holds:*

a. There is some $x \in \mathbb{R}^n_+$ with $Ax = a$.
b. There is some $u^* \in \mathbb{R}^m$ with $A^*u^* \in \mathbb{R}^n_-$ and $\langle u^*, a\rangle > 0$.

Proof. To show that (a) and (b) cannot both hold, suppose they did. Then

$$0 < \langle u^*, a\rangle = \langle u^*, Ax\rangle = \langle A^*u^*, x\rangle \le 0,$$

where the final inequality holds because $A^*u^* \in \mathbb{R}^n_-$ and $x \in \mathbb{R}^n_+$. This is a contradiction, so both cannot hold.

To complete the proof it suffices to show that if (a) fails then (b) holds. Write the columns of A as $a_1, \ldots, a_n$ and let

$$C = A(\mathbb{R}^n_+) = \mathrm{pos}\{a_1, \ldots, a_n\} = \mathrm{conv}\{0\} + \mathrm{pos}\{a_1, \ldots, a_n\}.$$

This C is a cone, and it is also finitely generated. By Proposition 1.1.22, C is closed. To say that (a) does not hold is to say that $a \notin C$. Let c_0 be the point in C closest to a, and let $u^* = a - c_0$. By construction, $u^* \neq 0$. Further, Lemma 2.2.1 says that for each $c \in C$,

$$0 \geq \langle a - c_0, c - c_0 \rangle = \langle u^*, c - c_0 \rangle. \tag{3.4}$$

By taking $c = (1/2)c_0$ and $c = 2c_0$ in (3.4) we find that $\langle u^*, c_0 \rangle = 0$. A point c belongs to C if and only if it is of the form $c = Ax$ for $x \in \mathbb{R}^n_+$; therefore for each $x \in \mathbb{R}^n_+$ we have

$$0 \geq \langle u^*, Ax \rangle = \langle A^* u^*, x \rangle. \tag{3.5}$$

Let $i \in \{1, \ldots, n\}$ and take x in (3.5) to be e_i, a point whose coordinates are 0 for every index except i and 1 for the index i. We obtain $(A^* u^*)_i \leq 0$, and as this holds for each i we have $A^* u^* \in \mathbb{R}^n_-$. Also, as $\langle u^*, c_0 \rangle = 0$ we have

$$\langle u^*, a \rangle = \langle u^*, a - c_0 \rangle = \|u^*\|^2 > 0,$$

so that (b) holds. □

Instead of using strong separation and the closedness of finitely generated convex sets to prove the Farkas lemma, one can do it with an induction on the dimension of the column space of A; for such a proof, see [11, Theorem 2.6]. The inductive proof does not provide the geometric perspective that the proof by separation gives.

One can prove a much more general version of the Farkas lemma, replacing $\mathbb{R}^n_+$ by a general convex cone K, if and only if $A(K)$ is closed. We will do this in Section 3.2, after establishing some results needed for the proof.

Another well known theorem of the alternative, slightly older than the Farkas lemma, is due to Gordan. In its statement we call an element of $\mathbb{R}^n$ *semipositive* if it belongs to $\mathbb{R}^n_+ \setminus \{0\}$. We also use the notation $x > 0$ for a *positive* vector in $\mathbb{R}^n$: that is, an element of $\operatorname{int} \mathbb{R}^n_+$.

Proposition 3.1.11 (Gordan, 1873). *Let A be an $m \times n$ matrix. Exactly one of the following is true:*

a. There is an $x \in \mathbb{R}^n$ with $Ax > 0$.
b. There is a semipositive $u^* \in \mathbb{R}^m$ with $A^* u^* = 0$.

Proof. If (a) and (b) were both true, then we would have

$$0 = \langle A^* u^*, x \rangle = \langle u^*, Ax \rangle > 0,$$

so that these alternatives cannot both hold. We complete the proof by showing that if (b) does not hold, then (a) does.

Any u^* appearing in (b) is semipositive, so we can scale it to make its components sum to 1 without affecting the truth or falsity of (b). Then saying that (b) does not hold is the same as saying that the origin does not belong to the convex hull C of the rows of A. We do not even need Proposition 1.1.22 to show that C is closed: as C is the convex hull of finitely many points, a simple compactness argument suffices. As

we are supposing that C does not contain the origin, we can project the origin onto C to obtain a point c_0 of C. By Lemma 2.2.1 we then have for each $c' \in C$,

$$\langle 0 - c_0, c' - c_0 \rangle \leq 0,$$

or equivalently,

$$\langle c_0, c' \rangle \geq \|c_0\|^2 > 0. \tag{3.6}$$

By setting $x = c_0$ and taking each row of A to be c' in (3.6), we obtain $Ax > 0$. □

The proofs of Propositions 3.1.10 and 3.1.11 are very similar; the only significant difference is that in the Gordan theorem we are dealing with projection onto the convex hull of a finite set, so we have less difficulty in determining that it is closed. As in the case of the Farkas lemma one can formulate a much more general version of the Gordan theorem. That version appears in Section 3.2.

3.1.4 Exercises for Section 3.1

Exercise 3.1.12. If C is a convex subset of $\mathbb{R}^n$, then $\operatorname{cone} C$ is convex.

Exercise 3.1.13. If K is a nonempty cone in $\mathbb{R}^n$, then

a. For any positive μ, $\mu K = K$;
b. $\operatorname{conv} K$ is a cone;
c. K° is a closed convex cone;
d. K is convex if and only if $K + K = K$.

Exercise 3.1.14. If K is a convex cone in $\mathbb{R}^n$ then so is $K \cup \{0\}$; one has $K \subset K \cup \{0\} \subset \operatorname{cl} K$, and all three are convex cones.

Exercise 3.1.15. For any set S in $\mathbb{R}^n$, $\operatorname{cone}\operatorname{conv} S = \operatorname{conv}\operatorname{cone} S$.

Exercise 3.1.16. If C and D are nonempty cones in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, then $(C \times D)^\circ = C^\circ \times D^\circ$.

Exercise 3.1.17. If L is a subspace of $\mathbb{R}^n$ then $L^\circ = L^\perp$.

Exercise 3.1.18. Consider the set

$$C = \{ x \in \mathbb{R}^3 \mid x_1 \geq 0,\ x_3 \geq 0,\ 2x_1x_3 \geq x_2^2 \}.$$

Show that C is the set of points in $\mathbb{R}^3$ making an angle not greater than $\pi/4$ with the halfline from the origin in the direction $(1,0,1)$. Using this result, show that C is a closed convex cone. Show further that $C^\circ = -C$. This cone is often useful for counterexamples.

Exercise 3.1.19. If S and S' are subsets of $\mathbb{R}^n$ then $\operatorname{cone}(S \cap S') \subset (\operatorname{cone} S) \cap (\operatorname{cone} S')$. If either (1) at least one of S and S' is a cone, or (2) each of S and S' is convex and contains the origin, then the reverse inclusion holds so that $\operatorname{cone}(S \cap S') = (\operatorname{cone} S) \cap (\operatorname{cone} S')$.

Exercise 3.1.20. If S is any subset of $\mathbb{R}^n$, then $\operatorname{cl}\operatorname{cone} S \supset \operatorname{cone}\operatorname{cl} S$. If S is a nonempty bounded subset of $\mathbb{R}^n$ whose closure does not contain the origin, then $\operatorname{cl}\operatorname{cone} S = (\operatorname{cone}\operatorname{cl} S) \cup \{0\}$.

Exercise 3.1.21. Exhibit an example of a closed convex subset S of $\mathbb{R}^2$ for which $\operatorname{cl}\operatorname{cone} S$ properly contains $\{0\} \cup \operatorname{cone}\operatorname{cl} S$.

3.2 Cones and linear transformations

It is often necessary to take forward or inverse images of cones under linear transformations. These images remain cones (Exercise 3.2.10); moreover, if the original cone was convex then so are the images. The first proposition below shows what happens if we take the polar of a forward image.

Proposition 3.2.1. *Let T be a nonempty cone in $\mathbb{R}^n$, and let L be a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then $L(T)^\circ = (L^*)^{-1}(T^\circ)$.*

Proof. An element y^* of $\mathbb{R}^m$ belongs to $L(T)^\circ$ if and only if, for each $t \in T$,

$$0 \geq \langle y^*, Lt \rangle = \langle L^* y^*, t \rangle.$$

This is equivalent to saying that $L^* y^* \in T^\circ$ and also to $y^* \in (L^*)^{-1}(T^\circ)$. □

Corollary 3.2.2. *Let $S_1, \ldots, S_k$ be nonempty cones in $\mathbb{R}^n$. Then*

$$(\sum_{i=1}^{k} S_i)^\circ = \bigcap_{i=1}^{k} S_i^\circ.$$

Proof. Let L be the linear transformation from $\mathbb{R}^{kn}$ to $\mathbb{R}^n$ given by $L(x_1, \ldots, x_k) = x_1 + \ldots, x_k$. Apply Proposition 3.2.1 to L and the set $U = \Pi_{i=1}^k S_i$, using Exercise 3.1.16. □

In proving the following technical result we use the fact that if C is a convex set whose closure is affine, then in fact C is closed and therefore it is affine (Exercise 1.2.20).

Proposition 3.2.3. *Let M and K be respectively a subspace and a convex cone in $\mathbb{R}^n$, both nonempty. The following are equivalent:*

a. M meets $\operatorname{ri} K$.
b. $M + K$ is a subspace.

c. $M^\perp \cap K^\circ$ is a subspace.

Proof. Let L be a nonempty convex cone; then if L is a subspace it is relatively open, so the origin belongs to its relative interior. On the other hand, if the origin belongs to $\operatorname{ri} L$ then L must contain a neighborhood of the origin in $\operatorname{aff} L$. As $\operatorname{aff} L$ is a subspace and L is a cone, L must coincide with $\operatorname{aff} L$. Therefore the origin belongs to $\operatorname{ri} L$ if and only if L is a subspace.

Now take $L = M + K$ and note that to say M meets $\operatorname{ri} K$ is the same as saying that the origin belongs to $(\operatorname{ri} K) - M$. But as M is a subspace (hence equal to $-M$) we have

$$(\operatorname{ri} K) - M = (\operatorname{ri} K) + M = (\operatorname{ri} K) + (\operatorname{ri} M) = \operatorname{ri}(K + M),$$

where we used Corollary 1.2.8. We conclude that $\operatorname{ri} K$ meets M if and only if the origin belongs to $\operatorname{ri}(K + M)$, and that in turn happens if and only if $K + M$ is a subspace. Therefore (a) and (b) are equivalent.

Corollary 3.2.2 says that$(M + K)^\circ = M^\perp \cap K^\circ$. Therefore (b) implies (c). On the other hand, if (c) holds then we have

$$\operatorname{cl}(M + K) = (M + K)^{\circ\circ} = [M^\perp \cap K^\circ]^\circ,$$

where the last equality uses Corollary 3.2.2. It follows that $\operatorname{cl}(M + K)$ is a subspace. Our observation just before the statement of this proposition then shows that $M + K$ is a subspace, so that (c) implies (b). □

Unfortunately, the linear image of a closed convex cone need not be closed, but Proposition 3.2.3 can help in finding conditions for it to be closed. Suppose $K \subset \mathbb{R}^n$ is such a cone and $L : \mathbb{R}^n \to \mathbb{R}^m$ is linear. Let $y \in \operatorname{cl} L(K)$. Then $L(K)$ will fail to be closed if for the inverse images

$$N(\varepsilon) := \{x \in K \mid Lx \in B(y, \varepsilon)\} = K \cap L^{-1}[B(y, \varepsilon)],$$

the distance from the origin to $N(\varepsilon)$ converges to $+\infty$ as ε converges to zero.

The cone C of Example 3.1.18 provides an example of this behavior. If we define the linear map $L : \mathbb{R}^3 \to \mathbb{R}^2$ by $L(x_1, x_2, x_3) = (x_1, x_2)$, then $L(C) = [(\operatorname{int} \mathbb{R}) \times \mathbb{R}] \cup \{(0,0)\}$. The point $(0,1)$ belongs to $\operatorname{cl} L(C)$ but not to $L(C)$. If we let $\delta > 0$ and ε be small numbers, then the point $(\delta, 1 + \varepsilon)$ belongs to $L(C)$ and is near $(0,1)$. The inverse image of this point is

$$\{(\delta, 1 + \varepsilon, x_3) \mid x_3 \geq 0,\ 2\delta x_3 \geq (1 + \varepsilon)^2\}.$$

Therefore x_3 must satisfy

$$x_3 \geq (1/2)\delta^{-1}(1 + \varepsilon)^2, \tag{3.7}$$

so as δ and ε become small, x_3 increases without bound and so then does the distance from the origin to $(\delta, 1 + \varepsilon, x_3)$.

The next result uses Proposition 3.2.3 to develop a condition preventing this bad behavior.

Proposition 3.2.4. *Let K be a closed convex cone in $\mathbb{R}^n$ and let L be a linear transformation from $\mathbb{R}^n \to \mathbb{R}^m$. If $\operatorname{im} L^*$ meets $\operatorname{ri} K^\circ$ then there is a nonnegative real number ρ such that for each $z \in L(K)$ there is some $x \in K$ with $Lx = z$ and $\|x\| \le \rho\|z\|$.*

Proof. Let $V = (\ker L) \cap K$. Under our hypotheses, Proposition 3.2.3 says that V is a subspace. Thus if $v \in V$ then $-v \in V$ too, so that in particular $-v \in K$.

We will first show that $L(K) = L(K \cap V^\perp)$, for which we only need to show that the inclusion $\subset$ holds. Let $z \in L(K)$ and find $x \in K$ with $Lx = z$. As $\mathbb{R}^n = V \oplus V^\perp$, we can write x uniquely as $x = x_V + x_N$, with $x_V \in V$ and $x_N \in V^\perp$. Then $x_N = x - x_V$, and as we saw that $-x_V \in K$, x_N is the sum of two elements of K and therefore it is in K. As $x_V \in V$ it belongs in particular to $\ker L$, so that $L(x_N) = L(x - x_V) = L(x) = z$. But x_N belongs to $K \cap V^\perp$, so $z \in L(K \cap V^\perp)$ and therefore $L(K) = L(K \cap V^\perp)$.

If $K \cap V^\perp = \{0\}$ we are finished, because then by what we have just shown $L(K)$ must consist only of the origin, so we may choose $x = 0$ and let $\rho = 0$. Therefore in the remainder of the proof suppose that $K \cap V^\perp \neq \{0\}$.

Let γ be the infimum of $\|Lx\|$ for points x belonging to the set $Q = K \cap V^\perp \cap S^{n-1}$, where S^{n-1} is the unit sphere in $\mathbb{R}^n$. As Q is nonempty and compact, this infimum must be attained at some x_0. If $Lx_0 = 0$ then $x_0 \in \ker L$, so that

$$x_0 \in (\ker L) \cap Q = V \cap V^\perp \cap S^{n-1} = \emptyset,$$

a contradiction; therefore $\gamma > 0$. Let $\rho = \gamma^{-1}$.

Now choose any $z \in L(K)$. If $z = 0$ then choose $x = 0 \in K$; then $\|x\| \le \rho\|z\|$. If $z \neq 0$ then as we have shown that $L(K) = L(K \cap V^\perp)$, there is some $x \in K \cap V^\perp$ with $Lx = z$. As x cannot be zero, we have $L(x/\|x\|) = z/\|x\|$, and $x/\|x\| \in Q$. It follows that $\|(z/\|x\|)\| \ge \rho^{-1}$, so that $\|x\| \le \rho\|z\|$ as required. □

In the hypothesis of Proposition 3.2.4 we could just as well have assumed either of the equivalent properties given in Proposition 3.2.3. Also, Proposition 3.2.4 does not by any means assert that its bound applies to *all* inverse images of z—it generally does not—but only that it applies to at least one.

Corollary 3.2.5. *Under the hypotheses of Proposition 3.2.4, $L(K)$ is closed.*

Proof. Suppose $\{z_j\}$ is a sequence of elements of $L(K)$ that converges to z. By choice of the z_j, for each j there is an element $x_j \in K$ with $Lx_j = z_j$. Use Proposition 3.2.4 to show that for some ρ we can select the x_j so that $\|x_j\| \le \rho\|z_j\|$. As the $\|z_j\|$ are bounded, the sequence $\{x_j\}$ is then also bounded; therefore it has a cluster point $x \in K$. Continuity then implies $Lx = z$ so that $z \in L(K)$, showing that $L(K)$ is closed.

We will obtain in the next chapter a criterion for the linear image of a closed convex set (not necessarily a cone) to be closed.

Proposition 3.2.6. *Let T be a nonempty closed convex cone in $\mathbb{R}^m$ and let L be a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then*

$$[L^{-1}(T)]^\circ = \operatorname{cl} L^*(T^\circ). \tag{3.8}$$

If also $\operatorname{im} L$ *meets* $\operatorname{ri} T$ *then* $L^*(T^\circ)$ *is closed, so that*

$$[L^{-1}(T)]^\circ = L^*(T^\circ). \tag{3.9}$$

Proof. For (3.8), we apply Proposition 3.2.1 to obtain

$$\operatorname{cl} L^*(T^\circ) = [L^*(T^\circ)]^{\circ\circ} = [L^{-1}(T^{\circ\circ})]^\circ = [L^{-1}(T)]^\circ. \tag{3.10}$$

Now assume in addition that $\operatorname{im} L$ meets $\operatorname{ri} T$, and apply Corollary 3.2.5 to T° and L^* to show that $L^*(T^\circ)$ is closed. Use this in (3.10) to obtain (3.9). □

Corollary 3.2.7. *Let* $S_1, \ldots, S_k$ *be nonempty closed convex cones in* $\mathbb{R}^n$. *Then*

$$(\bigcap_{i=1}^{k} S_i)^\circ = \operatorname{cl} \sum_{i=1}^{k} S_i^\circ. \tag{3.11}$$

If also the relative interiors of $S_1, \ldots, S_k$ *have a point in common, then* $\sum_{i=1}^{k} S_i^\circ$ *is closed, so that*

$$(\bigcap_{i=1}^{k} S_i)^\circ = \sum_{i=1}^{k} S_i^\circ.$$

Proof. Use Proposition 3.2.6, with the same technique used in the proof of Corollary 3.2.2. □

Here is the generalized version of the Farkas lemma mentioned in Section 3.1.

Proposition 3.2.8. *Let* A *be an* $m \times n$ *matrix and let* K *be a nonempty convex cone in* $\mathbb{R}^n$. *If* $A(K)$ *is closed, exactly one of the following is true:*

a. There is an $x \in K$ *such that* $Ax = a$.
b. There is a $u^* \in \mathbb{R}^m$ *such that* $A^*u^* \in K^\circ$ *and* $\langle u^*, a\rangle > 0$.

If $A(K)$ *is not closed, then there is an* $a \in \mathbb{R}^m$ *for which neither alternative (a) nor alternative (b) holds.*

Proof. Suppose $A(K)$ is closed; then alternative (b) says that

$$a \notin [(A^*)^{-1}(K^\circ)]^\circ = ([A(K)]^\circ)^\circ = \operatorname{cl} A(K) = A(K),$$

where we used Proposition 3.2.6. As alternative (a) says that $a \in A(K)$, exactly one of these alternatives holds.

If $A(K)$ is not closed take a to be a point of $[\operatorname{cl} A(K)] \backslash A(K)$. Then there is no $x \in K$ with $Ax = a$, but as $a \in \operatorname{cl} A(K)$ there is a (nonconvergent) sequence $\{x_i \in K\}$ with Ax_i converging to a. Then if $A^*u^* \in K^\circ$, we have for each i

$$\langle u^*, Ax_i\rangle = \langle A^*u^*, x_i\rangle \leq 0,$$

and as $\langle u^*, \cdot\,\rangle$ is continuous, this implies $\langle u^*, a\rangle \leq 0$. Therefore alternatives (a) and (b) both fail to hold for this a. □

As mentioned earlier, the Gordan theorem has a similar generalization. The condition $u^* \in K^\circ \backslash (\operatorname{lin} K^\circ)$ is the appropriate generalization of a semipositivity condition on u^*.

Proposition 3.2.9. *Let A be an $m \times n$ matrix and let K be a nonempty convex cone in $\mathbb{R}^m$. Exactly one of the following is true:*

a. There is an $x \in \mathbb{R}^n$ such that $Ax \in \operatorname{ri} K$.
b. There is a $u^* \in K^\circ \backslash (\operatorname{lin} K^\circ)$ such that $A^* u^* = 0$.

Proof. We have

$$(\operatorname{par} K)^\perp = [\operatorname{par}(\operatorname{cl} K)]^\perp = \operatorname{lin}[(\operatorname{cl} K)^\circ] = \operatorname{lin} K^\circ,$$

where the second equality comes from Proposition 3.1.2 and the third from Corollary 3.1.8. If both (a) and (b) were true, then the u^* from (b) would belong to K° but not to $(\operatorname{par} K)^\perp$, so we could find $v \in \operatorname{par} K$ with $\langle u^*, v \rangle > 0$. If we chose v to be small enough, then as $Ax \in \operatorname{ri} K$ we would have $Ax + v \in K$. Then

$$0 \geq \langle u^*, Ax + v \rangle = \langle A^* u^*, x \rangle + \langle u^*, v \rangle > 0,$$

and this contradiction shows that (a) and (b) cannot both be true.

Now suppose (a) does not hold; we prove (b). To say (a) is false is to say that $\operatorname{im} A$ does not meet $\operatorname{ri} K$, so we can separate $\operatorname{im} A$ and K properly by a hyperplane. From Lemma 3.1.6 we see that with no loss of generality we can suppose that the hyperplane passes through the origin and that its normal u^* belongs to K°. As the subspace $\operatorname{im} A$ lies in the upper closed halfspace $H_+(u^*, 0)$ it must in fact lie in the hyperplane $H_0(u^*, 0)$. This implies that $u^* \in (\operatorname{im} A)^\perp = \ker A^*$, so that $A^* u^* = 0$ and $u^* \in K^\circ$. But as the separation is proper and as $\operatorname{im} A \subset H_0(u^*, 0)$ we must have $K \not\subset H_0(u^*, 0)$. Therefore there is some $k \in K$ with $\langle u^*, k \rangle < 0$, so that u^* cannot be in the lineality space of K°. □

We stated Proposition 3.2.9 for an arbitrary nonempty convex cone, which not only may not be polyhedral but may not even be closed, and we were able to prove it using only standard separation tools. This is a significant difference from the situation with the Farkas lemma, which required that the image $A(K)$ be closed.

3.2.1 Exercises for Section 3.2

Exercise 3.2.10. Show that if L is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ and S and U are subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, then $\operatorname{cone} L(S) = L(\operatorname{cone} S)$ and $\operatorname{cone} L^{-1}(U) = L^{-1}(\operatorname{cone} U)$.

Exercise 3.2.11. Show that the closure operation in Corollary 3.2.7 is necessary by exhibiting two nonempty closed convex cones whose sum is not closed.

3.3 Tangent and normal cones

Two important cones associated with convex sets are the tangent and normal cones. Here are their definitions.

Definition 3.3.1. Let C be a convex set in $\mathbb{R}^n$, and let $x \in C$. A point $x^* \in \mathbb{R}^n$ is *normal* to C at x if for each $c \in C$, $\langle x^*, c-x\rangle \leq 0$. The *normal cone* to C at a point $x \in \mathbb{R}^n$ is the set of all x^* normal to C at x, if $x \in C$, and is empty otherwise. The *tangent cone* to C at x is $T_C(x) := N_C(x)^\circ$, if $x \in C$, and is empty otherwise. □

That the normal cone is indeed a cone follows at once from the definition of a cone; further, it is easy to prove that this cone is closed and convex. The tangent cone is also closed and convex because it is a polar cone. We then have

$$N_C(x) = N_C(x)^{\circ\circ} = T_C(x)^\circ. \tag{3.12}$$

The origin belongs to $N_C(x)$ whenever $x \in C$, but in interesting cases $N_C(x)$ contains other points too. It is easy to show from the definition that $N_C(x)$ is always a closed convex cone, and that it is invariant under translation: that is, for any element $y \in \mathbb{R}^n$, $N_{C+y}(x+y) = N_C(x)$. Another useful property of N_C is given in the following proposition. Here, as elsewhere, we denote the unit ball of $\mathbb{R}^n$ by B^n.

Proposition 3.3.2. *Let C be a convex subset of $\mathbb{R}^n$ and let $x \in C$. The normal cone $N_C(x)$ is a local object in the sense that if Q is any convex neighborhood of x then $N_C(x) = N_{C\cap Q}(x)$ and $T_C(x) = T_{C\cap Q}(x)$.*

Proof. Write $D := C \cap Q$. D is a nonempty convex subset of C, so the definition of normal cone shows that $N_C(x) \subset N_D(x)$. Suppose $x^* \in N_D(x)$ and let $c \in C$. Let μ be a positive number less than 1 and small enough so that $x + \mu(c-x) \in Q$. Then $x + \mu(c-x)$ also belongs to C by convexity, so it belongs to D. It follows that $0 \geq \langle x^*, [x+\mu(c-x)]-x\rangle = \mu\langle x^*, c-x\rangle$. As c was an arbitrary point of C and $\mu > 0$, we have $x^* \in N_C(x)$ and therefore $N_C(x) \supset N_D(x)$. Therefore $N_C(x) = N_{C\cap Q}(x)$, and as the tangent cone is the polar of the normal cone the second claim follows. □

Proposition 3.3.2 shows that if two convex sets U and V contain a common point x and if they "agree near x" in the sense that for some neighborhood N of x one has $U \cap N = V \cap N$, then the tangent and normal cones of U and V at x are the same. This is so because we can choose a convex neighborhood Q of x with $Q \subset N$; then $U \cap Q = V \cap Q$ and Proposition 3.3.2 says that the tangent and normal cones of U and of V are the same. This is often useful in allowing us to be sure that particular characteristics of the sets don't affect the normal cone.

Operators like $N_C(x)T_C(x)$ and $T_C(x)$ look like functions, but if they were functions then their values at a point x would have to be other points. That is not the case: rather, values of these operators are sets of points. An operator taking points to sets is called a *multifunction*, and as we will encounter these not only here but also in other parts of the book it will be helpful to have a description of multifunctions and some of their properties.

3.3.1 Multifunctions

Definition 3.3.3. A *multifunction* F from $\mathbb{R}^n$ to $\mathbb{R}^m$ is a correspondence assigning to each point $x \in \mathbb{R}^n$ a set $F(x) \subset \mathbb{R}^m$. The *graph* of F, which we write as $\operatorname{gph} F$ or more often just as F, is the set of ordered pairs (x,y) in $\mathbb{R}^n \times \mathbb{R}^m$ such that $y \in F(x)$. We say that F is *closed* if its graph is closed. The *effective domain* of F, written $\operatorname{dom} F$, is the set of x such that $F(x)$ is not empty, and the *image* of F, written $\operatorname{im} F$, is the set of y for which there exists an x with $y \in F(x)$. □

The sets $\operatorname{dom} F$ and $\operatorname{im} F$ are just the projections of the graph of F into $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively.

Definition 3.3.4. If F is a multifunction from $\mathbb{R}^m$ to $\mathbb{R}^n$, then the *inverse* of F is the multifunction F^{-1} whose graph is the set of all ordered pairs $(y,x) \in \mathbb{R}^n \times \mathbb{R}^m$ such that (x,y) belongs to the graph of F. □

We have $(x,y) \in F$ if and only if $(y,x) \in F^{-1}$. Therefore $\operatorname{dom} F^{-1} = \operatorname{im} F$ and $\operatorname{im} F^{-1} = \operatorname{dom} F$. Further, as we defined the property of closedness in terms of the graph, it holds for F if and only if it holds for F^{-1}.

3.3.2 Further properties of tangent and normal cones

If C is a closed convex subset of $\mathbb{R}^n$ then the normal-cone operator $N_C(x)$ is a multifunction from $\mathbb{R}^n$ to $\mathbb{R}^n$ that is always closed, and this is often useful in proofs. On the other hand, $T_C(x)$ is in general not closed.

Proposition 3.3.5. *If K is a nonempty closed convex cone in $\mathbb{R}^n$, then $N_K^{-1} = N_{K^\circ}$.*

Proof. By Exercise 3.3.13 (x,x^*) belongs to the graph of N_K if and only if $x \in K$, $x^* \in K^\circ$, and $\langle x^*, x\rangle = 0$. Apply the exercise to K°, using the fact that $K^{\circ\circ} = K$, to conclude that these properties hold if and only if (x^*,x) belongs to the graph of N_{K°. □

The following proposition develops two additional expressions for $T_C(x)$, one or the other of which will frequently be more useful than the definition.

Proposition 3.3.6. *Let C be a nonempty convex subset of $\mathbb{R}^n$, and let x be a point of C. Then*

$$\operatorname{cl}\operatorname{cone}(C-x) = T_C(x) = \{z \mid \text{for } \tau > 0, d_C(x+\tau z) = o(\tau)\}, \tag{3.13}$$

and

$$\operatorname{ri} T_C(x) = \operatorname{cone}[(\operatorname{ri} C) - x]. \tag{3.14}$$

Proof. $N_C(x)$ is the set of all y^* that make a nonpositive inner product with $c-x$ whenever $c \in C$. Hence $N_C(x) = [\operatorname{cone}(C-x)]^\circ$, so by using Corollary 3.1.9 and the definition of $T_C(x)$ we find that

$$T_C(x) = N_C(x)^\circ = [\operatorname{cone}(C-x)]^{\circ\circ} = \operatorname{cl}\operatorname{cone}(C-x).$$

The convex-hull operator and the union with the origin were not necessary because C contains x and the convexity of C implies that of $\operatorname{cone}(C-x)$. This proves the first equality.

For the second equality, suppose first that $z \in T_C(x)$; we have just shown that then $z \in \operatorname{cl}\operatorname{cone}(C-x)$, so for any positive ε there exist $c \in C$ and $\mu > 0$ such that $\|z - \mu^{-1}(c-x)\| < \varepsilon$. Rewriting this as $\|c-(x+\mu z)\| < \mu\varepsilon$, we see that $d_C(x+\mu z) < \mu\varepsilon$. Now let $\tau \in (0,\mu)$ and define $\lambda = \tau/\mu \in (0,1)$. As $d_C(x) = 0$ by choice of x, we have

$$\begin{aligned}\tau^{-1} d_C(x+\tau z) &= \tau^{-1} d_C[(1-\lambda)x + \lambda(x+\mu z)] \\ &\le \tau^{-1}[(1-\lambda)d_C(x) + \lambda d_C(x+\mu z)] \\ &= \tau^{-1}\lambda d_C(x+\mu z) \\ &= \mu^{-1} d_C(x+\mu z) < \varepsilon,\end{aligned}$$

which shows that $d_C(x+\tau z) = o(\tau)$, so that $T_C(x)$ is contained in the set on the right in (3.13).

Next, suppose z belongs to that set, and find positive numbers τ_n converging to zero and points $c_n \in C$ such that $c_n = x + \tau_n z + r_n$ and $\|\tau_n^{-1} r_n\|$ converges to zero. Then $z = \tau_n^{-1}(c_n - x) - \tau_n^{-1} r_n$. The first term on the right belongs to $\operatorname{cone}(C-x)$ and the second converges to zero, so $z \in \operatorname{cl}\operatorname{cone}(C-x)$ and therefore, by what we have already proved, $z \in T_C(x)$.

For (3.14), use what we have already proved to write

$$\operatorname{ri} T_C(x) = \operatorname{ri}\operatorname{cl}\operatorname{cone}(C-x) = \operatorname{ri}\operatorname{cone}(C-x) = \operatorname{cone}\operatorname{ri}(C-x) = \operatorname{cone}[(\operatorname{ri} C) - x],$$

where we used Theorem 3.1.7 as well as Proposition 1.2.6 and Corollary 1.2.8. □

Proper separation can also give us some insight into the structure of the normal cone, as the next theorem shows.

Theorem 3.3.7. *Let C be a convex subset of $\mathbb{R}^n$, and let x be a point of C. Then:*

a. $(\operatorname{par} C)^\perp = N_C(x) \cap [-N_C(x)]$;
b. $N_C(x) = (\operatorname{par} C)^\perp$ *if and only if* $x \in \operatorname{ri} C$;
c. $N_C(x) = \{0\}$ *if and only if* $x \in \operatorname{int} C$.

In particular, $N_C(x)$ is a subspace if and only if $x \in \operatorname{ri} C$.

Proof. Write L for $\operatorname{par} C$. If $x^* \in L^\perp$ then as $c - x \in L$ for each $c \in C$, we have

$$\langle x^*, c-x\rangle = 0. \tag{3.15}$$

Therefore both x^* and $-x^*$ belong to $N_C(x)$, so $L^\perp \subset N_C(x) \cap [-N_C(x)]$. Conversely, if $x^* \in N_C(x) \cap [-N_C(x)]$ then (3.15) holds for each $c \in C$, so x^* is orthogonal to L, which is $\operatorname{aff}(C-x)$. Therefore $N_C(x) \cap -N_C(x) \subset L^\perp$, which proves (a).

For (b), suppose first that $x \in \operatorname{ri} C$ and let $x^* \in N_C(x)$. Write $x^* = u^* + v^*$ with $u^* \in L$ and $v^* \in L^\perp$. For small positive ε we have $x + \varepsilon u^* \in C$, and therefore

$$0 \geq \langle x^*, (x+\varepsilon u^*) - x \rangle = \langle u^* + v^*, \varepsilon u^* \rangle = \varepsilon \|u^*\|^2,$$

implying that $u^* = 0$ and therefore $x^* \in L^\perp$. This shows that $N_C(x) \subset L^\perp$, but as $L^\perp \subset N_C(x)$ by part (a) we actually have $N_C(x) = L^\perp$.

Next suppose $x \notin \operatorname{ri} C$: then by Theorem 2.2.11 we can separate $\{x\}$ properly from C. This means there is a y^* such that for each $c \in C$, $\langle y^*, c \rangle \leq \langle y^*, x \rangle$ (so that $\langle y^*, c-x \rangle \leq 0$ and therefore $y^* \in N_C(x)$), but for some $\hat{c} \in C$, $\langle y^*, \hat{c} \rangle < \langle y^*, x \rangle$. Then $\langle -y^*, \hat{c} - x \rangle > 0$, so $-y^* \notin L^\perp$. Therefore $N_C(x) \neq L^\perp$, which proves (b).

To prove (c), note that if $x \in \operatorname{int} C$ then by (b) we have

$$N_C(x) = L^\perp = (\mathbb{R}^n)^\perp = \{0\}.$$

On the other hand, if $x \notin \operatorname{int} C$ we have two possibilities: first, if $\dim C < n$ then by (a) $N_C(x)$ must contain the nonzero subspace $L^\perp$; second, if $\dim C = n$ then $x \notin \operatorname{ri} C$, so by (b) $N_C(x)$ contains an element that is not in $L^\perp = \{0\}$. In either case $N_C(x)$ cannot be the origin.

For the final claim, if $x \in \operatorname{ri} C$ then (b) shows that $N_C(x)$ is the subspace $(\operatorname{par} C)^\perp$. On the other hand, if $N_C(x)$ is a subspace then $N_C(x) = -N_C(x) = N_C(x) \cap [-N_C(x)]$. In that case (a) shows that $N_C(x) = (\operatorname{par} C)^\perp$, and then from (b) we have $x \in \operatorname{ri} C$. □

The next theorem has a great many applications in optimization, because constraints in optimization problems often involve the intersection of sets.

Theorem 3.3.8. *Let $C_1, \ldots, C_k$ be nonempty convex subsets of $\mathbb{R}^n$, let C be their intersection and suppose $x \in C$. If the sets $\operatorname{ri} C_i$ for $i = 1, \ldots, k$ have a point in common, then*

$$T_C(x) = \bigcap_{i=1}^{k} T_{C_i}(x) \tag{3.16}$$

and

$$N_C(x) = \sum_{i=1}^{k} N_{C_i}(x), \tag{3.17}$$

the sum on the right then being closed.

Proof. For the tangent cone, we have

$$\begin{aligned} T_C(x) &= \mathrm{cl}\,\mathrm{cone}(C-x) \\ &= \mathrm{cl}\,\mathrm{cone}[(\cap_{i=1}^k C_i)-x] \\ &= \mathrm{cl}\,\mathrm{cone}[\cap_{i=1}^k (C_i-x)] \\ &= \mathrm{cl}\cap_{i=1}^k [\mathrm{cone}(C_i-x)] \\ &= \cap_{i=1}^k \mathrm{cl}\,\mathrm{cone}(C_i-x) \\ &= \cap_{i=1}^k T_{C_i}(x), \end{aligned}$$

where the fourth equality uses Exercise 3.1.19 and the fifth uses Proposition 1.2.12 and the fact that $\mathrm{ri}(C_i-x) = \mathrm{ri}\,C_i - \mathrm{ri}\{x\} = (\mathrm{ri}\,C_i)-x$, so that the relative interiors of the sets C_i-x have a point in common.

The form of $\mathrm{ri}\,T_C(x)$ shown in Proposition 3.3.6, together with the hypothesis, shows that the cones $\mathrm{ri}\,T_{C_i}(x)$ for $i=1,\ldots,k$ have a point in common. Using (3.12) and Corollary 3.2.7, we then find that

$$N_C(x) = T_C(x)^\circ = [\bigcap_{i=1}^k T_{C_i}(x)]^\circ = \sum_{i=1}^k T_{C_i}(x)^\circ = \sum_{i=1}^k N_{C_i}(x),$$

and that the sum on the right is closed. □

The formulas in (3.16) and (3.17) do not hold in general without the regularity condition, even if the sets involved are all closed: for an example, take $C_1 = \{(x_1,x_2) \mid x_2 \geq (1/2)x_1^2\}$ and $C_2 = -C_1$. They do, however, hold without that condition if all of the sets involved are polyhedral, as we will see later.

The normal cone is often used to express optimality conditions, as in the following proposition. We say a point x in a subset S of $\mathbb{R}^n$ is a *local minimizer* of a real-valued function f defined on some neighborhood N of x in S if there is a neighborhood $N' \subset N$ of x in S such that for each $x' \in N'$, $f(x) \leq f(x')$.

Proposition 3.3.9. *Suppose that C is a nonempty convex subset of $\mathbb{R}^n$ and f is a real-valued function on a neighborhood N of x. Assume that f has a Fréchet derivative $df(x)$ at x. If x is a local minimizer of $f|_C$, then*

$$0 \in df(x) + N_C(x). \tag{3.18}$$

Proof. Let c be any point of C and for $\tau \in (0,1)$ define $c_\tau = (1-\tau)x + \tau c$; then $c_\tau \in C$ by convexity. For all small positive values of τ the point c_τ belongs to N, and for such τ we have

$$0 \leq f(c_\tau) - f(x) = df(x)(c_\tau - x) + o(\tau) = \tau df(x)(c-x) + o(\tau).$$

Dividing by τ and letting $\tau \searrow 0$, we find that $df(x)(c-x) \geq 0$. As c was arbitrary in C, we have (3.18). □

We can combine this basic optimality condition with Theorem 3.3.8 to obtain a necessary optimality condition that is very important in practice.

Corollary 3.3.10. *Let $C_1, \ldots, C_k$ be nonempty convex subsets of $\mathbb{R}^n$ whose relative interiors have a point in common, and let $C = \cap_{i=1}^k C_i$. Let x be a point of C and f be a real-valued function on a neighborhood N of x. Assume that f has a Fréchet derivative $df(x)$ at x. If x is a local minimizer of $f|_C$, then*

$$0 \in df(x) + \sum_{i=1}^{k} N_{C_i}(x). \tag{3.19}$$

Proof. Use Proposition 3.3.9 together with Theorem 3.3.8. □

Many results about normal cones also hold for nonconvex sets, with a suitable extension of the definition of normal cone and, often, with additional restrictions. An exposition for the finite-dimensional case is in [36, Chapter 6].

3.3.3 Exercises for Section 3.3

Exercise 3.3.11. Let C be a nonempty closed convex subset of $\mathbb{R}^n$. Show that the normal-cone operator N_C is always closed. Give an example in $\mathbb{R}$ to show that the tangent-cone operator T_C may not be closed.

Exercise 3.3.12. If C is a nonempty closed convex set in $\mathbb{R}^n$ and x and y are two points of $\mathbb{R}^n$, show that the following are equivalent: (i) x is the projection of y on C; (ii) for each positive τ, x is the projection on C of $x + \tau(y - x)$; (iii) $y - x \in N_C(x)$.

Exercise 3.3.13. Let K be a nonempty convex cone in $\mathbb{R}^n$ and suppose $k \in K$. Show that $N_K(k) = \{k^* \in K^\circ \mid \langle k^*, k\rangle = 0\}$.

Exercise 3.3.14. Let C be a nonempty closed convex set in $\mathbb{R}^n$. Show that each point x of $\mathbb{R}^n$ has a unique decomposition $x = c + n$ where $c \in C$ and $n \in N_C(c)$. State the particular forms this decomposition takes when (a) C is a cone, (b) C is a subspace.

Exercise 3.3.15. Let C be a convex set in $\mathbb{R}^n$, and let S be a nonempty relatively open convex subset of C. Show that the normal-cone operator $N_C(\cdot)$ is constant on S: that is, show that for any x and y in S, $N_C(x) = N_C(y)$.

3.4 Notes and references

The Farkas lemma was apparently first published in 1896 [7], but with an incomplete proof. The first publication with a correct proof seems to have been in 1898 [8] in Hungarian. A paper of 1902 in German [9] is often cited. The names on the latter two papers appear to be different because the author's given name Gyula is used for the first, but the translation Julius for the second. The Gordan theorem appeared in [14]. Much more information on the history of the mathematics underlying optimization is in the excellent survey paper of Prékopa [30].

Chapter 4
Structure

This chapter concentrates on the structure of convex sets. Section 4.1 develops a way to describe *how* a convex set is unbounded, rather than just saying *that* it is unbounded. Then Section 4.2 shows how we can break up a convex set into pieces, called faces of the set, that have important geometric properties especially useful in the study of variational problems. Finally, Section 4.3 applies results from the earlier sections to identify a set of minimal structural elements that combine to make up a convex set.

4.1 Recession

Even for unbounded sets, convexity has important structural consequences. This section develops a usable definition of "directions in which a set is unbounded," and applies that information to find out more about the set's behavior. The key tool for this purpose is the recession cone, developed in the next section. This cone then allows us to make a very useful technical construction to homogenize a set. That construction exposes important structural properties of sets, and we will use it extensively.

4.1.1 The recession cone

Recession directions are directions in which we can move without leaving a set. Here is the definition.

Definition 4.1.1. Let S be a nonempty subset of $\mathbb{R}^n$. An element $w \in \mathbb{R}^n$ is a *recession direction* of S if $S + w \subset S$. We write $\operatorname{rc} S$ for the set of all recession directions of S. □

The definition implies that the origin is a recession direction of any nonempty set. Also, if w is a recession direction of S then for any $s \in S$ we have $s+w \in S$, which implies that S contains $(s+w)+w = s+2w$, and therefore $s+kw \in S$ for each nonnegative integer k. One can also show from the definition that the set of recession directions of a product $S \times S'$ is the product of the sets of recession directions of S and of S' respectively.

Suppose now that $s \in \mathrm{cl}\, S$ and $w \in \mathrm{cl}\,\mathrm{rc}\, S$. Then for any neighborhoods N of s and Q of w we can find points $s' \in N \cap S$ and $w' \in Q \cap \mathrm{rc}\, S$. Then $s'+w' \in S$, implying that $s+w \in \mathrm{cl}\, S$ and therefore $w \in \mathrm{rc}\,\mathrm{cl}\, S$. It follows that $\mathrm{cl}\,\mathrm{rc}\, S \subset \mathrm{rc}\,\mathrm{cl}\, S$, and if S is closed then

$$\mathrm{rc}\,\mathrm{cl}\, S \subset \mathrm{cl}\,\mathrm{rc}\,\mathrm{cl}\, S = \mathrm{cl}\,\mathrm{rc}\, S,$$

so that equality holds.

In general, the set of recession directions of S need not have any special properties such as convexity or positive homogeneity: for example, the set of integers coincides with its set of recession directions. We can produce more satisfactory behavior by assuming that S is convex.

Theorem 4.1.2. *Let C be a nonempty convex subset of $\mathbb{R}^n$. The set $\mathrm{rc}\, C$ of recession directions of C is then a nonempty convex cone containing the origin; it is closed if C is closed. Further, of the following statements (a) is equivalent to (b), (b) implies (c), and if C is closed then all three statements are equivalent.*

a. $w \in \mathrm{rc}\, C$.
b. For each $c \in C$, $c + w\mathbb{R}_+ \subset C$.
c. There exists $c \in C$ *with* $c + w\mathbb{R}_+ \subset C$.

Proof. We have already noted that the origin must be a recession direction of any nonempty set, so $\mathrm{rc}\, C$ is nonempty.

To show that (a) and (b) are equivalent, let $w \in \mathrm{rc}\, C$ and choose any $c \in C$. Then $c + kw \in C$ for each nonnegative integer k, and convexity of C implies that $c + w\mathbb{R}_+ \subset C$. On the other hand, if (b) holds then as $1 \in \mathbb{R}_+$ we also have (a). It is obvious that (b) implies (c).

The fact that $c + w\mathbb{R}_+ \subset C$ for any $c \in C$ and any recession direction w shows that $\mathrm{rc}\, C$ is a cone. It is also convex, as we can see by choosing recession directions w and w' and $\mu \in [0,1]$; then for $c \in C$ each of $c+w$ and $c+w'$ is in C, so by convexity we have

$$c + [(1-\mu)w + \mu w'] = (1-\mu)(c+w) + \mu(c+w') \in C,$$

so that $(1-\mu)w + \mu w' \in \mathrm{rc}\, C$.

Now assume C is closed. If $v \notin \mathrm{rc}\, C$ then there is some $x \in C$ with $x+v \notin C$. Then there is a neighborhood Q of v such that $x+Q$ does not meet C, so that Q does not meet $\mathrm{rc}\, C$, and this shows that $\mathrm{rc}\, C$ is closed.

Suppose that $c \in C$ and let $c + w\mathbb{R}_+ \subset C$. If $w = 0$ then $w \in \mathrm{rc}\, C$, so assume $w \neq 0$. Choose any $c' \in C$ and let $\alpha \geq 0$. For $k = 1, 2, \ldots$ let ε_k be positive numbers converging to zero and define

$$c_k = (1-\varepsilon_k)c' + \varepsilon_k(c + \alpha\varepsilon_k^{-1}w).$$

Then $c_k \in C$ because C is convex. But the c_k converge to $c' + \alpha w$, which must then be in C because C is closed. As α was any nonnegative number, we have $c' + w\mathbb{R}_+ \in C$, so that (c) is equivalent to (a) and to (b). □

An immediate consequence of Theorem 4.1.2 is that if C and C' are two nonempty closed convex sets with $C \subset C'$, then $\operatorname{rc} C \subset \operatorname{rc} C'$. If the sets are not closed, then it is possible for $\operatorname{rc} C$ to contain $\operatorname{rc} C'$ properly.

Remark 4.1.3. If C is convex and if $c \in \operatorname{ri} C$ and $w \in \operatorname{rc} C$, then the halfline $c + w\mathbb{R}_+$ also lies in $\operatorname{ri} C$. To see this, recall that we have already shown that $c + w\mathbb{R}_+ \subset C$. Choose any $\lambda > 0$ and let $\mu > \lambda$. We have $c \in \operatorname{ri} C$ and $c + \mu w \in C$, and $c + \lambda w$ lies in the relative interior of the segment $[c, c + \mu w]$. But then by Theorem 1.2.4, $c + \lambda w \in \operatorname{ri} C$. □

By combining the technique used in this remark with Proposition 1.2.6, one can show that $\operatorname{rc}\operatorname{ri} C = \operatorname{rc}\operatorname{cl} C$ (Exercise 4.1.22).

Theorem 4.1.2 has other very useful consequences. The next two corollaries illustrate some of these.

Corollary 4.1.4. *Let L be a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, and let C be a convex subset of $\mathbb{R}^m$ with $L^{-1}(C) \neq \emptyset$. Then $L^{-1}(\operatorname{rc} C) \subset \operatorname{rc} L^{-1}(C)$, and if C is closed then $L^{-1}(\operatorname{rc} C) = \operatorname{rc} L^{-1}(C)$.*

Proof. Suppose $x \in L^{-1}(\operatorname{rc} C)$; then $v := Lx \in \operatorname{rc} C$. If w is any point of $L^{-1}(C)$, then $c := Lw \in C$. Therefore the halfline $c + v\mathbb{R}_+$ is contained in C. But

$$c + v\mathbb{R}_+ = L(w + x\mathbb{R}_+),$$

so $w + x\mathbb{R}_+ \subset L^{-1}(C)$. As w was any point of $L^{-1}(C)$, we have $x \in \operatorname{rc} L^{-1}(C)$ and therefore $L^{-1}(\operatorname{rc} C) \subset \operatorname{rc} L^{-1}(C)$.

Now suppose that C is closed and let $v \in \operatorname{rc} L^{-1}(C)$; we will show that $Lv \in \operatorname{rc} C$, so that $v \in L^{-1}(\operatorname{rc} C)$ and therefore $\operatorname{rc} L^{-1}(C) \subset L^{-1}(\operatorname{rc} C)$. We assumed that $L^{-1}(C)$ was nonempty; let x belong to it, so that $y = Lx \in C$. Then $x + v\mathbb{R}_+ \subset L^{-1}(C)$, which implies that $y + (Lv)\mathbb{R}_+ \subset C$. Apply Theorem 4.1.2 to conclude that $Lv \in \operatorname{rc} C$. □

The next corollary shows the form of recession cones of intersections.

Corollary 4.1.5. *Let $\{ C_\alpha \mid \alpha \in \mathscr{A} \}$ be a nonempty collection of convex subsets of $\mathbb{R}^n$ with $\cap_{\alpha\in\mathscr{A}} C_\alpha \neq \emptyset$. Then $\cap_{\alpha\in\mathscr{A}} \operatorname{rc} C_\alpha \subset \operatorname{rc} \cap_{\alpha\in\mathscr{A}} C_\alpha$. If each C_α is closed, then*

$$\cap_{\alpha\in\mathscr{A}} \operatorname{rc} C_\alpha = \operatorname{rc} \cap_{\alpha\in\mathscr{A}} C_\alpha. \tag{4.1}$$

Proof. Let $w \in \cap_{\alpha\in\mathscr{A}} \operatorname{rc} C_\alpha$, and let x be any point of $\cap_{\alpha\in\mathscr{A}} C_\alpha$. For each α we have $x + w \in C_\alpha$, so $x + w \in \cap_{\alpha\in\mathscr{A}} C_\alpha$ and therefore $w \in \operatorname{rc} \cap_{\alpha\in\mathscr{A}} C_\alpha$.

If each C_α is closed then let $w \in \operatorname{rc} \cap_{\alpha\in\mathscr{A}} C_\alpha$. Choose some $c \in \cap_{\alpha\in\mathscr{A}} C_\alpha$; then $\cap_{\alpha\in\mathscr{A}} C_\alpha$ contains the halfline $c + w\mathbb{R}_+$. But that means that this halfline is in each of the C_α, and by applying Theorem 4.1.2 we conclude that $w \in \cap_{\alpha\in\mathscr{A}} \operatorname{rc} C_\alpha$, so that (4.1) holds. □

We can express the lineality space of a convex set in terms of its recession directions.

Proposition 4.1.6. *If C is a nonempty convex subset of $\mathbb{R}^n$, then* $\operatorname{lin} C = (\operatorname{rc} C) \cap (-\operatorname{rc} C)$.

Proof. Let $L = (\operatorname{rc} C) \cap (-\operatorname{rc} C)$; then if v and v' belong to L and γ and γ' are real numbers, then each of γv and $\gamma' v'$ belongs to both $\operatorname{rc} C$ and $-\operatorname{rc} C$. For any convex cone K we have $K = K + K$, so $\gamma v + \gamma' v' \in L$. Therefore L is a subspace of $\mathbb{R}^n$.

If $x \in L$ and $c \in C$ then $c + x \in C$, so $C + L \subset C$. As L contains the origin, we actually have $C + L = C$. To complete the proof we let K be any subspace of $\mathbb{R}^n$ with $C + K = C$, and show $K \subset L$. If $k \in K$ then $C + k \in C$, so $k \in \operatorname{rc} C$. As $-k \in K$, a similar argument shows that $k \in -\operatorname{rc} C$; therefore $k \in (\operatorname{rc} C) \cap (-\operatorname{rc} C) = L$. Accordingly, $K \subset L$, so $L = \operatorname{lin} C$. □

If C is closed and contains a line, that line must be of the form $c + v\mathbb{R}$ for some nonzero v. As then both v and $-v$ belong to $\operatorname{rc} C$ by Theorem 4.1.2, we have $v \in \operatorname{lin} C$. Conversely, if $v \in \operatorname{lin} C$ then for each $c \in C$ the line $c + v\mathbb{R}$ is contained in C. Suppose that for nonzero z we say that a line $x + z\mathbb{R}$ in $\mathbb{R}^n$ has *direction* z. Then each nonzero element of $\operatorname{lin} C$ indicates the direction of some line contained in C, and conversely the direction of every line contained in C belongs to $\operatorname{lin} C$. The latter fact need not be true if C is not closed.

Here is an alternate description of $\operatorname{rc} C$ that is valid when C is closed and convex. It shows how we can construct $\operatorname{rc} C$ from elements of C, and provides the justification for the notation 0^+C used for $\operatorname{rc} C$ in, e.g., [34].

Proposition 4.1.7. *Let C be a nonempty closed convex subset of $\mathbb{R}^n$. Then* $\operatorname{rc} C$ *is the set of all limits of sequences of the form $\{\alpha_k c_k\}$ in which $\alpha_k \searrow 0$ and $c_k \in C$.*

Proof. Let Q be the set of limits described in the proposition. If $w \in \operatorname{rc} C$ then let $c \in C$; choose $\{\alpha_k\} \downarrow 0$ and let $c_k = c + \alpha_k^{-1} w$. These c_k belong to C by hypothesis, and $\alpha_k c_k$ converges to w, so $\operatorname{rc} C \subset Q$.

If $w \in Q$ then there are elements $c_k \in C$ and $\alpha_k \searrow 0$ with $\alpha_k c_k$ converging to w. The numbers α_k are eventually in $(0, 1]$; then if c is any point of C we have $(1 - \alpha_k)c + \alpha_k c_k \in C$ by convexity. Taking the limit and recalling that C is closed, we find that $c + w \in C$, so that $w \in \operatorname{rc} C$ and therefore $Q \subset \operatorname{rc} C$. □

The following corollary provides a very useful test for boundedness.

Corollary 4.1.8. *Let C be a nonempty convex subset of $\mathbb{R}^n$. For C to be bounded it is necessary that* $\operatorname{rc} C = \{0\}$*, and if C is closed then this condition is also sufficient.*

Proof. If $c \in C$ and $w \in \operatorname{rc} C$ then Theorem 4.1.2 says that $c + w\mathbb{R}_+ \subset C$. If C is bounded then w must be zero, which proves necessity. For sufficiency, suppose that C is closed and unbounded. Let $c_k \in C$ and suppose that the sequence $\{\|c_k\|\}$ converges to $+\infty$. For large k let $w_k = c_k / \|c_k\|$; the norms of the w_k are all 1, so a subsequence converges to some w_0 with norm 1. Proposition 4.1.7 then shows that $w_0 \in \operatorname{rc} C$, so that $\operatorname{rc} C \neq \{0\}$. □

The hypothesis that C be closed is really needed here (Exercise 4.1.23).

4.1.2 The homogenizing cone

An application of recession in the convex case yields a homogenization construction, detailed in the following theorem, that provides the foundation for some key results in convex analysis. These include the fundamental conjugacy relation for convex functions.

Theorem 4.1.9. *Let Q be a convex cone contained in $\mathbb{R}^n \times \mathbb{R}_-$ but not in $\mathbb{R}^n \times \{0\}$. Define*

$$C := \{\, x \in \mathbb{R}^n \mid (x,-1) \in Q \,\}. \tag{4.2}$$

Then C is nonempty and one has

$$\operatorname{ri} Q = \operatorname{cone}[(\operatorname{ri} C) \times \{-1\}], \tag{4.3}$$

and

$$\operatorname{cl} Q = \{\operatorname{cone}[(\operatorname{cl} C) \times \{-1\}]\} \cup [(\operatorname{rc}\operatorname{cl} C) \times \{0\}]. \tag{4.4}$$

Proof. Throughout the proof we write $A(\xi)$ for the hyperplane $\mathbb{R}^n \times \{-\xi\}$. By hypothesis there is some $(x,-\xi) \in Q$ with $\xi > 0$. Then $(\xi^{-1}x,-1)$ belongs to C, which must therefore be nonempty. We have $\operatorname{ri} Q = \operatorname{ri}\operatorname{cone} Q = \operatorname{cone}\operatorname{ri} Q$, so $\operatorname{ri} Q$ is a convex cone. By Corollary 1.2.3, it cannot meet $A(0)$ because $Q \not\subset A(0)$; hence each point of $\operatorname{ri} Q$ is of the form $(x,-\xi)$ with $\xi > 0$. Therefore $\operatorname{ri} Q$ consists of all positive scalar multiples of points in $(\operatorname{ri} Q) \cap A(1)$. But by Proposition 1.2.12,

$$(\operatorname{ri} Q) \cap A(1) = \operatorname{ri}[Q \cap A(1)] = \operatorname{ri}(C \times \{-1\}) = (\operatorname{ri} C) \times \{-1\},$$

so we have the formula in (4.3).

If $\xi > 0$ then $A(\xi)$ meets $\operatorname{ri} Q$, so that

$$\begin{aligned}(\operatorname{cl} Q) \cap A(\xi) &= \operatorname{cl}[Q \cap A(\xi)] = \operatorname{cl}\{\xi[Q \cap A(1)]\} \\ &= \xi \operatorname{cl}[C \times \{-1\}] = \xi[(\operatorname{cl} C) \times \{-1\}].\end{aligned}$$

On the other hand, any point of $(\operatorname{cl} Q) \cap A(0)$ has the form $(x,0)$ and is a limit of some sequences $\{(x_k,-\xi_k)\}$ in $\operatorname{ri} Q$ in which $\xi_k \searrow 0$. We can write $x_k = \xi_k(\xi_k^{-1}x_k)$, and the definition of C implies that $\xi_k^{-1}x_k \in C$. As ξ_k converges to zero, the set of such x is contained in $\operatorname{rc}\operatorname{cl} C$ by Proposition 4.1.7. To show that each element x of $\operatorname{rc}\operatorname{cl} C$ can be obtained in this way, let $c \in \operatorname{ri} C = \operatorname{ri}\operatorname{cl} C$; then for each positive integer k, $c + kx \in \operatorname{ri}\operatorname{cl} C = \operatorname{ri} C$ (Remark 4.1.3). It follows from (4.3) that $(k^{-1}c + x, -k^{-1}) \in \operatorname{ri} Q$, and by taking the limit we find that $(x,0) \in \operatorname{cl} Q$. Therefore $(\operatorname{cl} Q) \cap A(0) = [(\operatorname{rc}\operatorname{cl} C) \times \{0\}]$, which establishes (4.4). □

In Theorem 4.1.9 we started with Q and derived C from it, because this allows us to consider a slightly more general situation. However, the usual procedure in using this construction is to start with C and then use the following definition.

Definition 4.1.10. Let S be a subset of $\mathbb{R}^n$, and define a cone $H_S \subset \mathbb{R}^{n+1}$ by

$$H_S = \operatorname{cone}(S \times \{-1\}) = \operatorname{cone}\{(s,-1) \mid s \in S\}. \tag{4.5}$$

This H_S is the *homogenizing cone* of S. □

The homogenizing cone will be useful here mainly when the set S is a nonempty convex set C. Then the homogenizing cone H_C fits the requirements of Theorem 4.1.9, and C satisfies (4.2), so the theorem applies.

There are important differences between the homogenizing cone H_C and the cone, written $\operatorname{cone} C$, that is generated by C. For one thing, they lie in different spaces, H_C in $\mathbb{R}^{n+1}$ and $\operatorname{cone} C$ in $\mathbb{R}^n$. For another, they have very different structure. In particular, $\operatorname{cone} C$ generally loses much of the information that was contained in C. For example, if $C = [1,2]$ then $\operatorname{cone} C = \operatorname{int} \mathbb{R}_+$, so one has no idea what the endpoints of the original interval were. On the other hand, one can always recover C from H_C, so the homogenization loses no information. This is one of the reasons why the homogenization construction is often a much more useful tool than is $\operatorname{cone} C$.

The intersection of H_C with the hyperplane $\mathbb{R}^n \times \{-1\}$ is the set C. If we look at the intersection of $(H_C)^\circ$ with the hyperplane $\mathbb{R}^n \times \{1\}$, a familiar set appears.

Proposition 4.1.11. *If C is a nonempty convex subset of $\mathbb{R}^n$, then $H_C^\circ \cap (\mathbb{R}^n \times \{1\}) = C^\circ \times \{1\}$.*

Proof. *(⊂)* Suppose $(x^*,1) \in H_C^\circ \cap (\mathbb{R}^n \times \{1\})$. If $x \in C$ then $(x,-1) \in H_C$ so

$$0 \geq \langle (x^*,1),(x,-1)\rangle = \langle x^*,x\rangle - 1.$$

Therefore $x^* \in C^\circ$, so $(x^*,1) \in C^\circ \times \{1\}$.

(⊃) If $x^* \in C^\circ$ and $(h,-\eta) \in H_C$ then $\eta > 0$ and $(\eta^{-1}h,-1) \in H_C$ so that $\eta^{-1}h \in C$. We have

$$0 \geq \eta[\langle x^*, \eta^{-1}h\rangle - 1] = \langle (x^*,1),(h,-\eta)\rangle,$$

and therefore $(x^*,1) \in H_C^\circ \cap (\mathbb{R}^n \times \{1\})$. □

4.1.3 When are linear images of convex sets closed?

Next we show how one can use the recession cone to derive a closedness condition for (forward) linear images of sets. The proof provides an illustration of how to use Corollary 4.1.8.

In order to state the main theorem, we need to develop three contributing results. The first of these defines a new cone closely related to the recession cone and develops of some of its properties. The second introduces a key condition involving a convex set and a linear operator, and applies Proposition 3.2.3 to derive some of its equivalent forms. The third applies that condition to extend Proposition 3.2.4 from closed convex cones to any closed convex sets.

Definition 4.1.12. Let S be a nonempty subset of $\mathbb{R}^n$. The *barrier cone* of S, written $\operatorname{bc} S$, is the set of all $x^* \in \mathbb{R}^n$ for which $\sup_{s\in S}\langle x^*, s\rangle < +\infty$. □

Direct use of the definition shows that $\operatorname{bc} S$ is a convex cone containing the origin. However, it need not be closed: if S is the set of points (x_1, x_2) in $\mathbb{R}^2$ satisfying $x_2 \geq x_1^2$, then $\operatorname{bc} S$ is the union of the origin and the open lower halfplane. By using Theorem 4.1.9 we can develop the relationship between the recession and barrier cones of a convex set.

Corollary 4.1.13. *Let C be a nonempty convex subset of $\mathbb{R}^n$. Then $(\operatorname{bc} C)^\circ = \operatorname{rc}\operatorname{cl} C$.*

Proof. Construct from C the cone H_C of Theorem 4.1.9. A point w^* belongs to $\operatorname{bc} C$ exactly when there is some finite number ϕ such that for each $c \in C$, $\langle w^*, c\rangle \leq \phi$. We can rewrite this inequality as $\langle (w^*, \phi), (c, -1)\rangle \leq 0$, and this shows that w^* belongs to $\operatorname{bc} C$ if and only if there is some finite ϕ with $(w^*, \phi) \in H_C^\circ$. If we define a linear transformation L from $\mathbb{R}^{n+1}$ to $\mathbb{R}^n$ by $L(x, \xi) = x$, then $\operatorname{bc} C = L(H_C^\circ)$. Now Proposition 3.2.1 and Corollary 3.1.9 yield $(\operatorname{bc} C)^\circ = (L^*)^{-1}(\operatorname{cl} H_C)$, which means that $(\operatorname{bc} C)^\circ$ is the set of $w \in \mathbb{R}^n$ such that $(w, 0) \in \operatorname{cl} H_C$. By Theorem 4.1.9, this set is $\operatorname{rc}\operatorname{cl} C$. □

Next, we introduce a key condition and use Proposition 3.2.3 to derive some equivalent forms in which we can express it.

Definition 4.1.14. Suppose that C is a nonempty convex set in $\mathbb{R}^n$ and L is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. We say that C satisfies the *recession condition* for L if C is closed and

$$(\ker L) \cap (\operatorname{rc} C) \subset \operatorname{lin} C. \tag{4.6}$$

□

Here are some equivalent forms of the recession condition.

Proposition 4.1.15. *Let C be a nonempty closed convex subset of $\mathbb{R}^n$, and L be a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. The following three conditions are equivalent:*

a. $(\ker L) \cap (\operatorname{rc} C) \subset \operatorname{lin} C$.
b. The intersection on the left in (a) is a subspace.
c. $\operatorname{im} L^* \cap \operatorname{ri} \operatorname{bc} C \neq \emptyset$.

Proof. Denote by K the set on the left in (a). If (a) holds and if x belongs to K, then x must belong to $-\operatorname{rc} C$ also, and therefore $-x \in K$. As K is a convex cone, it must be a subspace, so (b) holds. Now suppose (b) holds. As K is a subspace, if $x \in K$ then $-x \in K$ also, which in particular implies that $x \in -\operatorname{rc} C$ and therefore $x \in \operatorname{lin} C$, so (a) holds. Therefore (a) and (b) are equivalent; the equivalence of (b) and (c) follows from Proposition 3.2.3 together with the facts that $\operatorname{rc} C = (\operatorname{bc} C)^\circ$ and that $(\ker L)^\perp = \operatorname{im} L^*$. □

Next, we use the recession condition to extend Proposition 3.2.4 from closed convex cones to any closed convex sets.

Theorem 4.1.16. *Let L be a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, let C be a convex subset of $\mathbb{R}^n$ that satisfies the recession condition for L, and let $V := (\ker L) \cap (\operatorname{rc} C)$. Then*

a. $L(C) = L(C \cap V^\perp)$;
b. There are nonnegative real numbers α and β such that for each $y \in L(C)$ there is $x \in C \cap V^\perp$ such that

$$Lx = y, \qquad \|x\| \le \max\{\alpha, \beta\|y\|\}. \tag{4.7}$$

In particular, if $L(C)$ is bounded then we may take $\beta = 0$.

Proof. If C is empty there is nothing to prove. Otherwise, by hypothesis V is a subspace of $\operatorname{lin} C$. If we define $D := C \cap V^\perp$ then $L(D) \subset L(C)$. By Proposition A.6.23 we have $C = V + D$, so if $c \in C$ then $c = v + d$ with $v \in V$ and $d \in D$. Then $Lc = Ld$ and therefore $L(C) = L(D)$, which proves (a).

For (b), let ρ be a positive number such that $B(0,\rho)$ meets $L(D)$. In the special case in which $L(C)$ is bounded, we take ρ large enough so that $B(0,\rho) \supset L(D)$. Let $T := D \cap L^{-1}[B(0,\rho)]$. Corollary 4.1.5 shows that

$$\operatorname{rc} T = \operatorname{rc}\{C \cap V^\perp \cap L^{-1}[B(0,\rho)]\} = (\operatorname{rc} C) \cap V^\perp \cap \operatorname{rc} L^{-1}[B(0,\rho)],$$

and Corollaries 4.1.4 and 4.1.8 yield

$$\operatorname{rc} L^{-1}[B(0,\rho)] = L^{-1} \operatorname{rc} B(0,\rho) = L^{-1}(0) = \ker L,$$

so that

$$\operatorname{rc} T = V \cap V^\perp = \{0\}.$$

Then Corollary 4.1.8 shows that T is bounded, so we can choose α with $T \subset B(0,\alpha)$. Then for any $y \in L(C) \cap B(0,\rho)$ there is $x \in D$ with $Lx = y$ and $\|x\| \le \alpha$.

If $L(C) \subset B(0,\rho)$, we set $\beta = 0$ and are finished; this includes the case in which $L(C)$ is bounded. Otherwise, define a set S of real numbers by

$$S := \{\|x\|\|y\|^{-1} \mid x \in D,\ \|y\| > \rho,\ Lx = y\}.$$

The set S is nonempty because $L(C) = L(D)$ and we know $L(C) \not\subset B(0,\rho)$. If it had no upper bound then there would be sequences $\{y_k\} \subset L(C) \backslash B(0,\rho)$ and $\{x_k \in D \cap L^{-1}(y_k)\}$ such that for each k, $\|x\|\|y\|^{-1} > k$. Then $\|x_k\| \ge k\|y_k\| \ge k\rho$, so that the norms of the x_k converge to $+\infty$. By passing to a subsequence if needed, we can arrange that the elements $z_k := x_k/\|x_k\|$ converge to some z_0 having norm 1. By Proposition 4.1.7, $z_0 \in \operatorname{rc} D$. But for each k we have

$$\|Lz_k\| = \|y_k\|\|x_k\|^{-1} < k^{-1},$$

so that by continuity $Lz_0 = 0$. Then

$$z_0 \in (\ker L) \cap \operatorname{rc} D = V \cap V^\perp = \{0\},$$

contradicting $\|z_0\| = 1$. Therefore S has an upper bound β, so that whenever $y \in L(C)$ with $\|y\| > \rho$ there is $x \in D$ with $Lx = y$ and $\|x\| \leq \beta\|y\|$.

By combining the two cases we have considered, we find that whenever $y \in L(C)$ there is some $x \in D$ with $Lx = y$ and $\|x\| \leq \max\{\alpha, \beta\|y\|\}$, which proves (b). □

The bound in Theorem 4.1.16 differs in form from that in Proposition 3.2.4 because for small values of $\|y\|$ it is not positively homogeneous. To see why this is necessary when C is not a cone, let C be the closed line segment $[(1,0),(1,1)]$ in $\mathbb{R}^2$ and L be the linear map from $\mathbb{R}^2$ to $\mathbb{R}$ whose matrix in the standard basis is $[0,1]$. The procedure in the first part of the proof could choose any positive ρ, and for small $\rho > 0$ we could choose $\alpha = \|(1,\rho)\|$, so α could be any number greater than 1. On the other hand, if we tried to use $\alpha = 0$ then for the sequence $\{y_k\}$ with $y_k = k^{-1}$, the only point x_k in C with $Lx_k = y_k$ is $(1,k^{-1})$. Therefore no fixed β could satisfy $\|x_k\| \leq \beta\|y_k\|$ for all k.

Corollary 4.1.17. *Let L be a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ and let C be a convex subset of $\mathbb{R}^n$ that satisfies the recession condition for L. Then $L(C)$ is closed.*

Proof. Choose any $y_0 \in \operatorname{cl} C$ and let $\{y_k\}$ be a sequence in $L(C)$ converging to y_0. As $\{y_k\}$ converges, it is bounded. Apply Theorem 4.1.16 to produce a sequence x_k in C with $Lx_k = y_k$ and constants α and β with $\|x_k\| \leq \max\{\alpha, \beta\|y_k\|\}$ for each k. Then $\{x_k\}$ is bounded, so a subsequence must converge to some $x_0 \in C$. By continuity of L, $y_0 = Lx_0 \in L(C)$, so $L(C) = \operatorname{cl} L(C)$.

Corollary 4.1.18. *Let L be a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ and let C be a nonempty convex subset of $\mathbb{R}^n$. Then*

a. $L(\operatorname{rc} C) \subset \operatorname{rc} L(C)$*;*
b. If C satisfies the recession condition for L, then $\operatorname{rc} L(C)$ *is closed and*

$$L(\operatorname{rc} C) = \operatorname{rc} L(C) \tag{4.8}$$

Proof. Let $v \in \operatorname{rc} C$ and let y be any point of $L(C)$. There is then a point $c \in C$ with $Lc = y$. By choice of v we have $c + \mathbb{R}_+ v \subset C$, and therefore $L(c + \mathbb{R}_+ v) = y + \mathbb{R}_+ Lv \subset L(C)$. As y was any point of $L(C)$, $Lv \in \operatorname{rc} L(C)$, so (a) holds.

Now suppose that C satisfies the recession condition for L. Let $w \in \operatorname{rc} L(C)$. As $L(C)$ is closed by Corollary 4.1.17, $\operatorname{rc} L(C)$ is also closed by Theorem 4.1.2. Proposition 4.1.7 shows that w is the limit of a sequence $\gamma_k w_k$ where the γ_k are positive numbers converging to 0 and the w_k are elements of $L(C)$. Apply Theorem 4.1.16 to produce a sequence $\{c_k\}$ of points in C and real numbers α and β such that for each k, $Lc_k = w_k$ and

$$\|c_k\| \leq \max\{\alpha, \beta\|w_k\|\} \leq \alpha + \beta\|w_k\|.$$

Then for each k,

$$\|\gamma_k c_k\| \leq \gamma_k \alpha + \beta\|\gamma_k w_k\|. \tag{4.9}$$

The first term on the right in (4.9) converges to 0, and the second is bounded because $\gamma_k w_k$ converges to w. Therefore $\{\gamma_k c_k\}$ is bounded, and by taking a subsequence if

necessary we may suppose it converges to some v which must, by Proposition 4.1.7, be in $\operatorname{rc} C$. As $Lc_k = w_k$ for each k, by taking the limit in the equation $L(\gamma_k c_k) = \gamma_k w_k$ on the subsequence we obtain $Lv = w$. This shows that $w \in L(\operatorname{rc} C)$ and thereby proves (4.8). □

One might try to prove (4.8) under weaker hypotheses than those of Corollary 4.1.18. The following example from [34, p. 74] shows that at least one reasonable set of hypotheses—that both C and $L(C)$ be closed—will not work.

Example 4.1.19. Let $C = \{x \in \mathbb{R}^2 \mid x_2 \geq x_1^2\}$ and let $L : \mathbb{R}^2 \to \mathbb{R}$ be the linear map taking (x_1, x_2) to x_1 (the linear projector on $\mathbb{R}$). Then C is closed, and $L(C) = \mathbb{R}$ is also closed. However, $\operatorname{rc} C = \{0\} \times \mathbb{R}_+$ so that

$$L(\operatorname{rc} C) = \{0\} \neq \mathbb{R} = \operatorname{rc} L(C).$$

The recession condition does not hold here, because

$$(\ker L) \cap (\operatorname{rc} C) = \{0\} \times \mathbb{R}_+ \not\subset \{0\} = \operatorname{lin} C.$$

□

Corollary 4.1.20. *Let $C_1, \ldots, C_k$ be nonempty convex subsets of $\mathbb{R}^n$. Then*

$$\operatorname{rc} \sum_{i=1}^{k} C_i \supset \sum_{i=1}^{k} \operatorname{rc} C_i. \tag{4.10}$$

If also the C_i are closed and for any elements $v_1, \ldots, v_k$ with $v_i \in \operatorname{rc} C_i$ for each i and $\sum_{i=1}^{k} v_i = 0$ one has $v_i \in \operatorname{lin} C_i$ for each i, then $\sum_{i=1}^{k} C_i$ and its recession cone are closed and

$$\operatorname{rc} \sum_{i=1}^{k} C_i = \sum_{i=1}^{k} \operatorname{rc} C_i. \tag{4.11}$$

Proof. Let $D = \Pi_{i=1}^{k} C_i$. Then $\operatorname{rc} D = \Pi_{i=1}^{k} \operatorname{rc} C_i$ (Exercise 4.1.24). Define $L : \mathbb{R}^{kn} \to \mathbb{R}^n$ by $L(x_1, \ldots, x_k) = \sum_{i=1}^{k} x_i$, where each x_i is in $\mathbb{R}^n$. Then

$$L(D) = \sum_{i=1}^{k} C_i, \qquad L(\operatorname{rc} D) = \sum_{i=1}^{k} \operatorname{rc} C_i,$$

so (4.10) follows from part (a) of Corollary 4.1.18.

Now assume that the C_i are closed and that for any elements $v_1, \ldots, v_k$ with $v_i \in \operatorname{rc} C_i$ for each i and $\sum_{i=1}^{k} v_i = 0$ one has $v_i \in \operatorname{lin} C_i$ for each i. The kernel of L is the set of k-tuples $(x_1, \ldots, x_n)$ where the x_i belong to $\mathbb{R}^n$ and $\sum_{i=1}^{k} x_i = 0$. By Proposition 4.1.6 the lineality space of D is

$$\operatorname{lin} D = (\operatorname{rc} D) \cap (-\operatorname{rc} D) = \Pi_{i=1}^{k} [(\operatorname{rc} C_i) \cap (-\operatorname{rc} C_i)] = \Pi_{i=1}^{k} \operatorname{lin} C_i.$$

Our hypotheses then imply that $(\ker L)\cap(\operatorname{rc} D)\subset \operatorname{lin} D$, so that D satisfies the recession condition for L. Now the remaining assertions follow from Corollaries 4.1.17 and 4.1.18. □

For a convex set C, the set $\operatorname{cone} C$ may or may not be closed even if C is closed. The following corollary exhibits the structure of $\operatorname{cl}\operatorname{cone} C$.

Corollary 4.1.21. *If C is any convex subset of $\mathbb{R}^n$ whose closure does not contain the origin, then* $\operatorname{cl}\operatorname{cone} C = (\operatorname{cone}\operatorname{cl} C)\cup(\operatorname{rc}\operatorname{cl} C)$.

Proof. This is clear if C is empty, so we suppose that C is nonempty. Let $H_C = \operatorname{cone}(C\times\{-1\})$ and let L be the linear transformation from $\mathbb{R}^{n+1}$ to $\mathbb{R}^n$ defined by $L(x,\xi)=x$; then $\operatorname{cl}\operatorname{cone} C = \operatorname{cl} L(H_C)$. By Theorem 4.1.9 we have

$$\operatorname{cl} H_C = \{\operatorname{cone}[(\operatorname{cl} C)\times\{-1\}]\}\cup[(\operatorname{rc}\operatorname{cl} C)\times\{0\}].$$

Also, if $(z,\zeta)\in \ker L\cap \operatorname{rc}\operatorname{cl} H_C$ we have $z=0$ and $(0,\zeta)\in\operatorname{cl} H_C$. This means that there is a sequence of points of the form $(\tau_n c_n, -\tau_n)\in H_C$ converging to $(0,\zeta)$, with $\tau_n>0$ and $c_n\in C$ for each n. The norms of the c_n are bounded away from zero by hypothesis, so the τ_n must converge to zero and therefore $\zeta=0$. Now Corollary 4.1.17 tells us that $L(\operatorname{cl} H_C)=\operatorname{cl} L(H_C)$. Therefore

$$\operatorname{cl}\operatorname{cone} C = \operatorname{cl} L(H_C) = L(\operatorname{cl} H_C) = (\operatorname{cone}\operatorname{cl} C)\cup(\operatorname{rc}\operatorname{cl} C),$$

as was to be shown. □

4.1.4 Exercises for Section 4.1

Exercise 4.1.22. Show that for any nonempty convex set C in $\mathbb{R}^n$, $\operatorname{rc} C\subset \operatorname{rc}\operatorname{cl} C = \operatorname{rc}\operatorname{ri} C$. Exhibit a convex set C such that $\operatorname{rc}\operatorname{cl} C$ strictly contains $\operatorname{cl}\operatorname{rc} C$, and prove the inclusion.

Exercise 4.1.23. Exhibit an unbounded convex subset C of $\mathbb{R}^2$ such that $\operatorname{rc} C=\{0\}$, and prove the equality.

Exercise 4.1.24. For $i=1,\dots,n$ let n_i be a positive integer and let C_i be a nonempty convex subset of $\mathbb{R}^{n_i}$. Show that $\operatorname{rc}\Pi_{i=1}^n C_i = \Pi_{i=1}^n \operatorname{rc} C_i$.

Exercise 4.1.25. Let A be an $m\times n$ matrix, D be a $p\times n$ matrix, $a\in\mathbb{R}^m$, and $d\in\mathbb{R}^p$. Define a set $P\subset\mathbb{R}^n$ by

$$P=\{x\mid Ax=a,\ Dx\le d\},$$

where the symbol $\le$ expresses the partial order on $\mathbb{R}^p$ that defines $u\le v$ to mean that for each i, $u_i\le v_i$.

Show that if P is nonempty, then

$$\operatorname{rc} P=\{z\mid Az=0,\ Dz\le 0\},\qquad \operatorname{lin} P=\{z\mid Az=0,\ Dz=0\},$$

so that these sets do not depend on a and d.

Exercise 4.1.26. Let C be a nonempty convex cone in $\mathbb{R}^n$. Show that $C \subset \operatorname{rc} C$, and that if $0 \in C$ then $C = \operatorname{rc} C$.

Exercise 4.1.27. Let C be a convex set in $\mathbb{R}^n$ whose affine hull is a subspace. Then $(\operatorname{par} C)^\perp = \operatorname{lin}(C^\circ)$.

4.2 Faces

Convex sets have certain distinguished subsets, called *faces*, that appear frequently in convex analysis and especially in optimization. In this section we define faces and explain some of their special properties. We will use these properties in Section 4.3 to establish a fundamental representation theorem showing how a closed convex set can be written as a linear combination of certain simpler sets associated with it. One can think of that theorem as the internal analogue of the external representation given by Theorem 2.2.3.

Definition 4.2.1. Let C be a convex subset of $\mathbb{R}^n$. A convex subset F of C is called a *face* of C if whenever x and y belong to C, $\lambda \in (0,1)$, and $(1-\lambda)x + \lambda y \in F$, both x and y also belong to F. □

By applying the definition twice one sees that a face of a face of C is a face of C. The empty set is a face of every convex set, and any convex set is a face of itself. Sometimes we exclude the latter case by referring to *proper faces:* that is, faces that are proper subsets of the set in question.

The definition also shows that the intersection of any collection of faces of C is a face of C. Therefore for each pair of faces of C, say F_1 and F_2, there is a unique largest face of C contained in both (namely $F_1 \cap F_2$) and a unique smallest face that contains both (the intersection of all faces containing both; the collection over which that intersection is taken is not empty, because C is a face of itself).

Some faces of low dimension have special names. The following definition, which uses the operator pos introduced in Definition 1.1.17, explains these.

Definition 4.2.2. Let C be a convex subset of $\mathbb{R}^n$. A face of C that is a singleton is called an *extreme point* of C. If some face of C is a halfline of the form $\{x\} + \operatorname{pos}\{y\}$ where $y \neq 0$, then $\operatorname{pos}\{y\}$ is called an *extreme ray* of C and we say that y *generates* that extreme ray. □

We use the notation $\mathscr{F}(C)$ for the collection of all faces of a convex set C.

The definition of face said geometrically that if a face F met the relative interior of the line segment between x and y, then the whole segment had to lie in F. In fact, this property is not confined to line segments, as the following proposition shows. It gives two structural properties of faces that are often very useful in proofs.

Proposition 4.2.3. *Let C be a convex subset of $\mathbb{R}^n$ and let F be a face of C.*

a. If S is a convex subset of C whose relative interior meets F, then $S \subset F$.
b. $F = C \cap \operatorname{cl} F$. In particular, F is closed if C is closed.
c. $C \backslash F$ is convex.

Proof. For (a), let $f \in F \cap \operatorname{ri} S$ and let s be any point of S. We show that $s \in F$, which is immediate if $s = f$. If $s \neq f$, then Corollary 1.2.5 tells us that we can find $r \in S$ and $\lambda \in (0,1)$ with $f = (1-\lambda)r + \lambda s$. The definition of face now says that $s \in F$.

For (b), let $S = C \cap \operatorname{cl} F$. Then $F \subset S \subset \operatorname{cl} F$, so S and F have the same affine hull. But then

$$\operatorname{ri} F \subset \operatorname{ri} S \subset \operatorname{ri} \operatorname{cl} F = \operatorname{ri} F,$$

so F and S have the same relative interior. Then $F \supset S$ by (a), so in fact $F = S$.

For (c), suppose that $C \backslash F$ were not convex. Then F would be nonempty and there would be points x and x' of $C \backslash F$ and $\mu \in (0,1)$ such that $x_\mu := (1-\mu)x + \mu x' \notin C \backslash F$. As $x_\mu \in C$ we must have $x_\mu \in F$. But F is a face of C, so we would then have x and x' belonging to F as well as to $C \backslash F$, a contradiction. Therefore $C \backslash F$ is convex. □

If F is a singleton subset of a convex set C and $C \backslash F$ is convex, then F is an extreme point of C. Indeed, as F is a singleton $\{f\}$ we need only show that it is a face of C. Let c and c' be points of C, and let $\mu \in (0,1)$ with $f = (1-\mu)c + \mu c'$. Then $0 = (1-\mu)(c-f) + \mu(c'-f)$, so that $c - f = -\mu(1-\mu)^{-1}(c'-f)$. If either of $c-f$ or $c'-f$ were nonzero then so would be the other, and then each of c and c' would belong to the convex set $C \backslash F$. That would imply $f \in C \backslash F$, which is impossible. Therefore $c = f = c'$, so that F is an extreme point of C. However, for $\dim F > 0$ the convexity of $C \backslash F$ does not necessarily imply that F is a face: e.g., take $C = [0,1]$ and $F = [0,1)$.

Corollary 4.2.4. *If K is a closed convex cone in $\mathbb{R}^n$, then each face of K is a closed convex cone.*

Proof. Let F be a face of K. If $F = \emptyset$ or $F = \{0\}$, no proof is necessary. Otherwise, let $f \in F \backslash \{0\}$. Then $f(\operatorname{int} \mathbb{R}_+)$ is a relatively open halfline in K. Because it is relatively open and meets F, it must be contained in F. Therefore each positive multiple of a point in F belongs to F, so F is a cone. It is convex because it is a face of K, and closed by Proposition 4.2.3. □

Recession cones of faces of a set are faces of the set's recession cone, but the converse is not always true.

Proposition 4.2.5. *Let C be a nonempty closed convex subset of $\mathbb{R}^n$. If F is a nonempty face of C then* $\operatorname{rc} F$ *is a face of* $\operatorname{rc} C$.

Proof. Part (b) of Proposition 4.2.3 shows that F is closed, and as F is a convex subset of C Theorem 4.1.2 says that $\operatorname{rc} F$ is a convex subset of $\operatorname{rc} C$. Suppose v and v' belong to $\operatorname{rc} C$, let $\mu \in (0,1)$, and assume that $w := (1-\mu)v + \mu v' \in \operatorname{rc} F$. Choose any $f \in F$: then $f + w \in F$. As $F \subset C$, the points $f + v$ and $f + v'$ belong to C. But

$$f + w = (1-\mu)(f+v) + \mu(f+v'),$$

so as F is a face of C both $f+v$ and $f+v'$ belong to F. As f was an arbitrary point of F, both v and v' belong to $\operatorname{rc} F$, and therefore $\operatorname{rc} F$ is a face of $\operatorname{rc} C$. □

To see that the converse may not hold, let C be the subset of $\mathbb{R}^2$ defined by

$$C = \left\{ (x_1, x_2) \mid \begin{cases} x_2 \geq 0 & \text{if } x_1 \leq 0 \\ x_2 \geq x_1^2, & \text{if } x_1 > 0 \end{cases} \right\}.$$

Then $\operatorname{rc} C = \mathbb{R}_- \times \mathbb{R}_+$, and $G := \{0\} \times \mathbb{R}_+$ is a face of $\operatorname{rc} C$. However, G is not the recession cone of any nonempty face of C.

Proposition 4.2.6. *Let C be a convex subset of $\mathbb{R}^n$ and let F be any proper face of C. Then $\dim F < \dim C$, and F is entirely contained in the relative boundary of C.*

Proof. C is nonempty because F is proper. Further, F must lie entirely in the relative boundary of C because if F met $\operatorname{ri} C$ then we would have $C \subset F$ by Proposition 4.2.3, contradicting the hypothesis.

If the dimensions of F and C were the same, then F would be nonempty and, as $F \subset C$, $\operatorname{aff} F$ and $\operatorname{aff} C$ would be the same. Then $\operatorname{ri} F$ would be nonempty and we would have $\operatorname{ri} F \subset \operatorname{ri} C$, whereas we have already shown that F does not meet $\operatorname{ri} C$. Therefore $\dim F < \dim C$. □

It follows that any chain of faces of C, each properly contained in the preceding one, must be finite.

Proposition 4.2.7. *Let H be a hyperplane in $\mathbb{R}^n$ and let C be a nonempty convex set contained in one of the closed halfspaces of H. Then $C \cap H$ is a face of C, and it is proper if and only if H does not meet* $\operatorname{ri} C$.

Proof. If C and H are disjoint, we are finished. Therefore assume that H meets C. Write $F = C \cap H$ and let $H = H_0(y^*, \eta)$. If c_1 and c_2 are two points of C and $\lambda \in (0,1)$ with $(1-\lambda)c_1 + \lambda c_2 \in F$, then we must have $\langle y^*, c_1 \rangle = \eta = \langle y^*, c_2 \rangle$ so that both c_1 and c_2 belong to $C \cap H = F$. Therefore F is a face of C.

For the second claim, apply Corollary 1.2.3 to conclude that there are only two mutually exclusive possibilities: either $\operatorname{cl} C \subset H$ or H does not meet $\operatorname{ri} C$. The first occurs if and only if F is improper, and therefore the second occurs if and only if F is proper. □

Definition 4.2.8. Let C be a nonempty convex set in $\mathbb{R}^n$ and let F be a face of C. The face F is *exposed* if there is some $x^* \in \mathbb{R}^n$ such that $F = N_C^{-1}(x^*)$. □

An exposed face is therefore the set of maximizers on C of the linear functional $\langle x^*, \cdot \rangle$. It may be C itself (for $x^* = 0$), or the empty set, but otherwise it must be of the form $F = C \cap H_0(x^*, \mu)$, where $x^* \neq 0$ and μ is the finite maximum on C of $\langle x^*, \cdot \rangle$. As $C \subset H_-(x^*, \mu)$, Proposition 4.2.7 shows that F is a face of C.

For a simple example of a face that is not exposed, start with the nonnegative orthant $\mathbb{R}^2_+$ and obtain C by first deleting the closed square whose corners are $(0,0)$,

$(1,0)$, $(1,1)$, and $(0,1)$ and then taking the union of the result with a closed ball of radius 1 centered at $(1,1)$. The resulting C is closed and convex, but the points $(1,0)$ and $(0,1)$ are faces that are not exposed.

Exposed faces appear naturally in the following construction, which is very important in optimization.

Definition 4.2.9. Let C be a nonempty closed convex subset of $\mathbb{R}^n$ and let $z_0 \in \mathbb{R}^n$. Let $x_0 = \Pi_C(z_0)$ and $x_0^* = z_0 - x_0$. The *critical cone* induced by C and z_0 is

$$\kappa_C(z_0) := N_{T_C(x_0)}^{-1}(x_0^*).$$

□

Our previous work with the projector shows that in this situation x_0^* belongs to $N_C(x_0)$, which is the polar of $T_C(x_0)$. Therefore each point y of $T_C(x_0)$ satisfies, $\langle x_0^*, y\rangle \leq 0$, and as $T_C(x_0)$ is a closed set the maximum is zero. The set of maximizers includes the origin and possibly additional points of $T_C(x_0)$, and

$$\kappa_C(z_0) = T_C(x_0) \cap \{v \mid \langle x_0^*, v\rangle = 0\}.$$

The critical cone is a nonempty exposed face of $T_C(x_0)$, but it need not be a face of C. A companion concept, defined next, is useful when we want to deal only with points in C.

Definition 4.2.10. Let C be a nonempty closed convex subset of $\mathbb{R}^n$, and let $z_0 \in \mathbb{R}^n$. Let $x_0 = \Pi_C(z_0)$ and $x_0^* = z_0 - x_0$. The *critical face* of C corresponding to z_0 is the set

$$\phi_C(z_0) := N_C^{-1}(x_0^*).$$

□

$\phi_C(z_0)$ is an exposed face of C by definition, and it is nonempty because $(x_0, x_0^*) \in N_C$; if $z_0 \in C$ then $\phi_C(z_0)$ is all of C.

As $(x_0, x_0^*) \in N_C$ we have $x_0 \in \phi_C(z_0)$ and $\phi_C(z_0) = C \cap \{x \mid \langle x_0^*, x - x_0\rangle = 0\}$. Also $C - x_0 \subset T_C(x_0)$, so

$$\phi_C(z_0) - x_0 = (C - x_0) \cap \{v \mid \langle x_0^*, v\rangle = 0\} \subset T_C(x_0) \cap \{v \mid \langle x_0^*, v\rangle = 0\} = \kappa_C(z_0). \tag{4.12}$$

However, we generally do not have equality because the tangent cone may be much larger than the set C. In fact, the critical cone and the critical face need not even have the same dimension. For example, let C be the Euclidean unit ball in $\mathbb{R}^2$ and let $z_0 = (0,2)$. The projection of z_0 on C is $x_0 = (0,1)$ and $x_0^* = (0,1)$. The critical face $\phi_C(z_0)$ is the point $\{(0,1)\}$, whereas the critical cone $\kappa_C(z_0)$ is the line $\mathbb{R} \times \{0\}$. When the set in question is polyhedral then the situation is much better, as we will show later.

The next theorem gives a fundamental structural property of convex sets.

Theorem 4.2.11. *Let C be a convex set in $\mathbb{R}^n$. The relative interiors of the faces of C form a partition of C: that is, they are disjoint and their union is C. They are maximal in the sense that if D is any relatively open convex subset of C then D is contained in the relative interior of one of the faces of C.*

Proof. If C is empty there is nothing to prove. If C is nonempty and F_1 and F_2 are distinct faces of C, then $\operatorname{ri} F_1$ is disjoint from $\operatorname{ri} F_2$ by Proposition 4.2.3 (otherwise $F_1 = F_2$). Now let D be any relatively open convex subset of C. We show D is contained in the relative interior of some face of C. This implies the partition claim, because for any $c \in C$ we could take D to be the singleton $\{c\}$.

As D is relatively open, it must be completely contained in each face of C that it meets (Proposition 4.2.3). Take the intersection of all such faces, of which there is at least one (C), and call it F. Then F is a face of C, $D \subset F$, and D meets no proper face of F. We will show that $D \subset \operatorname{ri} F$ by assuming the contrary and producing a contradiction.

If $\operatorname{ri} F$ does not contain D then let d be a point of $D \setminus \operatorname{ri} F$. Separate d and F properly by a hyperplane H. As $D \subset F$, D must lie in one of the closed halfspaces of H and d in the other. But $d \in D$, so d must lie in H. As D is relatively open, $d \in \operatorname{ri} D$ and therefore H meets $\operatorname{ri} D$. Corollary 1.2.3 now shows that $D \subset H$, which implies $F \not\subset H$ because the separation is proper. Apply Proposition 4.2.7 to F to conclude not only that the set $G = F \cap H$ is a face of F but also, because $F \not\subset H$, that G is a proper face of F. But $D \subset G$, and this contradicts the construction of F. Therefore $D \subset \operatorname{ri} F$. □

Theorem 4.2.11 has many applications. One that frequently arises is the use of this theorem to associate particular faces of a set with particular relatively open subsets that are of special interest. The next three propositions use it in this way. They provide ways to infer the face structures of sets created by common operations from those of the original sets.

Proposition 4.2.12. *Let C and D be convex subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Then*

$$\mathscr{F}[C \times D] = \{F \times G \mid F \in \mathscr{F}(C),\ G \in \mathscr{F}(D)\}.$$

Proof. *($\supset$)* Choose $F \in \mathscr{F}(C)$ and $G \in \mathscr{F}(D)$. If either F or G is empty, then the product is the empty face of $C \times D$. Therefore suppose $(f,g) \in F \times G$. For $i = 1,2$ let (c_i, d_i) be points of $C \times D$ such that (f,g) belongs to the interval $((c_1,d_1),(c_2,d_2))$. Then f is in the interval (c_1,c_2), so the definition of face ensures that each of c_1 and c_2 belongs to F. Similarly, each of d_1 and d_2 belongs to G. Hence for $i = 1,2$ the pair (c_i,d_i) belongs to $F \times G$, so $F \times G$ is a face of $C \times D$.

($\subset$)) Suppose H is a face of $C \times D$. If it is empty then it is the product of the empty faces of C and D, so we may assume it is nonempty. Choose a pair $(c_0,d_0) \in \operatorname{ri} H$. Use Theorem 4.2.11 to find unique faces F of C and G of D such that $c_0 \in \operatorname{ri} F$ and $d_0 \in \operatorname{ri} G$. Now $F \times G$ is a convex subset of $C \times D$ whose relative interior contains (c_0,d_0). By Proposition 4.2.3 we have $F \times G \subset H$. However, we already saw in the first part of the proof that $F \times G$ was a face of $C \times D$, and as it meets the relative interior of H we must have $F \times G \supset H$, so that in fact $H = F \times G$. □

The second proposition deals with inverse images of convex sets under affine transformations.

Proposition 4.2.13. *Let C be a convex subset of $\mathbb{R}^n$ and let T be an affine transformation from $\mathbb{R}^m$ to $\mathbb{R}^n$. Then $\mathscr{F}[T^{-1}(C)] = T^{-1}[\mathscr{F}(C)]$.*

Proof. ($\supset$) Let $F \in \mathscr{F}(C)$ and define G to be $T^{-1}(F)$. We want to show that G is a face of $T^{-1}(C)$. If G is empty then it is the empty face of $T^{-1}(C)$. Therefore suppose it is nonempty; suppose that d_1 and d_2 belong to $T^{-1}(C)$ and that there is a point d_0 belonging to $G \cap (d_1, d_2)$. We show that d_1 and d_2 must then belong to G, so that G will be a face of $T^{-1}(C)$. For $i = 0, 1, 2$ let $c_i = T(d_i)$; then $c_0 \in F \cap (c_1, c_2)$, and as F is a face of C the points c_1 and c_2 must belong to F. But then d_1 and d_2 belong to G, as required.

($\subset$) Let G be a face of $T^{-1}(C)$. If G is empty then it is the inverse image of the empty face of C. Therefore assume G is nonempty and let g_0 be a point of its relative interior. Define f_0 to be $T(g_0)$, and let F be the unique face of C whose relative interior contains f_0. We show that $G = T^{-1}(F)$.

First choose any $g_1 \in G$. As $g_0 \in \operatorname{ri} G$ we can use Theorem 1.2.4 to find a point g_2 of G with $g_0 \in (g_1, g_2)$. For $i = 1, 2$ define c_i to be $T(g_i)$; then $f_0 \in (c_1, c_2)$. As F is a face of C, the points c_1 and c_2 must belong to F. But this means that $g_1 \in T^{-1}(F)$, and therefore $G \subset T^{-1}(F)$.

Next choose any point d_1 of $T^{-1}(F)$. We want to prove that d_1 belongs to the face G of $T^{-1}(C)$. Let $f_1 = T(d_1)$; as $f_1 \in F$ and $f_0 \in \operatorname{ri} F$ we can apply Corollary 1.2.5 to produce $f_2 \in F$ with $f_0 \in (f_1, f_2)$. Therefore there is some $\lambda \in (0, 1)$ with $f_0 = (1-\lambda) f_1 + \lambda f_2$. Define d_2 to be $(1-\lambda^{-1}) d_1 + \lambda^{-1} g_0$; then $T(d_2) = f_2$. Now g_0 belongs to G, a face of $T^{-1}(C)$, and each of d_1 and d_2 belongs to $T^{-1}(F)$ and hence to $T^{-1}(C)$. But also $g_0 = (1-\lambda) d_1 + \lambda d_2$, so each of d_1 and d_2 must belong to G. Therefore $T^{-1}(F) \subset G$, so in fact $G = T^{-1}(F)$. □

The next result deals with forward images under affine transformations. Unlike the last proposition, this one has only a partial statement: it says that any face of the image of a set is itself the image, under the affine transformation, of some face of the original set. It does not say that the transformation carries any face of the original set to a face of the image, because that is not generally true. To see this, consider the 2-simplex S in $\mathbb{R}^2$ that is the convex hull of $(0,0)$, $(2,0)$, and $(1,1)$. Let T be the linear projector taking the point $(x, y) \in \mathbb{R}^2$ to $x \in \mathbb{R}$. The set $T(S)$ has three nonempty faces; each of the two endpoints is the image under T of one of the vertices of S, while the entire set $T(S)$ is the image of both S and one of its one-dimensional faces. However, T carries one of the vertices of S and two of its one-dimensional faces into subsets of $T(S)$ that are not faces.

Proposition 4.2.14. *Let C be a convex subset of $\mathbb{R}^n$ and let T be an affine transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then $\mathscr{F}[T(C)] \subset T[\mathscr{F}(C)]$.*

Proof. Let G be a face of $T(C)$. If G is empty then it is the image of the empty face of C. If it is nonempty, choose $g_0 \in \operatorname{ri} G$; define $S = C \cap T^{-1}(g_0)$. This set S is

nonempty; let f_0 be any point in its relative interior. Use Theorem 4.2.11 to find the unique face F of C such that $f_0 \in \operatorname{ri} F$. We show that $G = T(F)$.

Let g_1 be any point of G, and and choose $g_2 \in G$ with $g_0 \in (g_1, g_2)$. For $i = 1, 2$ let $c_i \in C$ with $g_i = T(c_i)$. There is $\lambda \in (0,1)$ with $g_0 = (1-\lambda)g_1 + \lambda g_2$; define a point $f_a \in C$ by $f_a = (1-\lambda)c_1 + \lambda c_2$. As T is affine we have $T(f_a) = g_0$, so that $f_a \in S$. We chose f_0 to be in the relative interior of S, so we can use Theorem 1.2.4 to find some $f_b \in S$ with $f_0 \in (f_a, f_b)$. Each of f_a and f_b belongs to C, so as $f_0 \in F$ we have f_a and f_b in F also. But $f_a \in (c_1, c_2)$, so c_1 also belongs to F. Then $g_1 = T(c_1) \in T(F)$, so $G \subset T(F)$.

Next let d_1 be any point of $T(F)$ and let $f_1 \in F$ with $d_1 = T(f_1)$. As $f_0 \in \operatorname{ri} F$ we can use Corollary 1.2.5 to find some $f_2 \in F$ with $f_0 \in (f_1, f_2)$; let $d_2 = T(f_2)$. The points d_1 and d_2 are in $T(C)$, and as T is affine we have $g_0 = T(f_0) \in (d_1, d_2)$. As G is a face of $T(C)$ it follows that $d_1 \in G$, and therefore $T(F) \subset G$. □

Here are two consequences of the preceding three propositions. The first shows that the faces of an intersection of convex sets are precisely the intersections of faces of the individual sets.

Corollary 4.2.15. *Let $C_1, \ldots, C_k$ be convex subsets of $\mathbb{R}^n$, and define $C = \cap_{i=1}^k C_i$. A subset F of $\mathbb{R}^n$ is a face of C if and only if $F = \cap_{i=1}^k F_i$, where for each i the set F_i is a face of C_i.*

Proof. It suffices to prove the result for $k = 2$, as we can then extend it by induction. Let T be the linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^{2n}$ defined by $T(x) = (x, x)$. Then $C_1 \cap C_2 = T^{-1}(C_1 \times C_2)$. Accordingly, by Propositions 4.2.12 and 4.2.13 we have

$$\begin{aligned}\mathscr{F}(C_1 \cap C_2) &= \mathscr{F}[T^{-1}(C_1 \times C_2)] \\ &= T^{-1}(\{F \times G \mid F \in \mathscr{F}(C_1),\ G \in \mathscr{F}(C_2)\}) \\ &= \{F \cap G \mid F \in \mathscr{F}(C_1),\ G \in \mathscr{F}(C_2)\}.\end{aligned}$$

□

The second consequence is the fact that each face of a finite sum of convex sets is the sum of faces of individual sets appearing in the sum.

Corollary 4.2.16. *Let $C_1, \ldots, C_k$ be convex subsets of $\mathbb{R}^n$, and define $C = \sum_{i=1}^k C_i$. Each face F of C has the form $F = \sum_{i=1}^k F_i$, where for each i the set F_i is a face of C_i.*

Proof. Again it is enough to prove the result for $k = 2$. Let T^* be the adjoint of the linear transformation T defined in the proof of Corollary 4.2.15: that is, $T^*(x, y) = x + y$. Then $C_1 + C_2 = T^*(C_1 \times C_2)$. Applying Propositions 4.2.12 and 4.2.14 we have

$$\begin{aligned}\mathscr{F}(C_1 + C_2) &\subset T^*(\mathscr{F}(C_1 \times C_2)) \\ &= T^*(\{F \times G \mid F \in \mathscr{F}(C_1),\ G \in \mathscr{F}(C_2)\}) \\ &= \{F + G \mid F \in \mathscr{F}(C_1),\ G \in \mathscr{F}(C_2)\} \\ &= \mathscr{F}(C_1) + \mathscr{F}(C_2).\end{aligned}$$

□

In Corollary 4.2.16 we have $\mathscr{F}(C_1+C_2) \subset \mathscr{F}(C_1)+\mathscr{F}(C_2)$. The opposite inclusion may not hold: for example, consider the sum $[0,1]+[1,2]=[1,3]$ in $\mathbb{R}$. Each face of $[1,3]$ is a sum of faces of the two summands; however, $\{1\}+\{1\}=\{2\}$ is also such a sum, but is not a face of $[1,3]$.

If C is a closed convex set in $\mathbb{R}^n$, then the faces F of $\mathrm{cl}\,H_C$ fall into two disjoint classes: those contained in $\mathbb{R}^n \times \{0\}$, and those meeting $\mathbb{R}^n \times (-\infty, 0)$. The following theorem shows that these two classes correspond to the classes $\mathscr{F}(\mathrm{rc}\,C)$ and $\mathscr{F}(C)$ respectively.

Theorem 4.2.17. *If C is a nonempty closed convex subset of $\mathbb{R}^n$ and H_C is its homogenizing cone, then:*

a. The nonempty faces U of $\mathrm{cl}\,H_C$ that are contained in $\mathbb{R}^n \times \{0\}$ are in one-to-one correspondence with the nonempty faces V of $\mathrm{rc}\,C$, via the maps

$$\sigma(V) = V \times \{0\}, \qquad \tau(U) = \Pi(U),$$

where $\Pi(x_1, \ldots, x_{n+1}) = (x_1, \ldots, x_n)$.

b. The elements of the collection $\mathscr{H}$ of faces F of $\mathrm{cl}\,H_C$ that meet $\mathbb{R}^n \times (-\infty, 0)$ are in one-to-one correspondence with the nonempty faces G of C, via the maps

$$\phi(G) = \mathrm{cl}\,\mathrm{cone}(G \times \{-1\}), \qquad \gamma(F) = \Pi[F \cap (\mathbb{R}^n \times \{-1\})].$$

Proof. For (a), we apply Theorem 4.1.9 to show that $(\mathrm{cl}\,H_C) \cap (\mathbb{R}^n \times \{0\}) = (\mathrm{rc}\,C) \times \{0\}$. Proposition 4.2.12 shows that the faces U of $(\mathrm{rc}\,C) \times \{0\}$ are the sets of form $U = V \times \{0\}$ where V is a face of $\mathrm{rc}\,C$. Therefore these faces of $\mathrm{cl}\,H_C$ are in one-to-one correspondence with the faces of $\mathscr{F}(\mathrm{rc}\,C)$ through the equations $U = \sigma(V)$ and $V = \tau(U)$.

To prove (b), let G be a face of C. Define a subset F of $\mathbb{R}^{n+1}$ by

$$F = \phi(G) := \mathrm{cl}\,\mathrm{cone}(G \times \{-1\}).$$

This set F is closed and convex, and it meets $\mathbb{R}^n \times (-\infty, 0)$. As $G \subset C$, F is a subset of $\mathrm{cl}\,H_C$. To show that F is a face of $\mathrm{cl}\,H_C$, let $(h, -\alpha)$ and $(h', -\alpha')$ be points of $\mathrm{cl}\,H_C$ and suppose that $\mu \in (0,1)$ with

$$(h_\mu, -\alpha_\mu) := (1-\mu)(h, -\alpha) + \mu(h', -\alpha') \in F. \tag{4.13}$$

We show that both $(h, -\alpha)$ and $(h', -\alpha')$ belong to F. There are three cases, treated in separate paragraphs below.

Case 1 is that in which $\alpha_\mu = 0$: then both α and α' must be zero. Theorem 4.1.9 shows that h and h' belong to $\mathrm{rc}\,C$ and, as G is closed by Proposition 4.2.3, it also shows that h_μ belongs to $\mathrm{rc}\,G$. But $\mathrm{rc}\,G$ is a face of $\mathrm{rc}\,C$ by Proposition 4.2.5, so that both h and h' must belong to $\mathrm{rc}\,G$. As Theorem 4.1.9 says that

$$F \cap (\mathbb{R}^n \times \{0\}) = (\mathrm{rc}\,G) \times \{0\},$$

it follows that $(h,-\alpha)$ and $(h',-\alpha')$ are both in F.

The other possibility is that in which $\alpha_\mu > 0$, and here there are two subcases: in Case 2, exactly one of α and α' is zero, and with no loss we can assume it is α'; in Case 3 both are positive.

Suppose next that we are in Case 2. Then $\alpha' = 0$, and from (4.13) we have

$$(h_\mu, -\alpha_\mu) := [(1-\mu)h + \mu h', -(1-\mu)\alpha].$$

Then as $\alpha_\mu > 0$, we obtain $\alpha_\mu = (1-\mu)\alpha$ and

$$\alpha^{-1}h + \alpha^{-1}(1-\mu)^{-1}\mu h' = \alpha_\mu^{-1}h_\mu =: g \in G, \tag{4.14}$$

where the inclusion comes from Theorem 4.1.9. As $\alpha' = 0$ and $(h',-\alpha') \in \mathrm{cl}\, H_C$ we have $h' \in \mathrm{rc}\, C$. We also have $\alpha > 0$ and $(h,-\alpha) \in \mathrm{cl}\, H_C$ so that $\alpha^{-1}h =: c \in C$. Writing β for $\alpha^{-1}(1-\mu)^{-1}\mu$, we obtain from (4.14) the equation

$$g = c + \beta h'. \tag{4.15}$$

The halfline $c + h'\mathbb{R}_+$ is contained in C because $c \in C$ and $h' \in \mathrm{rc}\, C$. Moreover, (4.15) shows that the relative interior of this halfline meets G, so by Proposition 4.2.3 the entire halfline is in G. This tells us first that $c \in G$ and next, by Theorem 4.1.2, that $h' \in \mathrm{rc}\, G$. Then

$$(h,-\alpha) = (\alpha c, -\alpha) \in F, \qquad (h',-\alpha') = (h',0) \in F,$$

as required.

in Case 3, both α and α' are positive. If we write

$$c = \alpha^{-1}h, \qquad c' = \alpha'^{-1}h', \qquad g = \alpha_\mu^{-1}h_\mu,$$

then (4.13) yields

$$(1-\mu)\alpha(c,-1) + \mu\alpha'(c',-1) = \alpha_\mu(g,-1). \tag{4.16}$$

Define $\lambda = \mu\alpha'\alpha_\mu^{-1}$: then by using the second components of the quantities in (4.16) we find that $1-\lambda = (1-\mu)\alpha\alpha_\mu^{-1}$ and that $\lambda \in (0,1)$. Then (4.16) shows that

$$(1-\lambda)(c,-1) + \lambda(c',-1) = (g,-1). \tag{4.17}$$

As G is a face of C, (4.17) shows that both c and c' belong to G, and this implies that both $(h,-\alpha)$ and $(h',-\alpha')$ belong to F as required. Therefore $F \in \mathscr{H}$.

We have now shown that ϕ takes faces of G to elements of $\mathscr{H}$. The next step is to show that ϕ is injective, and for that purpose let G and G' belong to $\mathscr{F}(C)$ and suppose that $\phi(G) = F = F' = \phi(G')$. As $F \in \mathscr{H}$, by Theorem 4.1.9 there is a point in $\mathrm{ri}\, F$ of the form $(g,-1)$ where $g \in (\mathrm{ri}\, G) \cap (\mathrm{ri}\, G')$. But as G and G' are faces, this implies that $G = G'$, so that ϕ is injective.

Now suppose F is any element of $\mathscr{H}$. As $\operatorname{cl} H_C$ is a closed convex cone, so is F by Corollary 4.2.4. Let $G = \gamma(F) = \Pi[F \cap (\mathbb{R}^n \times \{-1\})]$. Then G is a closed convex subset of C. Let c and c' belong to C and choose $\mu \in (0,1)$ such that $(1-\mu)c + \mu c' =: g \in G$. Then $(c,-1)$ and $(c',-1)$ belong to $\operatorname{cl} H_C$, and we have

$$(1-\mu)(c,-1) + \mu(c',-1) = (g,-1) \in F.$$

As F is a face, both $(c,-1)$ and $(c',-1)$ are in F and so c and c' belong to G. This shows that G is a face of C. Theorem 4.1.9 then says that $F = \phi(G)$, so that ϕ is surjective and $\phi \circ \gamma = \mathrm{id}_{\mathscr{H}}$.

Let $G \in \mathscr{F}(C)$. By what we have just shown,

$$\phi(G) = (\phi \circ \gamma) \circ \phi(G) = \phi \circ (\gamma \circ \phi)(G).$$

If we let $G' = (\gamma \circ \phi)(G)$ then this shows that $\phi(G) = \phi(G')$. But ϕ is injective, so $G = G'$ and therefore $\gamma \circ \phi = \mathrm{id}_{\mathscr{F}(C)}$. This establishes the asserted one-to-one correspondence between $\mathscr{H}$ and $\mathscr{F}(C)$. □

Faces of a closed convex cone are closely related to faces of the polar cone. The following theorem explains this relationship.

Theorem 4.2.18. *Let K be a nonempty closed convex cone in $\mathbb{R}^n$, and let F be a nonempty face of K. Write L for* $\operatorname{par} F$. *Then $F = K \cap L$ and for each $x \in \operatorname{ri} F$, $N_K(x)$ has the constant value $F^* := K^\circ \cap L^\perp$. This F^* is a face of K°, and for each $x^* \in F^*$, $F \subset N_{K^\circ}(x^*)$. For $x^* \in \operatorname{ri} F^*$ one has $F = N_{K^\circ}(x^*)$ if and only if the face F is exposed.*

Proof. As K is a closed convex cone, so is F (Exercise 4.2.20). Then F contains the origin, so $L = \operatorname{par} F = \operatorname{aff} F \supset F$. The hypothesis says that $F \subset K$ and $\operatorname{par} F = L$, so $F \subset K \cap L$.

As F is a face of K, to show that $K \cap L \subset F$ we need only show that $\operatorname{ri}(K \cap L)$ meets F (part (a) of Proposition 4.2.3). If this did not happen, then as F and $K \cap L$ are both convex cones that are subsets of L, we could separate them properly by a hyperplane $H_0(y^*, 0)$ with $y^* \in L \backslash \{0\}$. (To see this, carry out the separation in L to produce y^*, then include y^* in $\mathbb{R}^n$ and use it to define H_0.) As $F \subset K \cap L$, F must be contained in the hyperplane. If $x \in \operatorname{ri} F$ then for small positive ε we have $x + \varepsilon y \in F$ because $L = \operatorname{aff} F$. Then as $\langle y^*, x \rangle = 0$ we have

$$0 = \langle y^*, x + \varepsilon y^* \rangle = \varepsilon \|y^*\|^2,$$

which contradicts $y^* \neq 0$. Therefore $\operatorname{ri}(K \cap L)$ meets F, so that $K \cap L \subset F$ and therefore $F = K \cap L$.

As $\operatorname{ri} F$ is a relatively open convex subset of K, the value of $N_K(x)$ for $x \in \operatorname{ri} F$ is some constant set F^* by Exercise 3.3.15. Let u^* be a point of $K^\circ \cap L^\perp$ and let $x \in \operatorname{ri} F$. Then $x - 0 \in \operatorname{par} F = L$, so $\langle u^*, x \rangle = 0$. For any $x' \in K$, as $u^* \in K^\circ$ we have

$$\langle u^*, x' - x \rangle = \langle u^*, x' \rangle - \langle u^*, x \rangle = \langle u^*, x' \rangle \leq 0,$$

so $u^* \in N_K(x) = F^*$, showing that $K^\circ \cap L^\perp \subset F^*$. Next, suppose that $x^* \in F^*$ and $x \in \operatorname{ri} F$, and let $v \in L$. For small positive ε we have $x + \varepsilon v \in \operatorname{ri} F$. The constancy of F^* then implies that

$$0 \geq \langle x^*, (x+\varepsilon v) - x \rangle, \qquad 0 \geq \langle x^*, x - (x+\varepsilon v) \rangle,$$

so $\langle x^*, v \rangle = 0$. As v was an arbitrary point of L, $x^* \in L^\perp$. But $x^* \in K^\circ$ because $N_K(x) = K^\circ \cap \{v^* \mid \langle v^*, x \rangle = 0\}$ (Exercise 3.3.13), so $x^* \in K^\circ \cap L^\perp$. Therefore $F^* \subset K^\circ \cap L^\perp$, so we have shown that $F^* = K^\circ \cap L^\perp$.

To show that F^* is a face of K°, let k_0^* and k_1^* be points of K° and suppose that $\mu \in (0,1)$ such that $(1-\mu)k_0^* + \mu k_1^* =: k^* \in F^*$. Choose $x \in \operatorname{ri} F$ and $v \in L$. For small positive ε we have $x + \varepsilon v \in \operatorname{ri} F$. As k_0^* and k_1^* belong to K°, both $\langle k_0^*, x \rangle$ and $\langle k_1^*, x \rangle$ are nonpositive. However, $k^* \in F^* \subset L^\perp$, so

$$0 = \langle k^*, x \rangle = (1-\mu)\langle k_0^*, x \rangle + \mu \langle k_1^*, x \rangle,$$

implying that both $\langle k_0^*, x \rangle$ and $\langle k_1^*, x \rangle$ are zero. A similar argument with $x + \varepsilon v$ in place of x shows that both $\langle k_0^*, x + \varepsilon v \rangle$ and $\langle k_1^*, x + \varepsilon v \rangle$ are zero, implying that $\langle k_0^*, v \rangle = 0 = \langle k_1^*, v \rangle$. As v was any element of L, both k_0^* and k_1^* belong to $L^\perp$ and therefore to F^*, which must then be a face of K°.

For the last assertions, let $x^* \in F^*$. The set $N_{K^\circ}(x^*)$ consists of those $x \in K$ such that $\langle x^*, x \rangle = 0$. As $x^* \in K^\circ \cap L^\perp$, if $x \in K \cap L$ then $x \in N_{K^\circ}(x^*)$, and therefore $F \subset N_{K^\circ}(x^*)$. If the face F is not exposed then there is no x^* such that $F = N_K^{-1}(x^*)$. But $N_K^{-1} = N_{K^\circ}$, so we cannot have $F = N_{K^\circ}(x^*)$. Now suppose F is exposed, and choose y^* such that $F = N_K^{-1}(y^*)$. Then for each $x \in F$ we have $y^* \in N_K(x)$, so that $y^* \in F^*$. Let V be the constant value of N_{K° on $\operatorname{ri} F^*$; we have already shown that $F \subset V$. Let v be any point of V. Choose a sequence $\{x_k^*\}$ of points in $\operatorname{ri} F^*$ converging to y^*; then for each k we have $v \in N_{K^\circ}(x_k^*)$. As N_{K° is a closed multifunction, $v \in N_{K^\circ}(x^*) = F$. As v was any point of V we conclude that $V \subset F$ and so $F \subset V \subset F$. Accordingly, F is the constant value of N_{K° on $\operatorname{ri} F^*$. □

If the cone K is polyhedral then $\operatorname{par} F^* = L^\perp$, as we show in Theorem 5.3.2 below. However, this is false in general as one can see by taking K to be the cone of Exercise 3.1.18 and L to be the z-axis. Then $K^\circ \cap L^\perp$ and $L^\perp$ have dimensions 1 and 2 respectively.

4.2.1 Exercises for Section 4.2

Exercise 4.2.19. Show that if C is a nonempty closed convex subset of $\mathbb{R}^n$ then the lineality space of each nonempty face of C is $\operatorname{lin} C$.

Exercise 4.2.20. If K is a convex cone in $\mathbb{R}^n$ and F is a face of K, then F is a convex cone; further, F is closed if K is closed.

Exercise 4.2.21. Let C be a nonempty convex subset of $\mathbb{R}^n$, and let L be a nonempty subspace contained in $\operatorname{lin} C$. Define C_0 to be $C \cap L^\perp$. Show that the faces F of C are in one-to-one correspondence with the faces F_0 of C_0 through the mappings $\phi(F_0) = F_0 + L$, $\psi(F) = F \cap L^\perp$.

Note. To establish a one-to-one correspondence between sets S and T one must show that there are functions $\sigma : S \to T$ and $\tau : T \to S$ such that $\tau \circ \sigma = \mathrm{id}_S$ and $\sigma \circ \tau = \mathrm{id}_T$, where id_X denotes the identity map of X.

4.3 Representation

We are going to develop for closed convex sets a representation of the form $\operatorname{conv} E + \operatorname{pos} D$, where E and D are specified sets. To do so, we will use the generalized simplex introduced and discussed in Definitions 1.1.18 and 1.1.19, and in Theorem 1.1.20.

We begin with the following proposition, which shows that a face inherits some of the structure of the original set.

Proposition 4.3.1. *Let X and Y be subsets of $\mathbb{R}^n$, and let $C = \operatorname{conv} X + \operatorname{pos} Y$. Let F be a nonempty face of C, and define $X_F = X \cap F$ and $Y_F = Y \cap \operatorname{rc} F$. Then $F = \operatorname{conv} X_F + \operatorname{pos} Y_F$.*

Proof. The equivalence of (a) and (b) in Theorem 4.1.2 shows that $F \supset \operatorname{conv} X_F + \operatorname{pos} Y_F$. Let $z \in F$; we complete the proof by showing that $z \in \operatorname{conv} X_F + \operatorname{pos} Y_F$.

As $z \in C$, by Theorem 1.1.20 we can write $z = \sum_{i=1}^{I} \lambda_i x_i + \sum_{j=1}^{J} \mu_j y_j$, with the x_i and y_j belonging to X and Y respectively, with the λ_i and μ_j strictly positive, and with $\sum_{i=1}^{I} \lambda_i = 1$. Let $E = \operatorname{conv}\{x_1, \ldots, x_I\} + \operatorname{pos}\{y_1, \ldots, y_J\}$ and define

$$U = \{u \in \mathbb{R}^I \mid u \geq 0,\ \sum_{i=1}^{I} u_i = 1\}, \qquad V = \mathbb{R}^J_+.$$

Then E is the image of $U \times V$ under the linear transformation represented by the matrix whose columns are $x_1, \ldots, x_I, y_1, \ldots, y_J$, and z is the image of the point $(\lambda_1, \ldots, \lambda_I, \mu_1, \ldots, \mu_J)$. As this point belongs to $\operatorname{ri}(U \times V)$ we have $z \in \operatorname{ri} E$ by Proposition 1.2.7. But $z \in F$, so by Proposition 4.2.3 $E \subset F$. It follows that $x_1, \ldots, x_I \in X_F$, and that the points $y_1, \ldots, y_J$ belong to $\operatorname{rc} \operatorname{cl} F$ (apply part (c) of Theorem 4.1.2 to $\operatorname{cl} F$). However, these points also belong to $\operatorname{rc} C$. By Proposition 4.2.3 we have $F = C \cap \operatorname{cl} F$, so in fact $y_1, \ldots, y_J$ belong to $\operatorname{rc} F$, and therefore to $Y \cap \operatorname{rc} F = Y_F$. Accordingly, $z \in \operatorname{conv} X_F + \operatorname{pos} Y_F$ as asserted. □

We now turn to the representation theorem mentioned at the beginning of this section. This theorem will permit us to represent a closed convex set containing no line as a certain combination of extreme points of the set and extreme rays of its recession cone. Corollary 4.3.5 extends this representation to general closed convex sets. Chapter 5 will show how to sharpen the version given here for a special class of sets called *polyhedral.*

As the theorem deals with sets that do not contain lines, we give first an equivalent way to describe such sets.

Lemma 4.3.2. *Let S be a nonempty subset of $\mathbb{R}^n$. If $\operatorname{bc} S$ has a nonempty interior then S contains no line. If S is closed and convex then the converse assertion holds also.*

Proof. Suppose that S contains a line. Let $x \in S$ and let y be a nonzero point of $\mathbb{R}^n$ such that $\{x+\mu y \mid \mu \in \mathbb{R}\} \subset S$. If $x^* \in \mathbb{R}^n$ and $\langle x^*, \cdot \rangle$ is bounded above on S, then $\langle x^*, y\rangle = 0$. Hence $\operatorname{bc} S$ lies in the $(n-1)$-dimensional subspace orthogonal to y, so $\operatorname{int}\operatorname{bc} S$ is empty, and this proves the first assertion.

For the converse, assume that S is closed and convex, and that it contains no line. Corollary 4.1.13 says that $(\operatorname{bc} S)^\circ = \operatorname{rc} S$. Taking polars, we find that $\operatorname{cl}\operatorname{bc} S = (\operatorname{rc} S)^\circ$. Let A be the affine hull of $\operatorname{bc} S$. Then A is a subspace, and it is also the affine hull of $\operatorname{cl}\operatorname{bc} S$. Therefore $(\operatorname{rc} S)^\circ \subset A$. If the dimension of A were $n-1$ or less, then by taking polars we could obtain

$$A^\perp = A^\circ \subset (\operatorname{rc} S)^{\circ\circ} = \operatorname{rc} S.$$

Then for each $x \in S$ we would have $x + A^\perp \subset S$. As the dimension of $A^\perp$ is at least 1, this contradicts the fact that S contained no line. Therefore A has dimension n, so the interior of $\operatorname{bc} S$ is nonempty. □

The next lemma provides the key step in the proof of the main theorem.

Lemma 4.3.3. *If C is a closed convex set in $\mathbb{R}^n$ having dimension at least 2 and containing no line, then each point of C is a convex combination of two points of $\operatorname{rb} C$.*

Proof. It suffices to show this for a point $c \in \operatorname{ri} C$, as any other points of C are already in $\operatorname{rb} C$. The hypothesis that C contains no line, together with Lemma 4.3.2, ensures that $\operatorname{bc} C$ has a nonempty interior. Let x^* be a nonzero point of this interior and define $E = C \cap H$, where H is the hyperplane consisting of all x with $\langle x^*, x\rangle = \langle x^*, c\rangle$. Suppose that $\operatorname{bc} C$ contains a ball of radius $\varepsilon > 0$ around x^*. If z is any nonzero point of $\operatorname{rc} C$ then $x^* + \varepsilon z/\|z\| \in \operatorname{bc} C$, so that we must have $\langle x^* + \varepsilon z/\|z\|, z\rangle \le 0$ because C is closed and therefore $\operatorname{rc} C = (\operatorname{bc} C)^\circ$ by Corollary 4.1.13. This implies $\langle x^*, z\rangle < 0$. Accordingly the recession cone $\operatorname{rc} H$ of H, which is the subspace M consisting of all z with $\langle x^*, z\rangle = 0$, meets $\operatorname{rc} C$ only in the origin. Corollary 4.1.5 now yields $\operatorname{rc} E = \operatorname{rc} C \cap \operatorname{rc} H = \{0\}$, which implies that E is bounded and therefore compact.

We have

$$[(\operatorname{par} C) \cap (\operatorname{par} H)]^\perp = (\operatorname{par} C)^\perp + (\operatorname{par} H)^\perp. \tag{4.18}$$

As $\operatorname{par} C$ has dimension at least 2 and $\operatorname{par} H$ has dimension $n-1$, the orthogonal complements on the right-hand side of (4.18) have dimensions $\gamma \le n-2$ and $\delta = 1$ respectively. The dimension of their sum is then not greater than $\gamma + \delta \le n-1$, so the subspace $M = (\operatorname{par} C) \cap (\operatorname{par} H)$ has dimension at least 1. Let w be any nonzero element of M. For each $\mu \in \mathbb{R}$ the points $c \pm \mu w$ belong to H. As $c \in \operatorname{ri} C$, for μ

close to zero they belong to C also, hence to E. But as E is compact, there are largest values μ_+ and μ_- such that $c_+ := c + \mu_+ w \in C$ and $c_- := c - \mu_- w \in C$. The points c_+ and c_- belong to the relative boundary of C (because otherwise we could increase either μ_+ or μ_- without leaving C), and c is a convex combination of these points. □

Here is the representation theorem.

Theorem 4.3.4. *Let C be a closed convex set in $\mathbb{R}^n$ containing no line. Let X be the set of extreme points of C, and let Y be a set of points of $\mathbb{R}^n$ such that* $\operatorname{pos} Y$ *is the set of extreme rays of C. Then $C = \operatorname{conv} X + \operatorname{pos} Y$.*

Proof. If the dimension of C is -1, 0, or 1 then the proof is immediate by enumeration of cases. Therefore assume that C has dimension $k > 1$ and, for induction, that the theorem holds for each closed convex set in $\mathbb{R}^n$ containing no line and having dimension less than k. We have $Y \subset \operatorname{rc} C$, so $C \supset \operatorname{conv} X + \operatorname{pos} Y$. To complete the proof we show that the opposite inclusion holds.

Lemma 4.3.3 shows that each point c of C can be written as a convex combination of two points of the relative boundary $\operatorname{rb} C$. Thus we need only show that any point w of $\operatorname{rb} C$ belongs to $\operatorname{conv} X + \operatorname{pos} Y$. To do so, apply Theorem 4.2.11 to show that w belongs to the relative interior of some proper face F of C. By Proposition 4.2.6 $\dim F \leq k-1$. Moreover, F is convex and by Proposition 4.2.3 it is closed. The induction hypothesis then says that F can be written as $\operatorname{conv} X' + \operatorname{pos} Y'$, where X' consists of the extreme points of F and Y' of the points that generate extreme rays of F. But $X' \subset X$ and $Y' \subset Y$, so $w \in \operatorname{conv} X + \operatorname{pos} Y$. □

Although we stated this theorem for sets containing no lines we can easily extend it to provide a structure result for general closed convex sets.

Corollary 4.3.5. *Let C be a closed convex subset of $\mathbb{R}^n$. Let $C' = C \cap (\operatorname{lin} C)^\perp$; define X to be the set of extreme points of C' and Y to be the set of elements y that generate extreme rays of C'. Then $C = \operatorname{conv} X + \operatorname{pos} Y + \operatorname{lin} C$.*

Proof. Using Proposition A.6.23, we can show that $C = C' + \operatorname{lin} C$. But by construction C' contains no line, so Theorem 4.3.4 shows that $C' = \operatorname{conv} X + \operatorname{pos} Y$. □

Corollary 4.3.6. *A compact convex subset of $\mathbb{R}^n$ is the convex hull of its extreme points.*

Proof. If the set is compact then it contains no line, so Theorem 4.3.4 applies; it also has no extreme rays, so the representation in the theorem reduces to $\operatorname{conv} X$. □

In infinite-dimensional spaces the statement of Corollary 4.3.6 has to be modified: a compact convex subset of a locally convex topological vector space is the *closed* convex hull of its extreme points (that is, the intersection of all closed convex sets containing the set of extreme points). This is the Krein-Milman theorem.

4.3.1 Exercises for Section 4.3

Exercise 4.3.7. Let X and Y be subsets of $\mathbb{R}^n$. Show that $\operatorname{conv} X + \operatorname{pos} Y$ is the smallest convex set C such that C contains X and the set of recession directions of C contains Y.

Exercise 4.3.8. Consider the sets

$$X = \left\{ \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\}, \qquad Y = \left\{ \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}.$$

Let $C = \operatorname{conv} X + \operatorname{pos} Y$. By considering the origin (a point of C), show that the representation of a point in the minimal expansion given by Theorem 4.3.4 need not be unique.

4.4 Notes and references

Chapter 5
Polyhedrality

This chapter analyzes a class of convex sets that is extremely important in applications, and that also has many nice properties not shared by general convex sets. It turns out that these sets can be described in two different ways, and the fact that these two descriptions really refer to the same class of sets is the central result of the chapter. The first section introduces these sets and examines some of their properties.

5.1 Polyhedral convex sets

Definition 5.1.1. A convex subset C of $\mathbb{R}^n$ is *polyhedral* if C is the intersection of a finite collection of closed halfspaces. □

A convenient way of representing a polyhedral convex set in $\mathbb{R}^n$ is as the solution set of some system of linear inequalities $Ax \leq a$, where A is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ and $a \in \mathbb{R}^m$. We use the symbol $\leq$ to express the partial order on $\mathbb{R}^m$ that defines $u \leq v$ to mean that for each i, $u_i \leq v_i$. For a matrix A and a subset I of the set $\mathscr{A}$ indexing the rows of A, A_I denotes the submatrix constructed of the rows of A whose indices belong to I. We write cI for the complement of I in $\mathscr{A}$.

Remark 5.1.2. If we translate a polyhedral convex set by some amount, the result is still polyhedral. To see this, suppose that the set is $P := \{x \mid Ax \leq a\}$, where $A \in \mathbb{R}^{m\times n}$. If $v \in \mathbb{R}^n$ then

$$P - v = \{x - v \mid Ax \leq a\} = \{y \mid Ay \leq a - Av\},$$

which is also polyhedral.

The same property holds for finitely generated sets, but it is easier to see. If X and Y are finite subsets of $\mathbb{R}^n$ with X nonempty, and if $F = \text{conv}(X,Y)$, then for each $v \in \mathbb{R}^n$ we have $F - v = \text{conv}(X - v, Y)$. □

Section 4.2 established general properties of faces, but we will need to use some special properties that for the most part hold only for faces of polyhedral convex sets. We first introduce some convenient tools for handling such faces.

Suppose we represent a nonempty polyhedral convex subset C of $\mathbb{R}^n$ by writing

$$C = \{x \mid Ax \leq a\}, \tag{5.1}$$

where A is an $m \times n$ matrix and $a \in \mathbb{R}^m$. If I is a subset of $\{1,\ldots,m\}$ then we write A_I and a_I for the matrix and vector formed from those rows of A and of a respectively, whose indices belong to I. We define cI to be $\{1,\ldots,m\}\backslash I$ and use the convention just introduced to define the matrix A_{cI} and a_{cI}. If either I or cI is empty then the associated matrix and vector are vacuous.

We will examine subsets of C having the form

$$C_I := \{x \mid A_I x = a_I, \quad A_{cI} x \leq a_{cI}\} \tag{5.2}$$

for some index set $I \subset \{1,\ldots,m\}$. We make the convention that if one of I or cI is empty, then the requirements pertaining to it in an expression like (5.2) are vacuous (that is, they do not apply).

Definition 5.1.3. Let C be a nonempty polyhedral convex subset of $\mathbb{R}^n$. Let $A \in \mathbb{R}^{m\times n}$ and $a \in \mathbb{R}^m$ be such that (5.1) holds, and suppose that I is a subset of $\{1,\ldots,m\}$. The set I is an *exact index set* for C with the representation (5.1) if there exists a point $x_I \in \mathbb{R}^n$ satisfying

$$A_I x_I = a_I, A_{cI} x_I < a_{cI}. \tag{5.3}$$

We write $\mathscr{E}_C$ for the collection of index sets that are exact for C with the representation (5.1). If I is exact then in particular $x_I \in C_I$.

Example 5.1.4. Suppose we take $C = \mathbb{R}_+$ and begin with the representation

$$C = \{x \mid [-1]x \leq 0\}. \tag{5.4}$$

For this representation there are two index sets, $\emptyset$ and $\{1\}$. If we take $I = \emptyset$ in (5.2) then

$$cI = \{1\}, \quad C_\emptyset = \{x \mid [-1]x \leq 0\} = \mathbb{R}_+,$$

and we can take any positive real number for $x_\emptyset$.

On the other hand, if $I = \{1\}$ then

$$cI = \emptyset, \quad C_{\{1\}} = \{x \mid [-1]x = 0\} = \{0\}, \quad x_{\{1\}} = 0.$$

Each of the two index sets is exact for C. In each case we used the vacuity of a system of equations or inequalities corresponding to the empty index set.

For additional perspective, we may represent the same set $C = \mathbb{R}_+$ in the form

$$C = \left\{x \mid \begin{bmatrix} -1 \\ 0 \end{bmatrix} x \leq \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right\}. \tag{5.5}$$

We now have four index sets: $\emptyset$, $\{1\}$,$\{2\}$,and $\{1,2\}$. These generate the following sets C_I. Accompanying each C_I is an example of x_I if one exists.

- $C_\emptyset = \mathbb{R}_+$; not exact; $x_\emptyset$ does not exist;
- $C_{\{1\}} = \{0\}$; not exact; $x_{\{1\}}$ does not exist;
- $C_{\{2\}} = \mathbb{R}_+$; exact; $x_{\{2\}}$ can be any positive number;
- $C_{\{1,2\}} = \{0\}$; exact; $x_{\{1,2\}} = 0$.

We see in both parts of this example that the exact index sets are in one-to-one correspondence with the nonempty faces of C, and that for exact sets the points x_I can be any elements of the relative interior of the face in question. The next theorem shows that this will be the case for each nonempty polyhedral convex C. □

Theorem 5.1.5. *Let C be a nonempty polyhedral convex subset of $\mathbb{R}^n$,and suppose that A is an $m \times n$ matrix such that for some $a \in \mathbb{R}^m$,*

$$C = \{x \mid Ax \leq a\}. \tag{5.6}$$

Let $\mathscr{E}$ be the collection of exact index sets for the representation (5.6) *of C. Define $\mathscr{F}'_C = \mathscr{F}_C \backslash \emptyset$. For $F \in \mathscr{F}'_C$ let*

$$\gamma(F) = \{i \mid \textit{for each } x \in F,\ A_i x = a_i\}, \tag{5.7}$$

and for $I \in \mathscr{E}$ let

$$\phi(I) = \{x \mid A_I x = a_I,\ A_{cI} x \leq a_{cI}\}. \tag{5.8}$$

Then the following results hold.

a. The functions γ and ϕ provide a one-to-one correspondence between the collections $\mathscr{F}'_C$ of nonempty faces of C and $\mathscr{E}$ of exact index sets for its representation (5.6).
b. If $F \in \mathscr{F}'_C$ and $I = \gamma(F)$ then

$$F = \{x \mid A_I x = a_I,\ A_{cI} x \leq a_{cI}\}, \tag{5.9}$$
$$\operatorname{aff} F = \{x \mid A_I x = a_I\}, \tag{5.10}$$
$$\operatorname{par} F = \ker A_I, \tag{5.11}$$
$$\operatorname{ri} F = \{x \mid A_I x = a_I,\ A_{cI} x < a_{cI}\}, \tag{5.12}$$
$$\operatorname{rc} F = \{z \mid A_I z = 0,\ A_{cI} z \leq 0\}, \tag{5.13}$$
$$\operatorname{lin} F = \ker A. \tag{5.14}$$

Thus each face of C is polyhedral, as is the recession cone of that face, and the lineality space of each face is that of C.

The correspondence in part (a) of Theorem 5.1.5 applies to nonempty faces only. The empty face of C cannot be of the form $\phi(I)$ for $I \in \mathscr{E}$: if it were, then no point x would satisfy

$$A_I x = a_I, \qquad A_{cI} x \leq a_{cI},$$

whereas the definition of $\mathscr{E}$ requires that there be a point x_I satisfying the stronger condition

$$A_I x_I = a_I, \qquad A_{cI} x_I < a_{cI}.$$

On the other hand, Example 5.1.4 showed that the empty index set may or may not be in $\mathscr{E}$, depending upon the representation used.

Proof. We establish the one-to-one correspondence by proving first that

$$\gamma : \mathscr{F}'_C \to \mathscr{E}, \qquad \phi : \mathscr{E} \to \mathscr{F}'_C \tag{5.15}$$

and then that

$$\gamma \circ \phi = \mathrm{id}_{\mathscr{E}}, \qquad \phi \circ \gamma = \mathrm{id}_{\mathscr{F}'_C}. \tag{5.16}$$

Proof of (5.15)**.**
Choose $F \in \mathscr{F}'_C$ and define $I := \gamma(F)$. F is nonempty, and for each point $x \in F$ and each $i \in I$ we have $A_i x = a_i$. If $I = \{1, \ldots, m\}$ then $I \in \mathscr{E}$ and $cI = \emptyset$; define x_I to be any point of F.

If $cI \neq \emptyset$, then for each $j \in cI$ there is a point $x_j \in F$ with $A_j x_j < a_j$: if not, then j would be in I. Define

$$x_I = |cI|^{-1} \sum_{j \in cI} x_j :$$

that is, x_I is the average of the points x_j. Since each x_j is in F and the average is a convex combination, $x_I \in F$. Then $A_I x_I = a_I$ and $A_{cI} x_I < a_{cI}$, so that I is exact and is therefore in $\mathscr{E}$. It follows that $\gamma : \mathscr{F}'_C \to \mathscr{E}$, so we have shown that for each $F \in \mathscr{F}'_C$ the set I:=$\gamma(F)$ is an exact index set.

We have also shown that for each such F not only is $I := \gamma(F)$ exact, but also there is a point x_I contained in F for which $A_I x_I = a_I$ and $A_{cI} x_I < a_{cI}$. This will be useful later in the proof.

Now let $I \in \mathscr{E}$ and let $F = \phi(I)$. The form of $\phi(I)$ shows that F is a convex subset of C. As I is exact, some point x_I satisfies (5.3), so that it belongs to $\phi(I)$ and therefore F is nonempty. If x and x' belong to C and $\mu \in (0,1)$ with $x'' := (1-\mu)x + \mu x' \in F$, then

$$0 = A_I x'' - a_I = (1-\mu)[A_I x - a_I] + \mu [A_I x' - a_I].$$

Each of the two quantities in square brackets on the right is nonpositive; as their convex combination is zero, each must be zero. Therefore x and x' belong to F, which must then be a face of C and hence in $\mathscr{F}'_C$. This shows that $\phi : \mathscr{E} \to \mathscr{F}'_C$, and completes the proof of (5.15).

Proof of (5.16)**.**
We need to show that $\gamma \circ \phi = \mathrm{id}_{\mathscr{E}}$. Let $I \in \mathscr{E}$ and define $G := \phi(I)$ and $H := \gamma(G) = (\gamma \circ \phi)(I)$. By what we have already shown, $G \in \mathscr{F}'_C$ and $H \in \mathscr{E}$. We will show that $H = I$. As I was any element of $\mathscr{E}$, this will prove that $\gamma \circ \phi = \mathrm{id}_{\mathscr{E}}$.

For each $i \in I$ and each $x \in G$ we have $A_i x = a_i$, so $i \in \gamma(G) = H$, and therefore $I \subset H$. If $I = \{1, \ldots, m\}$ then $H \subset I$ so $H = I$. Otherwise, cI is nonempty and, as I

is exact, G contains an element x_I with $A_{cI}x_I < a_{cI}$. Then no element of cI can be in $\gamma(G) = H$, so that $H \subset I$ and therefore $H = I$, Therefore $(\gamma \circ \phi) = id_{\mathscr{E}}$.

It remains to prove the second part of (5.16), which says that $\phi \circ \gamma = id_{\mathscr{F}'_C}$. Let $F \in \mathscr{F}'_C$; define $H = \gamma(F)$ and $J := \phi(H)$. We will show that $J = F$.

Let $y \in F$. For each index $i \in H$ we have $A_i y = a_i$, so $A_H y = a_H$. We also have $A_{cH}y \leq a_{cH}$. This means that $y \in \phi(H)$. As y was any point of F, we have $F \subset \phi(H) = (\phi \circ \gamma)(F)$.

Next, let $x \in \phi(H)$; then $A_H x = a_H$ and $A_{cH}x \leq a_{cH}$. As remarked earlier, because $H = \gamma(F)$ it is exact and there is a point $x_H \in F$ such that $A_H x_H = a_H$ and $A_{cH}x_H < a_{cH}$. Now for small positive ε let $x' := (1+\varepsilon)x_h - \varepsilon x$. We have

$$a - Ax' = a - A[(1+\varepsilon)x'_H - \varepsilon x] = (1+\varepsilon)[a - Ax_H] - \varepsilon[a - Ax].$$

If we recall that $A_H x = a_H = A_H x_H$ and use the definition of x', we see that $[a - Ax']_H = 0$. Also, $[a - Ax_H]_H > 0$ and $[a - Ax]_{cH} \geq 0$. Then if we make ε small enough, we have $[a - Ax']_{cH} > 0$, showing that then $x' \in C$. We have

$$F \ni x_H = (1+\varepsilon)^{-1}x' + \varepsilon(1+\varepsilon)^{-1}x;$$

both x' and x are in C; and F is a face of C; thus $x \in F$. As x was any element of $\phi(H)$, we see that $\phi(H) \subset F$ and therefore that

$$F = \phi(H) = (\phi \circ \gamma)(F),$$

so that $\phi \circ \gamma = id_{F'_C}$, which completes the proof of (5.16).

The representation (5.8) shows that each nonempty face of C is polyhedral, and the empty face is trivially polyhedral, so we have established part (a) of the theorem.

Proof of (5.10) **and** (5.11)**.**
Now let $F \in \mathscr{F}'_C$ and define $I := \gamma(F)$. We showed in part (a) that $\phi(I) = F$, so

$$F = \{x \mid A_I x = a_I, \quad A_{cI}x \leq a_{cI}\}. \tag{5.17}$$

The set $S := \{x \mid A_I x = a_I\}$ is affine and contains F, so it contains $\operatorname{aff} F$. Next, let $x \in S$. As I is exact there is a point $x_I \in F$ with $A_I x_I = a_I$ and $A_{cI}x_I < a_{cI}$. For small positive ε let $x' = x_I - \varepsilon(x - x_I)$. If ε is small enough then $x' \in F$. But then

$$x = \varepsilon^{-1}(1+\varepsilon)x_I - \varepsilon^{-1}x',$$

which shows that x is an affine combination of two points of F and therefore is in $\operatorname{aff} F$, so that $S \subset \operatorname{aff} F$. Then $S = \operatorname{aff} F$, which establishes (5.10). The expression shown for (5.11) follows directly because it defines a subspace that is parallel to $\operatorname{aff} F$.

Proof of (5.12)**.**
For (5.12), let $R = \{x \mid A_I x = a_I,\ A_{cI}x < a_{cI}\}$. Let $x \in R$ and $x' \in \operatorname{aff} F$. Then (5.10) shows that $A_I x' = a_I$. If x' is close enough to x then we have $A_{cI}x' < a_{cI}$, and then (5.17) shows that $x' \in F$. Therefore $x \in \operatorname{ri} F$, and as x was any point of R we have

$R \subset \operatorname{ri} F$. As $\operatorname{ri} F \subset F$, to show that $\operatorname{ri} F \subset R$ it suffices to show that if $x \in F \backslash R$ then $x \notin \operatorname{ri} F$. This choice of x implies that it satisfies (5.17) but that for some $j \in cI$ we have $A_j x = a_j$. Now for positive μ let $x' = x + \mu(x - x_I)$. This x' is an affine combination of x and x_I, so it is in $\operatorname{aff} F$. However,

$$(Ax' - a)_j = (1+\mu)(Ax - a)_j - \mu(Ax_I - a)_j = -\mu(Ax_I - a)_j > 0,$$

so that x' is not in C, hence certainly not in F. This shows that there are points of $\operatorname{aff} F$ arbitrarily close to x that do not belong to F, so that x cannot be in $\operatorname{ri} F$. Therefore $\operatorname{ri} F \subset R$, so that $\operatorname{ri} F = R$ and we have proved (5.12).

Proof of (5.13) **and** (5.14)**.**
To prove (5.13) and (5.14), it suffices to use (5.17) together with the fact that the recession cone of a nonempty set defined by a finite set of linear equations and inequalities is the set obtained by replacing the right-hand sides of the equations and inequalities by zero, and the lineality space is the set obtained by additionally changing all the inequalities to equations (Exercise 4.1.25). In particular, (5.13) shows that $\operatorname{rc} C$ is polyhedral, and so is $\operatorname{lin} C$ because any subspace is polyhedral. This establishes part (b) and completes the proof. □

Remark 5.1.6. The information developed in Theorem 5.1.5 can be very helpful in finding out where a given point is in relation to the structure of the set C. For example, suppose one has the coordinates of a point x and that $Ax \leq a$ so that this point belongs to C. Let I be the set of indices i for which $A_i x = a_i$, and define $cI := \{1, \ldots, m\} \backslash I$. For each $j \in cI$ we must have $A_j x < a_j$, so I is the exact index set associated with the unique face F whose relative interior contains x. Moreover, we can write F explicitly as $\{y \in \mathbb{R}^n \mid A_I y = a_I,\ A_{cI} y \leq a_{cI}\}$.

Theorem 5.1.5 is also very useful in establishing properties of polyhedral convex sets. The next theorem illustrates how one can use it to determine the form of the normal cone on faces of such a set.

Theorem 5.1.7. *Let C be a nonempty polyhedral convex subset of $\mathbb{R}^n$,and suppose that A is an $m \times n$ matrix such that for some $a \in \mathbb{R}^m$,*

$$C = \{x \mid Ax \leq a\}.$$

Let F be a nonempty face of C and define I to be the exact index set associated with F by part (a) of Theorem 5.1.5. If I is empty then $F = C$ and C has a nonempty interior, on which $N_C(\cdot)$ is identically zero. If I is nonempty then for each $f \in F$

$$A_I^*(\mathbb{R}_+^{|I|}) \subset N_C(f), \tag{5.18}$$

with equality if and only if $f \in \operatorname{ri} F$.

Proof. If I is empty then there is a point $x \in F$ with $Ax < a$. The strict inequality shows that $x \in \operatorname{int} C$, so C has a nonempty interior containing x. As $x \in F$, part (a) of Proposition 4.2.3 says that $C \subset F$. But $C \supset F$ so in fact $F = C$. The definition of the normal cone shows that it is zero on the interior of C.

Suppose then that I is nonempty, and write K for the cone $A_I^*(\mathbb{R}_+^{|I|})$. We show next that if $f \in F$ then $K \subset N_C(f)$, which will establish (5.18). Suppose that $k \in K$: then there is some $v \in \mathbb{R}_+^{|I|}$ with $k = vA_I$. For each $c \in C$ we have

$$\langle k, c-f\rangle = \langle vA_I, c-f\rangle = \langle v, A_Ic - A_If\rangle = \langle v, A_Ic - a_I\rangle \leq 0, \tag{5.19}$$

where we used the facts that $A_Ic \leq a_I$, $A_If = a_I$, and $v \geq 0$. As k was any element of K, (5.19) shows that $K \subset N_C(f)$.

To prove the final assertion, suppose first that $f \in \operatorname{ri} F$. Part (b) of Theorem 5.1.5 shows that $A_{cI}f < a_{cI}$. Suppose that $g \in \mathbb{R}^n \backslash K$. Then the Farkas lemma provides some $w \in \mathbb{R}^n$ with $A_Iw \leq 0$ and $\langle g, w\rangle > 0$. If ε is a sufficiently small positive number then

$$A_I(f+\varepsilon w) = a_I + \varepsilon A_Iw \leq a_I, \qquad A_{cI}(f+\varepsilon w) < a_{cI},$$

so that $f + \varepsilon w \in C$. However,

$$\langle g, (f+\varepsilon w) - f\rangle = \varepsilon\langle g, w\rangle > 0,$$

and this shows that $g \notin N_C(f)$. Accordingly, if $f \in \operatorname{ri} F$ then $K \supset N_C(f)$. The first part of the proof showed that $K \subset N_C(f)$, so $K = N_C(f)$.

In the remaining case f belongs to the relative boundary of F, and then by Part (b) of Theorem 5.1.5 there must be some index $j \in cI$ with $A_jf = a_j$. For each $c \in C$ we have $A_jc \leq a_j$, so that $\langle A_j, c-f\rangle \leq 0$ and therefore $A_j \in N_C(f)$.

Now suppose that A_j were a linear combination of the rows of A_I, so that for some $y \in \mathbb{R}^{|I|}$ $A_j = yA_I$. Then for each $f' \in F$ we would have

$$A_jf' = yA_If' = \langle y, a_I\rangle,$$

so that the value of A_jf' would not depend on the choice of $f' \in F$. But we know that $A_jf = a_j$, and Part (b) of Theorem 5.1.5 shows that whenever $f' \in \operatorname{ri} F$ we have $A_jf' < a_j$, so that no such y can exist. As every element of K is a linear combination of the rows of A_I, this shows that $A_j \notin K$ so that $K \not\supset N_C(f)$ and therefore $K \neq N_C(f)$. □

As each point of C belongs to the relative interior of some face of C (Theorem 4.2.11), Theorem 5.1.7 gives a complete description of the normal cone of C at each of its points as a finitely generated convex set.

The next proposition is useful in visualizing the polarity relationship for polyhedral sets. We will also need it for the proof of the Minkowski-Weyl theorem.

Proposition 5.1.8. *Let $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$, write e_p for the element of $\mathbb{R}^p$ having each component equal to 1, and let 0_n and 0_q be zero vectors of dimensions n and q respectively. Define*

$$P := \left\{ u \;\middle|\; \begin{bmatrix} 0_n^* \\ X^* \\ Y^* \end{bmatrix} u \leq \begin{bmatrix} 1 \\ e_p \\ 0_q \end{bmatrix} \right\} \tag{5.20}$$

and

$$F := \left\{ v \;\middle|\; \text{for some } \alpha \in \mathbb{R}_+,\ s \in \mathbb{R}^p_+, \text{ and } t \in \mathbb{R}^q_+,\ \begin{bmatrix} 0_n & X & Y \\ -1 & -e_p^* & 0_q^* \end{bmatrix} \begin{bmatrix} \alpha \\ s \\ t \end{bmatrix} = \begin{bmatrix} v \\ -1 \end{bmatrix} \right\}. \tag{5.21}$$

Then we have:

(a) A nonempty convex subset of $\mathbb{R}^n$ containing the origin can be written in the form (5.20) *if and only if it is polyhedral.*
(b) A nonempty convex subset of $\mathbb{R}^n$ containing the origin can be written in the form (5.21) *if and only if it is finitely generated.*
(c) For each instance of X and Y, the sets shown in (5.20) *and* (5.21) *contain the origin, are closed, and are mutually polar.*

Proof. Proof of (a): (if). Let $S := \{x \mid Ax \leq a\}$ be a nonempty polyhedral convex subset of $\mathbb{R}^n$ containing the origin. As $0 \in S$, no element of a can be negative. Let H be the set of indices i such that a_i is positive, and I the set for which a_i is zero. Let $p := |H|$ and $q := |I|$. For each $h \in H$ create a row of X^* whose content is $a_h^{-1} A_h$, and for each $i \in I$ create a row of Y^* whose content is A_i. Then $X^* \in \mathbb{R}^{p \times n}$, $Y^* \in \mathbb{R}^{q \times n}$, and

$$\left\{ u \;\middle|\; \begin{bmatrix} 0_n^* \\ X^* \\ Y^* \end{bmatrix} u \leq \begin{bmatrix} 1 \\ e_p \\ 0_q \end{bmatrix} \right\} = S.$$

Proof of(a): (only if). The set defined by (5.20) is clearly polyhedral; then so is any set equal to it.

Proof of (b): (if). Let M and N be finite subsets of $\mathbb{R}^n$ with M nonempty, and define a nonempty, convex, finitely generated subset of $\mathbb{R}^n$ by $T := \operatorname{conv} M + \operatorname{pos} N$. Assume that T contains the origin, and let $p := |M|$ and $q := |N|$. If we then define matrices $X \in \mathbb{R}^{p \times n}$ whose columns are the elements of M, and $Y \in \mathbb{R}^{q \times n}$ whose columns are the elements of N, we can express T as

$$T := \left\{ v \;\middle|\; \text{for some } s \in \mathbb{R}^p_+, \text{ and } t \in \mathbb{R}^q_+,\ \begin{bmatrix} X & Y \\ -e_p^* & 0_q^* \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = \begin{bmatrix} v \\ -1 \end{bmatrix} \right\}. \tag{5.22}$$

However, this is not quite what we want, as it lacks the first column of the matrix in (5.21). If we restore that column, we have the set

$$T' := \left\{ v \;\middle|\; \text{for some } \alpha \geq 0,\ s \in \mathbb{R}^p_+, \text{ and } t \in \mathbb{R}^q_+,\ \begin{bmatrix} 0 & X & Y \\ -1 & -e_p^* & 0_q^* \end{bmatrix} \begin{bmatrix} \alpha \\ s \\ t \end{bmatrix} = \begin{bmatrix} v \\ -1 \end{bmatrix} \right\}. \tag{5.23}$$

If we hold α at zero then (5.23) produces T, so T can be written in the form (5.21).

Proof of (b): (only if).

The form of T' shows that it is $\operatorname{conv}(X \cup \{0\}) + \operatorname{pos} Y$, so any set representable by it is finitely generated.

Proof of (c).

First suppose $u \in P$ and $v \in F$, and write D for the matrix in (5.21). As $u \in P$ we have from (5.20)

$$D^* \begin{bmatrix} u \\ 1 \end{bmatrix} \leq 0. \tag{5.24}$$

Using this with (5.21) we find that

$$\langle u^*, v \rangle - 1 = \begin{bmatrix} u^* & 1 \end{bmatrix} \begin{bmatrix} v \\ -1 \end{bmatrix} = \begin{bmatrix} u^* & 1 \end{bmatrix} D \begin{bmatrix} \alpha \\ s \\ t \end{bmatrix} = \left\langle D^* \begin{bmatrix} u \\ 1 \end{bmatrix}, \begin{bmatrix} \alpha \\ s \\ t \end{bmatrix} \right\rangle \leq 0,$$

because α, s and t are nonnegative. As u and v were arbitrary elements of the two sets, we have $F \subset P^\circ$.

To show that $P^\circ \subset F$, suppose that $v \notin F$; then the equation in (5.21) has no solution in nonnegative variables α, s and t. By the Farkas lemma there exist $r \in \mathbb{R}^{n+1}$ and $\rho \in \mathbb{R}$ such that

$$D^* \begin{bmatrix} r \\ \rho \end{bmatrix} \leq 0, \quad \begin{bmatrix} v^* & -1 \end{bmatrix} \begin{bmatrix} r \\ \rho \end{bmatrix} > 0. \tag{5.25}$$

The two inequalities in (5.25) yield

$$\rho \geq 0, \quad X^* r - e_p \rho \leq 0, \quad Y^* r \leq 0, \quad \langle v^*, r \rangle > \rho. \tag{5.26}$$

First suppose that $\rho = 0$; then the second and third inequalities in (5.26) show that $r \in \operatorname{rc} P$. As $0 \in P$, the halfline $\{0 + \alpha r \mid \alpha \in \mathbb{R}_+\}$ lies in P. However, the fourth inequality in (5.26) shows that then the inner product $\langle v^*, \alpha r \rangle$ takes arbitrarily large values for $\alpha \in \mathbb{R}_+$, so in this case $v \notin P^\circ$.

In the other case $\rho > 0$, and then as (5.25) is positively homogeneous in the pair (r, ρ) we can suppose that $\rho = 1$. The second and third inequalities in (5.26) then show that $r \in P$, but the fourth inequality shows that $\langle v^*, r \rangle > 1$, so that in this case also $v \notin P^\circ$. Therefore $P^\circ \subset F$, so in fact $F = P^\circ$. By taking polars we then find that $F^\circ = P^{\circ\circ} = P$, so that P and F are mutually polar. □

5.2 The Minkowski-Weyl theorem and consequences

We have seen that polyhedral convex sets containing the origin have finitely generated polars, and vice versa. This section proves more: namely, that a convex set is polyhedral if and only if it is finitely generated. This is often called the Minkowski-Weyl theorem, though the terminology varies with different authors. We prove the theorem in the following section, and then present some consequences in Section 5.2.2.

5.2.1 The Minkowski-Weyl theorem

Here is the main theorem about polyhedral and finitely generated convex sets.

Theorem 5.2.1. *If C is a convex subset of $\mathbb{R}^n$, the following are equivalent:*

a. C is polyhedral;
b. C is closed and has finitely many faces;
c. C is finitely generated.

Proof. If C is empty there is nothing to prove, so we assume that C is nonempty. If C does not contain the origin, then we translate it to a set that does, prove the equivalence for that set, and translate back to C. Remark 5.1.2 shows that the equivalence holds for C if and only if it holds for the translated set. Therefore there is no loss of generality in supposing that C contains the origin.

(a) implies (b). Suppose C is polyhedral. The definition of polyhedrality shows that C is closed, and Theorem 5.1.5 shows that it has only finitely many faces.

(b) implies (c). Suppose C is closed and has finitely many faces. Let $L = \operatorname{lin} C$ and $C' = C \cap L^{\perp}$; then Proposition A.6.23 shows that $C = L \oplus C'$. Corollary 4.3.5 shows that C is the sum of L and the generalized convex hull of the extreme points and extreme rays of C', and Exercise 4.2.21 showed that the faces of C' are in one-to-one correspondence with those of C. Extreme points are faces of dimension 0, and extreme rays are faces of dimension 1. As C has only finitely many faces, the sets of extreme points and extreme rays of C' are finite, so that C' is finitely generated and therefore so is $C' + L = C$.

(c) implies (a). Now suppose C is finitely generated; then by Proposition 1.1.22 it is closed. As it contains the origin, Proposition 5.1.8 shows that C° is polyhedral. As a polar set, it must contain the origin. We have already shown that hypothesis (a) implies (c) so as a polyhedral convex set containing the origin C° must be finitely generated. Another application of Proposition 5.1.8 shows that $C^{\circ\circ}$ is polyhedral. But as C is closed $C^{\circ\circ} = C$, so C is polyhedral. □

5.2.2 Applications

Theorem 5.2.1 has many useful consequences. Here are a few of them.

Corollary 5.2.2. *Let L be a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. If C is a polyhedral convex subset of $\mathbb{R}^n$ then $L(C)$ is polyhedral; if D is a polyhedral convex subset of $\mathbb{R}^m$ then $L^{-1}(D)$ is polyhedral.*

Proof. As C is polyhedral, Theorem 5.2.1 shows that it is finitely generated. It is immediate that the linear image of a finitely generated set is finitely generated. Therefore $L(C)$ is finitely generated and therefore polyhedral.

As D is polyhedral we can represent it as $\{y \mid E(y) \leq e\}$ for some matrix E and vector e. Then $L^{-1}(D) = \{x \mid EL(x) \leq e\}$, which is polyhedral. □

It follows in particular that the sum of a finite collection of polyhedral convex subsets of $\mathbb{R}^n$ is polyhedral, and that the projection of a polyhedral convex set into a subspace is also polyhedral, because the projector on a subspace is linear.

Corollary 5.2.3. *Let C be a nonempty polyhedral convex subset of $\mathbb{R}^n$ and let $L : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. Then* $\operatorname{rc} L(C) = L(\operatorname{rc} C)$.

Proof. Theorem 5.2.1 shows that C is finitely generated, so that $C = \operatorname{conv} X + \operatorname{pos} Y$ for finite subsets X and Y of $\mathbb{R}^n$ with $X \neq \emptyset$, and therefore $\operatorname{rc} C = \operatorname{pos} Y$ by Proposition 4.1.7. But $L(C) = \operatorname{conv} L(X) + \operatorname{pos} L(Y)$, so

$$\operatorname{rc} L(C) = \operatorname{pos} L(Y) = L(\operatorname{pos} Y) = L(\operatorname{rc} C).$$

Corollary 5.2.4. *Let C and D be polyhedral convex subsets of $\mathbb{R}^n$. Then C and D can be strongly separated by a hyperplane if and only if they are disjoint.*

Proof. For strong separation it is necessary that C and D be disjoint. To see that this is also sufficient, recall that by Theorem 2.2.7 strong separation is possible if and only if the origin does not belong to the closure of $C - D$. If C and D are disjoint then the set $C - D$ does not contain the origin, and it is closed by the remark just preceding Corollary 5.2.3. □

Corollary 5.2.5. *If C is a nonempty polyhedral convex subset of $\mathbb{R}^n$ then C° is polyhedral.*

Proof. Consider the sequence C, C°, $C^{\circ\circ}$, and $C^{\circ\circ\circ}$ obtained by successive application of the polarity operation. We know that C° contains the origin, and by Theorem 2.2.5 that $C^{\circ\circ} = \operatorname{cl}\operatorname{conv}(C \cup \{0\})$. As C is polyhedral it is also finitely generated. If $C = \operatorname{conv}(X, Y)$ then $\operatorname{conv}(C \cup \{0\}) = \operatorname{conv}(X \cup \{0\}, Y)$, which is finitely generated and therefore closed, so $C^{\circ\circ}$ is finitely generated and contains the origin. By Proposition 5.1.8, $C^{\circ\circ\circ}$ is polyhedral and contains the origin. But this set is the same as C°. □

Corollary 5.2.6. *If C is a nonempty polyhedral convex subset of $\mathbb{R}^n$, then* $\operatorname{cl} H_C$ *is polyhedral.*

Proof. Theorem 4.2.17 showed that $\mathscr{F}(\operatorname{cl} H_C)$ is in one-to-one correspondence with $\mathscr{F}(C) \cup \mathscr{F}(\operatorname{rc} C)$. As both C and $\operatorname{rc} C$ are polyhedral, Theorem 5.2.1 shows that each has only finitely many faces. Then $\mathscr{F}(\operatorname{cl} H_C)$ is also a finite set, and an application of Theorem 5.2.1 shows that $\operatorname{cl} H_C$ is polyhedral. □

Corollary 5.2.7. *Suppose that C is a nonempty polyhedral convex subset of $\mathbb{R}^n$ that is representable as $C = \{x \mid Ax \leq a\}$, where $A \in \mathbb{R}^{m \times n}$ and $a \in \mathbb{R}^m$. Then*

$$\operatorname{rc} C = A^{-1}(\mathbb{R}^m_-), \qquad \operatorname{bc} C = A^*(\mathbb{R}^m_+), \tag{5.27}$$

so that $\operatorname{rc} C$ *and* $\operatorname{bc} C$ *are mutually polar and each is polyhedral.*

Proof. We have already established the first equality in (5.27): e.g., in Theorem 5.1.5 (use (5.13) with the exact index set $\emptyset$). Also, as C is closed we have from Corollary 4.1.13 that $(\mathrm{bc}\,C)^\circ = \mathrm{rc}\,C$. Using the expression for $\mathrm{rc}\,C$ given in (5.27) and defining K to be $A^*(\mathbb{R}^m_+)$ we find that

$$\mathrm{cl}\,\mathrm{bc}\,C = (\mathrm{bc}\,C)^{\circ\circ} = (\mathrm{rc}\,C)^\circ = [A^{-1}(\mathbb{R}^m_-)]^\circ = K, \tag{5.28}$$

so that $\mathrm{bc}\,C \subset K$. Choose any $x^* \in K$; then there is a nonnegative $v^* \in \mathbb{R}^m$ such that $x^* = A^*v^*$. For each element c of C we have $Ac \le a$; as v^* is nonnegative we have

$$\langle x^*, c\rangle = \langle A^*v^*, c\rangle = \langle v^*, Ac\rangle \le \langle v^*, a\rangle. \tag{5.29}$$

It follows that $\langle v^*, a\rangle$ is an upper bound on the inner product of x^* with any element of C, so that $x^* \in \mathrm{bc}\,C$. As x^* was an arbitrary element of K this shows that $K \subset \mathrm{bc}\,C$, so $\mathrm{bc}\,C = K = A^*(\mathbb{R}^m_+)$ as asserted. Then $\mathrm{bc}\,C$ is polyhedral, hence closed; the same holds for $\mathrm{rc}\,C$ as given in (5.27). Their mutual polarity follows from Proposition 3.2.1. □

Next we look at a question that is very important for applications: given a nonzero z^* for which $\langle z^*, \cdot\rangle$ is bounded above on a polyhedral convex subset C of $\mathbb{R}^n$, are there any maximizers and, if so, what one can say about the structure of the set of all such maximizers?

Polyhedrality makes this question interesting, because in general no such maximizers need exist even when the set C is closed and convex. For example, the problem of maximizing the function $-x_2$ over the closed convex subset C of $\mathbb{R}^2$ defined by $C = \{x \mid x_1 > 0,\ x_1x_2 \ge 1\}$ has no solutions.

A very nice property of *polyhedral* convex sets—and a central result in the theory of linear programming—is that there are always maximizers; in fact, the set of all these maximizers forms a nonempty face of C, as we now show.

Proposition 5.2.8. *Let C be a nonempty polyhedral convex subset of $\mathbb{R}^n$, and let z^* be a nonzero element of $\mathbb{R}^n$. If $\langle z^*, \cdot\rangle$ is bounded above on C, then the set of maximizers of $\langle z^*, \cdot\rangle$ on C is a nonempty face of C.*

Proof. As C is polyhedral, by Theorem 5.2.1 it is finitely generated. Therefore there are finite subsets $X = \{x_1, \ldots, x_P\}$ and $Y = \{y_1, \ldots, y_Q\}$ of points of $\mathbb{R}^n$, with X nonempty, such that $C = \mathrm{conv}\,X + \mathrm{pos}\,Y$. As $\langle z^*, \cdot\rangle$ is bounded above on C, we must have

$$\langle z^*, y_q\rangle \le 0, \qquad q = 1, \ldots, Q, \tag{5.30}$$

because if that were not so then for some such q we could make $\langle z^*, \cdot\rangle$ increase without bound along the halfline $\{x_1 + \mu y_q \mid \mu \ge 0\}$, which is a subset of C. In turn, (5.30) implies that $\langle z^*, y\rangle \le 0$ for each $y \in \mathrm{pos}\,Y$.

We know that every element of C is the sum of an element x of $\mathrm{conv}\,X$ and an element y of $\mathrm{pos}\,Y$. For any given x we can maximize $\langle z^*, x+y\rangle$ by choosing $y = 0$, so under our boundedness hypothesis the supremum of $\langle z^*, \cdot\rangle$ on C is the same as its supremum on $\mathrm{conv}\,X$. The latter value, say ζ, is attained (in fact, at some point

$x_p \in X$). Therefore $\langle z^*, \cdot \rangle$ attains its maximum value of ζ on C. It follows that the hyperplane $H_0(z^*, \zeta)$ contains C in its lower closed halfspace, and moreover that the intersection $F := C \cap H$ is nonempty. By Proposition 4.2.7, F is a face of C. □

5.2.3 Exercises for Section 5.2

Exercise 5.2.9. Let P and Q be nonempty convex subsets of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. Show that $P \times Q$ is polyhedral if and only if each of P and Q is polyhedral.

The next exercise establishes an extremely important local property of polyhedral convex sets.

Exercise 5.2.10. Let P be a polyhedral convex subset of $\mathbb{R}^n$ that is representable as $P = \{x \mid Ax \leq a\}$ for some matrix A and vector a. Show that if $x_0 \in P$ then there is a neighborhood Q of the origin in $\mathbb{R}^n$ such that $Q \cap (P - x_0) = Q \cap T_P(x_0)$.

5.3 Polyhedrality and structure

For a polyhedral convex cone, we can sharpen the conclusions of Theorem 4.2.18.

Lemma 5.3.1. *Let C be a nonempty polyhedral convex subset of $\mathbb{R}^n$. Then each nonempty face of C is exposed.*

Proof. Let F be a nonempty face of C. Apply Theorem 5.1.5 to produce a representation of C as $\{x \mid Ax \leq a\}$ and an exact index set I such that $F = \{x \mid A_I x = a_I,\ A_{cI} x \leq a_{cI}\}$. If I is empty then $F = C = N_C^{-1}(0)$, so F is exposed. If I is nonempty, choose a strictly positive element $y^* \in \mathbb{R}^{|I|}$ and set $\mu = \langle y^*, a_I \rangle$.

For any $x' \in C$,

$$\langle A_I^* y^*, x' \rangle = \langle y^*, A_I x' \rangle \leq \langle y^*, a_I \rangle = \mu,$$

while equality holds whenever $x' \in F$ because there $\langle A_I^* y^*, x' \rangle = \langle y^*, A_I x' \rangle = \mu$. Therefore the maximum of $\langle A_I^* y^*, \cdot \rangle$ on C is μ and it is attained everywhere on F.

If $x \in C \backslash F$ then let G be the unique face of C having x in its relative interior, and let J be the exact index set for G. If $I \subset J$ then $F \supset G$, which is impossible because $x \in G \backslash F$. Therefore there is some index $i \in I \backslash J$. As $x \in \operatorname{ri} G$ and the index i does not belong to J, we must have $A_i x < a_i$. As $y^* > 0$ we then have

$$\langle A_I^* y^*, x \rangle = \langle y^*, A_I x \rangle < \langle y^*, a_I \rangle = \mu,$$

so x is not a maximizer of $\langle A_I^* y^*, \cdot \rangle$ on C. Therefore F is exactly the set of maximizers; as this set is $N_C^{-1}(A_I^* y^*)$, F is exposed. □

Theorem 5.3.2. *Let K be a nonempty polyhedral convex cone in $\mathbb{R}^n$, and let F be a nonempty face of K. Write L for $\operatorname{par} F$. Then $F = K \cap L$ and for each $x \in \operatorname{ri} F$, $N_K(x)$ has the constant value $F^* = K^\circ \cap L^\perp$. This F^* is a face of K° with $\operatorname{par} F^* = L^\perp$ and, for each $x^* \in \operatorname{ri} F^*$, $F = N_{K^\circ}(x^*)$.*

Proof. We proved most of the assertions in Theorem 4.2.18. Two things remain, one of which is the final statement. Theorem 4.2.18 establishes this under the hypothesis that F is an exposed face of K. Given Lemma 5.3.1, that must be true here.

The other remaining assertion is that $\operatorname{par} F^* = L^\perp$. As $F^* = K^\circ \cap L^\perp$, we have $\operatorname{par} F^* \subset L^\perp$. For the opposite inclusion, we show first that if $x_0 \in \operatorname{ri} F$ then $\operatorname{lin} T_K(x_0) = L$. Represent C as $\{x \mid Ax \le a\}$, and find an index set I such that $F = \{x \mid A_I x = a_I,\ A_{cI} x \le a_{cI}\}$ with strict inequality holding for all indices in cI whenever $x \in \operatorname{ri} F$. Then $A_{cI} x_0 < a_{cI}$, so $T_K(x_0) = A_I^{-1}(\mathbb{R}_-^{|i|})$ and $\operatorname{lin} T_K(x_0) = \ker A_I = \operatorname{par} F = L$.

Now Exercise 4.1.27 says that $[\operatorname{par} N_K(x_0)]^\perp = \operatorname{lin}[T_K(x_0)]$. As we have shown that $\operatorname{lin} T_K(x_0) = L$, we have $\operatorname{par} N_K(x_0) = L^\perp$ as required. □

For a polyhedral set we can demonstrate a much closer relationship between the critical cone and critical face than that for general convex sets. In Equation (4.12) we showed that $\phi_C(z_0) - x_0 \subset \kappa_C(z_0)$, but noted that $\phi_C(z_0)$ might have smaller dimension than $\kappa_C(z_0)$. The following proposition shows that in the polyhedral case, not only are the dimensions the same but in fact $\phi_C(z_0) - x_0$ and $\kappa_C(z_0)$ coincide on some neighborhood of the origin, and each is obtainable from the other.

Proposition 5.3.3. *Let P be a nonempty polyhedral convex subset of $\mathbb{R}^n$ and let $z_0 \in \mathbb{R}^n$. Let $x_0 = \Pi_P(z_0)$.*

a. We have

$$\kappa_P(z_0) = \operatorname{cone}[\phi_P(z_0) - x_0], \qquad \phi_P(z_0) = [x_0 + \kappa_P(z_0)] \cap P. \tag{5.31}$$

b. There is a neighborhood Q of the origin such that

$$Q \cap [\phi_P(z_0) - x_0] = Q \cap \kappa_P(z_0). \tag{5.32}$$

In particular, $\phi_P(z_0)$ and $\kappa_P(z_0)$ have the same dimension.

Proof. We first prove (b). Part (c) of Exercise 5.2.10 shows that there is a neighborhood Q of the origin such that $Q \cap (P - x_0) = Q \cap T_P(z_0)$. If we define $x_0^* = z_0 - x_0$ then

$$\begin{aligned} Q \cap [\phi_P(z_0) - x_0] &= Q \cap (P - x_0) \cap \{v \mid \langle x_0^*, v\rangle = 0\} \\ &= Q \cap T_P(x_0) \cap \{v \mid \langle x_0^*, v\rangle = 0\} \\ &= Q \cap \kappa_P(z_0), \end{aligned}$$

which proves (5.32). For the assertion about dimension, apply Corollary 1.1.16 to obtain

$$\dim[\phi_P(z_0) - x_0] = \dim(Q \cap [\phi_P(z_0) - x_0])\, \dim[Q \cap \kappa_P(z_0)] = \dim \kappa_P(z_0).$$

As translation of a convex set does not affect its dimension, $\dim \phi_P(z_0) = \dim \kappa_P(z_0)$.

To prove (a), we first apply Proposition 3.3.6 to obtain $\operatorname{cone}(P - x_0) \subset T_P(x_0)$. Next, define L to be the subspace of elements $v \in \mathbb{R}^n$ such that $\langle z_0 - x_0, v\rangle = 0$. The stronger part of Exercise 3.1.19 then yields $[\operatorname{cone}(P-x_0)] \cap L = \operatorname{cone}[(P-x_0)\cap L]$. Then

$$\operatorname{cone}[\phi_P(z_0) - x_0] = \operatorname{cone}[(P-x_0)\cap L] = [\operatorname{cone}(P-x_0)]\cap L \subset T_P(x_0)\cap L = \kappa_P(z_0).$$

For the opposite inclusion, start with $v \in \kappa_P(z_0)$; then $v \in T_P(x_0)$, so part (b) says that for some positive μ, $\mu v \in P - x_0$. But $v \in L$ so $\mu v \in (P-x_0)\cap L = \phi_P(z_0) - x_0$, and then $v \in \operatorname{cone}[\phi_P(z_0) - x_0]$, which establishes the first part of (a).

For the second part of (a), we use the first part to establish that $\phi_P(z_0) \subset x_0 + \kappa_P(z_0)$. But $\phi_P(z_0) \subset P$, so in fact $\phi_P(z_0) \subset [x_0 + \kappa_P(z_0)] \cap P$. For the opposite inclusion, choose $x \in [x_0 + \kappa_P(z_0)] \cap P$ and for nonnegative μ let $x_\mu = x_0 + \mu(x - x_0)$. For $\mu \in [0,1]$ we have $x_\mu \in P$, and as $x - x_0 \in \kappa_P(z_0)$, also $x_\mu \in x_0 + \kappa_P(z_0)$. If we choose a sufficiently small positive μ then

$$x_\mu - x_0 \in \kappa_P(z_0) \cap Q = [\phi_P(z_0) - x_0] \cap Q,$$

implying that $x_\mu \in \phi_P(z_0)$. As x and x_0 belong to P and $\phi_P(z_0)$ is a face of P, both x and x_0 must belong to $\phi_P(z_0)$. Therefore $[x_0 + \kappa_P(z_0)] \cap P \subset \phi_P(z_0)$, so the second part of (a) holds. □

Polyhedral convex sets have a useful property not shared by general closed convex sets: unless they are affine, they have facets. We use the following definition, which does not agree with some others in the literature.

Definition 5.3.4. Let C be a convex subset of $\mathbb{R}^n$. A *facet* of C is a nonempty face of C having dimension $(\dim C) - 1$. □

In general, a closed convex set need not have a facet: for example, the Euclidean unit ball in $\mathbb{R}^2$ has faces of dimensions 0 and 2, but not 1. No affine set has a facet, because such sets have no nonempty proper faces. However, if a nonempty polyhedral convex set is not affine (that is, if it has a nonempty relative boundary), then it has a facet.

Theorem 5.3.5. *Let P be a polyhedral convex subset of $\mathbb{R}^n$ Then each point of the relative boundary of P belongs to some facet of P.*

Proof. If P has no relative boundary then there is nothing to prove; this case arises if P is either affine or empty. Otherwise, suppose x_0 belongs to $\operatorname{rb} P$; then the dimension k of P must be at least 1. If $k < n$ then apply Theorem A.6.20 to find an affine homeomorphism that preserves inner products and that maps $\operatorname{aff} P$ onto $\mathbb{R}^k$. We may therefore suppose that we are working in $\mathbb{R}^k$ and that the set P has full dimension. As then $x_0 \notin \operatorname{int} P$, Exercise 2.2.15 shows that P has a supporting hyperplane at x_0, so $N_P(x_0) \neq \{0\}$.

Exercise 5.2.10 shows that the tangent and normal cone operators associated with P have polyhedral values. Therefore the cones $T := T_P(x_0)$ and $N := N_P(x_0)$ are

polyhedral convex cones dual to each other. The cone N contains no line (otherwise P could not have full dimension) and it is not the origin. The fundamental representation theorem (Theorem 4.3.4) says that N is the convex hull of its extreme points and extreme rays. As N is a cone its only extreme point is the origin, and there are only finitely many extreme rays because each is a face of N, which is polyhedral. But as N is not the origin, there must be at least one extreme ray R; suppose it is generated by n_0^*.

Now apply Theorem 5.3.2 with $K = N$, $K^\circ = T$, and $F = R$, and let $z_0 = x_0 + n_0^*$. The critical cone $\kappa_P(z_0)$ is the face $F^* = T \cap \operatorname{span}(n_0^*)^\perp$ of T that is identified in the theorem. The critical face $\phi_P(z_0)$ is a face of P containing x_0 and having the same dimension as $\kappa_P(z_0)$. But the theorem tells us that $\operatorname{par}\kappa_P(z_0) = \operatorname{span}(n_0^*)^\perp$, so that $\phi_P(z_0)$ has dimension $k-1$ and is therefore a facet of P containing x_0. □

We can use Theorem 5.3.5 recursively to show that P has faces of every dimension from $\dim\operatorname{lin} P$ to $\dim P$ inclusive.

Corollary 5.3.6. *Let P be a polyhedral convex subset of $\mathbb{R}^n$ having dimension k. If $M := \operatorname{lin} P$ has dimension $m \in [0,k]$, then for each $j \in [m,k]$ P has a nonempty face F_j with $\dim F_j = j$ and no face of P has dimension outside the interval $[m,k]$. Moreover, F_j has the form $M + G_{j-m}$ where G_{j-m} is a face of $P \cap M^\perp$ having dimension $j-m$; one has*

$$P \cap M^\perp := G_{k-m} \supset \ldots \supset G_0, \qquad P := F_k \supset \ldots \supset F_m. \tag{5.33}$$

Proof. The hypothesis requires $k \geq 0$, so P cannot be empty. P is affine if and only if $m = k$, in which case we take $F_m = M$ and G_0 to be any point of P, and we are finished.

In the remaining case, $m < k$. If we write M for $\operatorname{lin} P$ then M is also the lineality space of each face of P (Exercise 4.2.19), so no face of P can have dimension less than m, and certainly no face can have dimension greater than k.

We have $P = M + Q$ with $Q = P \cap M^\perp$, and Exercise 4.2.21 shows that the faces F of P are in one-to-one correspondence with the faces G of Q through the maps $F = G + M$ and $G = F \cap M^\perp$. As $G \subset M^\perp$, if we write $q = \dim M$ then we have $\dim F = m + \dim G$ for each such pair; taking $F = P$ and $G = Q$, we have $k = m + q$.

By construction, Q contains no lines and has positive dimension, so $\operatorname{rb} Q$ is nonempty and Theorem 5.3.5 shows that Q has a facet G_{q-1} of dimension $q-1$. As Q contains no lines, this facet can be affine only if $q - 1 = 0$, whereas if $q - 1 > 0$ we can apply the theorem to G_{q-1} to obtain a facet G_{q-2} of G_{q-1}. But a face of a face is a face, so G_{q-2} is a face of Q having dimension $q-2$. We can continue this process until the face G_0 produced at the qth stage has dimension 0 (so that it is an extreme point of Q), and we then have a sequence $\{G_0, \ldots, G_q\}$ of faces of Q with $\dim G_i = i$. The first inclusion in (5.33) follows from this construction. Corresponding to this sequence is the sequence $\{F_m, \ldots, F_k\}$ of faces of P obtained by setting $F_{i+m} = M + G_i$, and the second inclusion in (5.33) then follows from the first. □

5.4 Polyhedral multifunctions

Definition 5.4.1. A *multifunction* F from a set X to a set Y is an operator associating with each point $x \in X$ a subset $F(x)$ of Y. The *graph* of F is the subset $\operatorname{gph} F$ of $X \times Y$ defined by

$$\operatorname{gph} F = \{(x,y) \in X \times Y \mid y \in F(x)\}.$$

Very often we will omit the label gph and just write F for $\operatorname{gph} F$, saying for example that $(x,y) \in F$ instead of $(x,y) \in \operatorname{gph} F$.

Such a multifunction F could be empty for some (or every) x. The points of X where it is nonempty, if any, form a subset $\operatorname{dom} F$ of $\mathbb{R}^n$ called the *effective domain* of F, and the points of y that belong to $F(x)$ for some x form the *image* of F, written $\operatorname{im} F$. Thus $\operatorname{dom} F$ and $\operatorname{im} F$ are the projections of $\operatorname{gph} F$ on X and Y respectively.

Definition 5.4.2. If F is a multifunction from X to Y, the *inverse* of F is the multifunction $F^{-1} : Y \to X$ defined by $(y,x) \in F^{-1}$ if and only if $(x,y) \in F$.

This definition implies that $(F^{-1})^{-1} = F$.

If $S \subset X$ then the *forward image* of S under F is

$$F(S) := \cup_{x \in S} F(x), \tag{5.34}$$

and if $T \subset Y$ then the *inverse image* of T under F is

$$F^{-1}(T) := \cup_{y \in T} F^{-1}(y), \tag{5.35}$$

so that the inverse image of T under F is the forward image of T under the multifunction F^{-1}. We always have $(F \circ F^{-1})(T) \supset T \cap \operatorname{im} F$, and therefore by exchanging the roles of F and F^{-1} also $(F^{-1} \circ F)(S) \supset S \cap \operatorname{dom} F$, but in general we cannot replace $\supset$ by $=$. For example, if $X = \mathbb{R} = Y$ and

$$F(x) = \begin{cases} [0,1] & \text{if } x \in [0,1], \\ \emptyset & \text{otherwise,} \end{cases}$$

then for each $x \in [0,1]$

$$(F \circ F^{-1})(x) = [0,1] = (F^{-1} \circ F)(x).$$

The sets $F(x)$ for $x \in X$ are the *values* of F, and if Y is a topological space we say that F *has closed values* if for each x, $F(x)$ is a closed set.

In many applications one finds multifunctions whose graphs are either convex polyhedra or unions of such polyhedra. These multifunctions have particularly good continuity properties, some of which this section develops.

5.4.1 Elementary properties

Some properties of multifunctions can be easily visualized in terms of the graph. If X and Y are linear spaces we say that a multifunction is *positively homogeneous* if its graph is a (not necessarily convex) cone, and *homogeneous* if its graph is a union of lines through the origin.

Definition 5.4.3. If X and Y are topological spaces, a multifunction $F : X \to Y$ is *closed* if $\operatorname{gph} F$ is closed in $X \times Y$. It is *locally closed* at a point $(x,y) \in \operatorname{gph} F$ if there is a neighborhood N of (x,y) in $X \times Y$ such that $N \cap \operatorname{gph} F$ is closed in N.

As the definition of closedness involves only the graph, F is closed if and only if F^{-1} is closed.

If the space X is T_1 (in particular, if it is $\mathbb{R}^n$) then each closed multifunction has closed values, but the example of $F : \mathbb{R} \to \mathbb{R}$ with $F(x)$ defined to be $[0,1]$ if $0 < x < 1$ and empty otherwise shows that the converse need not hold.

Definition 5.4.4. If X and Y are linear spaces, a multifunction $F : \mathbb{R}^n \to \mathbb{R}^m$ is *convex* if its graph is convex.

Convex multifunctions are sometimes called *graph-convex.*

Definition 5.4.5. If X and Y are metric spaces, a multifunction $F : X \to Y$ is *bounded* if $\operatorname{im} F$ is a bounded set, It is *locally bounded* at a point $x \in \mathbb{R}^n$ if there is a neighborhood N of x such that $F(N)$ is bounded.

Definition 5.4.6. A multifunction is *polyhedral* if its graph is the union of a finite collection of polyhedral convex sets, called *components* of the graph. It is *polyhedral convex* if it is polyhedral and the graph is convex.

The graph of a polyhedral convex multifunction is by definition convex. On the other hand, the graph of a polyhedral multifunction is a union of convex sets, so it need be neither convex nor even connected.

One very common example of a polyhedral convex multifunction occurs in the case of a feasible set S in $\mathbb{R}^n$ defined by a set of linear inequalities:

$$S = \{x \in \mathbb{R}^n \mid Ax \leq a_0\},$$

where A is a matrix of dimension $k \times n$. For purposes of analysis, we could include linear equations in the definition of S simply by writing each equation as two opposite inequalities.

If we now want to consider such a feasible set not only for a fixed right-hand side a_0 but also for any $a \in \mathbb{R}^k$, then we can write this dependence in the form of a multifunction:

$$S(a) = \{x \in \mathbb{R}^n \mid Ax \leq a\}. \tag{5.36}$$

The graph of the function S defined by (5.36) is

$$\operatorname{gph} S = \left\{ (a,x) \in \mathbb{R}^k \times \mathbb{R}^n \mid \begin{bmatrix} -I & A \end{bmatrix} \begin{bmatrix} a \\ x \end{bmatrix} \leq 0 \right\},$$

which a polyhedral convex cone. So this example is very special, because the multifunction is actually positively homogeneous as well as being polyhedral convex.

Another familiar example of a polyhedral convex multifunction is a linear operator A: its graph, being a subspace, is also a convex cone so that the operator is homogeneous. A third familiar example is an affine transformation.

The following proposition shows that the normal-cone operator associated with the values of a multifunction like S is polyhedral. Thus, it is an example of a nontrivial polyhedral multifunction that is not convex. If we are interested in parametric variational problems like

$$0 \in f(a,x) + N_{S(a)}(x), \tag{5.37}$$

then the polyhedrality of the normal-cone operator may be useful to us.

Proposition 5.4.7. *Let S be a polyhedral convex multifunction from $\mathbb{R}^k$ to $\mathbb{R}^n$. Then the multifunction $G : \mathbb{R}^k \times \mathbb{R}^n \to \mathbb{R}^n$ defined by*

$$G(a,x) = N_{S(a)}(x) \tag{5.38}$$

is polyhedral.

Proof. The graph of G is

$$\operatorname{gph} G = \{(a,x,x^*) \mid (a,x) \in \operatorname{gph} S, x^* \in N_{S(a)}(x)\}.$$

As $\operatorname{gph} S$ is a polyhedral convex set there exist matrices V $(m \times k)$ and W $(m \times n)$, and an element $s \in \mathbb{R}^m$, such that

$$\operatorname{gph} S = \{(a,x) \mid Va + Wx \leq s\}.$$

For any subset A of $\{1,\ldots,m\}$, let cA be the complement of A in $\{1,\ldots,m\}$. For $Q \subset \{1,\ldots,m\}$ let V_Q, W_Q, and s_Q consist of those rows of V and W, and elements of s, whose indices are in Q. Define a (possibly empty) subset P_A of $\mathbb{R}^k \times \mathbb{R}^n \times \mathbb{R}^n$ by

$$P_A = \{(a,x) \mid V_A a + W_A x = s_A,\ V_{cA} a + W_{cA} x \leq s_{cA}\} \times W_A^T(\mathbb{R}_+^{|A|}),$$

where in the case $A = \emptyset$ we interpret the factor on the right to be the origin of $\mathbb{R}^n$. Evidently P_A is a polyhedral convex set. If there is no pair $(a,x) \in \operatorname{gph} S$ with $V_A a + W_A x = s_A$, then $P_A = \emptyset \subset \operatorname{gph} G$. On the other hand, if such a pair (a,x) exists then

$$V_A a + W_A x = s_A,\ V_{cA} a + W_{cA} x \leq s_{cA}.$$

A calculation shows that $W_A^T(\mathbb{R}_+^{|A|}) \subset N_{S(a)}(x)$, from which it follows that $P_A \subset \operatorname{gph} G$. Therefore

$$\cup_{A \subset \{1,\ldots,m\}} P_A \subset \operatorname{gph} G. \tag{5.39}$$

On the other hand, for any point (a,x,x^*) of $\operatorname{gph} G$ let A be the exact index set for the unique face of $S(a)$ containing x in its relative interior. Then Theorem 5.1.7 shows that $N_{S(a)}(x) = W_A^T(\mathbb{R}_+^{|A|})$, so that $(a,x,x^*) \in P_A$. Accordingly,

$$\cup_{A\subset\{1,\dots,m\}} P_A \supset \operatorname{gph} G. \tag{5.40}$$

From (5.39) and (5.40) we obtain

$$\operatorname{gph} G = \cup_{A\subset\{1,\dots,m\}} P_A,$$

which shows that G is a polyhedral multifunction. □

The class of polyhedral multifunctions is closed under several common operations.

Proposition 5.4.8. *Let $T : \mathbb{R}^n \to \mathbb{R}^m$, $U : \mathbb{R}^n \to \mathbb{R}^m$, and $V : \mathbb{R}^m \to \mathbb{R}^r$, and $W : \mathbb{R}^n \to \mathbb{R}^r$ be polyhedral multifunctions. Then the following multifunctions are also polyhedral:*

1. T^{-1};
2. $T \times W$, *where* $(T \times W)(x) := T(x) \times W(x)$;
3. $T + U$;
4. $T \cap U$, *where* $(T \cap U)(x) := T(x) \cap U(x)$;
5. $T \cup U$, *where* $(T \cup U)(x) := T(x) \cup U(x)$;
6. $V \circ T$.

Proof. Suppose that T is a polyhedral multifunction from $\mathbb{R}^n$ to $\mathbb{R}^m$ whose graph has components $\tau_k \subset \mathbb{R}^n \times \mathbb{R}^m$, for $k = 1,\dots,K$. Thus, the τ_k are polyhedral convex sets.

Then the graph of T^{-1} has components

$$\sigma_k = \begin{bmatrix} 0 & I_m \\ I_n & 0 \end{bmatrix} \tau_k, \qquad k = 1,\dots,K,$$

where I_m and I_n are the identity matrices of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. The σ_k are polyhedral convex sets, so T^{-1} is a polyhedral multifunction.

For the Cartesian product, let the components of $\operatorname{gph} W$ be β_i, $i = 1,\dots,I$. Then the components of $\operatorname{gph}(T \times W)$ are the sets of the form

$$\sigma_{ki} = \begin{bmatrix} I_n & 0 & 0 \\ 0 & I_m & 0 \\ I_n & 0 & 0 \\ 0 & 0 & I_r \end{bmatrix}^{-1} (\tau_k \times \beta_i), \qquad k = 1,\dots,K, \quad i = 1,\dots,I,$$

some of which may be empty, and these are polyhedral convex sets.

For addition, let the components of U be ρ_j, $j = 1,\dots,J$. Then the components of $\operatorname{gph} T + U$ are the sets of the form

$$\pi_{kj} = \begin{bmatrix} I_n & 0 & 0 \\ 0 & I_m & I_m \end{bmatrix} \begin{bmatrix} I_n & 0 & 0 \\ 0 & I_m & 0 \\ I_n & 0 & 0 \\ 0 & 0 & I_m \end{bmatrix}^{-1} (\tau_k \times \rho_j), \qquad k = 1,\dots,K, \quad j = 1,\dots,J,$$

which are polyhedral convex sets. Here we used the second assertion, already established.

For intersection, let U be as above; then the components of $\operatorname{gph} T \cap U$ are the sets of the form

$$\nu_{kj} = \begin{bmatrix} I_n & 0 \\ 0 & I_m \\ I_n & 0 \\ 0 & I_m \end{bmatrix}^{-1} (\tau_k \times \rho_j), \qquad k = 1,\dots,K, \quad j = 1,\dots,J,$$

which are polyhedral convex sets.

For union, write G for the graph of $T \cup U$. Any point of G must belong to some component τ_k of T or to some component ρ_j of U. Accordingly, $G \subset (\operatorname{gph} T) \cup (\operatorname{gph} U)$. If $(x,y) \in (\operatorname{gph} T) \cup (\operatorname{gph} U)$ then $y \in T(x)$ or $y \in U(x)$. Then $y \in (T \cup U)(x)$, so $(\operatorname{gph} T) \cup (\operatorname{gph} U) \subset G$ and therefore

$$\operatorname{gph} T \cup U = \operatorname{gph} T \cup \operatorname{gph} U;$$

this is a union of finitely many polyhedral convex sets, so $T \cup U$ is polyhedral.

Finally, for composition suppose that the components of V are α_h, $h = 1,\dots,H$. Then the components of $\operatorname{gph} V \circ T$ are the sets of the form

$$\mu_{kh} = \begin{bmatrix} I_n & 0 & 0 \\ 0 & 0 & I_r \end{bmatrix} \begin{bmatrix} I_n & 0 & 0 \\ 0 & I_m & 0 \\ 0 & I_m & 0 \\ 0 & 0 & I_r \end{bmatrix}^{-1} (\tau_k \times \alpha_h), \qquad k = 1,\dots,K, \quad h = 1,\dots,H,$$

which are polyhedral convex sets. □

One can apply Proposition 5.4.8 to show polyhedrality of other useful combinations of multifunctions. For example, suppose T is as in that proposition and L_A and L_B are affine transformations from $\mathbb{R}^m$ to $\mathbb{R}^q$ and from $\mathbb{R}^p$ to $\mathbb{R}^n$ respectively. As all three of these are polyhedral multifunctions, we can apply the result on composition twice using the expression

$$L_A \circ T \circ L_B = L_A \circ (T \circ L_B),$$

to conclude that $L_A \circ T \circ L_B$ is polyhedral.

5.4.2 Polyhedral convex multifunctions

We develop next several results about the behavior of polyhedral convex multifunctions. They are useful in applications, and in addition we will need some of them in establishing properties of the more general polyhedral multifunctions.

Lemma 5.4.9. *Let P be a polyhedral convex multifunction from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then for any bounded subset S of* $\operatorname{dom} P$ *there is a real number σ such that for each $x \in S$, $d_{P(x)}(0) \leq \sigma$.*

The property asserted in this lemma doesn't hold for general multifunctions, as you can see by considering the multifunction C from $\mathbb{R}^2$ to $\mathbb{R}$ whose graph is

$$\{(x,y,z) \mid x > 0,\ z \geq y^2/(2x)\} \cup \{(0,0,z) \mid z \geq 0\}.$$

Although its representation in this form looks complicated, the graph is simple: it is a right circular cone (ice-cream cone) centered on the 45° line in the positive quadrant of the xz-plane, and just touching the x and z axes.

This multifunction is often useful for finding counterexamples to plausible conjectures. In the present case, if for $k = 1,2,\ldots$ we take $x_k = .5 * k^{-3}$ and $y_k = k^{-1}$ then $C(x_k, y_k) = [k, +\infty)$. Therefore although the sequence $\{x_k, y_k\}$ converges to the origin, the sequence $d_{C(x_k,y_k)}(0)$ is $\{k\}$, which converges to $+\infty$.

Proof. Let S be a bounded subset of $\operatorname{dom} P$, and let Q be any bounded polyhedral convex subset of $\mathbb{R}^n$ containing S. If π_1 is the canonical projector from $\mathbb{R}^n \times \mathbb{R}^m$ to $\mathbb{R}^n$ taking (x,y) to x, then since π_1 is linear the set $\operatorname{dom} P$, which is $\pi_1(P)$, is polyhedral convex, and therefore so is the set $F = Q \cap \operatorname{dom} P$, which contains S. But F is bounded; hence Theorem 4.3.4 says it is the convex hull of its (finite) set of extreme points $f_1, \ldots, f_K$. For $k = 1, \ldots, K$ let p_k be the projection of the origin on the polyhedral convex set $P(f_k)$: that is, the smallest element of $P(f_k)$, and let σ be the maximum of the norms $\|p_k\|$ for $k = 1, \ldots, K$.

Choose any point x of F; then there are convex coefficients $\mu_1, \ldots, \mu_K$ such that $x = \sum_{k=1}^K \mu_k f_k$. Let $y := \sum_{k=1}^K \mu_k p_k$. As the pairs (f_k, p_k) for $k = 1, \ldots, K$ belong to P, we have

$$(x,y) = (\sum_{k=1}^K \mu_k f_k, \sum_{k=1}^K \mu_k p_k) = \sum_{k=1}^K \mu_k (f_k, p_k) \in P.$$

This shows that $y \in P(x)$, so if z is the smallest element of $P(x)$ then $\|z\| \leq \|y\|$. However, we also have

$$\|y\| = \|\sum_{k=1}^K \mu_k p_k\| \leq \sum_{k=1}^K \mu_k \|p_k\| \leq \sigma,$$

so that $\|z\| \leq \sigma$. □

It is sometimes useful to apply Lemma 5.4.9 to the inverse of the multifunction under consideration. Thus, suppose T is a polyhedral convex multifunction from $\mathbb{R}^n$

to $\mathbb{R}^m$, and let Q be any bounded subset of $\operatorname{im} T$. Apply the lemma to T^{-1} and Q to conclude that there is some μ such that for each $q \in Q$, $d_{T^{-1}(q)}(0) \leq \mu$. But this means that there is some $p \in \mathbb{R}^n$ with $q \in T(p)$ and $\|p\| \leq \mu$: in other words, that $Q \subset T(\mu B^n)$.

The following theorem gives an important bound for error in solutions of systems of linear inequalities. We use the notation x_+, where $x \in \mathbb{R}^n$, for the vector in $\mathbb{R}^n$ with

$$(x_+)_i = \begin{cases} x_i & \text{if } x_i \geq 0, \\ 0 & \text{if } x_i < 0. \end{cases}$$

That is, x_+ is the (Euclidean) projection of x on $\mathbb{R}^n_+$. We also define x_- by $x = x_+ + x_-$.

Theorem 5.4.10 (Hoffman, 1952). *Let A be an $m \times n$ matrix.Then there is a constant γ_A such that for each $a \in A(\mathbb{R}^n) + \mathbb{R}^m_+$ and for each $x \in \mathbb{R}^n$ there exists an $x' \in \mathbb{R}^n$ with $Ax' \leq a$ and*

$$\|x - x'\| \leq \gamma_A \|(Ax - a)_+\|. \tag{5.41}$$

An interpretation of this theorem is that if the system of inequalities $Ax \leq a$ has any solution at all, then the distance of the point x from the solution set of the inequalities is bounded above by a constant multiple of the "residual," $\|(Ax-a)_+\|$; furthermore, this constant can depend on A but not on a.

To include systems of linear equations along with the inequalities in this result, write each equation as two opposite inequalities.

Proof. Given A, define a multifunction Q from $\mathbb{R}^m$ to $\mathbb{R}^n$ by

$$Q(a) := \{x \in \mathbb{R}^n \mid Ax - a \in \mathbb{R}^m_-\}.$$

The graph of Q is the set

$$Q = \left\{ \begin{bmatrix} a \\ x \end{bmatrix} \mid \begin{bmatrix} -I & A \end{bmatrix} \begin{bmatrix} a \\ x \end{bmatrix} \in \mathbb{R}^m_- \right\} = \begin{bmatrix} -I & A \end{bmatrix}^{-1} (\mathbb{R}^m_-).$$

As this is a polyhedral convex set, Q is a polyhedral convex multifunction.

For each nonempty subset I of $\{1, \ldots, m\}$ define a positively homogeneous polyhedral convex multifunction P_I from $\mathbb{R}^n$ to $\mathbb{R}^m$ by

$$P_I(w) = \{y \mid A^* y = w,\ y \in \mathbb{R}^m_+,\ y_i = 0 \text{ for each } i \notin I\}.$$

By Lemma 5.4.9, the distance from the origin to $P_I(\cdot)$ is bounded above on the intersection of the unit ball B^n with $\operatorname{dom} P_I$. This intersection is nonempty because the graph of P_I is a polyhedral convex cone and therefore contains the origin. Let γ_A be the maximum of these bounds over all nonempty subsets I of $\{1, \ldots, m\}$.

Now choose any $a \in A(\mathbb{R}^n) + \mathbb{R}^m_+$; then $Q(a) := \{x \mid Ax \leq a\}$ is a nonempty polyhedral convex set. Let $x \in \mathbb{R}^n$ and let x' be the unique point in $Q(a)$ closest to x. If $x' = x$ then we certainly have (5.41).

If $x' \neq x$ then x' belongs to the relative interior of a unique face F of $Q(a)$. Let I be the exact index set associated with that face by Theorem 5.1.5 applied to $Q(a)$. The set I cannot be empty, because if it were then by Theorem 5.1.7 $Q(a)$ would have a nonempty interior. But then x' would have to lie in that interior, and as $x' \neq x$ x' could not be the closest point in $Q(a)$ to x. Therefore I is nonempty.

The difference $x - x'$ belongs to the normal cone of $Q(a)$ at x', and by Theorem 5.1.7 that normal cone is $A_I^*(\mathbb{R}_+^{|I|})$. Therefore we can find $y \in \mathbb{R}_+^m$ such that $A^*y = x - x'$ and $y_{cI} = 0$. The (nonempty) set of all such y is $P_I(x - x')$. Define $t := (x - x')/\|x - x'\|$; then $t \in (\operatorname{dom} P_I) \cap B^n$. Choose z to be the element of least norm in $P_I(t)$, so that $\|z\| \leq \gamma_A$, and define $y' := \|x - x'\|z$; then $y' \in P_I(x - x')$ and $\|y'\| \leq \gamma_A \|x - x'\|$. Then $x - x' = A^*y'$, and as $(y')_{cI} = 0$ and $(Ax' - a)_I = 0$ we have $\langle y', Ax' - a \rangle = 0$, while as $y' \geq 0$ and $(Ax - a)_- \leq 0$ we have $\langle y', (Ax - a)_- \rangle \leq 0$. Therefore

$$\begin{aligned} \|x - x'\|^2 &= \langle x - x', x - x' \rangle \\ &= \langle A^*y', x - x' \rangle \\ &= \langle y', A(x - x') \rangle \\ &= \langle y', (Ax - a) - (Ax' - a) \rangle \\ &= \langle y', (Ax - a) \rangle \\ &= \langle y', (Ax - a)_+ + (Ax - a)_- \rangle \\ &\leq \langle y', (Ax - a)_+ \rangle \\ &\leq \gamma_A \|x - x'\| \|(Ax - a)_+\|, \end{aligned} \tag{5.42}$$

where the last line follows from the Schwarz inequality. Now (5.41) follows from (5.42). □

One consequence of Hoffman's theorem is a very useful property of polyhedral convex multifunctions.

Corollary 5.4.11. *If F is a nonempty polyhedral convex multifunction from $\mathbb{R}^n$ to $\mathbb{R}^m$, then the restriction of F to* $\operatorname{dom} F$ *is Lipschitzian in the Hausdorff metric.*

Proof. Let the graph of F be the polyhedral convex set

$$G := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^m \mid Cx + Dy \leq a\},$$

where $C \in \mathbb{R}^{m \times n}$ and $D \in \mathbb{R}^{m \times m}$. Let γ_D be the number associated with D by Hoffman's theorem and define $\lambda := \gamma_D \|C\|$. Choose any points x' and x'' in $\operatorname{dom} F$.

If $y' \in F(x')$ we have $Dy' \leq a - Cx'$. Hoffman's theorem applied to the system $Dy \leq (a - Cx'')$ says that there is a point y'' with $Dy'' \leq a - Cx''$, so that $y'' \in F(x'')$, and with

$$\|y' - y''\| \leq \gamma_D \|(Dy' - [a - Cx''])_+\|. \tag{5.43}$$

However,

$$Dy' - [a - Cx''] = (Cx' + Dy' - a) + C[x'' - x'] \leq C[x'' - x'],$$

where the inequality holds because $y' \in F(x')$. Then

$$0 \le (Dy' - [a - Cx'])_+ \le (C[x'' - x'])_+,$$

and then 5.43 yields

$$\|y' - y''\| \le \gamma_D \|(Dy' - [a - Cx'])_+\| \le \gamma_D \|(C[x'' - x'])_+\|,$$

so that

$$e[F(x'), F(x'')] \le \gamma_D \|(C[x'' - x'])_+\|. \tag{5.44}$$

Reversing the roles of x' and x'' yields

$$e[F(x''), F(x')] \le \gamma_D \|(C[x' - x''])_+\|. \tag{5.45}$$

The Hausdorff distance $\rho[F(x'), F(x'')]$ is the maximum of the left sides of 5.44 and 5.45. The right side of each of those two inequalities is not greater than $\gamma_D \|C[x'' - x']\|$, so

$$\rho[F(x'), F(x'')] \le \gamma_D \|C[x'' - x']\| \le \lambda \|x'' - x'\|.$$

As x' and x'' were any points of $\operatorname{dom} F$, the multifunction F is Lipschitzian on $\operatorname{dom} F$ in the Hausdorff metric with modulus λ. □

5.4.3 General polyhedral multifunctions

Corollary 5.4.11 in the preceding section showed that polyhedral convex multifunctions are Lipschitzian in the Hausdorff metric on their effective domains. Such a strong property does not hold for general polyhedral multifunctions, but we can establish a kind of Lipschitz semicontinuity.

Definition 5.4.12. Let F be a multifunction from $\mathbb{R}^n$ to $\mathbb{R}^m$ and λ be a real number. F is *locally outer Lipschitz continuous* at a point $x \in \mathbb{R}^n$ with modulus λ if for some neighborhood N of x and each $x' \in N$ we have

$$F(x') \subset F(x) + \lambda \|x' - x\| B^m. \tag{5.46}$$

For (5.46) to hold we must have either $x \in \operatorname{dom} F$ or $x \notin \operatorname{cl} \operatorname{dom} F$. Also, if $F(x) \neq \emptyset$ then (5.46) implies in particular that the excess $e[F(x'), F(x)]$ is not more than $\lambda \|x' - x\|$, though it says nothing about the excess $e[F(x), F(x')]$, which could be infinite.

Theorem 5.4.13. *Let F be a polyhedral multifunction from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then there exists a constant λ such that F is everywhere locally outer Lipschitz continuous with modulus λ.*

Proof. Let the components of the graph of F be $G_1, \ldots, G_I$, and denote by P_i the multifunction from $\mathbb{R}^n$ to $\mathbb{R}^m$ whose graph is G_i. The G_i are polyhedral convex,

and therefore so are their projections $\pi_1(G_i)$, the union of which is $\operatorname{dom} F$. As there are only finitely many components, $\operatorname{dom} F$ is a closed set. If $x \notin \operatorname{dom} F$, then some neighborhood of x does not meet $\operatorname{dom} F$, and on that neighborhood the values of F all equal $\emptyset$, so F is trivially locally outer Lipschitz continuous at x. Therefore suppose $x \in \operatorname{dom} F$.

By Corollary 5.4.11, each P_i is Lipschitzian on its effective domain, with some modulus λ_i. Let λ be the maximum of $\lambda_1, \ldots, \lambda_I$ and choose any $x \in \mathbb{R}^n$. Let $\mathscr{I}$ be the set of indices i in $1, \ldots, I$ such that $x \in \pi_1(G_i)$. As the sets $\pi_1(G_i)$ are closed, for each $i \notin \mathscr{I}$ there is a neighborhood N_i of x that does not meet $\pi_1(G_i)$. Let N be the intersection of these neighborhoods. If $x' \in N$, then if $x' \notin \operatorname{dom} F$ the asserted bound holds because $F(x') = \emptyset$.

If $x' \in \operatorname{dom} F$ then for each $y' \in F(x')$ the pair (x', y') belongs to some G_i for which $\pi_1(G_i)$ meets N, and hence for which $i \in \mathscr{I}$. Then

$$\delta[P_i(x'), P_i(x)] \le \lambda_i \|x' - x\| \le \lambda \|x' - x\|,$$

and as the values of P_i are closed sets we have

$$P_i(x') \subset P_i(x) + \lambda \|x' - x\| B^m.$$

Then as $F(x) = \cup_{i \in \mathscr{I}} P_i(x)$ by definition of $\mathscr{I}$,

$$\begin{aligned} F(x') &= \cup \{P_i(x') \mid x' \in \pi_1(G_i)\} \\ &\subset \cup_{i \in \mathscr{I}} P_i(x') \\ &\subset \cup_{i \in \mathscr{I}} P_i(x) + \lambda \|x' - x\| B^m \\ &= F(x) + \lambda \|x' - x\| B^m, \end{aligned}$$

as claimed. □

The constant λ in Theorem 5.4.13 depends only on F, so the same Lipschitz modulus works for any $x \in \mathbb{R}^n$.

Theorem 5.4.13 implies that the polyhedral multifunctions satisfy a property that is similar, but not identical, to what is called in the literature *metric regularity.* The following proposition explains this property.

Proposition 5.4.14. *Let T be a polyhedral multifunction from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then there is some $\kappa \in \mathbb{R}_+$ such that for each $y \in \mathbb{R}^m$ there is a neighborhood Q of y with the property that if $x \in T^{-1}(Q)$ then*

$$d[x, T^{-1}(y)] \le \kappa d[y, T(x)]. \tag{5.47}$$

Proof. Theorem 5.4.13 shows that there is a nonnegative modulus κ such that T^{-1} is everywhere locally outer Lipschitz continuous with modulus κ. Choose $y \in \mathbb{R}^m$ and let Q be a ball $B(y, \varepsilon)$ such that for $y' \in Q$ we have

$$T^{-1}(y') \subset T^{-1}(y) + \kappa \|y' - y\| B^n. \tag{5.48}$$

If $T^{-1}(Q) = \emptyset$ there is nothing to prove. Otherwise, let $x \in T^{-1}(Q)$; as $T(x)$ is closed we can project y onto $T(x)$ to obtain $y' \in T(x)$ with $\|y' - y\| = d[y, T(x)]$. Then $x \in T^{-1}(y')$, so (5.48) implies $x \in T^{-1}(y) + \kappa\|y' - y\|B^n$, and therefore

$$d[x, T^{-1}(y)] \leq \kappa\|y' - y\| = \kappa d[y, T(x)].$$

□

In words, Proposition 5.4.14 says that if $T(x)$ is close enough to the point y then the bound (5.47) holds. To see why this closeness condition is necessary, consider the polyhedral multifunction T from $\mathbb{R}$ to $\mathbb{R}$ given by

$$T(x) = \begin{cases} -1 & \text{if } x < -1, \\ x & \text{if } x \in [-1, 1], \\ 1 & \text{if } x > 1. \end{cases}$$

If we choose $y = 0$, then for $Q \subset (-1, 1)$ the conclusions of the proposition hold with $\kappa = 1$. However, if Q contains either 1 or -1 then they fail with any κ, because for example $T^{-1}(1) = [1, +\infty) \not\subset T^{-1}(0) + \kappa\|1 - 0\|B^1$.

Lemma 5.4.9 established a bound on the distance to values of a polyhedral convex multifunction for arguments in a bounded set. A bound of the same kind holds for general polyhedral multifunctions.

Proposition 5.4.15. *Let T be a polyhedral multifunction from $\mathbb{R}^n$ to $\mathbb{R}^n$. For each bounded subset S of* $\operatorname{dom} T$ *there is a nonnegative real number γ_S such that for each $x \in S$, $d_{T(x)}(0) \leq \gamma_S$.*

Proof. Choose a bounded subset S of $\operatorname{dom} T$ and let $\{G_i \mid i = 1, \ldots, k\}$ be the components of $\operatorname{gph} T$. Let $\mathscr{I}$ be the subset of $\{1, \ldots, k\}$ consisting of those indices i for which $S \cap \operatorname{dom} G_i \neq \emptyset$. If $\mathscr{I} = \emptyset$ then S is empty and there is nothing to prove, so assume otherwise. Let γ_S be the maximum of the bounds γ_i provided by Lemma 5.4.9 for S and the polyhedral multifunctions F_i whose graphs are the G_i for $i \in \mathscr{I}$. Choose $x \in S$; then for some $i \in \mathscr{I}$ we have $x \in \operatorname{dom} F_i$, so that

$$d[0, T(x)] \leq d[0, F_i(x)] \leq \gamma_i \leq \gamma_S.$$

□

A consequence of Proposition 5.4.15 is that polyhedral multifunctions have the property that bounded subsets of their images are covered by the images of bounded subsets of their domains, as the following corollary shows.

Corollary 5.4.16. *Let G be a polyhedral multifunction from $\mathbb{R}^n$ to $\mathbb{R}^m$, and let Q be a bounded subset of* $\operatorname{im} G$. *Then there is a bounded subset P of* $\operatorname{dom} G$ *such that $G(P) \supset Q$.*

Proof. As G is polyhedral, the multifunction $T := G^{-1}$ is also polyhedral. Let Q be a bounded subset of $\operatorname{im} G = \operatorname{dom} T$, and apply Proposition 5.4.15 to produce $\gamma_Q \in \mathbb{R}_+$

such that for each $q \in Q$, the distance from the origin to $G^{-1}(q)$ is not more than γ_Q. This means that for any such q there is some $x_q \in \gamma_Q B^n$ with $q \in G(x_q)$, and therefore that $q \in G(\gamma_Q B^n)$. Then $G(\gamma_Q B^n) \supset Q$, so we can take $P = \operatorname{dom} G \cap (\gamma_Q B^n)$. □

5.4.4 Exercises for Section 5.4

Exercise 5.4.17. If P is a polyhedral convex subset of $\mathbb{R}^n$, then N_P is a polyhedral multifunction.

Exercise 5.4.18. Let A be an $m \times n$ matrix and $c^* \in \mathbb{R}^n$. For $a \in \mathbb{R}^m$ define

$$Q(a) = \{x \mid Ax = a,\ x \geq 0\}, \qquad \tau(a) = \inf_{x \in Q(a)} \langle c^*, x \rangle.$$

Assume the conventions that the infimum of the empty subset of $\mathbb{R}$ is $+\infty$, and the supremum of that set is $-\infty$.

1. What kind of multifunction is $Q(a)$? What is its effective domain?
2. What is the nature of the set on which $\tau(a) < +\infty$?
3. Could τ take the value $-\infty$? If so, specify the weakest possible condition on A and c^* that will ensure that $\tau > -\infty$, show that it is weakest, and assume that condition for the rest of this exercise.
4. Does τ have any convexity properties? If so, describe them.
5. What is the largest set on which you can prove that τ is Lipschitzian? Specify the set, and prove the Lipschitzian property.

Exercise 5.4.19. You are given a linear-programming model of a system that responds to exterior demands in the following way: a demand appears (a vector $d \in \mathbb{R}^m$), and the system must then solve

$$Dx \geq d, \qquad Fx = f, \qquad x \geq 0,$$

where D ($m \times n$), F ($k \times n$), and $f \in \mathbb{R}^k$ are given quantities. The constraints $Fx = f$ represent aspects of the system other than demand satisfaction. The columns of D and F represent activities of the system, whose use is costly. There is a given vector of costs $c^* \in \mathbb{R}^n$. For each demand vector d one then wants to solve the linear-programming problem

$$\inf\{\langle c^*, x \rangle \mid Dx \geq d,\ Fx = f,\ x \geq 0\}. \tag{5.49}$$

The system operators are concerned that small changes in demand might result in large increases in the cost of satisfying the demand. Show that:

a. For each demand d for which the infimum in (5.49) is finite, a solution x exists.
b. There is a constant α such that given a base case demand d_0 and a cost c_0 associated with that problem, the cost of satisfying demand d (from part (a) of this

exercise) will not be more than $c_0 + \alpha\|(d - d_0)_+\|$, and therefore will increase at most linearly with change in demand.

Comment. The assumed finiteness of the infimum guarantees feasibility of the problem, because if the problem were infeasible then the infimum would be taken over the empty set and therefore would be $+\infty$.

5.5 Polyhedrality and separation

Stronger separation theorems hold for polyhedral convex sets than for convex sets in general. These separation theorems often present themselves in forms involving alternative statements about the solvability of systems of linear equations and inequalities, and for that reason they are called theorems of the alternative. They are very useful in, e.g., the analysis of optimization problems.

It is possible to prove most theorems of the alternative either by direct separation arguments, such as we used in Lemma 3.1.10, or by using another theorem of the alternative together with substitutions to change the form of the equations and/or inequalities involved. However, these proofs can sometimes be rather tedious. One can avoid them by establishing a single theorem tailored to the polyhedral case, and then deriving the various theorems of the alternative as simple corollaries. The theorem also provides information useful for dealing with problems other than separation.

The next section presents such a theorem, which we call the CRF theorem. The subsequent section shows how to use it to derive an array of theorems of the alternative.

5.5.1 The CRF theorem

The result proved here comes from work of Camion, Rockafellar, and Fulkerson; see for example [34, Theorem 22.6] and the bibliographic comments on p. 428 of that work. Rather than use the name "Camion-Rockafellar-Fulkerson theorem," we just refer to it in what follows as the CRF theorem. The proof we give here is due to Minty [23].

The theorem uses *elementary subspaces* of a given subspace of $\mathbb{R}^n$. The basic idea is that with each one-dimensional subspace one can unambiguously associate a *support,* which is simply the set of indices for which elements of that subspace (except the origin) have nonzero entries. The elementary subspaces are those whose supports are minimal.

Definition 5.5.1. Let x be an element of $\mathbb{R}^n$. The *support* of x, written $\mathrm{supp}(x)$, is the set of indices i for which $x_i \neq 0$. □

The support depends on the basis one uses for the space in question.

Definition 5.5.2. Let S be a one-dimensional subspace of $\mathbb{R}^n$. The *support* of S, $\operatorname{supp}(S)$, is defined to be the set $\operatorname{supp}(x)$ where x is any nonzero element of S. □

This definition makes sense because all such x have the same support.

Definition 5.5.3. Let L be a subspace of $\mathbb{R}^n$. An *elementary subspace* of L is a one-dimensional subspace $S \subset L$ such that there is no one-dimensional subspace of L whose support is properly contained in that of S. □

The empty subspace and the origin have no elementary subspaces. However, any other subspace $L \subset \mathbb{R}^n$ has at least one elementary subspace. To find one, just list the supports corresponding to nonzero elements of L (a finite set), and find a support in that set that does not properly contain any other. Any element of L having that support will generate an elementary subspace.

Theorem 5.5.4. *Let L be a nonempty subspace of $\mathbb{R}^n$, of dimension $k \geq 1$.*

a. If S and T are elementary subspaces of L with $\operatorname{supp}(S) = \operatorname{supp}(T)$, then $S = T$.
b. L has only finitely many elementary subspaces.
c. L is the sum of its elementary subspaces.
d. The support of any elementary subspace of L has at most $n-k+1$ components.

Proof. To prove (a), let S and T be elementary subspaces having the same support, say $\mathscr{S}$. Choose elements $s \in S$ and $t \in T$, and let i be any index in $\mathscr{S}$. Define $u = s - (s_i/t_i)t$; then $\operatorname{supp}(u) \subset \mathscr{S}$, and moreover the containment is proper because $i \notin \operatorname{supp}(u)$. If u were not zero, it would generate a one-dimensional subspace whose support was strictly contained in $\mathscr{S}$, and this would contradict the fact that S and T are elementary. Therefore $u = 0$; hence s and t are proportional, so $S = T$.

Assertion (a) implies that there are no more elementary subspaces of L than there are possible supports. This is a finite number, which proves (b).

To prove (c) we have only to prove that L is contained in the sum of its elementary subspaces, because the sum of any finite collection of subspaces of L is contained in L. Call the sum of the elementary subspaces M; we have already observed that M is not empty.

Any vector in L whose support has only a single element must generate an elementary subspace; there may or may not be any such vectors. Now let $0 < J < n$ and assume that any vector in L whose support has no more than J elements belongs to M. We establish this assertion for elements of L having no more than $J+1$ elements, and thereby complete the proof of (c) by induction.

Let x be a vector of L whose support has no more than $J+1$ elements. If no nonzero element of L has support properly contained in $\operatorname{supp}(x)$, then x generates an elementary subspace of L, and then $x \in M$. Therefore suppose that y is a nonzero element of L whose support is properly contained in $\operatorname{supp}(x)$; then $y \in M$ by the induction hypothesis. Choose an index i with $y_i \neq 0$ and let $z = x - (x_i/y_i)y$. Then $\operatorname{supp}(z) \subset \operatorname{supp}(x)$, and the containment is proper since $z_i = 0$. Therefore the induction hypothesis implies that $z \in M$. Since $x = z + (x_i/y_i)y$, $x \in M$.

To prove (d), let S be a one-dimensional subspace of L and let $\mathscr{S}$ be its support. If $\mathscr{S}$ has more than $n-k+1$ components, form a set $\mathscr{T}$ consisting of the complement

of $\mathscr{S}$ in $\{1,\ldots,n\}$, together with exactly one element of $\mathscr{S}$. Let $v_1,\ldots,v_{n-k}$ be a basis for $L^\perp$, and consider the linear system

$$\langle v_i, x\rangle = 0, \quad (i = 1,\ldots,n-k), \qquad x_j = 0, \quad (j \in \mathscr{T}).$$

This is a system of fewer than $(n-k)+[n-(n-k+1)+1] = n$ linear equations in $\mathbb{R}^n$. Therefore it has a nontrivial solution, which spans a one-dimensional subspace of L whose support is properly contained in $\mathscr{S}$. Therefore S cannot be elementary, and it follows that the support of an elementary subspace has no more than $n-k+1$ elements. □

We now apply the concept of elementary subspace to state and prove the CRF theorem. This theorem deals with two objects: a nonempty subspace L of $\mathbb{R}^n$, and a collection of *n real intervals:* a real interval is any convex subset of the real line $\mathbb{R}$. In particular, these intervals could be open, closed, half-open, bounded, or unbounded. The theorem asserts that L meets the Cartesian product $\mathscr{I}$ of the intervals if and only if, for each elementary subspace S of $L^\perp$, $S^\perp$ meets $\mathscr{I}$. The "only if" part is immediate, since $S^\perp \supset L$, so the strength of the theorem lies in the "if" statement.

Theorem 5.5.5. *Let L be a nonempty subspace of $\mathbb{R}^n$ and let $I_1,\ldots,I_n$ be real intervals. Let $\mathscr{I} = \prod_{i=1}^n I_i$. Then L meets $\mathscr{I}$ if and only if, for each elementary subspace S of $L^\perp$, $S^\perp$ meets $\mathscr{I}$.*

Proof. (Only if) If L meets $\mathscr{I}$ and S is a subspace of $L^\perp$ then $S^\perp \supset L^{\perp\perp} = L$, so $S^\perp$ meets $\mathscr{I}$.

(If) This proof is easy in three special cases that we consider first.

a. If any of the intervals is empty then so is $\mathscr{I}$ and the theorem obviously holds. Therefore we assume $\mathscr{I} \neq \emptyset$.
b. If $\dim L = n$ then $L = \mathbb{R}^n$. In this case L must meet $\mathscr{I}$, so the first statement holds; so does the second because $L^\perp$ is $\{0\}$, which has no elementary subspaces.
c. If $\dim L = n-1$ then $L^\perp$ has a single elementary subspace (itself). The orthogonal complement of this elementary subspace is $L^{\perp\perp} = L$. Therefore the second statement holds exactly when the first does.

If $n = 1$ then the dimension of L is either n or $n-1$, so the theorem holds by observations (b) and (c) above. Therefore suppose that $n \geq 2$, that the dimension of L is not more than $n-2$, and that the theorem is true in spaces of dimension not more than $n-1$. Suppose also that for each elementary subspace S of $L^\perp$, $S^\perp$ meets $\mathscr{I}$. We will show that L meets $\mathscr{I}$.

For $i = 1,\ldots,n$ define

$$J_i = I_1 \times I_2 \times \ldots \times I_{i-1} \times \mathbb{R} \times I_{i+1} \times \ldots \times I_n.$$

We are going to prove next that for each i, L meets J_i. The proof for each i is similar, so with no loss of generality we prove this only for $i = n$. Define a subspace L' of $\mathbb{R}^{n-1}$ to be the image of L under the linear operator Q that takes $(x_1,\ldots,x_n) \in \mathbb{R}^n$ to

$(x_1,\ldots,x_{n-1}) \in \mathbb{R}^{n-1}$. For future reference, note that the adjoint of Q is the mapping Q^* from $\mathbb{R}^{n-1}$ to $\mathbb{R}^n$ that takes $(x_1,\ldots,x_{n-1})$ to $(x_1,\ldots,x_{n-1},0)$.

Suppose that S' is any elementary subspace of $(L')^\perp$, and let $S := S' \times \{0\}$. We first prove that S is an elementary subspace of $L^\perp$.

a. To see that $S \subset L^\perp$, note that

$$(L')^\perp = [Q(L)]^\perp = (Q^*)^{-1}(L^\perp), \tag{5.50}$$

We assumed that $S' \subset (L')^\perp$, so it follows from (5.50) that $S = S' \times \{0\} = Q^*(S')$ is contained in $L^\perp$.

b. The dimension of S must be that of S', which is 1, so S is a one-dimensional subspace of $L^\perp$.

c. Finally, if $\operatorname{supp}(S)$ were not minimal among the supports of one-dimensional subspaces of $L^\perp$, then we could find a one-dimensional subspace T of $L^\perp$ having strictly smaller support. Choose nonzero elements $s \in S$ and $t \in T$. By construction, s must have the form $(s',0)$ where $s' \in S'$. As $\operatorname{supp}(t)$ is strictly contained in $\operatorname{supp}(s)$, t must be of the form $(t',0)$, with the support of t' strictly contained in that of s'. Then $Q^*(t') = t \in L^\perp$, and (5.50) says that $t' \in (L')^\perp$. This contradicts the fact that S' was an elementary subspace of $(L')^\perp$, so in fact $\operatorname{supp}(S)$ is minimal among the supports of one-dimensional subspaces of $L^\perp$.

We now know that S is an elementary subspace of $L^\perp$, so we can apply the hypothesis that for each elementary subspace S of $L^\perp$, $S^\perp$ meets $\mathscr{I}$. But

$$S^\perp = (S' \times \{0\})^\perp = [(S')^\perp] \times \mathbb{R},$$

and this means that $(S')^\perp$ meets the set $\mathscr{I}' := I_1 \times \ldots \times I_{n-1}$. Therefore for each elementary subspace S' of $(L')^\perp$, $(S')^\perp$ meets $\mathscr{I}'$.

Now apply the induction hypothesis in $\mathbb{R}^{n-1}$ to conclude that L' meets $\mathscr{I}'$. Take any point x' of this intersection and use the fact that $L' = Q(L)$ to find a point $x \in L$ with $Qx = x'$. Then $x \in L \cap J_n$ as required.

To finish the proof, we use what we have already proved to find n points, say $x_1,\ldots,x_n$, with $x_i \in L \cap J_i$ for $i = 1,\ldots,n$. For $i = 2,\ldots,n$ define $v_i = x_i - x_1$. The v_i are all in L, which has dimension not more than $n-2$, so they are linearly dependent. Therefore there are $\mu_2,\ldots,\mu_n$, not all zero, with $\sum_{i=2}^n \mu_i(x_i - x_1) = 0$. If we let $\mu_1 := -\sum_{i=2}^n \mu_i$, then we have

$$\sum_{i=1}^n \mu_i x_i = 0, \qquad \sum_{i=1}^n \mu_i = 0.$$

Since the μ_i sum to zero yet are not all zero, they cannot all be of the same sign. By reordering the indices if necessary, we can suppose that $\mu_1,\ldots,\mu_k$ are nonnegative and $\mu_{k+1},\ldots,\mu_n$ are negative. We have $\sum_{i=1}^k \mu_i = \sum_{i=k+1}^n (-\mu_i)$; call this positive number σ. Now for $i = 1,\ldots,k$ define $\pi_i = \mu_i/\sigma$, and for $i = k+1,\ldots,n$ define $\nu_i = -\mu_i/\sigma$. We then have

$$\sum_{i=1}^{k} \pi_i x_i = \sum_{i=k+1}^{n} \nu_i x_i. \tag{5.51}$$

The left side of (5.51) is a convex combination of points, each of which belongs to

$$L \cap [\mathbb{R} \times \ldots \times \mathbb{R} \times I_{k+1} \times \ldots \times I_n];$$

the right side is another convex combination of points, each of which belongs to

$$L \cap [I_1 \times \ldots \times I_k \times \mathbb{R} \times \ldots \times \mathbb{R}].$$

Therefore the element of $\mathbb{R}^n$ expressed by (5.51) belongs to the intersection of these two sets, which is $L \cap \mathscr{I}$. □

5.5.2 Some applications of the CRF theorem

To illustrate a direct application of the CRF theorem we will use it to prove a complementarity theorem of A. W. Tucker. After that, we will rewrite it in the form of a theorem of the alternative, and illustrate how to prove other such theorems using that form.

Theorem 5.5.6 (Tucker, 1956). *Let L be any nonempty subspace of $\mathbb{R}^n$. There exist points $x \in L$ and $x^* \in L^\perp$ with $x \geq 0$, $x^* \geq 0$, and $x + x^* > 0$.*

The orthogonality of x and x^*, together with nonnegativity, requires that their supports be complementary, so this is sometimes referred to as Tucker's complementarity theorem.

Proof. Among the nonnegative elements of $L^\perp$, select x^* so that its support has maximum cardinality. If the support is $\{1,\ldots,n\}$ then take $x = 0 \in L$ to finish the proof. Otherwise, for $i = 1,\ldots,n$ define I_i to be $(0,+\infty)$ if $i \notin \text{supp}(x^*)$ and $I_i = \{0\}$ if $i \in \text{supp}(x^*)$. Define $\mathscr{I} = I_1 \times \ldots, \times I_n$. If $L \cap \mathscr{I} \neq \emptyset$ then let x be a point in this intersection. This x is nonnegative, belongs to L, and satisfies $x + x^* > 0$ as required.

To complete the proof we must show that L and $\mathscr{I}$ in fact have a nonempty intersection. If they did not, then by the CRF theorem there would be an elementary vector v^* of $L^\perp$ such that

$$\sum_{i=1}^{n} (v^*)_i I_i \subset (0,+\infty). \tag{5.52}$$

This implies that the coordinates of v^* corresponding to indices not in $\text{supp}(x^*)$ are nonnegative, and at least one of them is positive. Now by forming $v^* + \alpha x^*$ and taking α to be a sufficiently large positive number, we can produce a nonnegative element of $L^\perp$ whose support strictly contains that of x^*, which is impossible by our choice of x^*. Therefore the intersection of L and $\mathscr{I}$ must be nonempty. □

For many applications of the CRF theorem it helps to use the following reformulated version, in the form of a theorem of the alternative.

Corollary 5.5.7. *Let M be an $m \times n$ matrix of rank r, and let $I_1, \ldots, I_m$ be real intervals. Let $\mathscr{I} = I_1 \times \ldots \times I_m$. Then exactly one of the following statements holds.*

a. There is an $x \in \mathbb{R}^n$ with $Mx \in \mathscr{I}$.
b. There is a $y^* \in \mathbb{R}^m$ with $M^* y^* = 0$ and with

$$\sum_{i=1}^{m} y_i^* I_i \subset (0, +\infty). \tag{5.53}$$

Moreover, in case (a) the vector x can be chosen to have no more than r nonzero components, while in case (b) the vector y^* can be chosen to have no more than $r+1$ nonzero components and to span an elementary subspace of $\ker M^*$.

Proof. Let L be the image of M; then $L^\perp = \ker M^*$. Theorem 5.5.5 says that alternative (a) *fails* to hold if and only if there is an elementary subspace K of $\ker M^*$ such that $K^\perp$ does not meet $\mathscr{I}$. Let y^* be a vector spanning K; then $\langle y^*, x \rangle$ is nonzero for each $x \in \mathscr{I}$. As $\mathscr{I}$ is a convex set, this means that these inner products all have the same sign, so by choosing $y^* \in K$ to have the correct direction we can satisfy (5.53).

The image of M is spanned by some set of not more than r columns of M, so in case (a) no more than r components of x need be nonzero. In case (b), by construction y^* spans the elementary subspace K of $\ker M^*$. By Theorem 5.5.4, y^* has no more than $m-k+1$ nonzero components, where k is the dimension of $\ker M^*$. But we know from linear algebra that $k = m - r$, so y^* has no more than $r+1$ nonzero components. □

As an example of how one can use this reformulation, we prove a theorem of the alternative attributed to Gale. The exercises for this section ask for proofs of other theorems of the alternative.

Theorem 5.5.8. *For any $m \times n$ matrix A and any $a \in \mathbb{R}^m$, exactly one of the following systems is solvable:*

a. $Ax \geq a$.
b. $A^* u^* = 0$, $\langle u^*, a \rangle > 0$, $u^* \geq 0$.

Proof. It is immediate that (a) and (b) cannot both hold. We suppose (a) does not hold, and apply Corollary 5.5.7 with $M = A$ and $I_i = [a_i, +\infty)$ for $i = 1, \ldots, m$ to conclude that the second alternative in the corollary must hold. That means that there is some u^* with $A^* u^* = 0$ and such that

$$\sum_{i=1}^{m} (u^*)_i [a_i, +\infty) \subset (0, +\infty). \tag{5.54}$$

For each i we choose an arbitrarily large positive number in one of the I_i and the components of a in the others to conclude from (5.54) that $(u^*)_i \geq 0$. Finally, taking $a_i \in I_i$ for every i, we conclude that $\langle u^*, a \rangle > 0$. □

5.5.3 Exercises for Section 5.5

The following exercises ask for proofs of various theorems of the alternative. The names given here are often associated with the theorems in the literature.

Exercise 5.5.9. *(Motzkin)* If D, E, and F are matrices each having n columns, then exactly one of the following systems is solvable:

a. $Dx > 0$, $Ex \geq 0$, $Fx = 0$.
b. $D^*u^* + E^*v^* + F^*w^* = 0$, u^* and v^* nonnegative, and u^* not zero.

Exercise 5.5.10. *(Tucker)* If D, E, and F are matrices each having n columns, then exactly one of the following systems is solvable:

a. $Dx \geq 0$, $Ex \geq 0$, $Fx = 0$, with Dx not zero.
b. $D^*u^* + E^*v^* + F^*w^* = 0$, $u^* > 0$, $v^* \geq 0$.

Exercise 5.5.11. *(Ville)* For any matrix A, exactly one of the following systems is solvable:

a. $Ax < 0$, $x \geq 0$.
b. $A^*w \geq 0$, w nonnegative and not zero.

Exercise 5.5.12. *(Stiemke)* For any matrix A, exactly one of the following systems is solvable:

a. Ax nonnegative and not zero.
b. $A^*w = 0$, $w > 0$.

Exercise 5.5.13. *(A. Ben-Israel [3]):* Let A and D be linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively. Let K be a nonempty polyhedral convex cone in $\mathbb{R}^p$ and L a nonempty closed convex cone in $\mathbb{R}^q$. Show that the following are equivalent:

a. $Ax \in K$ implies $Dx \in L$.
b. $A^*(K^\circ) \supset D^*(L^\circ)$.

5.6 Notes and references

Polyhedral multifunctions were introduced in [31], where Theorem 5.4.13, Corollary 5.4.16, and a version of Proposition 5.4.14 appeared. The Hoffman theorem is from [16]. Instead of proving Corollary 5.4.11 from Hoffman's theorem as we did here, one can derive it from a more general theorem [44, Theorem 1] of Walkup and Wets, which actually characterizes polyhedrality.

The proof of Theorem 5.5.6 is from [33].

Part II
Convex Functions

Chapter 6
Basic properties

Convex functions are very important in both theory and practice, especially in the field of optimization. One of the reasons for their importance in optimization is that each local minimizer of a convex function is a global minimizer. As numerical algorithms usually produce only local minimizers, it is of great value to be able to identify cases in which such a local minimizer is in fact global.

Continuity is a basic property important in many areas of mathematical work. In order to establish properties and behavior of functions, we often require that the function be continuous. Nonetheless, many very important and useful convex functions are not continuous, but only *lower semicontinuous.* The lower semicontinuity property turns out to be the appropriate "niceness" condition for most operations with convex functions, especially when combined with a restriction on infinite function values to yield the class of *closed* convex functions.

6.1 Convexity

This section shows some reasons why convex functions are important in optimization and develops some of their most elementary properties.

6.1.1 Convex functions with extended real values

In applications using convex functions—particularly in optimization—we often want to define a function for points in a linear space such as $\mathbb{R}^n$, but not for all such points. Explicitly specifying the domain of definition will sometimes be inconvenient, so we will define functions on all of $\mathbb{R}^n$ but allow them to take values in the *extended real numbers*

$$\bar{\mathbb{R}} := \{-\infty\} \cup \mathbb{R} \cup \{+\infty\},$$

and we call such functions *extended-real-valued,* which we sometimes abbreviate as ERV.

For such a function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ the points of $\mathbb{R}^n$ at which f takes values strictly less than $+\infty$ correspond to the traditional domain of definition, and we call the set of all such points the *effective domain* of f, written $\operatorname{dom} f$. Points with values of $+\infty$ would in the traditional situation be outside the domain of definition. Points with values of $-\infty$ are in $\operatorname{dom} f$, but we will often restrict f so that there will be no such points.

The presence of the two special elements of $\bar{\mathbb{R}}$ does not cause any trouble in calculation as long as we observe three simple rules:

a. The product of any positive (negative) number with $\pm\infty$ is $\pm\infty$ ($\mp\infty$), and the product of zero with anything is zero.
b. The sum of any real number and $\pm\infty$ is $\pm\infty$.
c. The sum of either of the infinite elements with itself is itself.

We do not define the sums $(+\infty)+(-\infty)$ and $(-\infty)+(+\infty)$. Given rule (a), this implies that the differences $(+\infty)-(+\infty)$ and $(-\infty)-(-\infty)$ are also undefined. In calculating with ERV functions it is important to ensure that these undefined situations do not arise.

In order to have a concept of closeness for elements of $\bar{\mathbb{R}}$, we need to have a topology for it. Proposition C.3.1 in Section C shows how to construct such a topology. The base for that topology consists of all open sets of $\mathbb{R}$ together with all sets in $\bar{\mathbb{R}}$ of the form $[-\infty,\mu)$ or $(\mu,+\infty]$, where $\mu \in \mathbb{R}$, and the set

$$[-\infty,+\infty] := \{-\infty\} \cup \mathbb{R} \cup \{+\infty\},$$

which is $\bar{\mathbb{R}}$.

We associate with each function from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$ a set called the *epigraph* of the function.

Definition 6.1.1. If $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, its *epigraph* is the set $\operatorname{epi} f := \{\, (x,\mu) \in \mathbb{R}^{n+1} \mid f(x) \le \mu \,\}$. □

The epigraph is exceedingly useful, as it permits us to translate properties of functions into geometric properties of convex sets and vice versa. For example, the following definition of an extended-real-valued convex function rests on a geometric property of its epigraph.

Definition 6.1.2. A function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is *convex* if $\operatorname{epi} f$ is convex; f is *concave* if $-f$ is convex. □

The epigraph of f consists of those pairs $(x,\xi) \in \mathbb{R}^{n+1}$ having $f(x) \le \xi$. For x belonging to such a pair, $f(x)$ cannot be $+\infty$; on the other hand, if $f(x) < +\infty$ then there is a real ξ such that $(x,\xi) \in \operatorname{epi} f$. Therefore the projection of $\operatorname{epi} f$ onto $\mathbb{R}^n$ is the effective domain of f, so if f is convex then $\operatorname{dom} f$ is a convex set.

A simple example of a convex function that takes infinite values is the *indicator* of a convex subset C of $\mathbb{R}^n$; this is the function I_C taking the value 0 on C and $+\infty$

off C. If C is not all of $\mathbb{R}^n$ then I_C has to take $+\infty$ somewhere. The epigraph of I_C is just $C \times \mathbb{R}_+$, which will be convex exactly when C is convex.

In visualizing concave functions g one can use the *hypograph*, consisting of the pairs (x,μ) in $\mathbb{R}^{n+1}$ having $\mu \leq g(x)$. For example, one can define the *concave indicator* of a set C by specifying its hypograph to be $C \times \mathbb{R}_-$. However, operations involving the hypograph turn out to be just reflections of the corresponding operations for epigraphs, so that in most cases one need not keep track of separate definitions.

Sometimes we can use the definition to determine whether a function is convex, but often it is more convenient to have a test. The following proposition gives a test applicable to any ERV function.

Proposition 6.1.3. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$. The function f is convex if and only if for each pair $\{x,y\} \in \mathbb{R}^n$, each pair of real numbers ξ and η such that $f(x) < \xi$ and $f(y) < \eta$, and each $\lambda \in (0,1)$ one has*

$$f[(1-\lambda)x+\lambda y] < (1-\lambda)\xi + \lambda\eta.$$

Proof. (only if). Suppose that f is convex, and choose x, y, ξ, η, and λ as described. Find real numbers $\xi' \in (f(x),\xi)$ and $\eta' \in (f(y),\eta)$; then the pairs (x,ξ') and (y,η') belong to epi f, which is a convex set because f is convex. Then the pair $((1-\lambda)x+\lambda y,(1-\lambda)\xi'+\lambda\eta')$ also belongs to epi f, and so

$$f[(1-\lambda)x+\lambda y] \leq (1-\lambda)\xi'+\lambda\eta' < (1-\lambda)\xi+\lambda\eta.$$

(if). Assume that f satisfies the hypothesis. Choose any points (x,μ) and (y,ν) in epi f, and let $\lambda \in (0,1)$. For each positive ε we have $f(x) < \mu+\varepsilon$ and $f(y) < \nu+\varepsilon$. Therefore

$$f[(1-\lambda)x+\lambda y] < (1-\lambda)(\mu+\varepsilon)+\lambda(\nu+\varepsilon) = (1-\lambda)\mu+\lambda\nu+\varepsilon.$$

Now allowing ε to approach zero we obtain

$$f[(1-\lambda)x+\lambda y] \leq (1-\lambda)\mu+\lambda\nu,$$

which implies that $(1-\lambda)(x,\mu)+\lambda(y,\nu) \in$ epi f, so that f must be convex. □

Corollary 6.1.4. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be convex with $f(0) = 0$. Then for each $y \in \mathbb{R}^n$, $f(y) \geq -f(-y)$.*

Proof. Choose $y \in \mathbb{R}^n$ and suppose that $f(y) < -f(-y)$. Find real numbers α and β such that $f(y) < \alpha < \beta < -f(-y)$. Apply Theorem 6.1.3 using $f(y) < \alpha$ and $f(-y) < -\beta$ to show that

$$0 = f(0) = f[.5y+.5(-y)] < .5\alpha+.5(-\beta).$$

This implies that $\alpha > \beta$, which is false, so we must have $f(y) \geq -f(-y)$. □

Strict inequalities and the bounds ξ and η appear in (6.1.3) in order to avoid problems with extended real values. For example, let $f : \mathbb{R} \to \bar{\mathbb{R}}$ be be defined by

$$f(x) = \begin{cases} -\infty & \text{if } f < 0, \\ 0 & \text{if } f = 0, \\ +\infty & \text{if } f > 0. \end{cases}$$

If we were to apply a test using weak, instead of strict, inequalities and omitting the bounds we could obtain statements like

$$f(0) = f[.5(-1) + .5(+1)] \leq .5f(-1) + .5f(+1) = .5(-\infty) + .5(+\infty),$$

in which the right side is undefined. Proposition 6.1.8 below shows that if f never takes $-\infty$ then we can use weak inequalities. However, there are still situations in which the strict inequalities are helpful, e.g. in the proof of Theorem 9.3.1 below.

Here is a very important application of convexity to optimization. Although extremely simple, it yields one of the few available tests for global minimization of a function. It also provides a good example of the use of Proposition 6.1.3.

Definition 6.1.5. If $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, we say a point $x \in \mathbb{R}^n$ is a *local minimizer* of f if $x \in \operatorname{dom} f$ and there is some neighborhood N of x such that whenever $x' \in N$, $f(x') \geq f(x)$. The point x is a *global minimizer* of f if x is a local minimizer for which we can take $N = \mathbb{R}^n$.

Proposition 6.1.6. *If $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is convex, then each local minimizer of f is a global minimizer.*

Proof. If there are no local minimizers, the result is true. Therefore assume that x is a local minimizer of f. If $f(x) = -\infty$ then x is a global minimizer of f. Otherwise, as $x \in \operatorname{dom} F$, $f(x)$ is finite. Let y be any point of $\mathbb{R}^n$; we show that $f(y) \geq f(x)$. Assume the contrary: then $f(y) < f(x)$ and we can choose real η and δ such that

$$f(y) < \eta < \eta + \delta < f(x). \tag{6.1}$$

For small $\varepsilon \in (0,1)$ we have $f(x) \leq f[x + \varepsilon(y - x)]$ because x is a local minimizer. Then $f(x) < [f(x) + (1-\varepsilon)^{-1}\varepsilon\delta]$ and $f(y) < \eta$, so Proposition 6.1.3 shows that

$$f(x) \leq f[(1-\varepsilon)x + \varepsilon y] < (1-\varepsilon)[f(x) + (1-\varepsilon)^{-1}\varepsilon\delta] + \varepsilon\eta. \tag{6.2}$$

We can rewrite (6.2) as

$$f(x) < (1-\varepsilon)f(x) + \varepsilon\delta + \varepsilon\eta,$$

and since $\varepsilon > 0$ this reduces to $f(x) < \delta + \eta$, which contradicts (6.1). Therefore x is a global minimizer of f. □

The next definition produces a class of ERV functions that do not have certain inconvenient features, such as values of $-\infty$.

Definition 6.1.7. A function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is *proper* if f never takes $-\infty$ and f is not identically $+\infty$. If f is not proper it is *improper.* □

If f is a proper function then dom f is nonempty and f takes finite values everywhere on that set.

The following proposition simplifies the convexity criterion of Proposition 6.1.3 in the case of functions that do not take $-\infty$. In particular, this includes proper functions.

Proposition 6.1.8. *Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. The function f is convex if and only if for each pair $x, y \in \mathbb{R}^n$ and each $\lambda \in (0,1)$,*

$$f[(1-\lambda)x + \lambda y] \leq (1-\lambda)f(x) + \lambda f(y).$$

Proof. (only if). Choose x, y, and λ as described. If f takes $+\infty$ at x or at y, the conclusion is obvious. In the remaining case $f(x)$ and $f(y)$ are finite. Choosing any positive ε and applying Proposition 6.1.3 to $\xi := f(x) + \varepsilon$ and $\eta := f(y) + \varepsilon$ we obtain

$$f[(1-\lambda)x + \lambda y] < (1-\lambda)f(x) + \lambda f(y) + \varepsilon,$$

and by letting ε approach zero we obtain the desired result.

(if). If the inequality in the statement of the theorem holds for all x, y and λ as described, then if μ and ν are real numbers with $f(x) < \mu$ and $f(y) < \nu$ we have

$$f[(1-\lambda)x + \lambda y] \leq (1-\lambda)f(x) + \lambda f(y) < (1-\lambda)\mu + \lambda \nu.$$

Then f is convex by Proposition 6.1.3. □

The next result provides a useful result that is often called *Jensen's inequality,* though that term also applies to more general forms: e.g., with integrals.

Proposition 6.1.9. *Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then f is convex if and only if for each integer $k \geq 1$, each set of points $x_1, \ldots, x_k \in \mathbb{R}^n$ and each set of convex coefficients $\lambda_1, \ldots, \lambda_k$,*

$$f(\sum_{i=1}^{k} \lambda_i x_i) \leq \sum_{i=1}^{k} \lambda_i f(x_i).$$

Proof. For the "if" part we can just take $k = 2$ and use Proposition 6.1.8. For "only if," suppose f is convex and note that we can suppose all of the λ_i to be positive (if not, they contribute nothing to the sums). If k is 1 or 2, we have the result already. Therefore suppose that $k > 2$ and, for an induction, that the theorem is true for integers from 1 to $k-1$ inclusive. We prove it for k.

If $\lambda_k = 1$ then we are in the previous case since then there is only one nonzero λ_i. Therefore suppose $\lambda_k < 1$ and write $x' = \sum_{i=1}^{k-1} (1-\lambda_k)^{-1} \lambda_i x_i$. By induction we have

$$f(x') \leq \sum_{i=1}^{k-1} (1-\lambda_k)^{-1} \lambda_i f(x_i),$$

and also

$$f(\sum_{i=1}^{k} \lambda_i x_i) = f[(1-\lambda_k)x' + \lambda_k x_k] \leq (1-\lambda_k)f(x') + \lambda_k f(x_k).$$

Combining these two inequalities we have the assertion. □

Sometimes we are given a convex set $C \subset \mathbb{R}^n$ and a function $g : C \to \mathbb{R}$ that is convex on C in the traditional sense:

$$\begin{aligned} &\text{For each } c_1 \text{ and } c_2 \text{ in } C \text{ and each } \lambda \in [0,1], \\ &g[(1-\lambda)c_1 + \lambda c_2] \leq (1-\lambda)g(c_1) + \lambda g(c_2). \end{aligned} \tag{6.3}$$

We can extend such a function from the set C where it is defined to an ERV function on all of $\mathbb{R}^n$ by defining the $+\infty$ *extension* of g to be the function f that agrees with g on C and takes the value $+\infty$ off C. In such a case we have $\operatorname{dom} f = C$, and if C is nonempty then f is proper. The next proposition shows that the $+\infty$ extension f is convex in the sense of Definition 6.1.2 exactly when g is convex in the sense of (6.3).

Proposition 6.1.10. *Let C be a nonempty convex subset of $\mathbb{R}^n$ and let $g : C \to \mathbb{R}$. The function g is convex on C in the sense of* (6.3) *if and only if the $+\infty$ extension f of g is convex in the sense of Definition 6.1.2.*

Proof. (Only if.) Suppose g is convex in the sense of (6.3). The function f is proper, so the convexity criterion of Proposition 6.1.8 applies. Let x and y be points in $\mathbb{R}^n$ and fix $\lambda \in (0,1)$. If either x or y is outside C then the corresponding value of f is $+\infty$, so f trivially satisfies the inequality in Proposition 6.1.8. Otherwise f agrees with g at both x and y, which must then belong to C, so the inequality of Proposition 6.1.8 follows from that of (6.3). Therefore f is convex.

(If.) If f is convex, then the inequality of Proposition 6.1.8 holds for f everywhere on $\mathbb{R}^n$; in particular it holds on C, where g and f agree. Therefore g is convex in the sense of (6.3). □

One of the nice properties of convexity is that several functional operations preserve it. The next proposition develops one of these.

Proposition 6.1.11. *Let A be a nonempty set and let f be a function from $\mathbb{R}^n \times A$ to $\bar{\mathbb{R}}$ such that for each $a \in A$, $f(\cdot, a)$ is convex. Then the function*

$$g(x) := \sup_{a \in A} f(x,a)$$

is convex on $\mathbb{R}^n$.

Proof. $\operatorname{epi} g = \cap_{a \in A} \operatorname{epi} f(\cdot, a)$, and each of the sets in the intersection is convex by hypothesis. Therefore $\operatorname{epi} g$ is convex, so g is convex. □

Convex functions have very strong boundedness and continuity properties. The next two results establish some of the most important of these.

Lemma 6.1.12. *Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ be a convex function and let x be a point of* $\operatorname{ri}\operatorname{dom} f$. *Then f is bounded above on some neighborhood of x relative to* $\operatorname{dom} f$.

The qualification "relative to $\operatorname{dom} f$" means that we will construct a neighborhood of x in the relative topology that $\operatorname{dom} f$ inherits from $\mathbb{R}^n$, on which f will have a finite upper bound. We need the relative topology here because f takes $+\infty$ at all points not in $\operatorname{dom} f$, so no finite upper bound could hold for such points.

Proof. By hypothesis $\operatorname{dom} f$ is nonempty and f never takes $-\infty$, so it is proper. Let $\operatorname{dom} f$ have dimension $k \geq 0$ and use Lemma 1.1.14 to construct a k-simplex σ contained in $\operatorname{dom} f$ with x as its barycenter, and therefore with $x \in \operatorname{ri}\sigma$. Then σ is a neighborhood of x in the relative topology on $\operatorname{dom} f$. However, each point of σ is a convex combination of the $k+1$ vertices of σ, and therefore by Jensen's inequality f is bounded above on σ by the maximum value it attains on those vertices. □

The following theorem establishes local Lipschitz continuity of a proper convex function f, on $\operatorname{ri}\operatorname{dom} f$, relative to $\operatorname{dom} f$. Saying that a function is locally Lipschitz continuous(equivalently, *locally Lipschitzian*) at a point x means that it obeys a Lipschitz condition on some neighborhood of x. As f is finite only on $\operatorname{dom} f$, the Lipschitz condition here only holds with respect to points near x that are in $\operatorname{dom} f$: that is, on a relative neighborhood of x in $\operatorname{dom} f$. This continuity means in particular that f is bounded, rather than just bounded above, on $\operatorname{dom} f$ near a point $x \in \operatorname{ri}\operatorname{dom} f$.

Theorem 6.1.13. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a proper convex function. If $x \in \operatorname{ri}\operatorname{dom} f$ then f is locally Lipschitz continuous at x relative to* $\operatorname{dom} f$.

Proof. If $\operatorname{dom} f$ is a singleton the result is true, so assume that $\operatorname{dom} f$ contains more than one point. Let $x \in \operatorname{ri}\operatorname{dom} f$, and let ε be a positive number such that $N := B(x, 2\varepsilon) \cap \operatorname{aff}\operatorname{dom} f$ is contained in $\operatorname{dom} f$ and such that, by Lemma 6.1.12, f has an upper bound β on N. If y is any point of N then the point $w = 2x - y$ also belongs to N. We have $x = (1/2)y + (1/2)w$, so

$$f(x) \leq (1/2)f(y) + (1/2)f(w) \leq (1/2)f(y) + (1/2)\beta,$$

and therefore $f(y) \geq 2f(x) - \beta =: \gamma$. For later use we note that $\beta - \gamma = 2[\beta - f(x)] \geq 0$.

Let $M = B(x, \varepsilon) \cap \operatorname{aff}\operatorname{dom} f$, and let x' and x'' be any two distinct points of M. Let $t = \varepsilon / \|x' - x''\|$ and define $w' = x' - t(x'' - x')$. We have

$$\|w' - x\| \leq \|x' - x\| + t\|x'' - x'\| \leq 2\varepsilon,$$

so that $w' \in N$. Also, $x' = (1+t)^{-1}w' + t(1+t)^{-1}x''$. The convexity of f yields

$$\begin{aligned} f(x') &\leq (1+t)^{-1}f(w') + t(1+t)^{-1}f(x'') \\ &= f(x'') + (1+t)^{-1}[f(w') - f(x'')] \\ &\leq f(x'') + (1+t)^{-1}(\beta - \gamma), \end{aligned}$$

so that

$$f(x') - f(x'') \leq (1+t)^{-1}(\beta - \gamma).$$

By interchanging the roles of x' and x'' we obtain

$$f(x'') - f(x') \leq (1+t)^{-1}(\beta - \gamma),$$

and therefore

$$|f(x') - f(x'')| \leq (1+t)^{-1}(\beta - \gamma).$$

As $0 < t < 1+t$ we have $(1+t)^{-1} < t^{-1} = \varepsilon^{-1}\|x' - x''\|$, so that

$$|f(x') - f(x'')| \leq \varepsilon^{-1}(\beta - \gamma)\|x' - x''\|,$$

and this shows that f is Lipschitz continuous on M with modulus $\varepsilon^{-1}(\beta - \gamma)$. □

6.1.2 Some special results for finite-valued convex functions

Sometimes one has to deal with convex functions having the special properties that they are finite-valued and defined on some convex subset C of $\mathbb{R}^n$. They may have additional properties such as varying degrees of differentiability. In that case one can prove results stronger than those obtainable for general convex functions, and this section collects some of those results.

Definition 6.1.14. Let f be a real-valued function defined on a convex subset C of $\mathbb{R}^n$. A scalar ϕ is called a *modulus of convexity* for f on C if for each x and y in C and each $\lambda \in (0,1)$,

$$f[(1-\lambda)x + \lambda y] \leq (1-\lambda)f(x) + \lambda f(y) - (1/2)\phi\lambda(1-\lambda)\|x-y\|^2.$$

□

Definition 6.1.15. Let f be as in Definition 6.1.14 We say f is *strongly convex* if it has a positive modulus of convexity, *convex* if it has a zero modulus of convexity, and *weakly convex* if it has a negative modulus of convexity. In addition, we say f is *strictly convex* if for each x and y in C with $x \neq y$ and for each $\lambda \in (0,1)$, $f[(1-\lambda)x + \lambda y] < (1-\lambda)f(x) + \lambda f(y)$. □

If ϕ is a modulus of convexity for f then so is any smaller scalar. Accordingly, a function belonging to any of the categories weakly convex, convex, strictly convex, or strongly convex also belongs to each of the preceding categories.

It is often inconvenient to use the definition to verify that a number ϕ is a modulus of convexity for a function. If the function is differentiable, one can often obtain information about its convexity by testing the derivatives instead. Here are two tests, one using the first derivative and one using the second.

Theorem 6.1.16. *Let Ω be an open subset of $\mathbb{R}^n$, f a C^1 function from Ω to $\mathbb{R}$, and C a convex subset of Ω. For any real scalar ϕ, the following are equivalent:*

a. ϕ is a modulus of convexity for f on C.
b. For each x and y in C, $f(y) \geq f(x) + df(x)(y-x) + (\phi/2)\|y-x\|^2$.
c. For each x and y in C, $[df(y) - df(x)](y-x) \geq \phi\|y-x\|^2$.

Proof. ((a) implies (b)). Assume that ϕ is a modulus of convexity for f on C, and let x and y be points of C. For each small positive t, we have

$$(1-t)f(x) + tf(y) - f[(1-t)x+ty] \geq (\phi/2)t(1-t)\|y-x\|^2.$$

Therefore

$$t[f(y)-f(x)] \geq f[(1-t)x+ty] - f(x) + (\phi/2)t(1-t)\|y-x\|^2,$$

and so

$$f(y)-f(x) \geq t^{-1}\{f[x+t(y-x)] - f(x)\} + (\phi/2)(1-t)\|y-x\|^2.$$

Taking the limit as $t \downarrow 0$, we have (b).

((b) implies (a)). Suppose (b) holds. Let $\lambda \in (0,1)$ and write $w = (1-\lambda)x + \lambda y$, so that $y = w + (1-\lambda)(y-x)$ and $x = w + (-\lambda)(y-x)$. Using (b), we obtain

$$f(y) \geq f(w) + df(w)(1-\lambda)(y-x) + (\phi/2)(1-\lambda)^2\|y-x\|^2, \tag{6.4}$$

and

$$f(x) \geq f(w) + df(w)(-\lambda)(y-x) + (\phi/2)(-\lambda)^2\|y-x\|^2. \tag{6.5}$$

Multiplying (6.4) by λ and (6.5) by $(1-\lambda)$ and adding, we find that

$$\lambda f(y) + (1-\lambda)f(x) - f[(1-\lambda)x+\lambda y] \geq (\phi/2)\lambda(1-\lambda)\|y-x\|^2,$$

so that ϕ is a modulus of convexity for f on C.

((b) implies (c)). Assume that (b) holds, and choose any x and y in C. We have

$$f(y) - f(x) \geq df(x)(y-x) + (\phi/2)\|y-x\|^2,$$

and

$$f(x) - f(y) \geq df(y)(x-y) + (\phi/2)\|y-x\|^2,$$

so $[df(y) - df(x)](y-x) \geq \phi\|y-x\|^2$, which is (c).

((c) implies (b)). Let (c) hold, and suppose that x and y belong to C. Then writing $h = y-x$, we have

$$f(y) - f(x) = \int_0^1 df(x+th)h\,dt = df(x)h + \int_0^1 [df(x+th) - df(x)]h\,dt.$$

Now $[df(x+th) - df(x)](th) \geq \phi\|th\|^2$, so

$$\int_0^1 [df(x+th)-df(x)]h\,dt \geq \phi\|h\|^2 \int_0^1 t\,dt = (\phi/2)\|h\|^2.$$

Therefore we have

$$f(y) \geq f(x)+df(x)(y-x)+(\phi/2)\|y-x\|^2,$$

as required. □

Suppose f is a C^2 function from a neighborhood N of $x \in \mathbb{R}^n$ to $\mathbb{R}^m$. The second derivative d^2f, when evaluated at $x \in \mathbb{R}^n$, produces a linear operator $d^2f(x) : \mathbb{R}^n \to \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$, where $\mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ denotes the class of linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$. We will write $d^2f(x)yz$ for the result of applying $d^2f(x)$ to $y \in \mathbb{R}^n$ to produce the linear operator $J := d^2f(x)y : \mathbb{R}^n \to \mathbb{R}^m$, then applying J to $z \in \mathbb{R}^n$ to produce $J(z) \in \mathbb{R}^m$.

When $m = 1$ it is common to represent $d^2f(x)$ by an $n \times n$ symmetric matrix D, with the interpretation that $d^2f(x)yz = \langle Dy, z\rangle$. In this case we will say that $d^2f(x)$ is *positive semidefinite* or *positive definite* if the representation D has these properties. When $m > 1$ one can still use matrix representatives, but with m matrices $D_1, \ldots, D_m$. This can be somewhat awkward, and we will generally avoid this notation in favor of that explained in the preceding paragraph.

Theorem 6.1.17. *Let Ω be an open subset of $\mathbb{R}^n$; let C be a relatively open convex subset of Ω and L be the parallel subspace of C. Let f be a C^2 function from Ω to $\mathbb{R}$. A scalar ϕ is a modulus of convexity for f on C if and only if for each $x \in C$ and each $z \in L$, $d^2f(x)zz \geq \phi\|z\|^2$.*

Proof. If C is empty or a singleton there is nothing to prove, so we can assume that L contains a nonzero point.

(only if). Suppose ϕ is a modulus of convexity for f on C, and choose $z \in L$. If $z = 0$ then the assertion is true, so assume $z \neq 0$ and define $w = z/\|z\|$. Let $x \in C$ and let t be a real number small enough that $x+tw$ and $x-tw$ belong to C. We have

$$\begin{aligned} f(x) &= f[(1/2)(x+tw)+(1/2)(x-tw)] \\ &\leq (1/2)f(x+tw)+(1/2)f(x-tw)-(\phi/2)(1/4)\|2tw\|^2, \end{aligned}$$

so that

$$(\phi/2)t^2 \leq (1/2)f(x+tw)+(1/2)f(x-tw)-f(x). \tag{6.6}$$

We also have

$$f(x+tw)-f(x) = tdf(x)w+[(t^2)/2]d^2f(x)ww+o(t^2), \tag{6.7}$$

and

$$f(x-tw)-f(x) = -tdf(x)w+[(t^2)/2]d^2f(x)ww+o(t^2). \tag{6.8}$$

Multiplying (6.7) and (6.8) by $1/2$, adding, and combining the result with (6.6), we obtain

$$(\phi/2)t^2 \le [(t^2)/2]d^2f(x)ww + o(t^2),$$

so that

$$\phi \le d^2f(x)ww + t^{-2}o(t^2),$$

implying $d^2f(x)ww \ge \phi$. But then $d^2f(x)zz \ge \phi\|z\|^2$.

(if). Suppose that for each $x \in C$ and each $z \in L$ we have $d^2f(x)zz \ge \phi\|z\|^2$. Let x and y belong to C and suppose $\lambda \in (0,1)$. For each $z \in C$, each $v \in L$, and any scalar t such that $z+tv \in C$, we have for $\sigma \in [0,1]$

$$df(z+\sigma tv) - df(z) = \int_0^1 d^2f(z+\alpha\sigma tv)(\sigma tv)d\alpha, \tag{6.9}$$

and

$$f(z+tv) - f(z) = df(z)tv + \int_0^1 [df(z+\sigma tv) - df(z)](tv)d\sigma \tag{6.10}$$

Using (6.9) in (6.10), we find that

$$\begin{aligned} f(z+tv) - f(z) &= df(z)tv + \int_0^1 [\int_0^1 d^2f(z+\alpha\sigma tv)(\sigma tv)d\alpha](tv)d\sigma \\ &\ge df(z)tv + \int_0^1 \int_0^1 \phi\sigma t^2\|v\|^2 d\alpha d\sigma \\ &= df(z)tv + (\phi/2)t^2\|v\|^2. \end{aligned}$$

Now let $z = (1-\lambda)x + \lambda y$, so that $y = z + (1-\lambda)v$ and $x = z - \lambda v$, where $v = y - x$. Then

$$\begin{aligned} f(y) - f(z) &\ge (1-\lambda)df(z)(y-x) + (\phi/2)(1-\lambda)^2\|y-x\|^2, \\ f(x) - f(z) &\ge (-\lambda)df(z)(y-x) + (\phi/2)(-\lambda)^2\|y-x\|^2, \end{aligned}$$

and therefore

$$\begin{aligned} &\lambda f(y) + (1-\lambda)f(x) - f[(1-\lambda)x + \lambda y] \\ &\quad \ge (\phi/2)[\lambda(1-\lambda)^2 + (1-\lambda)(-\lambda)^2]\|y-x\|^2 \\ &\quad = (\phi/2)\lambda(1-\lambda)\|y-x\|^2. \end{aligned}$$

Therefore ϕ is a modulus of convexity for f on C. □

Corollary 6.1.18. *Let Ω, C, L, and f be as in Theorem 6.1.17. The function f is convex on C if and only if for each $x \in C$, $d^2f(x)$ is positive semidefinite on L.*

Proof. f is convex on C if and only if it has 0 as a modulus of convexity there. □

Corollary 6.1.18 shows, for example, that for an $n \times n$, not necessarily symmetric, matrix A, an n-vector a and a scalar α the quadratic function $\langle Ax,x\rangle + \langle a,x\rangle + \alpha$ is convex on $\mathbb{R}^n$ if and only if the symmetric matrix $A + A^*$ is positive semidefinite.

6.1.3 *Example: Convexity and an intractable function

Here is an illustration of a function occurring naturally in an application, but not having a tractable closed-form representation. Nevertheless, we will see that this function is convex, so that all of the theory that we have already developed and will develop below applies to it. In fact, some of that theory is central to effective computing methods for dealing with such functions. To state the example we need some technical terms from probability, but these do not affect the aspects of the example with which we are concerned.

Suppose that customers arrive at a service center according to a proper renewal process with interarrival times λ_i: that is, the process begins at time zero; the first customer arrives at epoch $a_1 = \lambda_1$, the second at epoch $a_2 = \lambda_1 + \lambda_2$, etc., where the λ_i have a common continuous distribution. A customer who arrives when no others are present is served immediately; if others are present the arriving customer joins a queue and is served after all prior customers have been served in the order of their arrival. The ith customer's service requires a time μ_i that is of the form $p + \gamma_i$, where p is a fixed positive real number and the γ_i are nonnegative, with a common continuous probability distribution, and are independent of each other and of the λ_i. A customer's *waiting time* is the time from arrival until the beginning of service; thus it is the total time in the system less the service time.

The management of the service center wants to avoid long waiting times because they make customers unhappy. Under appropriate technical assumptions on p and on the λ_i and γ_i, the managers can be sure that the expected value of the waiting time is finite. Moreover, by reducing p they can reduce the service time, which should reduce waiting time, but at a cost. The contribution of p to the total daily cost of operating the system is given by a function $\pi(p)$ that is convex, continuous, and, for values of p in the interval $[p_L, p_U]$, decreasing; here $0 < p_L < p_U < +\infty$ and the managers are able to set p at any value in $[p_L, p_U]$. Thus if p decreases, the daily cost increases.

From past experience the managers believe that the contribution of the average waiting time w_* to the average daily revenue from customers is given by a real-valued function $\delta(w^*)$ that is negative, concave, continuous, and decreasing: thus, if average waiting time increases, then the average daily revenue will decrease. Accordingly, the managers would like to find the best tradeoff of p against average waiting time. Specifically, if for a given value of p the average waiting time is $w_*(p)$, then they would like to set p so as to minimize the function $\pi(p) - \delta[w_*(p)]$ over values of p in $[p_L, p_U]$. However, in order to implement such a plan they need to know the relationship of $w_*(p)$ to p.

To understand how $w_*(p)$ (or an approximation to it) is related to p, we will investigate the way in which individual waiting times are determined. All these times depend on p, but as for the moment p is fixed, to simplify the notation we suppress p in the part of the analysis immediately following.

- Suppose that the system starts operating at time zero, and that the first customer arrives at time $a_1 = \lambda_1 > 0$. This customer finds the system empty, and

immediately enters service, departing as soon as service is completed at time $d_1 = \lambda_1 + p + \gamma_1$. The waiting time w_1 for this first customer is zero.

- Now suppose that $k \geq 1$. The $k+1$st customer arrives at $a_{k+1} = a_k + \lambda_{k+1}$. We have

$$\begin{aligned} d_k - a_{k+1} &= (a_k + w_k + p + \gamma_k) - (a_k + \lambda_{k+1}) \\ &= w_k + p + \gamma_k - \lambda_{k+1}. \end{aligned} \tag{6.11}$$

If $d_k - a_{k+1} > 0$ then this customer has to wait to enter service until d_k, departing at $d_{k+1} = d_k + p + \gamma_{k+1}$. The waiting time is then

$$\begin{aligned} w_{k+1} &= d_{k+1} - p - \gamma_{k+1} - a_{k+1} \\ &= (d_k + p + \gamma_{k+1}) - p - \gamma_{k+1} - a_{k+1} \\ &= d_k - a_{k+1} \\ &= w_k + p + \gamma_k - \lambda_{k+1}, \end{aligned}$$

where the last line came from (6.11).

On the other hand, if $d_k - a_{k+1} < 0$ then this customer enters service immediately and $w_{k+1} = 0$. In this case, again from (6.11), we have $w_k + p + \gamma_k - \lambda_{k+1} < 0$. Therefore

$$w_{k+1} = \max\{0, w_k + p + \gamma_k - \lambda_{k+1}\}. \tag{6.12}$$

We have $w_1 = 0$, while (6.12) gives w_{k+1} for each $k \geq 1$.

With this expression in hand, the managers decide to attack their problem by simulating w_k for a long run of n arrivals, then to estimate $w_*(p)$ by the sample average

$$m_*(p) = n^{-1} \sum_{k=1}^{n} w_k(p).$$

If n is chosen to be large, there is *a priori* a high probability that $m_*(p)$ will be close to $w_*(p)$. For a given sample path produced by simulation, the managers plan to regard $m_*(p)$ as a deterministic function of p, then optimize it. This is the method of *sample-path optimization,* and it can be shown that under suitable assumptions, for large n an optimizer of $m^*(p)$ will with probability 1 be close to an optimizer of $w_*(p)$ [32].

However, there is a difficulty: the preceding paragraph discussed an optimizer of $m^*(p)$, but how are we to know that one exists? It certainly is infeasible to write out an expression for $m^*(p)$ for n of the size (in the tens of thousands) typically used in sample-path optimization. If we cannot even write down the function, how are we to answer questions about the existence of an optimizer, or about whether it is a local or global optimizer?

Here the structure of $w_k(p)$, shown in (6.12), is of crucial importance. Recall that once the sample path is generated, the γ_k and λ_k are no longer random variables but are fixed, known numbers. Then beginning with $k = 2$ we argue that $w_2(p)$, as the maximum of a constant function and an affine function of p, is convex and continuous. Then for $k \geq 2$, w_{k+1} will be the maximum of a constant and the convex,

continuous function $w_k(p)$, and therefore it will be convex and continuous too. In general we would be using lower semicontinuity instead of continuity, but in this case the expressions are simple enough that continuity holds.

Finally, since we are dealing with finite convex functions on the interval $[p_L, p_U]$, the sum of all the $w_k(p)$, and therefore the sample mean $m_*(p)$ will be a convex continuous function of p. Then the tradeoff function $\pi(p) - \delta[m_*(p)]$ to be minimized will also be convex and continuous. As $[p_L, p_U]$ is compact and convex, this means that first, we are guaranteed that a minimizer exists, and second, we are guaranteed that any local minimizer will be a global minimizer.

This example is highly simplified in order to provide some insight into the methodology without introducing more complications and technical details than were necessary. However, the reasoning used here applies to much more complex problems as well. For example, the paper [29] reports computational solution of two classes of such problems: optimizing buffer sizes in manufacturing transfer lines with random breakdown and repair, and cost optimization in stochastic PERT networks. At the time that work was done, the problems reported were larger than any of their kind for which solutions had been reported. In those cases the tools of convexity were important not only for analyzing the minimization problem but also for providing some of the computational methods for obtaining the numerical solutions.

6.1.4 Exercises for Section 6.1

Exercise 6.1.19. Suppose that g is a convex function from $\mathbb{R}^n \times \mathbb{R}^m$ to $\bar{\mathbb{R}}$. Show that the function

$$h(x) := \inf_{y \in \mathbb{R}^m} g(x,y)$$

is convex on $\mathbb{R}^n$.

Exercise 6.1.20. Show that if f is a convex function from $\mathbb{R}^n$ to $(-\infty, +\infty]$ and g is a nondecreasing convex function from $(-\infty, +\infty]$ to $(-\infty, +\infty]$ such that $g(+\infty) = +\infty$, then the composition $g \circ f$ is convex.

Exercise 6.1.21. Let f be a proper convex function on $\mathbb{R}^n$. We say that f is *positively homogeneous* if for each $\tau > 0$ and each $x \in \mathbb{R}^n$, $f(\tau x) = \tau f(x)$. Show that f is positively homogeneous if and only if epi f is a convex cone in $\mathbb{R}^{n+1}$.

Exercise 6.1.22. Let C be a convex subset of $\mathbb{R}^{n+1}$ and for each $x \in \mathbb{R}^n$ define $\phi(x) := \inf\{\xi \mid (x,\xi) \in C\}$. Then ϕ is a convex function from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$.

6.2 Lower semicontinuity and closure

Sometimes convex functions may not look the way one might expect them to. Here is an example.

Example 6.2.1. Let ψ be any function from S^{n-1} to the non-negative halfline $\mathbb{R}_+$. Define a function $f : \mathbb{R}^n \to (-\infty, +\infty]$ by:

$$f(x) = \begin{cases} 0 & \text{if } \|x\| < 1, \\ \psi(x) & \text{if } \|x\| = 1, \\ +\infty & \text{if } \|x\| > 1. \end{cases}$$

This f is then a proper convex function (verify!).

Functions like f in Example 6.2.1 do not fit one's mental image of what convex functions ought to be like. If we were to ask ourselves why this f seems bad, the answer would be that it takes arbitrary non-negative values on S^{n-1}.

What if we were to redefine it to be zero there, to match its values on the interior of the unit ball B^n? Then we would simply have the indicator I_{B^n} of the unit ball, which is a perfectly nice convex function. It is lower semicontinuous, not continuous, but no function g taking finite values on a closed proper subset F of $\mathbb{R}^n$ and values of $+\infty$ on $\mathbb{R}^n \backslash F$ can be continuous, because the inverse image under g of the open set $\mathbb{R}$ is F, which is not open. Lower semicontinuity is the best we can expect from a function like I_B. However, f as defined above is in general not lower semicontinuous.

Section 6.2.1 introduces semicontinuity and provides some useful technical tools. Following that, Section 6.2.2 combines lower semicontinuity with convexity.

6.2.1 Semicontinuous functions

The functions with which we work in this section are not necessarily convex.

Definition 6.2.2. Let $S \subset \mathbb{R}^n$ and let $f : S \to \bar{\mathbb{R}}$. f is *lower semicontinuous* (lsc) at $x_0 \in S$ if for each real $\mu < f(x_0)$ the set $\{x \mid f(x) > \mu\}$ is a neighborhood of x_0 in S. f is *upper semicontinuous* (usc) at x_0 if $-f$ is lsc at x_0. f is *lsc (usc) on S* if it is lsc (usc) at each point of S. When $S = \mathbb{R}^n$ we just say f is lsc (usc). □

An equivalent definition of lower semicontinuity says that f is lower semicontinuous at x_0 if $f(x_0) \leq \liminf_{x \to_S x_0} f(x)$ (and upper semicontinuous if $f(x_0) \geq \limsup_{x \to_S x_0} f(x)$). It will often be convenient for us to use this definition instead of Definition 6.2.2, but before doing so we need to clarify some differences in terminology used in the literature.

Suppose f is a function from a topological space X to the real line $\mathbb{R}$. For any point x of X let $\mathcal{N}(x)$ be the neighborhood system at x: that is, the collection of open sets in X that contain x. The traditional definition of the limit inferior of f as x approaches a point $x_0 \in X$ is

$$\liminf_{x \to x_0} f(x) = \sup_{N \in \mathcal{N}(x_0)} \inf_{x \in N \backslash \{x_0\}} f(x); \tag{6.13}$$

see for example [37, p.48] for the case in which X is a subspace of $\mathbb{R}$. A different definition that has been used recently in variational analysis is

$$\liminf_{x\to x_0} f(x) = \sup_{N\in\mathscr{N}(x_0)} \inf_{x\in N} f(x); \tag{6.14}$$

see [34, p.51] and [36, p.8]. The exclusion of $\{x_0\}$ in (6.13) matters, as one can see by defining $f:\mathbb{R}\to\mathbb{R}$ to be -1 at the point $x_0 := 0$ and zero everywhere else. The following lemma shows the relationship between these two definitions.

Lemma 6.2.3. *Let X be a topological space containing a point x_0 and let $f: X\to\bar{\mathbb{R}}$. Define*

$$\alpha := \sup_{N\in\mathscr{N}(x_0)} \inf_{x\in N} f(x), \tag{6.15}$$

and

$$\beta := \sup_{N\in\mathscr{N}(x_0)} \inf_{x\in N\setminus\{x_0\}} f(x). \tag{6.16}$$

Then $\alpha = \min\{f(x_0),\beta\}$.

Proof. If $N\in\mathscr{N}(x_0)$ then $x_0\in N$ so that $\inf_{x\in N} f(x)\le f(x_0)$. Take the supremum over $N\in\mathscr{N}(x_0)$ in this inequality to get $\alpha\le f(x_0)$. Considering N again we see that as $N\setminus\{x_0\}\subset N$, by taking the supremum over $N\in\mathscr{N}(x_0)$ we obtain $\alpha\le\beta$. We have thus proved that $\alpha\le\min\{f(x_0),\beta\}$. To complete the proof it suffices to show that if $\alpha < f(x_0)$ then $\alpha=\beta$.

If $\alpha < f(x_0)$ then choose $\varepsilon>0$ so that $\alpha < f(x_0)-\varepsilon$. The definition of α shows that then for each $N\in\mathscr{N}(x_0)$ there must be some x_N with $f(x_N) < f(x_0)-\varepsilon$. But then removing x_0 from N will not affect the infimum of $f(x)$ on N, so $\inf_{x\in N} f(x) = \inf_{x\in N\setminus\{x_0\}} f(x)$ Taking the supremum over $N\in\mathscr{N}(x_0)$ yields $\alpha=\beta$. □

With this lemma we can show that the choice between the definitions of limit inferior in (6.13) and in (6.14) has no effect on the determination of lower semicontinuity at x_0.

Proposition 6.2.4. *Let X, x_0, and f be as in Lemma 6.2.3. The inequality*

$$f(x_0)\le\liminf_{x\to x_0} f(x) \tag{6.17}$$

holds with the definition of limit inferior in (6.13) *if and only if it holds with the definition of limit inferior in* (6.14). □

Proof. We use the definitions of α and β given in (6.15) and (6.16). Suppose that (6.17) holds using β for the limit inferior. Then $f(x_0)\le\beta$, so Lemma 6.2.3 shows that $f(x_0)=\alpha$ and therefore (6.17) holds using α for the limit inferior. On the other hand, if (6.17) holds using α for the limit inferior then $f(x_0)\le\alpha$, but the lemma shows that $\alpha\le\beta$ so that (6.17) holds using β for the limit inferior. □

As the choice between (6.13) and (6.14) does not affect the determination of lower semicontinuity, we are free to use either definition. In the remainder of this book we use (6.14).

If f is both lsc and usc at x_0 then it is continuous there (Exercise 6.2.18). Therefore the idea of semicontinuity in effect splits the definition of continuity into two parts. That refinement is useful because, as we see next, some functional operations commonly occurring in optimization preserve semicontinuity although they do not preserve continuity.

Proposition 6.2.5. *Let $S \subset \mathbb{R}^n$, let $x_0 \in S$, and let $\{ f_\alpha \mid \alpha \in A \}$ be a collection of functions from S to $\bar{\mathbb{R}}$. If each f_α is lsc at x_0, then the function f defined by $f(x) = \sup_{\alpha \in A} f_\alpha(x)$ is lsc at x_0. The corresponding statement holds for usc functions with "infimum" replacing "supremum."*

Proof. We prove only the first statement, as the second follows immediately from it. If A is empty then we are dealing with the constant function $-\infty$ which is clearly lsc. Therefore suppose that $A \neq \emptyset$.

If $f(x_0) = -\infty$ then f is certainly lsc at x_0. Therefore suppose otherwise, and let μ be a real number with $\mu < f(x_0)$. There must be some $\alpha \in A$ with $f_\alpha(x_0) > \mu$, and by hypothesis the set $N = \{ x \mid f_\alpha(x) > \mu \}$ is a neighborhood of x_0 in S. But $\{ x \mid f(x) > \mu \} \supset N$, so this set is also a neighborhood of x_0 in S and therefore f is lsc at x_0. □

Proposition 6.2.5 shows one reason why semicontinuity is useful, since the supremum and infimum operations certainly do not preserve continuity even when the index set is countable (Exercise 6.2.19). It also shows that statements about lsc functions have mirror images for usc functions. For simplicity, in the remainder of this section we deal only with lsc functions, leaving it to the reader to derive the corresponding statements for usc functions.

Sometimes it is simpler to deal with semicontinuity on $\mathbb{R}^n$ than with semicontinuity on a subset. If the subset in question is closed, we can easily pass back and forth between the two by using the $+\infty$ extension.

Proposition 6.2.6. *Let S be a closed subset of $\mathbb{R}^n$ and let $f : S \to \bar{\mathbb{R}}$. Then f is lsc on S if and only if the $+\infty$ extension of f is lsc on $\mathbb{R}^n$.*

Proof. Denote the $+\infty$ extension by g, and let $x_0 \in \mathbb{R}^n$.

(only if) Assume that f is lsc on S. If $g(x_0) = -\infty$ then g is lsc at x_0. Therefore suppose that $g(x_0) > -\infty$ and let μ be a real number with $\mu < g(x_0)$. If $x_0 \notin S$ then there is some neighborhood of x_0 that is disjoint from S, and on that neighborhood g takes the value $+\infty$, which certainly is greater than μ. On the other hand, if $x_0 \in S$ then by hypothesis there is a neighborhood N_S of x_0 in S on which f (hence g) takes values greater than μ. This N_S is of the form $S \cap N$, where N is a neighborhood of x_0 in $\mathbb{R}^n$. As g takes $+\infty$ everywhere on $N \backslash S$, the values of g everywhere on N are greater than μ, so g is lsc at x_0.

(if) Assume g is lsc on $\mathbb{R}^n$ and choose $x_0 \in S$. If $f(x_0) = -\infty$ there is nothing more to prove, so let μ be any real number with $\mu < f(x_0)$. Then $\mu < g(x_0)$, since f and g agree on S. Find a neighborhood N of x_0 in $\mathbb{R}^n$ such that for each $x \in N$ $g(x) > \mu$, and let $N_S = N \cap S$. Then N_S is a neighborhood of x_0 in S, and for each $x \in N_S$, $f(x) = g(x) > \mu$. Therefore f is lsc at x_0. □

The assumption that S is closed is necessary in the "only if" part of Proposition 6.2.6 (Exercise 6.2.20).

It is not always easy to prove that a function is lsc using Definition 6.2.2, so it is helpful to have a general criterion for a function defined on $\mathbb{R}^n$ to be lsc. The next theorem develops such a criterion. In it we use the notation $\text{lev}_\alpha f$ to denote the *lower level set* $\{x \mid f(x) \leq \alpha\}$, where f is any extended-real-valued function on $\mathbb{R}^n$ and $\alpha \in \bar{\mathbb{R}}$.

Theorem 6.2.7. *Let f be an extended-real-valued function on $\mathbb{R}^n$. The following are equivalent:*

a. f is lsc.
b. epi f *is closed.*
c. For each real μ, $\text{lev}_\mu f$ *is closed.*

Proof. If $f \equiv -\infty$ then each of the three assertions is true, so there is nothing to prove. Now assume that $f \not\equiv -\infty$.

(a) implies (b) Suppose f is lsc. As $f \not\equiv -\infty$ we can find $x \in \mathbb{R}^n$ and $\mu \in \mathbb{R}$ such that $(x, \mu) \notin \text{epi}\, f$. Then $\mu < f(x)$, so there is $\nu \in \mathbb{R}$ with $\mu < \nu < f(x)$. As f is lsc there is a neighborhood N of x such that for each $x' \in N$, $f(x') > \nu$. Let $\lambda \in \mathbb{R}$ with $\lambda < \mu$; then the Cartesian product of N with the open interval (λ, ν) is a neighborhood of (x, μ) that is disjoint from epi f. But (x, μ) was an arbitrary point of the complement of epi f, so epi f is closed and this proves (b).

(b) implies (c) The set $(\text{lev}_\mu f) \times \{\mu\}$ is the intersection of epi f with $\mathbb{R}^n \times \{\mu\}$. That intersection is closed, so $\text{lev}_\mu f$ is closed, which proves (c).

(c) implies (a) We show that if (a) does not hold then neither does (c). Let $x_0 \in \mathbb{R}^n$ and assume that f is not lsc at x_0. Definition 6.2.2 shows that if $f(x_0)$ were $-\infty$ then f would be lsc at x_0, so our assumption implies $f(x_0) > -\infty$. It also implies that there is some real μ with $\mu < f(x_0)$, such that the set $\{x \mid f(x) > \mu\}$ is not a neighborhood of x_0. Accordingly, there are points arbitrarily close to x_0 at which the value of f is less than or equal to μ. These points are therefore in $\text{lev}_\mu f$, so x_0 is a limit point of $\text{lev}_\mu f$. However, $x_0 \notin \text{lev}_\mu f$ so $\text{lev}_\mu f$ cannot be closed. Therefore (c) fails, and this completes the proof. □

Here is an important consequence of Theorem 6.2.7 for optimization.

Proposition 6.2.8. *Let f be a lsc extended-real-valued function on $\mathbb{R}^n$, and let S be a nonempty, compact subset of $\mathbb{R}^n$. Then f attains a minimum on S.*

Proof. For each real μ let $S_\mu = S \cap \text{lev}_\mu f$. Then S_μ is compact. If all of the S_μ are empty then f is identically $+\infty$ on S, so it attains a minimum of $+\infty$ there. Suppose therefore that some S_μ is not empty; then the value $I := \inf_{x \in S} f(x)$ is not $+\infty$, and for each $\mu > I$ the set S_μ is nonempty. As the S_μ are nested, Proposition C.1.3 ensures that the set $S_* := \cap_{\mu > I} S_\mu$ is nonempty and compact. On this set f cannot take any value greater than I; therefore it takes the constant value I, which must then be the minimum of f on S. □

6.2.2 Closed convex functions

As we have seen that lower semicontinuity can be helpful in optimization, we can try to find ways in which to regularize functions as we did in Example 6.2.1. For general functions this leads to the definition of lower semicontinuous hull; for convex functions it is helpful to adjust that definition in cases where the function takes $-\infty$, to obtain the closure operation.

Definition 6.2.9. Let f be an extended-real-valued function on $\mathbb{R}^n$. The *lower semicontinuous hull* of f is the function $\operatorname{lsc} f$ whose epigraph is $\operatorname{cl}\operatorname{epi} f$. The *closure* of f is the function $\operatorname{cl} f$ that is the lower semicontinuous hull of f if f never takes $-\infty$, and is identically $-\infty$ otherwise. f is *closed* if $f = \operatorname{cl} f$. □

The definition implies that $\operatorname{epi}\operatorname{cl} f \supset \operatorname{epi} f$ and therefore for each x, $\operatorname{cl} f(x) \leq f(x)$. We say that a function *g minorizes f*, or equivalently that *g is a minorant of f*, if $g \leq f$: that is, for each $x \in \mathbb{R}^n$ one has $g(x) \leq f(x)$. Accordingly, the closure of f always minorizes f.

We will see in Theorem 6.2.14 below that the two functions actually agree everywhere except possibly on the relative boundary of $\operatorname{dom} f$. In Example 6.2.1 both $\operatorname{lsc} f$ and $\operatorname{cl} f$ equal the indicator I_B. The closure operation removes all of the bad behavior of f on S^{n-1}, which in that case is the relative boundary of $\operatorname{dom} f$.

We can see that in a certain sense the closure operation regularizes a convex function. In the remainder of this section we investigate the relationship between such a function and its closure, developing an *internal* representation for the closure in terms of limits of function values along line segments. We will later establish an *external* representation in terms of pointwise suprema of affine minorants.

To develop the first of these results we need two technical tools, which we prove next. The first shows that an improper convex function can assume finite values only on the relative boundary of its effective domain, though it is not guaranteed to be finite even there.

Proposition 6.2.10. *Let f be an improper convex function on $\mathbb{R}^n$. Then for each $x \in \operatorname{ri}\operatorname{dom} f$ one has $f(x) = -\infty$.*

Proof. If $f \equiv +\infty$ then the claim is true. Otherwise, $\operatorname{dom} f$ is nonempty and there must (since f is improper) be some z with $f(z) = -\infty$. Choose any $x \in \operatorname{ri}\operatorname{dom} f$; then by part (c) of Corollary1.2.5 there is some $y \in \operatorname{dom} f$ and some $\lambda \in (0,1)$ with $x = (1-\lambda)z + \lambda y$. Let α be a fixed real number with $f(y) < \alpha$ and let β be any real number. As $\beta > f(z)$, Proposition 6.1.3 tells us that $f(x) < (1-\lambda)\beta + \lambda\alpha$. But β was arbitrary, so we must have $f(x) = -\infty$. □

Proposition 6.2.11. *Let f be a convex function on $\mathbb{R}^n$. Then*

$$\operatorname{ri}\operatorname{epi} f = \{\, (x,\xi) \in \mathbb{R}^{n+1} \mid x \in \operatorname{ri}\operatorname{dom} f,\ f(x) < \xi \,\}.$$

Proof. If $f \equiv +\infty$ the claim is surely true. Otherwise, define a linear transformation $L : \mathbb{R}^{n+1} \to \mathbb{R}^n$ by $L(x,\xi) = x$. Using Proposition 1.2.7, we find that

$$L(\operatorname{ri}\operatorname{epi} f) = \operatorname{ri} L(\operatorname{epi} f) = \operatorname{ri}\operatorname{dom} f,$$

so that the points x of $\operatorname{ri}\operatorname{dom} f$ are exactly those such that for some ξ one has $(x,\xi) \in \operatorname{ri}\operatorname{epi} f$. Let x be such a point and let $V = \{x\} \times \mathbb{R}$. V is relatively open because it is affine, and it must meet $\operatorname{ri}\operatorname{epi} f$ since $x \in L(\operatorname{ri}\operatorname{epi} f)$. By Proposition 1.2.12,

$$V \cap \operatorname{ri}\operatorname{epi} f = \operatorname{ri}(V \cap \operatorname{epi} f) = \operatorname{ri}\{\,(x,\xi) \mid \xi \geq f(x)\,\} = \{x\} \times (f(x), +\infty),$$

which proves the assertion. □

Proposition 6.2.11, together with separation, yields an extremely important result about the existence of subgradients. We define these first, then prove the existence result.

Definition 6.2.12. Let f be a convex function from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$, and let $x_0 \in \mathbb{R}^n$. An element $x^* \in \mathbb{R}^n$ is a *subgradient* of f at x_0 if for each $x \in \mathbb{R}^n$,

$$f(x) \geq f(x_0) + \langle x^*, x - x_0 \rangle. \tag{6.18}$$

□

If x^* is a subgradient of f at x_0 and $f(x_0) = +\infty$, then $f \equiv +\infty$, whereas if $f(x_0) = -\infty$ then every point of $\mathbb{R}^n$ is a subgradient of f at x_0. These cases are not very interesting compared to the case in which $f(x_0)$ is finite, when the existence of a subgradient yields the affine function on the right in (6.18). This function minorizes f and agrees with f at the point x_0, so its graph is a hyperplane $H[(x^*,-1), \langle (x^*,-1), (x_0, f(x_0))\rangle]$ in $\mathbb{R}^{n+1}$ that supports $\operatorname{epi} f$ at the point $(x_0, f(x_0))$.

The next proposition gives the fundamental existence result for subgradients.

Proposition 6.2.13. *Let f be a convex function from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$. If $x_0 \in \operatorname{ri}\operatorname{dom} f$ then f has a subgradient at x_0.*

Proof. Let $x_0 \in \operatorname{ri}\operatorname{dom} f$. If f is improper then $f(x_0) = -\infty$ (Proposition 6.2.10), and then each element of $\mathbb{R}^n$ is a subgradient of f at x_0. Therefore suppose f is proper; then $f(x_0)$ is finite, and Proposition 6.2.11 tells us that the pair $(x_0, f(x_0))$ does not belong to $\operatorname{ri}\operatorname{epi} f$. By Theorem 2.2.11 we can then properly separate this point from $\operatorname{epi} f$. This means that there exists some nonzero (x^*, ξ^*) such that whenever $(x,\xi) \in \operatorname{epi} f$ we have

$$\langle x^*, x \rangle + \xi^* \xi \leq \langle x^*, x_0 \rangle + \xi^* f(x_0),$$

with strict inequality for some $(\hat{x}, \hat{\xi}) \in \operatorname{epi} f$.

Because of the form of $\operatorname{epi} f$ we cannot have $\xi^* > 0$. If $\xi^* = 0$ then for each $x \in \operatorname{dom} f$ we have the inequality $\langle x^*, x \rangle \leq \langle x^*, x_0 \rangle$, with strict inequality for $x = \hat{x}$.

So we have properly separated $\{x_0\}$ from $\operatorname{dom} f$, which is impossible (Theorem 2.2.11) because $x_0 \in \operatorname{ri}\operatorname{dom} f$. Therefore $\xi^* < 0$, and there is no loss of generality in scaling x^* so that $\xi^* = -1$. Then for each $(x,\xi) \in \operatorname{epi} f$ we have

$$\xi \geq f(x_0) + \langle x^*, x - x_0 \rangle,$$

which shows that x^* is a subgradient of f at x_0. □

The next theorem states several important facts about the relationship between a convex function not taking $-\infty$ and its closure. If f ever takes $-\infty$ then its closure is identically $-\infty$, so that case is uninteresting. As we do elsewhere, we use here the notation $C \cong D$ to mean that C and D have the same closure and the same relative interior.

Theorem 6.2.14. *Let f be a convex function from $\mathbb{R}^n$ to $(-\infty,+\infty]$. Then $\operatorname{cl} f$ is a closed convex function, and it is proper if and only if f is proper. One has $\operatorname{cl} f(x) = f(x)$ everywhere except perhaps at points x on the relative boundary of $\operatorname{dom} f$. Further,*

$$\operatorname{cl}\operatorname{dom} f \supset \operatorname{dom}\operatorname{cl} f \supset \operatorname{dom} f,$$

and therefore in particular $\operatorname{dom} f \cong \operatorname{dom}\operatorname{cl} f$.

Proof. If f is improper, then f and $\operatorname{cl} f$ are both identically $+\infty$, so $\operatorname{cl} f$ is closed and improper, and the claims about domains are certainly true. Therefore suppose that f is proper, and note that $\operatorname{epi}\operatorname{cl} f = \operatorname{cl}\operatorname{epi} f$. As this set is convex, $\operatorname{cl} f$ must be a convex function. Let L be the linear transformation from $\mathbb{R}^{n+1}$ to $\mathbb{R}^n$ taking (x,ξ) to x. Then we have

$$\operatorname{cl} L(\operatorname{epi} f) \supset L(\operatorname{cl}\operatorname{epi} f) \supset L(\operatorname{epi} f);$$

that is,

$$\operatorname{cl}\operatorname{dom} f \supset \operatorname{dom}\operatorname{cl} f \supset \operatorname{dom} f,$$

which establishes the inclusion claim about domains. This inclusion also implies that the two domains must have the same affine hull, so that in fact $\operatorname{dom}\operatorname{cl} f \cong \operatorname{dom} f$. Now let x be in the common relative interior of $\operatorname{dom} f$ and $\operatorname{dom}\operatorname{cl} f$, and let $V = L^{-1}(x)$. Then V is a line which, by Proposition 6.2.11, must meet $\operatorname{ri}\operatorname{epi} f$. Then

$$\begin{aligned} V \cap \operatorname{epi}\operatorname{cl} f &= V \cap \operatorname{cl}\operatorname{epi} f = \operatorname{cl}(V \cap \operatorname{epi} f) \\ &= \operatorname{cl}[\{x\} \times [f(x),+\infty)] = \{x\} \times [f(x),+\infty) \\ &= V \cap \operatorname{epi} f, \end{aligned}$$

where we used Proposition 1.2.12. This shows that f and $\operatorname{cl} f$ agree on $\operatorname{ri}\operatorname{dom} f$. As f is finite there, so is $\operatorname{cl} f$, but as this set is also $\operatorname{ri}\operatorname{dom}\operatorname{cl} f$ it follows from Proposition 6.2.10 that $\operatorname{cl} f$ is proper. But then $\operatorname{cl} f$ never takes $-\infty$, and as its epigraph is closed it must be a closed convex function. We have shown that f and $\operatorname{cl} f$ agree on $\operatorname{ri}\operatorname{dom} f$ and off $\operatorname{cl}\operatorname{dom} f$, so that they can differ (if at all) only at relative boundary points of $\operatorname{dom} f$. □

Here is the first representation result for closures. It says that we can calculate the value of $\operatorname{lsc} f$ at any point by simply taking limits of the values of f along a line segment from the relative interior of $\operatorname{dom} f$ to that point, and it asserts that such limits will exist.

Proposition 6.2.15. *Let f be a convex function on $\mathbb{R}^n$. If $x \in \mathbb{R}^n$ and $x' \in \operatorname{ri}\operatorname{dom} f$, then*

$$(\operatorname{lsc} f)(x) = \lim_{\lambda \downarrow 0} f[(1-\lambda)x + \lambda x'].$$

Proof. Let $x' \in \operatorname{ri}\operatorname{dom} f$ and $x \in \mathbb{R}^n$. For each $\lambda \in (0,1]$, define $x_\lambda = (1-\lambda)x + \lambda x'$.

Recall that $\operatorname{epi}\operatorname{lsc} f = \operatorname{cl}\operatorname{epi} f$. Therefore $\operatorname{lsc} f \leq f$, and we have

$$(\operatorname{lsc} f)(x) \leq \liminf_{\lambda \downarrow 0} (\operatorname{lsc} f)(x_\lambda) \leq \liminf_{\lambda \downarrow 0} f(x_\lambda). \tag{6.19}$$

If $(\operatorname{lsc} f)(x) = +\infty$ then this completes the proof. Therefore suppose that $(\operatorname{lsc} f)(x) < +\infty$, and choose any real number $\mu \geq (\operatorname{lsc} f)(x)$. Let β be a real number greater than $f(x')$. The pair (x', β) then belongs to $\operatorname{ri}\operatorname{epi} f$ by Proposition 6.2.11, and (x, μ) belongs to $\operatorname{cl}\operatorname{epi} f$. Using Theorem 1.2.4, we find that for each $\lambda \in (0,1]$, $(1-\lambda)(x,\mu) + \lambda(x',\beta) \in \operatorname{ri}\operatorname{epi} f$, so that $f(x_\lambda) < (1-\lambda)\mu + \lambda\beta$ by Proposition 6.2.11. It follows that

$$\limsup_{\lambda \downarrow 0} f(x_\lambda) \leq \lim_{\lambda \downarrow 0} (1-\lambda)\mu + \lambda\beta = \mu.$$

As μ was any real number greater than or equal to $(\operatorname{lsc} f)(x)$, $\limsup_{\lambda \downarrow 0} f(x_\lambda) \leq \operatorname{lsc} f(x)$. Using this inequality together with (6.19), we find that $\lim_{\lambda \downarrow 0} f(x_\lambda)$ exists and equals $(\operatorname{lsc} f)(x)$ as claimed. □

If in Proposition 6.2.15 we do not permit f to take $-\infty$ then we obtain a formula for $\operatorname{cl} f$.

Corollary 6.2.16. *If f is a convex function from $\mathbb{R}^n$ to $(-\infty, +\infty]$, then for each $x \in \mathbb{R}^n$ and each $x' \in \operatorname{ri}\operatorname{dom} f$, one has*

$$(\operatorname{cl} f)(x) = \lim_{\lambda \downarrow 0} f[(1-\lambda)x + \lambda x'].$$

Proof. For such f, $\operatorname{cl} f = \operatorname{lsc} f$. □

Corollary 6.2.17. *Let $f_1, \ldots, f_k$ be proper convex functions on $\mathbb{R}^n$. If*

$$\cap_{i=1}^k \operatorname{ri}\operatorname{dom} f_i \neq \emptyset, \tag{6.20}$$

then

$$\operatorname{cl} \sum_{i=1}^k f_i = \sum_{i=1}^k \operatorname{cl} f_i. \tag{6.21}$$

Proof. Proposition 1.2.12 together with (6.20) yields

$$\operatorname{ri} \cap_{i=1}^{k} \operatorname{dom} f_i = \cap_{i=1}^{k} \operatorname{ri} \operatorname{dom} f_i.$$

As the f_i are proper we have

$$\operatorname{dom} \sum_{i=1}^{k} f_i = \cap_{i=1}^{k} \operatorname{dom} f_i,$$

so

$$\operatorname{ri} \operatorname{dom} \sum_{i=1}^{k} f_i = \cap_{i=1}^{k} \operatorname{ri} \operatorname{dom} f_i.$$

Let x' belong to this intersection. As the f_i are proper we can apply Corollary 6.2.16 to obtain for $x \in \mathbb{R}^n$

$$(\operatorname{cl} \sum_{i=1}^{k} f_i)(x) = \lim_{\lambda \downarrow 0} \sum_{i=1}^{k} f_i[(1-\lambda)x + \lambda x'] = \sum_{i=1}^{k} \lim_{\lambda \downarrow 0} f_i[(1-\lambda)x + \lambda x'] = \sum_{i=1}^{k} \operatorname{cl} f_i(x),$$

which proves (6.21). □

6.2.3 Exercises for Section 6.2

Exercise 6.2.18. Let f be an extended-real-valued function on a subset S of $\mathbb{R}^n$, and let $x_0 \in S$. Show that if f is usc and lsc at x_0, then it is continuous there.

Exercise 6.2.19. Exhibit a countable collection of continuous, uniformly bounded functions from $[-1,1]$ to $\mathbb{R}$ whose supremum is not continuous.

Exercise 6.2.20. Give an example in $\mathbb{R}$ to show that the "only if" part of Proposition 6.2.6 need not be true if the set S is not closed.

Exercise 6.2.21. Show that the function f of Definition 6.2.2 is lower semicontinuous at x_0 if and only if $f(x_0) \leq \liminf_{x \to x_0} f(x)$.

Exercise 6.2.22. Consider the cone K of Exercise 3.1.18, namely

$$K = \{x \in \mathbb{R}^3 \mid x_1 \geq 0,\ x_3 \geq 0,\ 2x_1x_3 \geq x_2^2\}.$$

Show that K is the epigraph of the function

$$f(x_1,x_2) = \begin{cases} x_2^2/(2x_1) & \text{if } x_1 > 0, \\ 0 & \text{if } (x_1,x_2) = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

(f is clearly proper and positively homogeneous, but this shows that it is also a closed convex function.)

Exercise 6.2.23. Consider the function f of Exercise 6.2.22. Let $\alpha > 0$ and define three functions from $\mathbb{R}_+$ to $\mathbb{R}^2$ by

$$x_0(t) = (.5t, t), \qquad x_\alpha(t) = ((2\alpha)^{-1}t^2, t), \qquad x_\infty(t) = (.5t^3, t).$$

Compute the limits of the composite functions $f \circ x_0$, $f \circ x_\alpha$, and $f \circ x_\infty$ as $t \downarrow 0$. What does this show about continuity properties of closed proper convex functions like f?

Exercise 6.2.24. The *Cobb-Douglas production function,* used in economics, is a function from $\operatorname{int}\mathbb{R}^n_+$ to $\mathbb{R}$ defined by

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} x_i^{\lambda_i},$$

where the λ_i are fixed, positive numbers with $\sum_{i=1}^n \lambda_i = 1$. This function is positively homogeneous. Show how to extend this function to a closed, proper, positively homogeneous concave function defined on all of $\mathbb{R}^n$. (A function g is closed proper concave if $-g$ is closed proper convex.)

Chapter 7
Conjugacy and recession

One of the most useful aspects of convex functions is the fact that each closed convex function f has a *conjugate* function f^* that is also a closed convex function. If we take the conjugate of f^*, we recover f. The conjugacy operation provides substantial insight into properties of the functions involved, and it is also very useful for computation.

The relationship between a function and its conjugate depends strongly upon the behavior of the function "at infinity": that is, its behavior as the argument becomes large. *Recession functions* permit us to describe and study that behavior.

7.1 Conjugate functions

This section introduces one of the principal tools of convex analysis, the conjugacy correspondence for functions. We define this correspondence and develop a few of its basic properties, as well as an extension of the conjugate called the *adjoint.* Later in the section we illustrate a powerful application of conjugacy using the support function as a vehicle.

7.1.1 Definition and properties

Definition 7.1.1. Let f be a function from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$. The *convex conjugate* of f is the function $f^* : \mathbb{R}^n \to \bar{\mathbb{R}}$ defined by

$$f^*(x^*) = \sup_x \{\langle x^*, x\rangle - f(x)\}. \tag{7.1}$$

□

To understand geometrically what we are doing in computing the conjugate, note that (7.1) is equivalent to the expression

$$f^*(x^*) = \sup_{(x,\xi)\in \operatorname{epi} f} \langle (x^*,-1),(x,\xi)\rangle,$$

where the inner product in $\mathbb{R}^{n+1}$ is defined by $\langle (a,\alpha),(b,\beta)\rangle = \langle a,b\rangle + \alpha\beta$. The vector $(x^*,-1)$ is the normal to a collection of parallel hyperplanes in $\mathbb{R}^{n+1}$; the fact that the last component is -1 simply means that such a hyperplane is not "vertical" (that is, its parallel subspace does not contain the line $\{0\}^n \times \mathbb{R}$, where $\{0\}^n$ is the origin of $\mathbb{R}^n$), and that the normal vector points downward. Each hyperplane in the collection consists of all pairs (x,ξ) that yield a particular numerical value of the inner product $\langle (x^*,-1),(x,\xi)\rangle$; as this value increases, the hyperplanes move downward. If f is not identically $+\infty$, the formula in (7.1) gives the least upper bound–but not necessarily the maximum, which may not exist–of all such values for which the hyperplane meets epi f; otherwise it gives $-\infty$. The operation of computing the conjugate therefore amounts to moving the hyperplanes as far down as possible while still maintaining contact with epi f.

By contrast, the epigraph of f^* consists of those (x^*,ξ^*) for which $f(x) \geq \langle x^*,x\rangle - \xi^*$ for each $x \in \mathbb{R}^n$: that is, for which the affine function $\langle x^*,\cdot\rangle - \xi^*$ minorizes the function f. Here the requirement that $\xi^* \geq f^*(x^*)$ means that the affine function has to stay on or below the graph of f everywhere.

This geometric viewpoint suggests that, if f were convex and if the supremum in (7.1) were attained at some point x, $f^*(x^*)$ should be associated with a hyperplane supporting the epigraph of f at the point $(x,f(x))$. This is a geometric statement of the basic relationship among functions, conjugates, and subgradients.

Under normal circumstances we just refer to the convex conjugate as the "conjugate," and no other kind of conjugate is required. However, for calculations involving duality relationships we sometimes need the *concave conjugate,* defined by substituting "inf" for "sup" in (7.1). By rearranging the formula in (7.1) we can then see that the concave conjugate of f takes the value $-(-f)^*(-x^*)$ at the point x^*. This permits us to use properties of the convex conjugate to infer corresponding properties of the concave conjugate.

As the conjugate in (7.1) is the supremum of the collection (indexed by x) of affine functions $\langle \cdot,x\rangle - f(x)$, its epigraph is empty (that is, $f^* \equiv +\infty$) if f ever takes $-\infty$, it is $\mathbb{R}^{n+1}$ (that is, $f^* \equiv -\infty$) if f is identically $+\infty$, and it is the intersection of a collection of closed halfspaces otherwise. Therefore, f^* must be a lower semicontinuous convex function no matter what kind of function f was. However, we can say more.

Proposition 7.1.2. *Let f be a function from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$. Then f^* is a closed convex function, and if f is convex then f^* is proper if and only if f is proper.*

Proof. We just observed that f^* must be lower semicontinuous and convex. It can take $-\infty$ only if $f \equiv +\infty$, and in that case $f^* \equiv -\infty$. Therefore f^* is closed.

If f is improper then either it is identically $+\infty$, in which case $f^* \equiv -\infty$ as just noted, or else f takes $-\infty$ somewhere, in which case $f^* \equiv +\infty$. In either case f^* is improper. If f is convex and proper then it cannot be everywhere $+\infty$, so f^* can never take $-\infty$. Also, there is a point x_0 in $\operatorname{ri}\operatorname{dom} f$ at which f is finite, and

by Proposition 6.2.13 f has a subgradient x_0^* at x_0. Therefore, for each $x \in \mathbb{R}^n$ one has $f(x) \geq f(x_0) + \langle x_0^*, x - x_0 \rangle$. By rearranging this inequality we see that $f^*(x_0^*) = \langle x_0^*, x_0 \rangle - f(x_0)$, a finite number. Then f^* cannot be identically $+\infty$, so it is proper. □

Much of the importance of conjugate functions rests on the fact that it is possible not only to pass from f to f^*, but also from f^* back to f provided f is closed and convex. The lemma and theorem that follow develop this relationship and give some of its properties.

Lemma 7.1.3. *If $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is convex, then* $\operatorname{cl} f$ *is the pointwise supremum of the set $\mathscr{A}$ of all affine functions on $\mathbb{R}^n$ that minorize f.*

Proof. If f is improper, either it takes $-\infty$ somewhere, in which case $\mathscr{A} = \emptyset$ and its supremum is everywhere $-\infty$, or $f \equiv +\infty$ so that $\operatorname{cl} f \equiv +\infty$ also. In the latter case $\mathscr{A}$ contains every affine function, and its supremum is everywhere $+\infty$. In either case the lemma is proved. Therefore suppose f is proper, in which case we have $\operatorname{epi}\operatorname{cl} f = \operatorname{cl}\operatorname{epi} f$. For $x \in \mathbb{R}^n$ let $\sigma(x) := \sup\{t(x) \mid t \in \mathscr{A}\}$. We must show that $\sigma = \operatorname{cl} f$.

If $t \in \mathscr{A}$ and $t \leq f$ then $\operatorname{epi} t \supset \operatorname{epi} f$ and, as $\operatorname{epi} t$ is closed,

$$\operatorname{epi} t \supset \operatorname{cl}\operatorname{epi} f \supset \operatorname{epi}\operatorname{cl} f. \tag{7.2}$$

The epigraph of σ is the intersection of the epigraphs of those $t \in \mathscr{A}$ that minorize f, so (7.2) shows that $\operatorname{epi}\sigma \supset \operatorname{epi}\operatorname{cl} f$ and therefore $\sigma \leq \operatorname{cl} f$.

To complete the proof we must show $\sigma \geq \operatorname{cl} f$, and for that it suffices to show that for each pair $(x, \mu) \in \mathbb{R}^{n+1}$ such that $\mu < \operatorname{cl} f(x)$ there is an affine function $t \in \mathscr{A}$ such that $t(x) > \mu$. This is easy if $x \in \operatorname{ri}\operatorname{dom} f$, as then Proposition 6.2.13 says that f has a subgradient x^* at x. Define an affine function t on $\mathbb{R}^n$ by setting $t(y) := f(x) + \langle x^*, y - x \rangle$; then as x^* is a subgradient, $t \leq f$ and so $t \in \mathscr{A}$. At the point $y = x$ we also have

$$t(x) = f(x) \geq \operatorname{cl} f(x) > \mu,$$

as required.

The remaining case is that in which x belongs to the relative boundary of $\operatorname{dom} f$. By choice of (x, μ) we have $(x, \mu) \notin \operatorname{epi}\operatorname{cl} f = \operatorname{cl}\operatorname{epi} f$. The strong separation theorem then shows that there exist a nonzero pair $(y^*, \eta^*) \in \mathbb{R}^{n+1}$ and a real number $\delta > 0$ such that for each pair $(z, \zeta) \in \operatorname{cl}\operatorname{epi} f = \operatorname{epi}\operatorname{cl} f$,

$$\langle (y^*, \eta^*), (z, \zeta) \rangle + \delta \leq \langle (y^*, \eta^*), (x, \mu) \rangle. \tag{7.3}$$

Since (z, ζ) belongs to an epigraph, ζ can be arbitrarily large, so (7.3) shows that η^* cannot be positive.

If η^* were zero then y^* would have to be nonzero. Recalling that the orthogonal projection of an epigraph into the space of the variable is the effective domain, we would then see that for each $z \in \operatorname{dom}\operatorname{cl} f$,

$$\langle y^*, z\rangle + \delta \le \langle y^*, x\rangle,$$

and then the Schwarz inequality would show that the distance of x from $\operatorname{dom}\operatorname{cl} f$ was at least the positive number $\delta\|y^*\|^{-1}$. But Theorem 6.2.14 shows that $\operatorname{cl}\operatorname{dom} f \supset \operatorname{dom}\operatorname{cl} f \supset \operatorname{dom} f$, so the distance of x from $\operatorname{dom} f$ would be at least $\delta\|y^*\|^{-1}$. This would contradict our specification that x be in the relative boundary of $\operatorname{dom} f$, so η^* cannot be zero and therefore it must be negative. By multiplying (7.3) by $(-\eta^*)^{-1}$ and defining $v^* := (-\eta^*)^{-1}y^*$ and $\varepsilon := (-\eta^*)^{-1}\delta$ we obtain for each $(z,\zeta) \in \operatorname{epi}\operatorname{cl} f$ the inequality

$$\langle (v^*, -1), (z,\zeta)\rangle + \varepsilon \le \langle (v^*, -1), (x,\mu)\rangle. \tag{7.4}$$

Now for $z \in \mathbb{R}^n$ define $t(\,\cdot\,)$ to be the affine function on $\mathbb{R}^n$ whose value at $z \in \mathbb{R}^n$ is $\langle v^*, z-x\rangle + \varepsilon + \mu$. From (7.4) we find that whenever $(z,\zeta) \in \operatorname{cl}\operatorname{epi} f$, $t(z) \le \zeta$, so we have $t \le \operatorname{cl} f \le f$; then $t \in \mathscr{A}$. But

$$t(x) = \varepsilon + \mu > \mu,$$

so we have shown that $\sigma \ge \operatorname{cl} f$, which completes the proof of the lemma. □

By using the concave conjugate instead of the convex conjugate one can prove a result parallel to Lemma 7.1.3, but phrased in terms of concave majorants and infima of affine functions, rather than convex minorants and suprema.

The next theorem derives a fundamental property of the conjugation operation.

Theorem 7.1.4. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$.*

a. f^{**} is the greatest closed convex minorant of f.
b. If f is convex then $f^{**} = \operatorname{cl} f$.

Proof. We prove (b) first, then use it to prove (a).

To prove (b), suppose f is convex. If f is improper then the enumeration of cases at the beginning of the proof of Lemma 7.1.3 shows that (b) holds. Therefore assume that f is proper, in which case $\operatorname{epi}\operatorname{cl} f = \operatorname{cl}\operatorname{epi} f$. By Lemma 7.1.3, $\operatorname{cl} f$ is the pointwise supremum of the collection $\mathscr{A}$ of all affine functions t from $\mathbb{R}^n$ to $\mathbb{R}$ such that $t \le f$. To see that this $\mathscr{A}$ is not empty, recall that for any $x_0 \in \operatorname{ri}\operatorname{dom} f$ there is a subgradient x_0^* satisfying $f(x) \ge f(x_0) + \langle x_0^*, -x_0\rangle$ for all x. Rearranging this and writing $\xi^* := \langle x_0^*, x_0\rangle - f(x_0)$ we find that for each x, $\langle x_0^*, x\rangle - \xi^* \le f(x)$.

Choose some member of $\mathscr{A}$ and write it as $t(x) = \langle x^*, x\rangle - \xi^*$. As $t(\cdot) \in \mathscr{A}$ we must have $f(x) \ge t(x)$ for each $x \in \mathbb{R}^n$. We can rewrite that requirement as

$$\xi^* \ge \langle x^*, x\rangle - f(x) \text{ for each } x \in \mathbb{R}^n. \tag{7.5}$$

The requirement in (7.5) will be met if and only if

$$\xi^* \ge \sup_x\{\langle x^*, x\rangle - f(x)\} = f^*(x^*).$$

As ξ^* must be a real number, this imposes the implicit requirement that $f^*(x^*) < +\infty$: that is, that $x^* \in \operatorname{dom} f^*$. Affine functions $t(\cdot)$ for which the x^* does not satisfy this requirement will not minorize f and therefore will not appear in $\mathscr{A}$.

If we now ask, for a given $x^* \in \operatorname{dom} f^*$, what value of ξ^* will produce the largest values of t while obeying the constraint that $t \le f$, then the answer is that we should find the smallest allowable ξ^*. From (7.5) we see that that value will be

$$\xi^* = \sup_x \{\langle x^*, x\rangle - f(x)\} = f^*(x^*).$$

Values of ξ^* larger than $f^*(x^*)$ do not disqualify t from membership in $\mathscr{A}$, so we can see from what we have just found that the pairs (x^*, ξ^*) associated with affine functions t appearing in $\mathscr{A}$ are exactly those pairs that belong to $\operatorname{epi} f^*$. However, for $x^* \in \operatorname{dom} f^*$ the affine function $\langle x^*, x\rangle - f^*(x^*)$ majorizes all functions of the form $\langle x^*, x\rangle - \xi^*$ with $\xi^* > f^*(x^*)$.

Now choose any $x \in \mathbb{R}^n$. Applying Lemma 7.1.3 and the above discussion we find that

$$\operatorname{cl} f(x) = \sup_{t \in \mathscr{A}} t(x) = \sup_{(x^*, \xi^*) \in \operatorname{dom} f^*} \{\langle x^*, x\rangle - \xi^*\} = \sup_{x^*} \{\langle x^*, x\rangle - f^*(x^*)\} = f^{**}(x),$$

where the last supremum need not be restricted to $\operatorname{dom} f^*$ because any x^* not in $\operatorname{dom} f^*$ will generate a value of $+\infty$ for $f^*(x^*)$ and therefore will not influence the supremum. This proves (b).

To prove (a), recall that f^* and f^{**} are closed convex functions no matter what kind of function f is. Choose any $x \in \mathbb{R}^n$. For each $x^* \in \mathbb{R}^n$ the definition of f^* shows that $f(x) \ge \langle x^*, x\rangle - f^*(x^*)$. Then

$$f(x) \ge \sup_{x^*} \{\langle x^*, x\rangle - f^*(x^*)\} = f^{**}(x),$$

so that f^{**} minorizes f.

Now let g be any closed convex minorant of f. As $g \le f$ we have $g^* \ge f^*$ (from the definition of the conjugate function) and then $g^{**} \le f^{**}$. But we have shown in the proof of (b) that $g^{**} = g$, so $g \le f^{**}$. It follows that f^{**} is the greatest closed convex minorant of f. □

Remark 7.1.5. As f^* is closed and $f^{**} = \operatorname{cl} f$, we have

$$(cl\, f)^* = (f^{**})^* = (f^*)^{**} = f^*,$$

so that f and $\operatorname{cl} f$ have the same conjugate. □

7.1.2 Adjoints

For applications to duality in convex optimization it is useful to have an extension of the conjugacy concept to functions of two groups of variables.

Definition 7.1.6. Suppose that F is a function from $\mathbb{R}^n \times \mathbb{R}^m$ to $\bar{\mathbb{R}}$. The *adjoint of F in the convex sense* is the function $F^A : \mathbb{R}^m \times \mathbb{R}^n \to \bar{\mathbb{R}}$ defined by

$$F^A(y^*, x^*) = \inf_{(x,y)} \{-\langle x^*, x\rangle + \langle y^*, y\rangle + F(x,y)\}. \tag{7.6}$$

The *adjoint of F in the concave sense* is defined by replacing infimum by supremum in (7.6). □

The hypograph of the function F^A defined by (7.6) is the intersection of the hypographs of a collection of affine functions of (x^*, y^*) (indexed by (x,y)). Therefore the adjoint of F in the convex sense is an upper semicontinuous *concave* function, and in fact it is closed because it can only take $+\infty$ at a point if $F \equiv +\infty$, and in that case $F^A \equiv +\infty$ too. Similarly, the adjoint in the concave sense is a closed *convex* function. We use the same notation for these two different operations because we will use the adjoint in the convex sense when the function F with which we begin is convex, and in the concave sense when F is concave. Therefore the only possible ambiguity might be when F is both convex and concave, and in such cases the context will make it clear which adjoint is intended.

Suppose that F is a closed convex function, and that we compute the adjoint F^A in the convex sense using (7.6). This F^A is then closed and concave, and we could take its adjoint in the concave sense to obtain a function F^{AA}. However, note that (7.6) expresses $F^A(y^*, x^*)$ as $-F^*(x^*, -y^*)$. Therefore

$$\begin{aligned} F^{AA}(x,y) &= \sup_{(y^*,x^*)} \{-\langle y^*, y\rangle + \langle x^*, x\rangle + F^A(y^*, x^*)\} \\ &= \sup_{(x^*,-y^*)} \left\{ \left\langle \begin{pmatrix} x^* \\ -y^* \end{pmatrix}, \begin{pmatrix} x \\ y \end{pmatrix} \right\rangle - F^*(x^*, -y^*) \right\} \\ &= F^{**}(x,y) = F(x,y), \end{aligned} \tag{7.7}$$

where we used Theorem 7.1.4. Accordingly, if F is closed and convex then $F^{AA} = F$, and a similar argument shows that this formula holds also if F is closed and concave.

The term "adjoint" comes from the fact that if F is the convex indicator of the graph of a linear transformation $A : \mathbb{R}^n \to \mathbb{R}^m$ (so that $F(x,y)$ is zero if $y = Ax$ and $+\infty$ otherwise), then $F^A(y^*, x^*)$ is zero if $x^* = A^* y^*$ and $-\infty$ otherwise, so that F^A is the concave indicator of the graph of the adjoint transformation A^*.

7.1.3 Support functions

The support function provides a powerful tool for dealing with properties of closed convex sets.

Definition 7.1.7. For any subset S of $\mathbb{R}^n$, the *support function* of S is the function I_S^* taking the value $\sup_{s\in S}\langle x^*, s\rangle$ at the point x^*. □

As

$$\sup_{s\in S}\langle x^*, s\rangle = \sup_{s\in\mathbb{R}^n}\{\langle x^*, s\rangle - I_S(s)\},$$

where $I_S(s)$ is zero if $s \in S$ and $+\infty$ otherwise, the notation I_S^* for the support function is compatible with the definition of conjugacy. The definition shows that I_S^* is positively homogeneous ($I_S^*(\sigma x^*) = \sigma I_S^*(x^*)$ for each positive σ and each $x^* \in \mathbb{R}^n$), and Proposition 7.1.2 shows that I_S^* is a closed convex function. Furthermore, if I_S^* is improper then it is identically $-\infty$, which occurs exactly when S is empty; I_S^* cannot be identically $+\infty$ because if S is nonempty then $I_S^*(0) = 0$.

If S is a nonempty subset of $\mathbb{R}^n$ then

$$\begin{aligned}\operatorname{epi} I_S^* &= \{(x^*, \xi^*) \in \mathbb{R}^{n+1} \mid \xi^* \geq I_S^*(x^*)\}\\ &= \{(x^*, \xi^*) \mid \text{for each } x \in S,\ \xi^* \geq \langle x^*, x\rangle\}\\ &= \{(x^*, \xi^*) \mid \text{for each } x \in S,\ \langle (x^*, \xi^*), (x, -1)\rangle \leq 0\}\\ &= H_S^\circ,\end{aligned}$$

where H_S is the homogenizing cone of S. Thus *the epigraph of the support function is the polar of the homogenizing cone.* This polar is often called the "dual cone" of S. The support function in fact furnishes us with a *one-to-one correspondence* between the class $\mathscr{C}$ of nonempty closed convex subsets of $\mathbb{R}^n$ and the class $\mathscr{F}$ of positively homogeneous closed proper convex functions on $\mathbb{R}^n$. We establish this correspondence in the following lemma and theorem.

Lemma 7.1.8. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a proper positively homogeneous convex function.*

a. $\operatorname{cl} f(0) = 0$*, and $f(0)$ is either zero or $+\infty$.*
b. $f^* = I_{\operatorname{dom} f^*}$*, and* $\operatorname{dom} f^*$ *is nonempty and closed.*

Proof. As f is proper it is not identically $+\infty$, so $\operatorname{dom} f$ is nonempty and f is finite there. Let $x_0 \in \operatorname{ri}\operatorname{dom} f$. As f never takes $-\infty$ we can use Corollary 6.2.16 and the positive homogeneity of f to obtain

$$\operatorname{cl} f(0) = \lim_{\tau\downarrow 0} f[(1-\tau)0 + \tau x_0] = \lim_{\tau\downarrow 0} \tau f(x_0) = 0.$$

Now suppose $f(0) \neq +\infty$; then $0 \in \operatorname{dom} f$ so that $f(0)$ is finite. For any positive real α, positive homogeneity requires that $f(0) = f([\alpha][0]) = \alpha f(0)$. As $f(0)$ is a real number it must be zero, and this proves assertion (a).

To prove (b), recall from Proposition 7.1.2 that f^* is proper because f is, so that $\operatorname{dom} f^*$ is nonempty. Remark 7.1.5 shows that f and $\operatorname{cl} f$ have the same conjugate, so for each $x^* \in \mathbb{R}^n$,

$$f^*(x^*) = (\operatorname{cl} f)^*(x^*) \geq \langle x^*, 0\rangle - (\operatorname{cl} f)(0) = 0,$$

so that f^* is nonnegative on $\mathbb{R}^n$. Suppose now that for some $x^* \in \mathbb{R}^n$ we have $f^*(x^*) > 0$. Then there is some x_1 with $f^*(x^*) > \langle x^*, x_1\rangle - f(x_1) =: \alpha > 0$. Then for any $\tau > 0$, positive homogeneity of f requires that $\langle x^*, \tau x_1\rangle - f(\tau x_1) = \tau\alpha$, which can be arbitrarily large. Therefore $f^*(x^*) = +\infty$, so that f^* takes only two values on $\mathbb{R}^n$: zero and $+\infty$.

The set of points where it takes zero is the set where it does not take $+\infty$: that is, $\operatorname{dom} f^*$. Then f^* takes zero on $\operatorname{dom} f^*$ and $+\infty$ everywhere else, so it is the indicator function $I_{dom f^*}$ of $\operatorname{dom} f^*$.

We already know that $\operatorname{dom} f^*$ is nonempty, and the properties we have just shown imply that $\operatorname{dom} f^* = \operatorname{lev}_0 f^*$. Theorem 6.2.7 shows that as f^* is closed and proper, the latter set is closed, and this establishes (b). □

Here is the theorem on correspondence.

Theorem 7.1.9. *Let $\mathscr{C}$ be the class of nonempty closed convex subsets of $\mathbb{R}^n$ and $\mathscr{F}$ be the class of closed proper positively homogeneous convex functions on $\mathbb{R}^n$. The formulas*

$$\phi(C) = I_C^*, \qquad \gamma(f) = \operatorname{dom} f^*$$

establish a one-to-one correspondence between $\mathscr{C}$ and $\mathscr{F}$.

Proof. We prove that $\phi : \mathscr{C} \to \mathscr{F}$, that $\gamma : \mathscr{F} \to \mathscr{C}$, and that $\phi \circ \gamma$ and $\gamma \circ \phi$ are the identity maps of $\mathscr{F}$ and of $\mathscr{C}$ respectively.

If $C \in \mathscr{C}$ then $\phi(C) = I_C^*$. The comments following Definition 7.1.7 show that I_C^* is closed, convex, and positively homogeneous; it is also proper because $C \neq \emptyset$. Therefore $\phi : \mathscr{C} \to \mathscr{F}$.

To show that $\gamma : \mathscr{F} \to \mathscr{C}$, suppose $f \in \mathscr{F}$. Lemma 7.1.8 says that $f(0) = 0$, $f^* = I_{\operatorname{dom} f^*}$, and $\operatorname{dom} f^*$ is closed. As f is proper, so is f^*; therefore $\operatorname{dom} f^*$ is nonempty, and it is certainly convex. Therefore it belongs to $\mathscr{C}$ and so $\gamma : \mathscr{F} \to \mathscr{C}$.

Next, we observe that

$$(\phi \circ \gamma)(f) = I^*_{\operatorname{dom} f^*} = (I_{\operatorname{dom} f^*})^* = (f^*)^* = f, \tag{7.8}$$

where the third equality comes from Lemma 7.1.8 and the fourth holds because f is closed. This shows that $\phi \circ \gamma = \mathrm{id}_{\mathscr{F}}$.

Finally, if $c \in \mathscr{C}$ then

$$(\gamma \circ \phi)(C) = \operatorname{dom}[(I_C^*)^*] = \operatorname{dom} I_C = C,$$

so that $\gamma \circ \phi = \mathrm{id}_{\mathscr{C}}$. □

This correspondence between nonempty closed convex sets and their support functions has very useful properties. The following proposition gives some of these.

Here as elsewhere in this book, if f and g are functions from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$, the notation $f \leq g$ means that $f(x) \leq g(x)$ for each $x \in \mathbb{R}^n$, and similarly for $f \geq g$. In part (c) we use the Pompeiu-Hausdorff distance $\rho(C,D)$ from Definition 1.1.11.

Proposition 7.1.10. *Let C and D be nonempty closed convex subsets of $\mathbb{R}^n$.*

a. If γ and δ are nonnegative real numbers, then $I^*_{\gamma C+\delta D} = \gamma I^*_C + \delta I^*_D$, where on the right-hand side we use the convention that $0(+\infty) = 0$.
b. $I^*_C \leq I^*_D$ if and only if $C \subset D$.
c. $\rho(C,D) = \sup\{|I^*_C(y^*) - I^*_D(y^*)| \mid \|y^*\| = 1\}$, where here we use the convention that $(+\infty) - (+\infty) = 0$.

Proof. Claim (a). If either or both of γ and δ are zero then the result is immediate; therefore we assume that both are positive. Let $x^* \in \mathbb{R}^n$ and choose $c \in C$ and $d \in D$. Then

$$\langle x^*, \gamma c + \delta d\rangle = \gamma\langle x^*, c\rangle + \delta\langle x^*, d\rangle \leq \gamma I^*_C(x^*) + \delta I^*_D(x^*).$$

Taking the supremum on the left over $c \in C$ and $d \in D$ yields $I^*_{\gamma C+\delta D}(x^*) \leq \gamma I^*_C(x^*) + \delta I^*_D(x^*)$, and as x^* was arbitrary in $\mathbb{R}^n$ we have $I^*_{\gamma C+\delta D} \leq I^*_{\gamma C} + I^*_{\delta D}$.

The opposite inequality holds whenever the left-hand side is $+\infty$. Let $x^* \in \mathbb{R}^n$ with $I^*_{\gamma C+\delta D}(x^*) < +\infty$. If either of $I^*_C(x^*)$ and $I^*_D(x^*)$ were $+\infty$ then we could take a sequence in that set with the support function increasing without bound, and a fixed element of the other set, to contradict the finiteness of $I^*_{\gamma C+\delta D}(x^*)$. Therefore both are finite. For $c \in C$ and $d \in D$ we have

$$\gamma\langle x^*, c\rangle + \delta\langle x^*, d\rangle = \langle x^*, \gamma c + \delta d\rangle \leq I^*_{\gamma C+\delta D}(x^*).$$

Now by taking the supremum in the left-hand side over $c \in C$ and $d \in D$ we obtain $\gamma I^*_C(x^*) + \delta I^*_D(x^*) \leq I^*_{\gamma C+\delta D}(x^*)$, and as x^* is any point with $I^*_{\gamma C+\delta D}(x^*) < +\infty$ we have $I^*_{\gamma C+\delta D} \geq \gamma I^*_C + \delta I^*_D$.

Claim (b). The "if" direction follows directly from the definition of the support function. For the "only if" direction, suppose that $I^*_C \leq I^*_D$. Take conjugates to obtain

$$I_C = I^{**}_C \geq I^{**}_D = I_D,$$

where the equalities follow from Theorem 7.1.4 because $C \in \mathscr{C}$ and therefore I_C is closed, and similarly for D. The indicator inequality $I_C \geq I_D$ implies $C \subset D$.

Claim (c). First suppose that there is no $\mu \geq 0$ such that $C \subset D + \mu B^n$ and $D \subset C + \mu B^n$, where B^n is the unit ball. Then $\rho(C,D) = +\infty$. Also, in this case for any positive μ we have either $C \not\subset D + \mu B^n$ or $D \not\subset C + \mu B^n$. Suppose it is the former; then as the support function of B^n is the norm $\|\cdot\|$, $I^*_C \not\leq I^*_D + \mu\|\cdot\|$ by part (b). Therefore there is some x^* with $\|x^*\| = 1$ and

$$I^*_C(x^*) > I^*_D(x^*) + \mu,$$

so that

$$\sup\{|I^*_C(y^*) - I^*_D(y^*)| \mid \|y^*\| = 1\} \geq |I^*_C(x^*) - I^*_D(x^*)| \geq I^*_C(x^*) - I^*_D(x^*) > \mu.$$

In the case $D \not\subset C + \mu B^n$ we reason similarly to obtain the same inequality $\sup\{|I_C^*(y^*) - I_D^*(y^*)| \mid \|y^*\| = 1\} > \mu$. As μ was arbitrary, we have

$$\sup\{|I_C^*(y^*) - I_D^*(y^*)| \mid \|y^*\| = 1\} = +\infty = \rho(C,D).$$

The remaining case is that in which there is $\mu \geq 0$ such that

$$C \subset D + \mu B^n, \qquad D \subset C + \mu B^n. \tag{7.9}$$

Then

$$I_C^* \leq I_D^* + \mu\|\cdot\|, \qquad I_D^* \leq I_C^* + \mu\|\cdot\|.$$

For each such μ and for each x^* with $\|x^*\| = 1$,

$$|I_C^*(x^*) - I_D^*(x^*)| \leq \mu,$$

so that

$$\sup\{|I_C^*(y^*) - I_D^*(y^*)| \mid \|y^*\| = 1\} \leq \mu. \tag{7.10}$$

Taking the infimum in (7.10) over all $\mu \geq 0$ satisfying (7.9), we find that

$$\sup\{|I_C^*(y^*) - I_D^*(y^*)| \mid \|y^*\| = 1\} \leq \rho(C,D). \tag{7.11}$$

If the inequality in (7.11) were strict, then we could find μ with

$$\sup\{|I_C^*(y^*) - I_D^*(y^*)| \mid \|y^*\| = 1\} < \mu < \rho(C,D).$$

Then we would have for each y^* with norm 1 the inequality $I_C^*(y^*) \leq I_D^*(y^*) + \mu$, and therefore for each $z^* \in \mathbb{R}^n$ $I_C^*(z^*) \leq I_D^*(z^*) + \mu\|z^*\|$. Then $I_C^* \leq I_D^* + \mu\|\cdot\|$, which by part (b) would imply $C \subset D + \mu B^n$. Reasoning in the same way we could also conclude that $D \subset C + \mu B^n$. But then μ satisfies (7.9) but $\mu < \rho(C,D)$, a contradiction. Therefore (7.11) holds as an equation. □

It is often possible to use the information in Proposition 7.1.10 to establish facts about sets by manipulating the corresponding support functions, or *vice versa,* when a direct approach would have been much more difficult.

The next definition introduces an operation to produce from a given function f a positively homogeneous function $\operatorname{cone} f$ by applying the operation described in Exercise 6.1.22 to the cone generated by $\operatorname{epi} f$. For that reason we call it the *conical hull of* f. As the function is explicitly mentioned in this description, there should be no confusion with the conical hull of a set.

Definition 7.1.11. Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a convex function. The function $\operatorname{cone} f$ from $\mathbb{R}^n \to \bar{\mathbb{R}}$ is defined by

$$\operatorname{cone} f(y) = \inf_{\tau > 0} \tau^{-1} f(\tau y). \tag{7.12}$$

□

As mentioned before the definition, $\operatorname{cone} f(y)$ is the infimum of those η for which $(y,\eta) \in \operatorname{cone}\operatorname{epi} f$, and therefore in particular $\operatorname{cone} f$ is a convex function. To see

this, observe that $(y,\eta) \in \operatorname{cone}\operatorname{epi} f$ if and only if there is some positive τ such that $(\tau y, \tau\eta) \in \operatorname{epi} f$, which in turn is equivalent to

$$\eta \geq \tau^{-1} f(\tau y). \tag{7.13}$$

Taking the infimum shown in (7.12) is equivalent to taking the infimum of η in (7.13) and thus also to taking the infimum of the η such that $(y,\eta) \in \operatorname{cone}\operatorname{epi} f$.

This function can behave in an unpleasant way; for example, suppose that the origin is in the interior of $\operatorname{dom} f$ and that $f(0) < 0$. Then $f(x)$ is negative for all x near the origin (Theorem 6.1.13), so $\operatorname{epi} f$ contains a ball about the origin in $\mathbb{R}^{n+1}$. It follows that $\operatorname{cone} f \equiv -\infty$. However, if f is closed and proper with $f(0)$ being finite and positive, then $\operatorname{cone} f$ is closed and proper. We will prove this in Theorem 7.2.15 below. In the meantime, we show that this function lets us identify the support function of a level set in case the function involved is closed and proper.

Theorem 7.1.12. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a closed proper convex function and let ϕ be a real number. The support function of* $\operatorname{lev}_\phi f$ *is* $\operatorname{cl}\operatorname{cone}(f^* + \phi)$.

Proof. It suffices to prove the result for $\phi = 0$, because for any real ϕ we then have $\operatorname{lev}_\phi f = \operatorname{lev}_0(f - \phi)$, so the support function will be $\operatorname{cl}\operatorname{cone}(f-\phi)^* = \operatorname{cl}\operatorname{cone}(f^* + \phi)$, as required. Corollary 7.1.3 says that $\operatorname{cl}\operatorname{cone} f^*$ is the pointwise supremum of the affine minorants of $\operatorname{cone} f^*$. Therefore in particular a linear function of the form $\langle x^*, \cdot\rangle$ minorizes $\operatorname{cl}\operatorname{cone} f^*$ if and only if it minorizes $\operatorname{cone} f^*$.

The function $\operatorname{cl}\operatorname{cone} f^*$ is closed, positively homogeneous, convex, and not identically $+\infty$, so there are two cases: either it is identically $-\infty$ (the support function of the empty set) or else it is proper.

In the first case $\operatorname{cone} f^*$ takes $-\infty$ somewhere, so that there is some y^* such that for each real μ there is a positive τ such that $\tau^{-1} f^*(\tau y^*) < \mu$. If $x \in \operatorname{lev}_0 f$, then since $f = f^{**}$ we see that for each x^* we have $\langle x^*, x\rangle - f^*(x^*) \leq f(x) \leq 0$. Setting $\mu = \langle y^*, x\rangle$ and taking x^* to be of the form τy^* for positive τ, we have $\tau^{-1} f^*(\tau y^*) \geq \langle y^*, x\rangle = \mu$, a contradiction. Therefore the occurrence of the first case means that $\operatorname{lev}_0 f = \emptyset$, so the assertion of the theorem is true in that case.

In the second case $\operatorname{cl}\operatorname{cone} f^*$ is not identically $-\infty$; then it is proper, and then by Theorem 7.1.9 it is the support function of the set S consisting of x such that $\langle \cdot, x\rangle$ minorizes $\operatorname{cl}\operatorname{cone} f^*$. These x are, by our previous comment, the x such that $\langle \cdot, x\rangle$ minorizes $\operatorname{cone} f^*$. But the latter statement means that for each x^* and each positive τ, $\langle x^*, x\rangle \leq \tau^{-1} f^*(\tau x^*)$. Introducing a new variable $y^* = \tau x^*$ we have the equivalent statement that $x \in S$ if and only if for each $y^* \in \mathbb{R}^n$, $\langle y^*, x\rangle - f^*(y^*) \leq 0$. That inequality is equivalent to $0 \geq f^{**}(x) = f(x)$, so that $S = \operatorname{lev}_0 f$. Therefore $\operatorname{cl}\operatorname{cone} f^*$ is the support function of $\operatorname{lev}_0 f$ in the second case also. □

By evaluating the support function of a convex set C we can recover information about the position of the origin with respect to C. In fact, we can do this for any designated point x_0 of $\mathbb{R}^n$: the support function of $\{x_0\}$ is $\langle \cdot, x_0\rangle$, and C and $\operatorname{cl} C$ have the same support function (Exercise 7.1.20). Using these facts together with Proposition 7.1.10, we find that $I^*_{C-x_0} = I^*_{(\operatorname{cl} C)-x_0} = I^*_{\operatorname{cl} C} - \langle \cdot, x_0\rangle$. So by investigating

the position of the origin with respect to $C - x_0$ we can obtain information about the position of x_0 with respect to C.

Theorem 7.1.13. *Let C be a convex set in $\mathbb{R}^n$. Then:*

a. $0 \in \mathrm{cl}C$ *if and only if* I_C^* *is everywhere nonnegative.*
b. $0 \in \mathrm{int}C$ *if and only if* I_C^* *is positive everywhere except at* $x^* = 0$.
c. $0 \in \mathrm{ri}C$ *if and only if* I_C^* *is positive everywhere except at points* x^* *for which* $I_C^*(x^*) = I_C^*(-x^*) = 0$.

Proof. Since the closure, interior, and relative interior are the same for C and for $\mathrm{cl}C$, and since by Exercise 7.1.20 C and $\mathrm{cl}C$ have the same support function, we may assume in what follows that C is closed.

For (a), note that to say $I_{\mathrm{cl}C}^*$ is nonnegative is equivalent to saying that it majorizes the support function of the origin, and by our earlier remarks that is true if and only if C contains $\{0\}$.

To establish (b) and (c), we start by observing that if the origin is in C then $N_C(0)$ consists of those points x^* at which $I_C^*(x^*) = 0$. Now for (b), note that if $0 \in \mathrm{int}C$ then C contains a ball about the origin. The definition of the support function then shows it must be positive at any nonzero x^*. On the other hand, if I_C^* is positive everywhere except at the origin, then

$$I_C(0) = I_C^{**}(0) = \sup_{x^*}\{\langle x^*, 0\rangle - I_C^*(x^*)\} = 0,$$

so that $0 \in C$; further, the observation that we just made shows that $N_C(0) = \{0\}$. Now part (c) of Theorem 3.3.7 shows that 0 is actually in the interior of C.

For (c) note first that the exceptional points x^* mentioned in (c) may consist of two types. First, the origin always qualifies as long as C is nonempty. Second, if any such x^* are nonzero then they are normals to hyperplanes through the origin that contain C, and the set of all such normals, together with the origin, is exactly $(\mathrm{par}C)^\perp$. To see this, note that C must be contained in the lower closed halfspaces associated with hyperplanes through the origin with normals x^* and $-x^*$.

Now if the origin belongs to $\mathrm{ri}C$ then clearly I_C^* is everywhere nonnegative. By part (b) of Theorem 3.3.7 we have $N_C(0) = (\mathrm{par}C)^\perp$, which means that I_C^* takes the value 0 at exactly those points that are mentioned in (c). On the other hand, if I_C^* is positive everywhere but at the exceptional points, then the origin belongs to C by part (a), and then our observation shows that $N_C(0) = (\mathrm{par}C)^\perp$. By part (b) of Theorem 3.3.7 this is equivalent to saying that $0 \in \mathrm{ri}C$. □

7.1.4 Gauge functions

In looking at the support function of a convex set C we found that its epigraph consisted of the polar H_C° of the homogenizing cone of C. We know from Section 4.1.2 that this polar cone is closely connected to the polar set C° of C. That suggests

looking for symmetry in this construction: suppose that we look at $H_C^{\circ\circ}$, which is the closure of the homogenizing cone H_C. Suppose further that we turn the entire picture upside down by multiplying by the linear *reflection operator* $J : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ defined by $J(x,\xi) = (x,-\xi)$. Now $C \times \{1\}$ is at the top and $C^\circ \times \{-1\}$ is at the bottom. If we assume, as we will in this section, that C is closed and contains the origin, then it is $C^{\circ\circ}$.

From what we saw in the previous section the cone $\mathrm{cl}\, J(H_C)$, which is the polar of $J(H_C^\circ)$, should be the epigraph of the support function of C°. This support function will be a closed proper positively homogeneous convex function. It is called the *gauge* of C, and this section develops a few of its useful properties.

Suppose that C is nonempty. A point (x,ξ) belongs to $J(H_C)$ exactly when $\xi > 0$ and $\xi^{-1}x \in C$. The function we want to examine is therefore

$$\gamma_C(x) := \inf\{\xi \mid (x,\xi) \in J(H_C)\} = \inf\{\xi > 0 \mid \xi^{-1}x \in C\}. \tag{7.14}$$

This function is positively homogeneous because, for fixed $\alpha > 0$ and for any x and $\xi > 0$, one has $\xi^{-1}x \in C$ if and only if $(\alpha\xi)^{-1}(\alpha x) \in C$, and therefore $\gamma_C(\alpha x) = \alpha\gamma_C(x)$. As $J(H_C)$ is convex, γ_C is convex by Exercise 6.1.22. Its values are always nonnegative, so it cannot take $-\infty$, and if $x \in C$ then $(1)^{-1}x \in C$ and therefore $\gamma_C(x) \le 1$, so it is not identically $+\infty$. Therefore it is proper. To investigate conditions for it to be closed we need a preliminary lemma.

Lemma 7.1.14. *Let C be a closed convex subset of $\mathbb{R}^n$ containing the origin. Then*

$$\{x \mid \gamma_C(x) \le 0\} = \mathrm{rc}\, C, \qquad \{x \mid \gamma_C(x) \le 1\} = C. \tag{7.15}$$

Proof. First suppose $v \in \mathrm{rc}\, C$. As $0 \in C$, C must contain the ray $v\mathbb{R}_+$, so for $\alpha > 0$ we have $\alpha^{-1}x \in C$. The infimum of such α is zero, so $\gamma_C(x) = 0$. Conversely, if $x \notin \mathrm{rc}\, C$ then by Theorem 4.1.2 C cannot contain the ray $x\mathbb{R}_+$. Accordingly, there is some positive β such that if $\alpha > 0$ and $\alpha^{-1}x \in C$ then $\alpha^{-1} \le \beta$ and therefore $\alpha \ge \beta^{-1}$. Then $\gamma_C(x)$ is at least as large as β^{-1}, so it cannot be zero.

For the other assertion, let $x \in C$; then $(1)^{-1}x \in C$, so $\gamma_C(x) \le 1$. Now suppose $x \notin C$. Then $x \ne 0$ and $(1)^{-1}x \notin C$. As C is closed there is $\beta > 1$ with $\beta^{-1}x \notin C$. But then if $\alpha^{-1}x \in C$ we must have $\alpha^{-1} < \beta^{-1}$, so that $\alpha > \beta$. Then $\gamma_C(x) \ge \beta > 1$, so $\gamma_C(x) \ne 1$. □

Proposition 7.1.15. *Let C be a closed convex subset of $\mathbb{R}^n$ containing the origin. Then* $\mathrm{epi}\, \gamma_C = J(\mathrm{cl}\, H_C) = \mathrm{cl}\, J(H_C)$.

Proof. The second equality holds because J is nonsingular. To prove the first, write E for $J(\mathrm{cl}\, H_C)$. E is convex and is closed by the second equality; it is a cone because H_C is. As $0 \in C$ the ray $\{0\}^n \times \mathbb{R}_-$ lies in H_C and therefore $\{0\}^n \times \mathbb{R}_+ \subset E$. As E is a closed convex cone this ray belongs to $\mathrm{rc}\, E$, and it follows that E is an epigraph.

If $(x,\xi) \in E$ and $\xi > 0$ then $(\xi^{-1}x, -1) \in C \times \{-1\}$ by Theorem 4.1.9. Then $\gamma_C(x) \le \xi$, which implies $(x,\xi) \in \mathrm{epi}\, \gamma_C$. Therefore $E \subset \mathrm{epi}\, \gamma_C$.

Now suppose that $(x,\xi) \notin E$ and consider two cases.

a. If $\xi > 0$ then $(x, -\xi) \notin \operatorname{cl} H_C$ and therefore for sufficiently small positive ε and $\rho = \xi + \varepsilon$, also $(x, -\rho) \notin \operatorname{cl} H_C$ and therefore $(\rho^{-1}x, -1) \notin C \times \{-1\}$ by Theorem 4.1.9. Then $\rho^{-1}x \notin C$ and, as $0 \in C$, whenever $\alpha \leq \rho$ we also have $\alpha^{-1}x \notin C$. Then $\gamma_C(x) \geq \rho > \xi$, so that $(x, \xi) \notin \operatorname{epi} \gamma_C$.
b. If $\xi = 0$ then $(x, 0) \notin \operatorname{cl} H_C$ and therefore $x \notin \operatorname{rc} C$ by Theorem 4.1.9. Then Lemma 7.1.14 requires $\gamma_C(x) > 0$ so that $(x, \xi) \notin \operatorname{epi} \gamma_C$.

The results of these two cases show that $\operatorname{epi} \gamma_C \subset E$. □

If we require that the origin lie in the relative interior of C rather than just in C, we can get better behavior from γ_C.

Proposition 7.1.16. *Let C be a closed convex subset of $\mathbb{R}^n$ with $0 \in \operatorname{ri} C$. Then*

a. $\operatorname{dom} \gamma_C = \operatorname{par} C$.
b. There is a positive β such that for each $x \in \mathbb{R}^n$, $0 \leq \gamma_C(x) \leq \beta \|x\|$.
c. γ_C is locally Lipschitz continuous relative to $\operatorname{par} C$.
d. If C is compact, then there is a positive δ such that for each $x \in \mathbb{R}^n$, $\delta \|x\| \leq \gamma_C(x)$.

Proof. Let $L = \operatorname{par} C$ and write $B_L(x, \xi) = L \cap B(x, \xi)$. Find a positive β such that $B_L(0, \beta^{-1}) \subset C$. If $x \notin L$ then there is no positive α with $\alpha^{-1}x \in L$ and therefore $\gamma_C(x) = +\infty$ and $x \notin \operatorname{dom} \gamma_C$.

If $x \in L$ then if $x \in \operatorname{rc} C$ we have $\gamma_C(x) = 0$ (Lemma 7.1.14. If $x \notin \operatorname{rc} C$ then $\beta^{-1}(x/\|x\|) \in C$ by choice of β and it follows that $0 \leq \gamma_C(x) \leq \beta \|x\|$. In either case $x \in \operatorname{dom} \gamma_C$. This proves (a) and (b).

For (c) we use the fact that $\operatorname{dom} \gamma_C$ is the subspace L, hence relatively open. Then Theorem 6.1.13 shows that γ_C is locally Lipschitz continuous relative to L.

To prove (d), suppose that C is compact and choose a positive δ small enough so that $C \subset B_L(0, \delta^{-1})$. Let $x \in \mathbb{R}^n$. If α is positive with $\alpha^{-1}\|x\| > \delta^{-1}$ then $\alpha^{-1}x \notin C$. It follows that if $\alpha < \delta \|x\|$ it cannot appear in the infimum defining $\gamma_C(x)$, and therefore $\gamma_C(x) \geq \delta \|x\|$. □

Proposition 7.1.16 shows that if C is closed and $0 \in \operatorname{ri} C$ then γ_C is a fairly nice function if we remain in L. We will give two examples of how a gauge can be useful under these conditions. The first shows that we have already been making heavy use of gauges even though they have not appeared before this section. To state it we need a definition: a subset S of $\mathbb{R}^n$ is *symmetric* if whenever $x \in S$ then $-x \in S$ also.

Theorem 7.1.17. *Let C be a compact convex subset of $\mathbb{R}^n$ containing the origin in its interior. Then γ_C is a norm on $\mathbb{R}^n$ if and only if C is symmetric.*

Proof. (Only if.) If C were not symmetric there would be some $x \in C$ with $-x \notin C$. Then $\gamma_C(x) \leq 1$ and $\gamma_C(-x) > 1$ by Lemma 7.1.14, so γ_C could not be a norm.

(If.) Suppose C is symmetric. We show that γ_C satisfies the conditions given in Definition A.2.1, namely that for each x and y in $\mathbb{R}^n$ and each $\alpha \in \mathbb{R}$,

a. $\gamma_C(x) \geq 0$, and $\gamma_C(x) = 0$ if and only if $x = 0$.
b. $\gamma_C(\alpha x) = |\alpha| \gamma_C(x)$.

c. $\gamma_C(x+y) \le \gamma_C(x) + \gamma_C(y)$.

For (a), we know that γ_C is nonnegative, and is finite everywhere on $\mathbb{R}^n$ by Proposition 7.1.16 because the hypotheses require $\dim C = n$. As C is compact its recession cone is the origin, and then by Lemma 7.1.14 it is zero at that point only.

To prove (b) we choose $x \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$. If $\alpha > 0$ we have (b) by positive homogeneity. If $\alpha = 0$ then (b) holds because $\gamma_C(0) = 0$. If $\alpha < 0$ then we write

$$\gamma_C(\alpha x) = \gamma_C(-|\alpha|x) = \gamma_C(|\alpha|x) = |\alpha|\gamma_C(x),$$

where for the second and third equalities we used symmetry and positive homogeneity respectively.

Finally, let x and y be points of $\mathbb{R}^n$; then

$$\begin{aligned}\gamma_C(x+y) &= \gamma_C[2(.5x+.5y)] \\ &= 2\gamma_C(.5x+.5y) \\ &\le 2[.5\gamma_C(x)+.5\gamma_C(y)] \\ &= \gamma_C(x)+\gamma_C(y),\end{aligned}$$

where we used in turn positive homogeneity, convexity, and again positive homogeneity. This proves (c). □

If $|\cdot|$ is a norm on $\mathbb{R}^n$ let $Q = \{x \mid |x| \le 1\}$ be the unit ball of $|\cdot|$ (although it may not look like a ball). Then for $x \in \mathbb{R}^n$,

$$\gamma_Q(x) = \inf\{\alpha > 0 \mid \alpha^{-1}x \in Q\} = \inf\{\alpha > 0 \mid |x| \le \alpha\} = |x|,$$

so that $|\cdot|$ is the gauge of its unit ball. In particular, the Euclidean norm $\|\cdot\|$ is the gauge of the standard unit ball B^n. But Theorem 7.1.17 shows how one can create a norm from any symmetric compact convex set containing the origin in its interior, and Corollary A.2.9 shows that all of these norms are equivalent.

The following theorem establishes a fact that would be difficult to prove without the assistance of gauges.

Theorem 7.1.18. *If C and D are compact convex subsets of $\mathbb{R}^n$, each having dimension n, then they are homeomorphic.*

Proof. We must show that there are continuous maps $g : C \to D$ and $h : D \to C$ such that $g \circ h = \mathrm{id}_C$ and $h \circ g = \mathrm{id}_D$. Begin by choosing points $c_0 \in \operatorname{int} C$ and $d_0 \in \operatorname{int} D$ and defining translations $\tau_C : C \to \mathbb{R}^n$ and $\tau_D : D \to \mathbb{R}^n$ by $\tau_C(c) = c - c_0$ and $\tau_D(d) = d - d_0$. Let $E = \tau_C(C)$ and $F = \tau_D(D)$. We will show that E and F are homeomorphic via maps $e : E \to F$ and $f : F \to E$, and it will then follow that C and D are homeomorphic via $\tau_D^{-1} \circ e \circ \tau_C$ and $\tau_C^{-1} \circ f \circ \tau_D$ respectively.

By construction, E and F are compact convex sets, each containing the origin in its interior, so that all of the conclusions of Proposition 7.1.16 apply to the gauges γ_E and γ_F. We define the maps e and f by

$$\begin{aligned} e(x) &= [\gamma_E(x)/\gamma_F(x)]x \quad (0 \neq x \in E), \qquad e(0) = 0, \\ f(y) &= [\gamma_F(y)/\gamma_E(y)]y \quad (0 \neq y \in F), \qquad f(0) = 0. \end{aligned} \tag{7.16}$$

For $x \in E$ we have $\gamma_F[e(x)] = \gamma_E(x) \leq 1$, so $e(x) \in F$ and therefore $e : E \to F$, and a similar calculation shows that $f : F \to E$.

Also, $e(x)$ is surely continuous when $x \neq 0$, because parts (b) and (d) of Proposition 7.1.16 show that there are positive β and δ, and positive η and θ, such that

$$\delta\|x\| \leq \gamma_E(x) \leq \beta\|x\|, \qquad \theta\|x\| \leq \gamma_F(x) \leq \eta\|x\|,$$

so for any nonzero $x \in E$ we have

$$\delta\eta^{-1} \leq \gamma_E(x)/\gamma_F(x) \leq \beta\theta^{-1}$$

and therefore

$$\|e(x)\| \leq \beta\theta^{-1}\|x\|.$$

As $e(0) = 0$ this shows that e is also continuous at 0, and a similar argument shows that f is continuous on F.

We have for $0 \neq x \in E$

$$(f \circ e)(x) = \{\gamma_F[e(x)]/\gamma_E[e(x)]\}e(x). \tag{7.17}$$

However,

$$\begin{aligned} \gamma_F[e(x)] &= [\gamma_E(x)/\gamma_F(x)]\gamma_F(x), \\ \gamma_E[e(x)] &= [\gamma_E(x)/\gamma_F(x)]\gamma_E(x), \\ e(x) &= [\gamma_E(x)/\gamma_F(x)]x, \end{aligned}$$

and by substituting these values into (7.17) we find that $(f \circ e)(x) = x$. This equation also holds if $x = 0$, so we conclude that $f \circ e = \mathrm{id}_E$, and a similar calculation shows that $e \circ f = \mathrm{id}_F$. □

7.1.5 Exercises for Section 7.1

Exercise 7.1.19. Compute the closure and the conjugate of the function defined by

$$f(x) = \begin{cases} x\ln x & \text{if } x > 0, \\ +\infty & \text{if } x \leq 0. \end{cases}$$

In computing the closure, explain how you know that the values you have prescribed on $(-\infty, 0]$ are correct.

Exercise 7.1.20. Show that for any subset S of $\mathbb{R}^n$, whether convex or not, $I_S^* = I_{\mathrm{cl}\,\mathrm{conv}\,S}^*$.

Exercise 7.1.21. a. Let $g(x^*)$ be the function of Exercise 6.2.22, namely

$$g(x_1^*, x_2^*) = \begin{cases} x_2^{*2}/(2x_1^*) & \text{if } x_1^* > 0, \\ 0 & \text{if } (x_1^*, x_2^*) = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Identify a closed convex set C such that g is the support function of C, and show that your specification of C is correct. *Suggestion.* Start by using g to calculate some supporting hyperplanes to C, then sketch them in order to get an idea of the shape of that set.

b. If you were asked to prove that the cone K of Exercise 3.1.18 is closed and convex, what would be a relatively easy way to do so given the information that you now have?

7.2 Recession functions

This section develops several results dealing with the way in which convex functions vary as one moves along halflines in $\mathbb{R}^n$. An intuitive, though oversimplified, way to summarize these results is to say that the asymptotic behavior of f along such a halfline depends only on the direction of the halfline and not on its location. This is a very strong property of convex functions.

A convenient way to present these results is through the *recession function,* whose epigraph is just the recession cone of the epigraph of the function under consideration.

Definition 7.2.1. Let f be a proper convex function on $\mathbb{R}^n$. The *recession function* of f is the function $\operatorname{rec} f : \mathbb{R}^n \to \bar{\mathbb{R}}$ defined by $\operatorname{epi} \operatorname{rec} f = \operatorname{rc} \operatorname{epi} f$. □

The next proposition shows that this definition makes sense, and gives some elementary properties of $\operatorname{rec} f$.

Proposition 7.2.2. *If f is a proper convex function on $\mathbb{R}^n$ then so is* $\operatorname{rec} f$*, which is positively homogeneous with* $\operatorname{rec} f(0) = 0$*. For each $z \in \mathbb{R}^n$ one has*

$$\operatorname{rec} f(z) = \sup_{x \in \operatorname{dom} f} [f(x+z) - f(x)]. \tag{7.18}$$

Proof. We first show that $\operatorname{rec} f$ is well defined, convex, and positively homogeneous. The set $\operatorname{rc}(\operatorname{epi} f)$ is convex, and as it contains $\{(0) \times \mathbb{R}_+\}$ it is an epigraph. As that epigraph is a cone, $\operatorname{rec} f$ is positively homogeneous.

To show that $\operatorname{rec} f$ is proper and that $\operatorname{rec} f(0) = 0$, choose any $z \in \mathbb{R}^n$. The vertical line $J := \{z\} \times \mathbb{R}$ through z either meets $\operatorname{rc} \operatorname{epi} f$ or fails to do so.

If it fails, then for each real η the point (z, η) fails to belong to $\operatorname{rc} \operatorname{epi} f$, and therefore there is some $(x, \xi) \in \operatorname{epi} f$ with $(x+z, \xi+\eta) \notin \operatorname{epi} f$. This implies $x \in \operatorname{dom} f$, so $f(x)$ is finite, and it also implies that

$$f(x+z) > \xi + \eta \geq f(x) + \eta,$$

so that $f(x+z) - f(x) > \eta$. As η was arbitrary, we see that in this case the supremum in (7.18) is $+\infty$.

If J meets $\operatorname{rc}\operatorname{epi} f$, then for some real η $(z,\eta) \in \operatorname{rc}\operatorname{epi} f = \operatorname{epi}\operatorname{rec} f$. Then $\operatorname{rec} f(z) \leq \eta < +\infty$, so $\operatorname{rec} f$ is not always $+\infty$, and $z \in \operatorname{dom}\operatorname{rec} f$. If $\operatorname{rec} f(z) = -\infty$ then $J \subset \operatorname{epi}\operatorname{rec} f = \operatorname{rc}\operatorname{epi} f$, so for any $x \in \operatorname{dom} f$ and any real number β, $(x+z, f(x)+\beta) \in \operatorname{epi} f$. This implies that $f(x+z) = -\infty$ so that f is improper, contradicting the hypothesis. So $\operatorname{rec} f(z)$ is finite, and as we could take z to be any element of $\operatorname{dom}\operatorname{rec} f$, $\operatorname{rec} f$ is proper.

The origin of $\mathbb{R}^{n+1}$ belongs to $\operatorname{rc}\operatorname{epi} f = \operatorname{epi}\operatorname{rec} f$, so that $\operatorname{rec} f(0) \leq 0$. Lemma 7.1.8 then shows that $\operatorname{rec} f(0) = 0$.

It remains to show that (7.18) holds. We have already shown that it holds in the case where $\operatorname{rec} f(z) = +\infty$. In the case where $z \in \operatorname{dom}\operatorname{rec} f$, $\operatorname{rec} f(z)$ was a real number bounded above by η, and was not $-\infty$. In that case $\operatorname{rec} f(z)$ is the finite real number $\min\{\mu \in \mathbb{R} \mid (z,\mu) \in \operatorname{rc}\operatorname{epi} f\}$. We have $(z,\mu) \in \operatorname{rc}\operatorname{epi} f$ if and only if for each $x \in \operatorname{dom} f$, $(x+z, f(x)+\mu) \in \operatorname{epi} f$: that is, $f(x+z) \leq f(x)+\mu$, or equivalently $\mu \geq f(x+z) - f(x)$. The least such μ is $\sup_{x \in \operatorname{dom} f}[f(x+z) - f(x)]$, and this is then the value of $\operatorname{rec} f(z)$. □

If f is closed then we can say more. We need a small lemma about the behavior of a difference quotient.

Lemma 7.2.3. *Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ be a convex function. If f has a finite value at $x \in \mathbb{R}^n$, then for any $y \in \mathbb{R}^n$ the difference quotient*

$$Q(x,y,\tau) = \tau^{-1}[f(x+\tau y) - f(x)] \tag{7.19}$$

is nondecreasing for $\tau > 0$.

Proof. Suppose that $0 < \tau_1 < \tau_2$; then

$$x + \tau_1 y = [1 - (\tau_1/\tau_2)](x) + (\tau_1/\tau_2)(x + \tau_2 y),$$

so by convexity

$$f(x+\tau_1 y) \leq [1 - (\tau_1/\tau_2)]f(x) + (\tau_1/\tau_2) f(x+\tau_2 y).$$

Rearranging this inequality yields $Q(x,y,\tau_1) \leq Q(x,y,\tau_2)$ as required. □

Proposition 7.2.4. *Let f be a closed proper convex function on $\mathbb{R}^n$. Then* $\operatorname{rec} f$ *is closed, proper and convex. If we define $Q(x,y,\tau)$ by* (7.19)*, then for any $x \in \operatorname{dom} f$ and any $y \in \mathbb{R}^n$ one has*

$$\operatorname{rec} f(y) = \sup_{\tau > 0} Q(x,y,\tau) = \lim_{\tau \to +\infty} Q(x,y,\tau).$$

Proof. Lemma 7.2.3 shows that $Q(x,y,\tau)$ is nondecreasing for $\tau > 0$, so the supremum and the limit in the assertion of the theorem are the same. Also, as f is closed

its epigraph is closed, so $\operatorname{rc}\operatorname{epi} f$ is closed and convex by Theorem 4.1.2. This means $\operatorname{rec} f$ is lower semicontinuous (Theorem 6.2.7). But $\operatorname{rec} f$ is proper (Proposition 7.2.2), so it is actually a closed proper convex function.

Now let $x \in \operatorname{dom} f$ and $y \in \mathbb{R}^n$. The point $(x, f(x))$ belongs to $\operatorname{epi} f$, which is a closed set. Therefore, for any real number η, saying that (y,η) belongs to $\operatorname{epi}\operatorname{rec} f$ is equivalent to saying that $\operatorname{epi} f$ contains the halfline $(x+\tau y, f(x)+\tau\eta)$ for $\tau \geq 0$ (Theorem 4.1.2). In turn, this is the same as saying that $\eta \geq Q(x,y,\tau)$ for all positive τ. Therefore $\operatorname{rec} f(y)$ is the minimum of the η for which the latter inequality holds: that is, it is the supremum in the assertion of the theorem. □

To illustrate the usefulness of recession functions we establish some properties of separable convex functions that we will need later.

Proposition 7.2.5. *For $i = 1,\ldots,k$ let $g_i : \mathbb{R}^{n_i} \to (0,+\infty]$ be a proper convex function. Let $n = \sum_{i=1}^k n_i$; for $i = 1,\ldots,k$ and $x_i \in \mathbb{R}^{n_i}$ define*

$$g(x_1,\ldots,x_k) = \sum_{i=1}^k g_i(x_i).$$

The function g is then a proper convex function on $\mathbb{R}^n$, and one has for $i = 1,\ldots,k$ and $x_i \in \mathbb{R}^{n_i}$

$$\operatorname{cl} g(x_1,\ldots,x_k) = \sum_{i=1}^k \operatorname{cl} g_i(x_i), \tag{7.20}$$

and for $y_i \in \mathbb{R}^{n_i}$

$$\operatorname{rec}\operatorname{cl} g(y_1,\ldots,y_k) = \sum_{i=1}^k \operatorname{rec}\operatorname{cl} g_i(y_i). \tag{7.21}$$

This function g is not the usual sum of functions on a given space: rather, it is a separable function consisting of the sum of several functions on different spaces. We will establish recession (and other) properties of the usual sum in a later result whose proof will use this proposition.

Proof. The function g is convex by Proposition 6.1.8. As each g_i is proper, g cannot take $-\infty$. Further, if for $i = 1,\ldots,n$ g_i we take $\hat{x}_i$ to be a point of $\operatorname{ri}\operatorname{dom} g_i$, then g_i is finite there; so then is $g(\hat{x}_1,\ldots,\hat{x}_k)$, and therefore g is proper.

For the closure formula, let $x_i \in \mathbb{R}^{n_i}$ for $i = 1,\ldots,k$. We have $\operatorname{dom} g = \Pi_{i=1}^k \operatorname{dom} g_i$, and therefore $\operatorname{ri}\operatorname{dom} g = \Pi_{i=1}^k \operatorname{ri}\operatorname{dom} g_i$ by Exercise 1.2.15. Then Proposition 6.2.15 says that

$$\begin{aligned}
\operatorname{cl} g(x_1,\ldots,x_k) &= \lim_{\mu\downarrow 0} \sum_{i=1}^k g_i[(1-\mu)x_i + \mu\hat{x}_i] \\
&= \sum_{i=1}^k \lim_{\mu\downarrow 0} g_i[(1-\mu)x_i + \mu\hat{x}_i] \\
&= \sum_{i=1}^k \operatorname{cl} g_i(x_1,\ldots,x_k),
\end{aligned}$$

which proves (7.20).

The claim in (7.21) involves only closures, so for brevity let $h_i = \operatorname{cl} g_i$ and $h = \operatorname{cl} g$. We have already shown that $h(x_1,\dots,x_k) = \sum_{i=1}^k h_i(x_i)$. Let $H = \operatorname{epi} h$ and for $i = 1,\dots,k$ let $H_i = \operatorname{epi} h_i$. If we define L to be the linear operator from $\mathbb{R}^{n+k}$ to $\mathbb{R}^{n+1}$ given by

$$L(x_1,\xi_1,\dots,x_k,\xi_k) = (x_1,\dots,x_k,\sum_{i=1}^k \xi_i),$$

where for each i $x_i \in \mathbb{R}^{n_i}$ and $\xi_i \in \mathbb{R}$, then

$$H = L(\Pi_{i=1}^k H_i). \tag{7.22}$$

The kernel of L consists of points $z = (0,\tau_1,\dots,0,\tau_k) \in \mathbb{R}^n$ such that $\sum_{i=1}^k \tau_i = 0$. The recession cone of $\Pi_{i=1}^k H_i$ is $\Pi_{i=1}^k \operatorname{rc} H_i$ by Exercise 4.1.24, so if $z \in (\ker L) \cap (\operatorname{rc} \Pi_{i=1}^k H_i)$ then each τ_i is nonnegative because otherwise $\operatorname{cl} g_i$ would be identically $-\infty$. As the τ_i must sum to zero, they are all zero and therefore $z = 0$. This shows that the closed set $\Pi_{i=1}^k H_i$ satisfies the recession condition for L. By Corollary 4.1.18,

$$\operatorname{rc} H = L(\operatorname{rc} \Pi_{i=1}^k H_i) = L(\Pi_{i=1}^k \operatorname{rc} H_i) :$$

that is,

$$\operatorname{epi} \operatorname{rec} \operatorname{cl} g = L(\Pi_{i=1}^k \operatorname{epi} \operatorname{rec} \operatorname{cl} h_i),$$

so that (7.21) holds. □

Points y at which $\operatorname{rec} f$ takes nonpositive values are of special interest, so they have special names.

Definition 7.2.6. Let f be a proper convex function on $\mathbb{R}^n$. The *recession cone* of f is the set $\operatorname{rc} f$ of points y at which $\operatorname{rec} f(y) \le 0$, and the *constancy space* of f is $\operatorname{cs} f = (\operatorname{rc} f) \cap (-\operatorname{rc} f)$. □

Note that $\operatorname{rec} f(0) = 0$ (Proposition 7.2.2), so that each of the sets in Definition 7.2.6 contains the origin. To see why they are important – and why they have the names given in that definition – we need to develop more information about the way in which f behaves.

Theorem 7.2.7. *Let f be a proper convex function on $\mathbb{R}^n$. A point $y \in \mathbb{R}^n$ belongs to* $\operatorname{rc} f$ *if and only if for each $x \in \mathbb{R}^n$ the function $\phi_x(\tau) = f(x+\tau y)$ is nonincreasing on $\mathbb{R}$, and y belongs to* $\operatorname{cs} f$ *if and only if for each $x \in \mathbb{R}^n$ ϕ_x is constant on $\mathbb{R}$.*

Proof. If $y = 0$ the assertion is trivial, so we can assume that $y \neq 0$. Also, we need only prove the first assertion, since then by applying it to y and $-y$ we can establish the second.

First suppose that for each $x \in \mathbb{R}^n$ ϕ_x is nonincreasing. Let $x \in \operatorname{dom} f$; then $\phi_x(1) \le \phi_x(0)$, so $f(x+y) - f(x) \le 0$. Taking the supremum over $x \in \operatorname{dom} f$ and applying Proposition 7.2.2, we conclude that $\operatorname{rec} f(y) \le 0$, so that $y \in \operatorname{rc} f$.

Next, suppose that $y \in \operatorname{rc} f$, and choose $x \in \mathbb{R}^n$. If the line $L = \{x + \tau y \mid \tau \in \mathbb{R}\}$ does not meet $\operatorname{dom} f$, then $\phi_x \equiv +\infty$, so the assertion is true. If L meets $\operatorname{dom} f$ (a convex set), the set $\{\tau \in \mathbb{R} \mid x + \tau y \in \operatorname{dom} f\}$ is an interval T in $\mathbb{R}$. This interval must be unbounded on the right, because $y \in \operatorname{rc} f$ implies $f(x+y) \le f(x)$ by Proposition 7.2.2, and therefore $f(x+ky) \le f(x)$ for each positive integer k. Any $\tau \notin T$ must have $\phi_x(\tau) = +\infty$ by definition of ϕ_x. Therefore it suffices to show that for each τ_1 and τ_2 in T with $\tau_1 < \tau_2$, $\phi_x(\tau_1) \ge \phi_x(\tau_2)$. As $x + \tau_1 y \in \operatorname{dom} f$, we have by Proposition 7.2.2 and the hypothesis

$$\begin{aligned} 0 &\ge (\tau_2 - \tau_1) \operatorname{rec} f(y) = \operatorname{rec} f[(\tau_2 - \tau_1)y] \\ &\ge f[(x + \tau_1 y) + (\tau_2 - \tau_1)y] - f(x + \tau_1 y) \\ &= \phi_x(\tau_2) - \phi_x(\tau_1), \end{aligned}$$

which shows that ϕ_x is nonincreasing. □

Remark 7.2.8. If f is closed and proper, and if for some particular $x \in \operatorname{dom} f$ and $y \in \mathbb{R}^n$ the quantity $f(x+\tau y)$ does not converge to $+\infty$ as $\tau \to +\infty$, then the difference quotient $Q(x, y, \tau)$ must converge to a nonpositive value. This in turn means that $\operatorname{rec} f(y) \le 0$ by Proposition 7.2.4 and so by Theorem 7.2.7 $f(x+\tau y)$ is actually *nonincreasing* as a function of $\tau \in \mathbb{R}$.

The recession cone of a closed convex function f can be used to gain information about other quantities associated with f. We first look at the situation for level sets, which we write for $\phi \in \mathbb{R}$ as $\operatorname{lev}_\phi f = \{x \mid f(x) \le \phi\}$.

Theorem 7.2.9. *Let f be a closed proper convex function on $\mathbb{R}^n$. and let $\phi \in \mathbb{R}$. If $\operatorname{lev}_\phi f \ne \emptyset$ then $\operatorname{rc} \operatorname{lev}_\phi f = \operatorname{rc} f$ and $\operatorname{lin} \operatorname{lev}_\phi f = \operatorname{cs} f$.*

Proof. The second assertion follows from the first, so we prove only the first. The set $\operatorname{lev}_\phi f$ is closed, so a vector y belongs to its recession cone only if the level set contains a halfline from some x in the direction of y. Remark 7.2.8 then shows that $y \in \operatorname{rc} f$. Conversely, if $y \in \operatorname{rc} f$ then let $x \in \operatorname{lev}_\phi f$. By Theorem 7.2.7 the values of f along the halfline from x in the direction of y must be nonincreasing, so this halfline lies in $\operatorname{lev}_\phi f$. Then Theorem 4.1.2 says that $y \in \operatorname{rc} \operatorname{lev}_\phi f$. □

Theorem 7.2.9 is a very strong statement about level sets: it says that, roughly speaking, all of the nonempty level sets have to be unbounded in the same way. Therefore we can use information about $\operatorname{rc} f$, if we have it, to predict properties of the level sets. If the set of minimizers of f is nonempty, it is a level set, and so in particular we can get information about minimizers in this way.

To enhance our ability to compute $\operatorname{rc} f$ we now look at a way of determining the recession function $\operatorname{rec} f$. It turns out to be easier to work with the conjugate, which is always closed and whose recession function is therefore closed. As this recession function is closed, proper, positively homogeneous, and convex, Theorem 7.1.9 tells us that it must then be the support function of some nonempty convex set. In fact the support function uniquely specifies the closure of this set, but if we do not demand that the set be closed then there may be many candidates. The following theorem identifies this set as the effective domain of the original function.

Theorem 7.2.10. *Let f be a proper convex function on $\mathbb{R}^n$. Then* $\operatorname{rec} f^* = I^*_{\operatorname{dom} f}$. *If f is closed then* $\operatorname{rec} f = I^*_{\operatorname{dom} f^*}$.

Proof. The second statement follows from the first because $f^{**} = \operatorname{cl} f$ (Theorem 7.1.4). To prove the first statement, recall that $\operatorname{epi} f^*$ is the intersection over $x \in \operatorname{dom} f$ of the closed halfspaces $\{(x^*, \xi^*) \mid \langle (x^*, \xi^*), (x, -1)\rangle \le f(x)\}$. By Exercise 4.1.25 the recession cones of these halfspaces are the halfspaces $\{(x^*, \xi^*) \mid \langle (x^*, \xi^*), (x, -1)\rangle \le 0\}$. Because the halfspaces involved are closed, Corollary 4.1.5 tells us that the recession cone of $\operatorname{epi} f^*$ is the set of (x^*, ξ^*) such that for each $x \in \operatorname{dom} f$, $\langle x^*, x\rangle \le \xi^*$: that is, $\xi^* \ge I^*_{\operatorname{dom} f}$. Therefore the epigraphs of $\operatorname{rec} f^*$ and of $I^*_{\operatorname{dom} f}$ are the same, which proves the assertion. □

By combining Theorems 7.2.9 and 7.2.10 we obtain very strong results about the level sets of a function f in terms of properties of the effective domain of f^*.

Theorem 7.2.11. *Let f be a closed proper convex function on $\mathbb{R}^n$, and let ϕ be a real number with* $\operatorname{lev}_\phi f \neq \emptyset$. *Then:*

a. $\operatorname{lev}_\phi f = \operatorname{cs} f + K_\phi$, *where* $K_\phi = [(\operatorname{lev}_\phi f) \cap (\operatorname{cs} f)^\perp]$.
b. K_ϕ is compact if and only if $0 \in \operatorname{ri} \operatorname{dom} f^*$.
c. $\operatorname{lev}_\phi f$ *is compact if and only if* $0 \in \operatorname{int} \operatorname{dom} f^*$.

Proof. By Theorem 7.2.9 the lineality space and the recession cone of $\operatorname{lev}_\phi f$ are $\operatorname{cs} f$ and $\operatorname{rc} f$ respectively. Then assertion (a) follows from Proposition A.6.23.

For assertion (b), we apply Proposition A.6.23 again, using the fact that $\operatorname{cs} f = \operatorname{lin} \operatorname{rc} f$, to obtain

$$\operatorname{rc} f = \operatorname{cs} f + (\operatorname{rc} f) \cap (\operatorname{cs} f)^\perp. \tag{7.23}$$

As K_ϕ is nonempty, we have

$$\operatorname{rc} K_\phi = \operatorname{rc}[(\operatorname{lev}_\phi f) \cap (\operatorname{cs} f)^\perp] = (\operatorname{rc} \operatorname{lev}_\phi f) \cap (\operatorname{cs} f)^\perp = (\operatorname{rc} f) \cap (\operatorname{cs} f)^\perp,$$

where we used Corollary 4.1.5 and Theorem 7.2.9. If K_ϕ is compact, its recession cone is the origin (Corollary 4.1.8), which means that the second term in (7.23) is zero. Therefore $\operatorname{rc} f$ is just the constancy space $\operatorname{cs} f$. This means that the support function $I^*_{\operatorname{dom} f^*}$ can take nonpositive values only at those x where it is nonpositive also at $-x$. But since it takes the value zero at the origin, convexity then implies that it takes zero at both x and $-x$. Theorem 7.1.13 then tells us that $0 \in \operatorname{ri} \operatorname{dom} f^*$. Conversely, if $0 \in \operatorname{ri} \operatorname{dom} f^*$ then $\operatorname{rec} f$ is positive everywhere except on $\operatorname{cs} f$, so $\operatorname{rc} f = \operatorname{cs} f$ and therefore by the above analysis $\operatorname{rc} K_\phi$ is the origin. By Corollary 4.1.8, K_ϕ is then compact.

For assertion (c), note that $\operatorname{lev}_\phi f$ is compact if and only if its recession cone is the origin (Corollary 4.1.8), which is equivalent to saying that $\operatorname{rc} f$ is the origin (Theorem 7.2.9). In turn, by Theorem 7.2.10 this is equivalent to saying that $I^*_{\operatorname{dom} f^*}$ is positive everywhere except at the origin, and by Theorem 7.1.13 this is equivalent to the statement that the origin belongs to the interior of $\operatorname{dom} f^*$. □

One of the consequences of Theorem 7.2.11 is an important criterion for attainment of the minimum of a closed proper convex function.

Corollary 7.2.12. *Let f be a closed proper convex function on $\mathbb{R}^n$. If $0 \in \operatorname{ri}\operatorname{dom} f^*$ then f attains its minimum, which is finite.*

Proof. The minimum, if attained, must be finite because f is proper. If the origin belongs to $\operatorname{ri}\operatorname{dom} f^*$ then, according to Theorem 7.2.11, each nonempty level set $\operatorname{lev}_\phi f$ is of the form $\operatorname{cs} f + K_\phi$, with $K_\phi = (\operatorname{lev}_\phi f) \cap (\operatorname{cs} f)^\perp$. Moreover, K_ϕ is compact. This means that each element of $\operatorname{lev}_\phi f$ yields the same function value as some element of K_ϕ (because f is constant along any line in the direction of an element of $\operatorname{cs} f$). Therefore a particular value of f is attained on $\operatorname{lev}_\phi f$ if and only if it is attained on K_ϕ, and accordingly the minimum of f, which is also its minimum on $\operatorname{lev}_\phi f$, is the same as its minimum on K_ϕ. But K_ϕ is a compact subset of $\operatorname{lev}_\phi f$, so by Proposition 6.2.8 the minimum is attained on K_ϕ, and hence on $\operatorname{lev}_\phi f$. □

The criterion in Corollary 7.2.12 is sufficient, but not necessary (Exercise 7.2.22).

There is a close connection between the recession properties of a function and the effective domain of its conjugate. The next two results express some aspects of that connection.

Corollary 7.2.13. *Let f be a proper convex function on $\mathbb{R}^n$. Then* $\operatorname{rc} f^* = (\operatorname{cone}\operatorname{dom} f)^\circ$.

Proof. As f is proper, $\operatorname{dom} f$ is nonempty and so the polar operation is well defined. The polar of $\operatorname{cone}\operatorname{dom} f$ consists of all $y^* \in \mathbb{R}^n$ such that, for each $x \in \operatorname{dom} f$, $\langle y^*, x\rangle \le 0$, or equivalently the set of y^* for which $I^*_{\operatorname{dom} f}(y^*) \le 0$. But $I^*_{\operatorname{dom} f} = \operatorname{rec} f^*$ by Theorem 7.2.10, so this set is $\operatorname{rc} f^*$. □

Corollary 7.2.14. *Let f be a proper convex function on $\mathbb{R}^n$ and let L be a nonempty subspace of $\mathbb{R}^n$. Then $L \cap \operatorname{rc} f^*$ is a subspace if and only if $L^\perp$ meets* $\operatorname{ri}\operatorname{dom} f$.

Proof. By Proposition 3.2.3, $L \cap \operatorname{rc} f^*$ is a subspace if and only if $L^\perp$ meets $\operatorname{ri}[(\operatorname{rc} f^*)^\circ]$. Using Corollary 7.2.13, Corollary 3.1.9, Theorem 3.1.7, and Proposition 1.2.6, we have

$$\begin{aligned}\operatorname{ri}[(\operatorname{rc} f^*)^\circ] &= \operatorname{ri}[(\operatorname{cone}\operatorname{dom} f)^{\circ\circ}] = \operatorname{ri}[\operatorname{cl}\operatorname{cone}\operatorname{dom} f] \\ &= \operatorname{ri}\operatorname{cone}\operatorname{dom} f = \operatorname{cone}\operatorname{ri}\operatorname{dom} f.\end{aligned}$$

But a subspace like $L^\perp$ meets $\operatorname{cone}\operatorname{ri}\operatorname{dom} f$ if and only if it meets $\operatorname{ri}\operatorname{dom} f$. □

As mentioned after Definition 7.1.11 above, the recession function lets us establish a criterion for the conical hull $\operatorname{cone} f$ to be closed and proper.

Theorem 7.2.15. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a closed proper convex function and suppose that $f(0)$ is finite and positive. Then* $\operatorname{cone} f$ *is a closed proper convex function.*

Proof. We already know $\operatorname{cone} f$ is a convex function. Under our hypotheses its epigraph is nonempty, so it is not identically $+\infty$. The origin in $\mathbb{R}^{n+1}$ does not belong to the (closed) epigraph of f, so we can separate strongly to obtain a pair $(x^*, \xi^*) \in \mathbb{R}^{n+1}$ and a positive δ such that for each $(x,\xi) \in \operatorname{epi} f$, $0 > -\delta > \langle x^*, x\rangle + \xi^*\xi$. The form of $\operatorname{epi} f$ ensures that $\xi^* \le 0$, and if $\xi^* = 0$ then we obtain a contradiction

because $(0, f(0)) \in \operatorname{epi} f$. So $\xi^* < 0$, and we can scale (x^*, ξ^*) so that $\xi^* = -1$. Then for each pair $(y, \eta) \in \operatorname{cone}\operatorname{epi} f$ we have $\langle x^*, y\rangle - \eta \leq 0$, so the affine function $\langle x^*, \cdot\rangle$ minorizes $\operatorname{cone} f$. Then $\operatorname{cone} f$ cannot take $-\infty$, so it is proper. Therefore $\operatorname{epi}\operatorname{cl}\operatorname{cone} f = \operatorname{cl}\operatorname{epi}\operatorname{cone} f$.

The comment following the definition of $\operatorname{cone} f$ implies that

$$\operatorname{cone}\operatorname{epi} f \subset \operatorname{epi}\operatorname{cone} f \subset \operatorname{cl}\operatorname{cone}\operatorname{epi} f,$$

so that

$$\operatorname{epi}\operatorname{cl}\operatorname{cone} f = \operatorname{cl}\operatorname{epi}\operatorname{cone} f = \operatorname{cl}\operatorname{cone}\operatorname{epi} f.$$

Now Corollary 4.1.21 tells us that, since the origin does not belong to the (closed) epigraph of f, we have

$$\operatorname{epi}\operatorname{cl}\operatorname{cone} f = \operatorname{cl}\operatorname{cone}\operatorname{epi} f = (\operatorname{cone}\operatorname{epi} f) \cup (\operatorname{rc}\operatorname{epi} f).$$

As $\operatorname{rc}\operatorname{epi} f = \operatorname{epi}\operatorname{rec} f$, we see that for each y, $\operatorname{cl}\operatorname{cone} f(y)$ is the smaller of $\operatorname{cone} f(y)$ and $\operatorname{rec} f(y)$.

The rest of the proof consists of showing that $\operatorname{rec} f(y)$ is not required, and to do so we recall that as f is closed and $f(0)$ is finite, we have

$$\operatorname{rec} f(y) = \sup_{\tau > 0} \tau^{-1}[f(0 + \tau y) - f(0)].$$

Now $f(0)$ is finite and positive, so for any fixed positive δ we can find τ large enough so that

$$\operatorname{rec} f(y) \geq \tau^{-1} f(\tau y) - \tau^{-1} f(0) \geq \tau^{-1} f(\tau y) - \delta \geq \operatorname{cone} f(y) - \delta.$$

It follows that $\operatorname{rec} f(y) \geq \operatorname{cone} f(y)$, so in fact $\operatorname{cl}\operatorname{cone} f(y) = \operatorname{cone} f(y)$, and therefore $\operatorname{cone} f$ is closed. □

The next definition creates two new functions by a kind of composition operation with a scalar, using the recession function in the process. As we then show, if f is a closed proper convex function these compositions have interesting properties. In particular, they permit us to calculate the support function of an epigraph.

Definition 7.2.16. Suppose $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is a proper convex function. For $x \in \mathbb{R}^n$ and $\mu \in \mathbb{R}$ we define a function μf by

$$(\mu f)(x) := \begin{cases} \mu f(x) & \text{if } \mu > 0, \\ I_{\operatorname{cl}\operatorname{dom} f}(x) & \text{if } \mu = 0, \\ +\infty & \text{otherwise,} \end{cases} \tag{7.24}$$

and a function $f\mu$ by

$$(f\mu)(x) := \begin{cases} \mu f(\mu^{-1}x) & \text{if } \mu > 0, \\ (\operatorname{rec} f)(x) & \text{if } \mu = 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{7.25}$$

As the notation for the function μf is the same as that for the ordinary product of μ with f, there is a possibility of confusion. For that reason we usually write (μf) for the function of Definition 7.2.16 when there is danger of ambiguity.

We can regard these functions either as functions of x for fixed μ or as functions of (x,μ). In the first case, when f is closed we have the following relationship.

Proposition 7.2.17. *Let f be a closed proper convex function from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$ and let $\mu \in [0,+\infty)$. Then (μf) and $(f\mu)$ are closed proper convex functions of x, and one has*

$$(\mu f)^* = (f^*\mu), \qquad (f\mu)^* = (\mu f^*). \tag{7.26}$$

Proof. Both (μf) and $(f\mu)$ are closed proper convex for $\mu > 0$ because f is closed proper convex. For $\mu = 0$, (μf) is closed proper convex because $\operatorname{cl}\operatorname{dom} f$ is nonempty, closed, and convex, and $(f\mu)$ is closed proper convex because $\operatorname{rec} f$ is closed proper convex by Proposition 7.2.4.

If $\mu = 0$ then (7.26) says that $I^*_{\operatorname{cl} dom f} = \operatorname{rec} f^*$ and that $(\operatorname{rec} f)^* = I_{\operatorname{cl}\operatorname{dom} f^*}$. These follow from Theorem 7.2.10 because f is closed and because the support functions of a set and of its closure are the same.

If $\mu > 0$ then

$$\begin{aligned} (\mu f)^*(x^*) &= \sup_x \{\langle x^*, x\rangle - \mu f(x)\} \\ &= \mu \sup_x \{\langle \mu^{-1}x^*, x\rangle - f(x)\} \\ &= \mu f^*(\mu^{-1}x^*) = (f^*\mu)(x^*). \end{aligned}$$

establishing the first equation in (7.26). For the second, if we introduce the variable $y = \mu^{-1}x$ then we have

$$\begin{aligned} (f\mu)^*(x^*) &= \sup_x \{\langle x^*, x\rangle - \mu f(\mu^{-1}x)\} \\ &= \mu \sup_y \{\langle x^*, y\rangle - f(y)\} \\ &= \mu f^*(x^*) = (\mu f^*)(x^*). \end{aligned}$$

□

Suppose now that we were to consider (μf) and $(f\mu)$ to be functions of (x,μ). Then even when $f(x) = x \in \mathbb{R}$ and $\mu \geq 0$ the function $g(x,\mu) := (\mu f)(x)$ would not be convex, because $g(0,1) = 0 = g(1,0)$ but $g(.5,.5) = .25$. However, the function $h(x,\mu) := (f\mu)(x)$ is not only proper convex but is even closed if f is. That is a consequence of the next proposition, which establishes a very useful connection with the epigraph of f.

Proposition 7.2.18. *Let f be a closed proper convex function from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$. Then*

$$(f^*\mu^*)(x^*) = I^*_{\operatorname{epi} f}(x^*, -\mu^*). \tag{7.27}$$

Proof. If $\mu^* < 0$ then (7.27) holds because then the structure of the epigraph requires its support function to have the value $+\infty$. If $\mu^* = 0$ then the right side of (7.27) is the support function of $\operatorname{dom} f$ (equivalently, $\operatorname{cl}\operatorname{dom} f$) evaluated at x^*. This is $(\mu f)^*(x^*)$, which by Proposition 7.2.17 is $(f^*\mu)(x^*)$, so (7.27) holds also in this case.

Now suppose $\mu^* > 0$ and write $I^*_{\operatorname{epi} f}(x^*, -\mu^*)$ as

$$\sup_{x,\mu}\{\langle x^*, x\rangle - \mu^*\mu - I_{\operatorname{epi} f}(x,\mu)\} = \sup_{x,\mu}\{\langle x^*, x\rangle - \mu^*\mu \mid \mu \geq f(x)\}. \tag{7.28}$$

The best choice of μ that we can make in (7.28) to make the supremum on the right side as large as possible is $f(x)$. Then we have

$$I^*_{\operatorname{epi} f}(x^*, -\mu^*) = \sup_x\{\langle x^*, x\rangle - \mu^* f(x)\} = (\mu^* f)^*(x^*) = (f^*\mu^*)(x^*), \tag{7.29}$$

where we used Proposition 7.2.17 again. This proves (7.27). □

Corollary 7.2.19. *If c is a closed proper convex function on $\mathbb{R}^n$, then $(c\rho)(x)$ is a closed proper convex function of (x,ρ).*

Proof. We have assumed c to be closed, so in Proposition 7.2.18 take $f^* = c$, $f = c^*$, $x^* = x$, and $\mu^* = \rho$. Then the right side of (7.29) is $c\rho(x)$ and the left side is $I^*_{\operatorname{epi} c^*}(x, -\rho)$. Here c^* is proper because c is proper, so $\operatorname{dom} c^*$ is nonempty and therefore its support function is closed proper convex. □

7.2.1 Exercises for Section 7.2

Exercise 7.2.20. Let $f = I_{[\mathbb{R}_+ \times (0,+\infty)]\cup\{(0,0)\}}$. Find $\operatorname{rec} f$, and determine whether or not it is closed.

Exercise 7.2.21. Let f be a convex function on $\mathbb{R}^n$. Show that $\inf f = -f^*(0)$. If $\inf f = -\infty$, then what can you say about the position of the origin with respect to $\operatorname{dom} f^*$?

Exercise 7.2.22. By considering the function $f(x) = \sup\{0, -x\}$ on $\mathbb{R}$, show that the converse of Corollary 7.2.12 does not hold.

Exercise 7.2.23. *(Stable minimum property.)* Let f be a closed proper convex function on $\mathbb{R}^n$. Show that the origin belongs to the interior of $\operatorname{dom} f^*$ if and only if there is a positive ε such that, for each x^* having norm less than ε, the infimum of $f(x) - \langle x^*, x\rangle$ is finite and attained.

Exercise 7.2.24. Let f be a closed proper convex function on $\mathbb{R}^n$. Suppose you are told that that there exist elements $x \in \operatorname{dom} f$ and $y \in \mathbb{R}^n$, and a negative number μ, such that for all $\lambda \geq 0$, $f(x+\lambda y) \leq f(x) + \mu\lambda$. Give the best lower bound that you can determine for the distance from the origin to $\operatorname{dom} f^*$.

Chapter 8
Subdifferentiation

We have already seen subgradients of functions in Definition 6.2.12. They allow us to describe supporting hyperplanes to the epigraph of a convex function f defined on $\mathbb{R}^n$. If $x \in \operatorname{dom} f$, then a hyperplane with normal (z^*, ζ^*) supports $\operatorname{epi} f$ at $(x, f(x))$ if for each $(x', \xi') \in \operatorname{epi} f$ we have

$$\langle \begin{bmatrix} z^* \\ \zeta^* \end{bmatrix}, \begin{bmatrix} x' \\ \xi' \end{bmatrix} \rangle \leq \langle \begin{bmatrix} z^* \\ \zeta^* \end{bmatrix}, \begin{bmatrix} x \\ f(x) \end{bmatrix} \rangle,$$

or equivalently,

$$\langle z^*, x' \rangle + \zeta^* \xi' \leq \langle z^*, x \rangle + \zeta^* f(x). \tag{8.1}$$

The properties of an epigraph prevent the ξ^* in (8.1) from being positive, because if it were then we could take a large positive value of ξ' and contradict the inequality. Therefore it can be negative or else zero; in the latter case we say this hyperplane is *vertical.*

If $\zeta^* < 0$ then we can divide (8.1) by $-\zeta^*$, write $x^* = z^*/(-\zeta^*)$ and substitute $f(x')$ for ξ'. In this way we scale the normal (z^*, ζ^*) to $(x^*, -1)$ and obtain for each $x' \in \mathbb{R}^n$,

$$f(x') \geq f(x) + \langle x^*, x' - x \rangle. \tag{8.2}$$

The element x^* is a subgradient of f at x if and only if it satisfies (8.2).

From this we can see that subgradients represent scaled normals of non-vertical supporting hyperplanes: x^* is a subgradient of f at x exactly when $(x^*, -1)$ is the normal to a hyperplane supporting $\operatorname{epi} f$ at $(x, f(x))$. For some f the form of the epigraph may also permit vertical supporting hyperplanes, but these do not correspond to subgradients.

8.1 Subgradients and the subdifferential

To study the existence and properties of subgradients in a systematic way, it helps to introduce a multifunction called the *subdifferential.*

Definition 8.1.1. Let f be a convex function from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$. The *subdifferential* $\partial f(x)$ of the function f at a point $x \in \mathbb{R}^n$ is the set of all subgradients of f at x. The same definition holds for a concave function g on $\mathbb{R}^n$, except that then a subgradient x^* is defined to be a point such that for each $z \in \mathbb{R}^n$,

$$g(z) \leq g(x) + \langle x^*, z - x \rangle.$$

□

One indication of the importance of the subdifferential is the fact that x is a global minimizer of a convex f, or a global maximizer of a concave f, if and only if the origin belongs to $\partial f(x)$. Another important case is that of the normal cone of a convex set $C \subset \mathbb{R}^n$. If we take the convex function f to be the indicator I_C, then for $x \in C$ $\partial I_C(x)$ consists of those x^* such that for each $z \in \mathbb{R}^n$,

$$I_C(z) \geq I_C(x) + \langle x^*, z - x \rangle. \tag{8.3}$$

For $z \notin C$ (8.3) holds for any x^*, so the important case is that of $z \in C$. Then (8.3) says that for each $z \in C$, $\langle x^*, z - x \rangle \leq 0$, so that $\partial I_C(x)$ is the normal cone $N_C(x)$ of C at x.

Here are some equivalent ways of describing the subdifferential.

Proposition 8.1.2. *Let f be a convex function from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$. For each x and x^* in $\mathbb{R}^n$, $\partial f(x)$ is a closed convex set, and the following are equivalent:*

a. $x^* \in \partial f(x)$.
b. $f^*(x^*) = \langle x^*, x \rangle - f(x)$.
c. $\langle x^*, \cdot \rangle - f(\cdot)$ *attains its maximum on* $\mathbb{R}^n$ *at* x.

If $\operatorname{cl} f(x) = f(x)$ *then the preceding statements are all equivalent to the following:*

d. $x \in \partial f^*(x^*)$.
e. $\langle \cdot, x \rangle - f^*(\cdot)$ *attains its maximum on* $\mathbb{R}^n$ *at* x^*.

Proof. $\partial f(x)$ is the set of x^* for which, for each $z \in \mathbb{R}^n$, $f(z) \geq f(x) + \langle x^*, z - x \rangle$. A calculation then shows that the set $\partial f(x)$ is closed and convex.

((a) implies (b)). If (a) holds, then for each $z \in \mathbb{R}^n$, $f(z) \geq f(x) + \langle x^*, z - x \rangle$. Rearranging this, we obtain $\langle x^*, z \rangle - f(z) \leq \langle x^*, x \rangle - f(x)$, and this is (b).

((b) implies (c)). The supremum of $\langle x^*, \cdot \rangle - f(\cdot)$ is $f^*(x^*)$ by definition. If (b) holds then this is a maximum, attained at x, which is the statement in (c).

(c) implies (a)). If (c) holds, then for each $z \in \mathbb{R}^n$ one has $\langle x^*, z \rangle - f(z) \leq \langle x^*, x \rangle - f(x)$. By rearranging this we obtain $f(z) \geq f(x) + \langle x^*, z - x \rangle$, which is (a).

Now suppose $\operatorname{cl} f(x) = f(x)$ and write statements (a) to (c) for f^* and x^*. By what we have already shown, these must be equivalent. The first and third of these will

be (d) and (e) respectively, while the second statement is identical to (b) because $f^{**}(x) = \operatorname{cl} f(x) = f(x)$. Therefore in this case (d) and (e) are equivalent to (a), (b), and (c). □

It is certainly possible for $\partial f(x)$ to be empty. We are particularly interested in points x for which it is nonempty: that is, points of $\operatorname{dom} \partial f$. .

Proposition 8.1.3. *Let f be a proper convex function on $\mathbb{R}^n$. Then* $\operatorname{ri}\operatorname{dom} f \subset \operatorname{dom} \partial f \subset \operatorname{dom} f$.

Proof. Proposition 6.2.13 says that f has a subgradient at each point of $\operatorname{ri}\operatorname{dom} f$, and this proves the first inclusion. However, if f has a subgradient at a point x, then since f is not always $+\infty$ the value $f(x)$ cannot be $+\infty$ since otherwise the subgradient inequality could not hold. This proves the second inclusion. □

Next, we see that the regularizing operation that replaces f by $\operatorname{cl} f$ does not affect the subdifferential of f at any point x for which that subdifferential was nonempty. The symbol "$\subset$" in the last statement of the proposition denotes graph inclusion.

Proposition 8.1.4. *Let f be a proper convex function on $\mathbb{R}^n$. For each $x \in \operatorname{dom} \partial f$, one has $f(x) = (\operatorname{cl} f)(x)$ and $\partial f(x) = \partial(\operatorname{cl} f)(x)$. In particular, one has $\partial f \subset \partial \operatorname{cl} f$.*

Proof. Let $x \in \operatorname{dom} \partial f$ and $x^* \in \partial f(x)$; then we have

$$f(x) \geq (\operatorname{cl} f)(x) = f^{**}(x) \geq \langle x^*, x\rangle - f^*(x^*) = f(x),$$

where the last equality came from Proposition 8.1.2. Therefore $f(x) = (\operatorname{cl} f)(x)$. If $y^* \in \partial(\operatorname{cl} f)(x)$ then for each z one has

$$f(z) \geq \operatorname{cl} f(z) \geq \operatorname{cl} f(x) + \langle y^*, z - x\rangle = f(x) + \langle y^*, z - x\rangle,$$

and therefore $\partial(\operatorname{cl} f)(x) \subset \partial f(x)$. Now let $z_0 \in \operatorname{ri}\operatorname{dom} f$. For each $z \in \mathbb{R}^n$, we have by Corollary 6.2.16

$$(\operatorname{cl} f)(z) = \lim_{\tau \downarrow 0} f[(1-\tau)z + \tau z_0].$$

However,

$$f[(1-\tau)z + \tau z_0] \geq f(x) + \langle x^*, [(1-\tau)z + \tau z_0] - x\rangle,$$

and by taking the limit we find that

$$(\operatorname{cl} f)(z) \geq f(x) + \langle x^*, z - x\rangle = (\operatorname{cl} f)(x) + \langle x^*, z - x\rangle.$$

It follows that $\partial f(x) \subset \partial(\operatorname{cl} f)(x)$, so the two sets are the same.

Finally, for any $x \in \mathbb{R}^n$ either $x \in \operatorname{dom} \partial f$, in which case we have shown that $\partial f(x) = \partial \operatorname{cl} f(x)$, or else $\partial f(x)$ is empty, in which case it certainly is contained in $\partial \operatorname{cl} f(x)$. Therefore the graph of ∂f is contained in that of $\partial \operatorname{cl} f$. □

We now show that the subdifferentials of a closed convex function f and of its conjugate are inverses of each other, and when f is proper they are not only monotone operators but actually obey a stronger property called *cyclic monotonicity,* defined as follows.

Definition 8.1.5. A multifunction F defined from $\mathbb{R}^n$ to $\mathbb{R}^n$ is *cyclically monotone* if for any K and for any pairs $(x_0, f_0), \ldots, (x_K, f_K)$ in the graph of F, one has

$$\sum_{k=0}^{K} \langle f_k, x_{k+1} - x_k \rangle \leq 0,$$

where we use the convention that $x_{K+1} = x_0$. □

Monotonicity is the special case of cyclic monotonicity with $K = 1$.

Proposition 8.1.6. *Let f be a closed convex function on $\mathbb{R}^n$. Then ∂f and ∂f^* are closed multifunctions with $\partial f^* = (\partial f)^{-1}$. If f is proper then ∂f and ∂f^* are cyclically monotone.*

Proof. The equivalence of (a) and (d) in Proposition 8.1.2 shows that $(x, x^*) \in \partial f$ if and only if $(x^*, x) \in \partial f^*$, which is a rephrasing of $\partial f^* = (\partial f)^{-1}$. Knowing this, we need to prove closedness only for ∂f; as this is a graph property it will then hold also for ∂f^*.

First suppose that (x_k, x_k^*) belong to ∂f and converge to (x, x^*). For each $z \in \mathbb{R}^n$ and each k we have $f(z) \geq f(x_k) + \langle x_k^*, z - x_k \rangle$. Letting k approach ∞ and recalling that f is closed, we have

$$f(z) \geq [\liminf_{k \to \infty} f(x_k)] + \langle x^*, z - x \rangle \geq f(x) + \langle x^*, z - x \rangle,$$

so that $(x, x^*) \in \partial f$ and therefore ∂f is closed.

Now assume f is proper, choose a positive integer K, and suppose that for $k = 0, \ldots, K$ $(x_k, x_k^*) \in \partial f$. By Proposition 8.1.3, for each k $f(x_k)$ must be finite. Write the $K + 1$ subgradient inequalities

$$\begin{aligned}
f(x_1) &\geq f(x_0) + \langle x_0^*, x_1 - x_0 \rangle, \\
f(x_2) &\geq f(x_1) + \langle x_1^*, x_2 - x_1 \rangle, \\
&\ldots, \\
f(x_K) &\geq f(x_{K-1}) + \langle x_{K-1}^*, x_K - x_{K-1} \rangle, \\
f(x_0) &\geq f(x_K) + \langle x_K^*, x_0 - x_K \rangle,
\end{aligned}$$

and add these. As the function values cancel, we have the inequality defining cyclic monotonicity. The same method of proof works for f^*. □

As already noted, cyclic monotonicity implies monotonicity, so these subdifferentials are also monotone.

8.1.1 Exercises for Section 8.1

Exercise 8.1.7. Let $A \in \mathbb{R}^{n\times n}$ be a nonzero skew matrix (that is, $0 \neq A = -A^*$). Show that the map F taking $x \in \mathbb{R}^n$ to Ax is a monotone operator but is not the subdifferential of any closed proper convex function on $\mathbb{R}^n$.

8.2 Directional derivatives

It turns out that we can develop much information about ∂f by studying the *directional derivative* of f. We define this directional derivative next, and study some of its properties. Then, later in this section, we return to ∂f and use the properties of the directional derivative to find out more about the subdifferential.

Definition 8.2.1. Let f be a convex function from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$, and let x be a point at which f is finite. For each $z \in \mathbb{R}^n$ we define the *directional derivative* $f'(x;z)$ to be

$$f'(x;z) = \lim_{\tau\downarrow 0} \tau^{-1}[f(x+\tau z) - f(x)] = \inf_{\tau>0} \tau^{-1}[f(x+\tau z) - f(x)].$$

The limit exists and is equal to the infimum because, by Lemma 7.2.3, the difference quotient in the definition is nondecreasing in τ. The limit may, however, be $+\infty$ or $-\infty$. A calculation shows that $f'(x;0) = 0$. Also, for any $\alpha > 0$ we have

$$f'(x;\alpha z) = \lim_{\tau\downarrow 0} \alpha(\tau\alpha)^{-1}\{f[x+(\tau\alpha)z] - f(x)\} = \alpha f'(x;z)$$

because $\tau\alpha \downarrow 0$ when $\tau \downarrow 0$, so that for this fixed x the function $f'(x;\cdot)$ is positively homogeneous. In the rest of this section, if x is fixed we will frequently write $g(\cdot)$ instead of $f'(x;\cdot)$.

Shapiro [39] provides an excellent review of directional derivatives for general (not necessarily convex) functions.

Proposition 8.2.2. *Let f be a convex function from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$, and let x be a point at which f is finite. Then $f'(x;\cdot)$ is convex, and for each $h \in \mathbb{R}^n$ $f'(x;h) \geq -f'(x;-h)$. If $x \in \operatorname{ri}\operatorname{dom} f$ then $f'(x;\cdot)$ is closed and proper, and its effective domain is* $\operatorname{par}\operatorname{dom} f$.

Proof. Choose h_1 and h_2 in $\mathbb{R}^n$ and write $g(\cdot)$ for $f'(x;\cdot)$. Let α_1 and α_2 be any real numbers such that $g(h_i) < \alpha_i$ for $i = 1,2$ (if any such exist). The fact that $g(h_i)$ is an infimum shows that for all sufficiently small positive τ, $f(x+\tau h_i) < f(x) + \tau\alpha_i$ for $i = 1,2$. For $\lambda \in (0,1)$ the convexity of f then implies

$$\begin{aligned} f(x+\tau[(1-\lambda)h_1 + \lambda h_2]) &< (1-\lambda)[f(x)+\tau\alpha_1] + \lambda[f(x)+\tau\alpha_2] \\ &= f(x) + \tau[(1-\lambda)\alpha_1 + \lambda\alpha_2], \end{aligned}$$

so that

$$g[(1-\lambda)h_1+\lambda h_2] < (1-\lambda)\alpha_1+\lambda\alpha_2.$$

Therefore $g(\cdot)$ is convex. As $g(0)=0$ the inequality $g(h)\geq -g(-h)$ holds by Corollary 6.1.4.

Now suppose $x\in \operatorname{ri}\operatorname{dom} f$. If $h\in \operatorname{par}\operatorname{dom} f$ then for sufficiently small $\tau>0$ we have $f(x+\tau h)<+\infty$, so that $x+\tau h\in \operatorname{dom} f$; therefore $g(h)<+\infty$, so $\operatorname{par}\operatorname{dom} f\subset \operatorname{dom} g(\cdot)$. On the other hand, if $h\notin \operatorname{par}\operatorname{dom} f$ then for each positive τ we have $x+\tau h\notin \operatorname{dom} f$. Then $g(h)=+\infty$, so in fact $\operatorname{dom} g(\cdot)=\operatorname{par}\operatorname{dom} f$. As $\operatorname{par}\operatorname{dom} f$ is relatively open, $g(\cdot)$ must be closed (Theorem 6.2.14), and as it takes the value zero and not $-\infty$ at the origin, which is in $\operatorname{ri}\operatorname{dom} g$, we find from Proposition 6.2.10 that it is proper. □

The next theorem develops an important connection between the directional derivative and the subdifferential. We have shown that if f is convex and $f(x)$ is finite then the directional derivative $f'(x;\cdot)$ is a positively homogeneous convex function taking the value 0 at the origin. The theorem shows that its closure is the support function of $\partial f(x)$.

Theorem 8.2.3. *Let f be a convex function on $\mathbb{R}^n$ and let x be a point at which f is finite. Then $\partial f(x)=\operatorname{dom} f'(x;\cdot)^*$ and $\operatorname{cl} f'(x;\cdot)$ is the support function of $\partial f(x)$. Further, if $\partial f(x)$ is nonempty then $f'(x;\cdot)$ is proper.*

Proof. Fix $x\in\mathbb{R}^n$ and, for brevity, write $g(\cdot)$ for $f'(x;\cdot)$.

If $\partial f(x)=\emptyset$ then $\partial f(x)\subset \operatorname{dom} g^*$. Otherwise choose $x^*\in\partial f(x)$. Then for each $y\in\mathbb{R}^n$ and each $\tau>0$, $f(x+\tau y)\geq f(x)+\langle x^*,\tau y\rangle$. We can rewrite this as $\tau^{-1}[f(x+\tau y)-f(x)]\geq\langle x^*,y\rangle$, and therefore

$$g(y)\geq\langle x^*,y\rangle. \tag{8.4}$$

Then $\langle x^*,y\rangle-g(y)\leq 0$, so that $g^*(x^*)\leq 0$ and therefore $x^*\in\operatorname{dom} g^*$, showing that $\partial f(x)\subset\operatorname{dom} g^*$. As (8.4) shows that g never takes $-\infty$ and we know that $g(0)=0$, g is proper.

If $\operatorname{dom} g^*=\emptyset$ then $\operatorname{dom} g^*\subset\partial f(x)$. Otherwise, let $x^*\in\operatorname{dom} g^*$. As $g(x^*)<+\infty$ there is some real α such that for each $y\in\mathbb{R}^n$, $\langle x^*,y\rangle-g(y)\leq\alpha$. As then g cannot take $-\infty$ and as $g(0)=0$, g is proper. It is also positively homogeneous, so Lemma 7.1.8 shows that $g^*=I_{\operatorname{dom} g^*}$ and that $\operatorname{dom} g^*$ is closed. As $x^*\in\operatorname{dom} g^*$ we must have $g^*(x^*)=0$, so for each $y\in\mathbb{R}^n$ $0=g^*(x^*)\geq\langle x^*,y\rangle-g(y)$ and therefore

$$f(x+y)-f(x)\geq g(y)\geq\langle x^*,y\rangle.$$

Rearranging this yields $x^*\in\partial f(x)$, so $\operatorname{dom} g^*\subset\partial f(x)$.

We have now shown that $\partial f(x)=\operatorname{dom} g^*$ and that if this set is nonempty then g is proper. To show that $\operatorname{cl} g$ is the support function of $\partial f(x)$ we consider two cases. If $\partial f(x)$ is nonempty, then $g^*=I_{\operatorname{dom} g^*}$ so that

$$\operatorname{cl} g=g^{**}=I^*_{\operatorname{dom} g^*}=I^*_{\partial f(x)}.$$

If $\partial f(x)$ is empty, then $\operatorname{dom} g^*$ is empty so that $g^*\equiv+\infty$ and then

$$\operatorname{cl} g = g^{**} \equiv -\infty \equiv I_{\emptyset}^* = I_{\partial f(x)}^*.$$

□

We can use these results to investigate the connection between local properties of the function and nonemptiness of the subdifferential at a given point.

Corollary 8.2.4. *Let f be a convex function on $\mathbb{R}^n$ and let x be a point at which f is finite. If $\partial f(x)$ is nonempty then f is proper. If $\partial f(x)$ is empty then for each h in* $\operatorname{cone}[(\operatorname{ri}\operatorname{dom} f) - x]$ *one has $f'(x;h) = -\infty$ and $f'(x;-h) = +\infty$.*

Proof. If $f(x)$ is finite and $\partial f(x)$ is nonempty then f cannot take $-\infty$ anywhere, and since we know f is not always $+\infty$ we see that f is proper.

Now suppose $\partial f(x)$ is empty. Then by Theorem 8.2.3, the closure of $f'(x;\cdot)$ is the support function of the empty set, namely the function that takes $-\infty$ everywhere. Therefore $f'(x;h) = -\infty$ for some h, so $f'(x;\cdot)$ is improper and therefore it must take $-\infty$ at each h in the relative interior of its effective domain. Further, for each such h, as $f'(x;h) \geq -f'(x;-h)$ we have $f'(x;-h) = +\infty$. Now the effective domain of $f'(x;\cdot)$ is the set of v such that for some positive τ, $x + \tau v \in \operatorname{dom} f$: that is, $\operatorname{cone}[(\operatorname{dom} f) - x]$. Therefore the relative interior of $\operatorname{dom} f'(x;h)$ is

$$\operatorname{ri}\operatorname{cone}[(\operatorname{dom} f) - x] = \operatorname{cone}\operatorname{ri}[(\operatorname{dom} f) - x] = \operatorname{cone}[(\operatorname{ri}\operatorname{dom} f) - x],$$

where we used Theorem 3.1.7 and Corollary 1.2.8. □

For an example of the situation when $\partial f(x)$ is empty, consider the function $f : \mathbb{R} \to \bar{\mathbb{R}}$ defined by

$$f(x) = \begin{cases} +\infty & \text{if } x \in (-\infty, -1) \\ -(1-x^2)^{1/2} & \text{if } x \in [-1, 1] \\ +\infty & \text{if } x \in (1, +\infty). \end{cases}$$

To visualize the epigraph of f, take the lower half of the unit ball in $\mathbb{R}^2$ and adjoin to it the set $[-1,1] \times \mathbb{R}_+$. This f has a subgradient everywhere on $(-1,1)$ but no subgradients elsewhere, although it takes the finite value 0 at -1 and at 1. As $\partial f(-1)$ is empty, we expect—and get—the behavior predicted by Corollary 8.2.4: $f'(-1;1) = -\infty$ and $f'(-1;-1) = +\infty$.

8.2.1 Exercises for Section 8.2

Exercise 8.2.5. Suppose that f is an extended-real-valued, closed proper convex function on $\mathbb{R}^n$, and that x_0 is a point in the relative interior of $\operatorname{dom} f$. We will call a direction $v \neq 0$ a *descent direction* for f at x_0 if there is some $\mu < 0$ such that for all small positive t we have $f(x_0 + tv) \leq f(x_0) + \mu t \|v\|$.

Assume that someone has proposed the following idea for finding a descent direction: "If x_0^* is any nonzero element of the subdifferential $\partial f(x_0)$, the direction $-x_0^*$ is a descent direction for f at x_0." This idea extends a known property of (Gâteaux or Fréchet) derivatives.

a. Show by counterexample that this proposal is wrong (that is, produce a function f and a point x_0 for which it doesn't work).
b. Prove that if x_0 is not a minimizer of f and if, instead of any element of $\partial f(x_0)$, one chooses the element x_0^* closest to the origin in the Euclidean norm, then $-x_0^*$ is a descent direction for f at x_0.

Exercise 8.2.6. Let g be the Euclidean norm $\|\cdot\|$ on $\mathbb{R}^n$, defined by

$$\|x\| = \sup_{x^* \in B^n} \langle x^*, x \rangle,$$

where B^n is the unit ball of $\mathbb{R}^n$. Show that g is a closed proper convex function. Determine its subdifferential ∂g, and prove that your answer is correct.

8.3 Structure of the subdifferential

We now have enough information to describe the subdifferential $\partial f(x)$ in terms of other quantities associated with f.

Theorem 8.3.1. *Let f be a proper convex function on $\mathbb{R}^n$, and let $x \in \operatorname{dom} \partial f$. Then:*

a. $\operatorname{rc} \partial f(x) = N_{\operatorname{dom} f}(x)$.
b. $\operatorname{lin} \partial f(x) = (\operatorname{par} \operatorname{dom} f)^\perp$.
c. $\partial f(x) = (\operatorname{par} \operatorname{dom} f)^\perp + K$, *where* $K = \partial f(x) \cap (\operatorname{par} \operatorname{dom} f)$.
d. K is compact if and only if $x \in \operatorname{ri} \operatorname{dom} f$.
e. $\partial f(x)$ *is compact if and only if* $x \in \operatorname{int} \operatorname{dom} f$.

Proof. Proposition 8.1.2 says that $\partial f(x)$ is the set of x^* for which $f^*(x^*) + f(x) - \langle x^*, x \rangle = 0$. Defining $g(y)$ to be $f(x+y) - f(x)$, we find by computation that $g^*(x^*) = f^*(x^*) + f(x) - \langle x^*, x \rangle$. Noting that we always have $f^*(x^*) \geq \langle x^*, x \rangle - f(x)$, we can express $\partial f(x)$ as $\operatorname{lev}_0 g^*$. The hypothesis says that f is proper; therefore so is g. It follows that g^* is closed proper convex. Applying Theorem 7.2.9 to g^*, we find that $\operatorname{rc} \partial f(x) = \operatorname{rc} g^*$ and $\operatorname{lin} \partial f(x) = \operatorname{cs} g^*$. Now Corollary 7.2.13, applied to g, tells us that $\operatorname{rc} g^* = (\operatorname{cone} \operatorname{dom} g)^\circ$. The effective domain of g is the set of y for which $x + y \in \operatorname{dom} f$: that is, it is $\operatorname{dom} f - x$. This shows that $\operatorname{rc} g^* = (\operatorname{cone}[\operatorname{dom} f - x])^\circ$. But a set and its closure have the same polar, so $\operatorname{rc} g^*$ is the polar of the set $\operatorname{cl} \operatorname{cone}(\operatorname{dom} f - x)$. The latter set is $T_{\operatorname{dom} f}(x)$ by Proposition 3.3.6, so

$$\operatorname{rc} \partial f(x) = \operatorname{rc} g^* = [T_{\operatorname{dom} f}(x)]^\circ = N_{\operatorname{dom} f}(x),$$

and this is assertion (a).

Assertion (b) now follows because

$$\begin{aligned}\operatorname{lin}\partial f(x) &= \operatorname{rc}\partial f(x)\cap[-\operatorname{rc}\partial f(x)]\\ &= N_{\operatorname{dom} f}(x)\cap[-N_{\operatorname{dom} f}(x)]\\ &= (\operatorname{par}\operatorname{dom} f)^{\perp},\end{aligned}$$

where the last equality comes from part (a) of Theorem 3.3.7. To obtain (c), (d), and (e) we apply Theorem 7.2.11 to g^* and recall that $g^{**} = \operatorname{cl} g$, so that the conditions in Theorem 7.2.11 amount to requirements that the origin lie in the relative interior or interior of $\operatorname{dom}\operatorname{cl} g$. But by Theorem 6.2.14 this set is the same as the relative interior or interior of $\operatorname{dom} g$, to which the origin belongs exactly when x belongs to the relative interior or interior of $\operatorname{dom} f$. □

One might wonder how the subgradients and directional derivatives that we have introduced here relate to ordinary derivatives when the latter exist. For the case of Fréchet derivatives there is a simple answer.

Theorem 8.3.2. *Let f be a convex function on $\mathbb{R}^n$ and let x be a point at which f is finite. Then f is Fréchet differentiable at x if and only if $\partial f(x)$ is a singleton $\{x^*\}$. In that case $x^* = df(x)$.*

Proof. (only if). Suppose that f is F-differentiable at x. Then for each nonzero $h \in \mathbb{R}^n$ we have for small positive τ,

$$f(x+\tau h) = f(x) + df(x)(\tau h) + o(\tau),$$

which implies that $f'(x;h) = df(x)h$. Therefore $f'(x;\cdot)$ is the linear functional $df(x)(\cdot)$; this is closed, so Theorem 8.2.3 tells us that it is the support function of $\partial f(x)$. Taking conjugates, we find that the indicator of $\partial f(x)$ is the indicator of $\{df(x)\}$, which proves the assertion.

(if). Suppose that $\partial f(x) = \{x^*\}$. Applying Theorem 8.2.3 again, we find that the closure of $f'(x;\cdot)$ is $\langle x^*, \cdot\rangle$. As the effective domain of this function is all of $\mathbb{R}^n$ (hence is relatively open) the closure operation is irrelevant (Theorem 6.2.14), so in fact we must have $f'(x;\cdot) = \langle x^*, \cdot\rangle$. This is enough for Gâteaux differentiability of f, but to establish Fréchet differentiability we have to show that the quantity

$$\|v\|^{-1}[f(x+v) - f(x) - \langle x^*, v\rangle]$$

converges to zero as v does (not just for v restricted to be of the form τh for fixed h).

By part (e) of Theorem 8.3.1, x must lie in the interior of $\operatorname{dom} f$. Therefore we can find an n-simplex S containing the origin in its interior and such that $x + S \subset \operatorname{dom} f$. Then $S \supset \delta B$ for some positive δ, where B is the unit ball. Let the vertices of S be h_i for $i = 0, \ldots, n$, and for $h \in S$ and for fixed $\tau \in (0,1]$ define

$$g_\tau(h) = \tau^{-1}[f(x+\tau h) - f(x)] - \langle x^*, h\rangle.$$

This function g_τ is convex, nonnegative (because $x^* \in \partial f(x)$), and finite. Moreover, as each point $h \in S$ is a convex combination of the vertices, the value of $g_\tau(h)$ is

bounded above by the maximum of the values $g_\tau(h_i)$; call this maximum $\mu(\tau)$. Note that for each i $g_\tau(h_i)$ converges to zero as $\tau \to 0+$ because $f'(x;h_i) = \langle x^*, h_i \rangle$. Therefore $\mu(\tau)$ also converges to zero as $\tau \to 0+$.

Now let $0 \neq v \in \delta B$, and let

$$r(v) = \|v\|^{-1}[f(x+v) - f(x) - \langle x^*, v \rangle].$$

Write

$$h = \delta v/\|v\|, \qquad \tau = \delta^{-1}\|v\|,$$

so that $h \in S$ and $v = \tau h$. By substitution in the formula for $r(v)$ we find that

$$r(v) = \delta^{-1} g_\tau(h) \leq \delta^{-1}\mu(\tau).$$

Further, as $\|v\|$ converges to zero so does τ; therefore $\lim_{v\to 0} r(v) = 0$. It follows that f is in fact F-differentiable at x with $df(x) = x^*$. □

We showed in Proposition 8.1.6 that the subdifferential of a closed proper convex function is cyclically monotone. The following theorem shows that the graph of any cyclically monotone multifunction from $\mathbb{R}^n$ to $\mathbb{R}^n$ is contained in the graph of one of these subdifferentials.

Theorem 8.3.3. *Let F be a multifunction from $\mathbb{R}^n$ to $\mathbb{R}^n$. Then F is cyclically monotone if and only if there exists a closed proper convex function f on $\mathbb{R}^n$ with $F \subset \partial f$.*

Proof. (If). If $F \subset \partial f$ for some closed proper convex function F, then F must be cyclically monotone because ∂f is.

(Only if). If $F = \emptyset$ the assertion is true with any choice of f. Suppose $F \neq \emptyset$, choose any $(x_0, y_0) \in F$ and for each $x \in \mathbb{R}^n$ define

$$\begin{aligned} f(x) = \sup\{ &\langle y_m, x - x_m \rangle + \langle y_{m-1}, x_m - x_{m-1} \rangle + \ldots + \langle y_0, x_1 - x_0 \rangle \\ &\mid (x_1, y_1), \ldots, (x_m, y_m) \in F,\ m \text{ finite} \}. \end{aligned}$$

This f is the supremum of a collection of affine functions; therefore it is closed and convex. None of the affine functions can take $-\infty$ anywhere, and as f is the supremum of this collection it never takes $-\infty$. As F is cyclically monotone we always have $f(x_0) \leq 0$, so f is not identically $+\infty$. Accordingly, f is proper.

Now let $(x, y) \in F$. We must show that for each $z \in \mathbb{R}^n$, $f(z) \geq f(x) + \langle y, z - x \rangle$, so that $F \subset \partial f$. Another way of stating this inequality is to say that for each finite subset $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ of F we have

$$f(z) \geq \langle y, z - x \rangle + \langle y_m, x - x_m \rangle + \ldots + \langle y_0, x_1 - x_0 \rangle,$$

but this is true by the definition of f, because the subset

$$(x_1, y_1), \ldots, (x_m, y_m), (x, y)$$

will have been included in the supremum operation. Therefore $F \in \partial f$. □

8.3.1 Exercises for Section 8.3

Exercise 8.3.4. Let f be a proper convex function, and let x be a point of the boundary of $\operatorname{dom} f$ at which f is subdifferentiable. Show that there is a nonzero $v^* \in N_{\operatorname{dom} f}(x)$ such that for each $x^* \in \partial f(x)$ the halfline $x^* + v^*\mathbb{R}_+$ is contained in $\partial f(x)$. Show also that if x is in the relative boundary of $\operatorname{dom} f$ then v^* can be chosen to lie in $\operatorname{par}\operatorname{dom} f$.

8.4 The epsilon-subdifferential

Subgradients are invaluable tools for working with, and determining the properties of, convex functions. However, they may be hard to calculate and, without special attention, they may not provide some of the useful features of derivatives. For example, the function $f(x) = |x|$ on $\mathbb{R}$ has a subdifferential that equals $\{-1\}$ on $(-\infty, 0)$, $[-1, 1]$ at the origin, and $\{+1\}$ on $(0, +\infty)$. Therefore at any nonzero point, even one very close to the origin, the subdifferential consists of a single element, either $+1$ or -1. It does not give any indication that the function's behavior may change violently in a very small neighborhood of that point, and for many purposes—particularly in computation—such a warning can be helpful.

For that reason it is useful to introduce another object, the *ε-subdifferential,* that extends the ordinary subdifferential. If we relax the requirements for a subgradient we can obtain *ε-subgradients* that are sometimes easier to handle and can also be very useful in applications. The ε-subdifferential of a function f at x is then the set of all ε-subgradients of f at x.

8.4.1 Definition and properties

Definition 8.4.1. Suppose $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ and let $\varepsilon \geq 0$. An element x^* of $\mathbb{R}^n$ is an ε-subgradient of f at $x \in \mathbb{R}^n$ if for each $x' \in \mathbb{R}^n$, $f(x') \geq f(x) + \langle x^*, x' - x\rangle - \varepsilon$. The *$\varepsilon$-subdifferential* $\partial_\varepsilon f(x)$ of f at x is the set of all ε-subgradients of f at x. □

We have $\partial_\varepsilon f(x) \subset \partial_\eta f(x)$ when $\varepsilon \leq \eta$, and when ε is zero the definition of ε-subgradient reduces to that of the ordinary subgradient. If $\varepsilon > 0$ then an ε-subgradient gives information about minimization similar to what a subgradient provides: $x \in \mathbb{R}^n$ is a global *epsilon-minimizer* of f (that is, for each $x' \in \mathbb{R}^n$ one has $f(x') \geq f(x) - \varepsilon$) if and only if the origin belongs to $\partial_\varepsilon f(x)$. Here are some properties of $\partial_\varepsilon f$ analogous to those shown for ∂f in Theorem 8.3.1.

Theorem 8.4.2. *Let f be a proper convex function on $\mathbb{R}^n$, and let x be a point at which f is finite. If $\varepsilon > f(x) - \operatorname{cl} f(x)$, then*

a. $\partial_\varepsilon(x)$ is a nonempty closed convex set.

b. $\operatorname{rc} \partial_\varepsilon f(x) = N_{\operatorname{dom} f}(x)$.
c. $\operatorname{lin} \partial_\varepsilon f(x) = (\operatorname{par} \operatorname{dom} f)^\perp$.
d. $\partial_\varepsilon f(x) = (\operatorname{par} \operatorname{dom} f)^\perp + K$, *where* $K = \partial_\varepsilon f(x) \cap (\operatorname{par} \operatorname{dom} f)$.
e. K *is compact if and only if* $x \in \operatorname{ri} \operatorname{dom} f$.
f. $\partial_\varepsilon f(x)$ *is compact if and only if* $x \in \operatorname{int} \operatorname{dom} f$.

Proof. If we define $g(y) = f(x+y) - f(x)$ as we did in the proof of Theorem 8.3.1, then $g^*(x^*) = f^*(x^*) + f(x) - \langle x^*, x\rangle$ and $\partial_\varepsilon f(x)$ is the set of x^* such that, for each z,

$$[\langle x^*, z\rangle - f(z)] + f(x) - \langle x^*, x\rangle \le \varepsilon.$$

By taking the supremum over z we see that this is the set of x^* for which $g^*(x^*) \le \varepsilon$: that is, $\operatorname{lev}_\varepsilon g^*$. It is therefore a closed convex set. By hypothesis we have

$$f(x) = [f(x) - \operatorname{cl} f(x)] + f^{**}(x) < \varepsilon + \sup_{x^*}\{\langle x^*, x\rangle - f^*(x^*)\},$$

so that for some x^* we have $f(x) < \langle x^*, x\rangle - f^*(x^*) + \varepsilon$, and therefore $\partial_\varepsilon f(x)$ is nonempty, which proves the claim in (a). The proof of (b)–(f) is just the same as the proof for $\partial f(x)$ in Theorem 8.3.1, as we used there only the fact that we had a nonempty level set of g^*. □

The level-set representation also helps us to identify $\partial_\varepsilon f^*$.

Theorem 8.4.3. *Let* $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ *be a closed proper convex function, and let* $\varepsilon \ge 0$. *Then* $\partial_\varepsilon f^* = (\partial_\varepsilon f)^{-1}$.

Proof. Choose any x^* and x in $\mathbb{R}^n$. We have to show that $(x^*, x) \in \partial_\varepsilon f^*$ if and only if $(x, x^*) \in \partial_\varepsilon f(x)$.

First suppose that $f^*(x^*) = +\infty$. Then as f^* is proper $\partial_\varepsilon f^*(x^*)$ must be empty, so $(x^*, x) \notin \partial_\varepsilon f^*$. If x^* were to belong to $\partial_\varepsilon f(x)$ then as f is proper $f(x)$ would be finite, and for each $x' \in \mathbb{R}^n$ we would have $f(x') \ge f(x) + \langle x^*, x' - x\rangle - \varepsilon$. Rearranging this yields

$$\langle x^*, x\rangle - f(x) + \varepsilon \ge \langle x^*, x'\rangle - f(x'),$$

so that the finite quantity on the left must be an upper bound for $f^*(x^*)$, contradicting $f^*(x^*) = +\infty$. Therefore $(x, x^*) \notin \partial_\varepsilon f(x)$.

The remaining case is that in which $f^*(x^*)$ is finite. Applying the function g used in the proof of Theorem 8.4.2 to f^* instead of to f, we see that $(x^*, x) \in \partial_\varepsilon f^*$ if and only if

$$f^{**}(x) + f^*(x^*) - \langle x^*, x\rangle \le \varepsilon.$$

As $f^{**} = f$ under our hypotheses, this is equivalent to $g^*(x^*) \le \varepsilon$ and therefore also equivalent to

$$x^* \in \operatorname{lev}_\varepsilon g = \partial_\varepsilon f(x),$$

which is the same as saying $(x, x^*) \in \partial_\varepsilon f$. □

As indicated earlier, one of the reasons for using the ε-subdifferential in practice is that it can give information about the behavior of the function at nearby points. The next theorem explains this.

Theorem 8.4.4. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a proper convex function and let $x \in \operatorname{ri}\operatorname{dom} f$. For each positive ε there is a neighborhood U of x in $\operatorname{dom} f$ such that*

$$\partial_\varepsilon f(x) \supset \cup_{w\in U} \partial f(w). \tag{8.5}$$

Proof. Choose $\varepsilon > 0$ and write L for $\operatorname{par}\operatorname{dom} f$. Theorem 6.1.13 says that there exist a constant λ and a neighborhood N of x in $\operatorname{dom} f$ such that whenever w and w' belong to N, $|f(w) - f(w')| \le \lambda \|w - w'\|$. Let $B_L(x,\rho)$ be a ball $\{x + h \mid h \in L,\ \|h\| \le \rho\}$ with center x and positive radius $\rho \le \varepsilon/(2\lambda)$, contained in $\operatorname{int} N$. Let $U = \operatorname{ri} B_L(x,\rho)$, a neighborhood of x in $\operatorname{dom} f$, choose $w \in U$, and suppose $w^* \in \partial f(w)$. We will show that $w^* \in \partial_\varepsilon f(x)$, and that will establish (8.5).

The first step is to estimate the size of the component of w^* in L. Write $w^* = u^* + v^*$, with $u^* \in L$ and $v^* \in L^\perp$. Choose $\delta \in (0, \rho - \|w\|)$ and consider the ball $D_L := \{h \in L \mid \|h\| \le \delta\}$ about the origin. If $h \in D_L$ then $w + h \in \operatorname{int} B(x,\rho)$, and then by using the Lipschitz continuity together with the facts that $w^* \in \partial_\varepsilon f(x)$ and $\langle v^*, h\rangle = 0$ we find that

$$\lambda \|h\| \ge f(w+h) - f(w) \ge \langle w^*, h\rangle = \langle u^*, h\rangle,$$

so that

$$\lambda \ge \langle u^*, h/\|h\|\rangle. \tag{8.6}$$

Taking the supremum in (8.6) over all nonzero $h \in D_L$, we have $\|u^*\| \le \lambda$.

Next, use the Lipschitz continuity again to obtain $f(x) - f(w) \le \lambda \|x - w\| \le \varepsilon/2$, which we can rewrite as

$$f(w) \ge f(x) - \varepsilon/2. \tag{8.7}$$

As $w - x \in L$ we also have

$$\langle w^*, w - x\rangle = \langle u^*, w - x\rangle \le \|u^*\|\|w - x\| \le \lambda[\varepsilon/(2\lambda)] = \varepsilon/2,$$

and we can rewrite this as

$$0 \ge \langle w^*, w - x\rangle - \varepsilon/2. \tag{8.8}$$

Adding (8.7) and (8.8), we obtain

$$f(w) \ge f(x) + \langle w^*, w - x\rangle - \varepsilon,$$

which shows that $w^* \in \partial_\varepsilon f(x)$. □

The support function of $\partial_\varepsilon f(x)$ turns out to be an object analogous to the ordinary directional derivative, which when closed was the support function of $\partial f(x)$.

Definition 8.4.5. Let f be a convex function on $\mathbb{R}^n$, and let x be a point at which f is finite. Let $\varepsilon \ge 0$. For any $h \in \mathbb{R}^n$ the *ε-directional derivative* of f at x in the direction of h is

$$f'_\varepsilon(x;h) = \inf_{\tau>0} \tau^{-1}[f(x + \tau h) - f(x) + \varepsilon].$$

□

For $\varepsilon = 0$ this reduces to our previous definition. However, when $\varepsilon > 0$ the equivalent limit form used in Definition 8.2.1 no longer applies.

Theorem 8.4.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a closed proper convex function that is finite at $x \in \mathbb{R}^n$, and let $\varepsilon > 0$. Then $f'_\varepsilon(x;\cdot)$ is the support function of $\partial_\varepsilon f(x)$.*

Proof. Let $g(y) = f(x+y) - f(x)$; we showed in the proof of Theorem 8.4.2 that $\partial_\varepsilon f(x) = \operatorname{lev}_\varepsilon g^*$. The support function of this set is, by Theorem 7.1.12, $\operatorname{cl}\operatorname{cone}(g+\varepsilon)$, since the fact that f is closed means that g is also closed. However, as $g(0) = 0$ we see that $g+\varepsilon$ is positive at the origin, so we can use Theorem 7.2.15 to remove the closure symbol. Then the support function of $\partial_\varepsilon f(x)$ takes, at any $h \in \mathbb{R}^n$, the value

$$\operatorname{cone}(g+\varepsilon)(h) = \inf_{\tau>0} \tau^{-1}[f(x+\tau h) - f(x) + \varepsilon] = f'_\varepsilon(x;h),$$

as required. □

For a positive ε, a convex function f on $\mathbb{R}^n$, and a point $x \in \mathbb{R}^n$ we always have $\partial f(x) \subset \partial_\varepsilon f(x)$. However, the second set may be much larger than the first: for example, if $g(y) = f(x+y) - f(x)$ and if the function g^* is rather flat, then the sets $\partial_\varepsilon f(x) = \operatorname{lev}_\varepsilon g^*$ can be expected to grow rapidly as ε increases. However, if we are willing to consider distance in the graph space $\mathbb{R}^n \times \mathbb{R}^n$ then we can bound the distance in terms of ε.

Theorem 8.4.7. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a closed proper convex function and let $\varepsilon \geq 0$. Suppose that $(x,x^*) \in \partial_\varepsilon f$. Then for each positive β there is a unique $y \in \mathbb{R}^n$ such that*

$$\|y\| \leq \varepsilon^{1/2}, \qquad (x+\beta y, x^* - \beta^{-1} y) \in \partial f. \tag{8.9}$$

In particular, if we use the Euclidean norm on $\mathbb{R}^n \times \mathbb{R}^n$ then we have

$$e(\partial_\varepsilon f, \partial f) \leq (2\varepsilon)^{1/2}. \tag{8.10}$$

Proof. Choose $\varepsilon \geq 0$ and $\beta > 0$, and define $h(z) = f(x+\beta z) + (1/2)\|z - \beta x^*\|^2$. This function h is closed, proper, and convex. Moreover, it has compact level sets: to see this, use Proposition 6.2.13 to find an affine function minorizing f; then the sum of this affine function and the quadratic term in h has compact level sets and minorizes h. The level sets of h must then be compact too, and so since h is lower semicontinuous it has a minimizer y by Proposition 6.2.8. The minimizer is in fact unique because of the presence of the quadratic term in h.

As y minimizes h we have $0 \in \partial h(y)$ (use the definition). We are going to show that

$$0 \in \beta \partial f(x+\beta y) + y - \beta x^*; \tag{8.11}$$

this would be immediate from some calculus rules for subdifferentials developed in the next chapter, but in this case we can give a direct proof. It is easy to show from the definition that the subdifferential of $f(x+\beta y)$ as a function of y is $\beta \partial f(x+\beta y)$.

Now as y minimizes h we have $h'(y;z) \geq 0$ for each z. We can calculate directly that this directional derivative is

$$h'(y;z) = \beta f'(x+\beta y;z) + \langle y - \beta x^*, z\rangle.$$

Therefore the closed proper convex function $-\langle y-\beta x^*, \cdot\,\rangle$ minorizes $\beta f'(x+\beta y;\,\cdot\,)$ and therefore also minorizes $\mathrm{cl}\,\beta f'(x+\beta y;\,\cdot\,)$. This closure is the support function of the set $\partial k(y)$, where $k(y) = f(x+\beta y)$. We saw above that $\partial k(y) = \beta \partial f(x+\beta y)$. By part (b) of Proposition 7.1.10 we then have

$$\beta \partial f(x+\beta y) \supset \{\beta x^* - y\},$$

which is the claim in (8.11).

From (8.11) we see that $(x+\beta y, x^* - \beta^{-1} y) \in \partial f$. Therefore

$$f(x) \geq f(x+\beta y) + \langle x^* - \beta^{-1} y, x - (x+\beta y)\rangle.$$

However, we also have by hypothesis

$$f(x+\beta y) \geq f(x) + \langle x^*, (x+\beta y) - x\rangle - \varepsilon,$$

and by adding these two inequalities we find that $\|y\|^2 \leq \varepsilon$, so that $\|y\| \leq \varepsilon^{1/2}$, and this proves (8.9). If we take $\beta = 1$ then the second assertion in (8.9) shows that

$$e(\partial_\varepsilon f, \partial f) \leq (\|\beta y\|^2 + \|\beta^{-1} y\|^2)^{1/2} \leq (2\varepsilon)^{1/2},$$

which establishes (8.9). □

Chapter 9
Functional operations

Up to now we have developed numerous properties of individual convex functions, but with few exceptions we have not done anything about combining different functions. This chapter shows how to use functional calculus to create new convex functions from existing ones.

9.1 Convex functions and linear transformations

Definition 9.1.1. Let g and h be convex functions from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$ and let G and H be linear transformations, G from $\mathbb{R}^m$ to $\mathbb{R}^n$ and H from $\mathbb{R}^n$ to $\mathbb{R}^m$. The functions gG and Hh are defined from $\mathbb{R}^m$ to $\bar{\mathbb{R}}$ by

$$(gG)(y) = g(Gy), \qquad (Hh)(y) = \inf\{h(x) \mid Hx = y\}.$$

A standard calculation using Proposition 6.1.3 shows that gG and Hh are convex functions. If g does not take $-\infty$ then neither does gG, so gG is closed if g is closed and proper. However, gG need not be proper: for example, it might happen that the effective domain of g did not meet the image of G, in which case gG would be identically $+\infty$. The function Hh, however, need be neither closed nor proper, even when h is both closed and proper. For example, if h is the indicator of the set of x in $\mathbb{R}^2$ having $x_1 > 0$ and $x_1 x_2 \geq 1$, and if $H(x_1, x_2) = x_1$, then Hh is the indicator of $(0, +\infty)$, which is not closed. Again, if h is defined on $\mathbb{R}^2$ by $h(x) = x_2$ and if $H(x_1, x_2) = x_1$, then Hh is identically $-\infty$, which is improper.

Even with no additional assumptions, we can establish a simple relationship between these two operations.

Proposition 9.1.2. *Let f be a convex function on $\mathbb{R}^n$ and let A be a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then $(Af)^* = f^* A^*$.*

Proof. Let $y^* \in \mathbb{R}^m$. Then

$$\begin{aligned}(Af)^*(y^*) &= \sup_y\{\langle y^*, y\rangle - Af(y)\}\\ &= \sup_y\{\langle y^*, y\rangle - \inf_x\{f(x) \mid Ax = y\}\}\\ &= \sup_{x,y}\{\langle y^*, y\rangle - f(x) \mid Ax = y\}\\ &= \sup_x\{\langle y^*, Ax\rangle - f(x)\}\\ &= \sup_x\{\langle A^*y^*, x\rangle - f(x)\}\\ &= (f^*A^*)(y^*),\end{aligned}$$

so that $(Af)^* = f^*A^*$. □

Proposition 9.1.2 explains the reason for using the notation Hh for the infimum operation shown in Definition 9.1.1. It also shows in particular that fA is closed whenever f is closed, because we can then write fA as $(A^*f^*)^*$. Still, it does not do anything to improve the possible bad behavior that we pointed out earlier. For that we have to assume a regularity condition.

To motivate this condition, remember that fA is identically $+\infty$ (hence improper) whenever the image of A fails to meet $\operatorname{dom} f$. Therefore, in order to show good behavior we certainly have to require that these sets meet. A slight strengthening of this requirement yields a regularity condition that in fact ensures much more. We first use it to establish a closure formula, then to prove a much more comprehensive result

Theorem 9.1.3. *Let f be a proper convex function on $\mathbb{R}^n$ and let A be a linear transformation from $\mathbb{R}^m$ to $\mathbb{R}^n$. If* $\operatorname{im} A$ *meets* $\operatorname{ri}\operatorname{dom} f$*, then fA is a proper convex function, and one has* $\operatorname{cl}(fA) = (\operatorname{cl} f)A$ *and* $\operatorname{rec}\operatorname{cl}(fA) = (\operatorname{rec}\operatorname{cl} f)A$.

Proof. We have already seen that fA is convex. It never takes $-\infty$ because f is proper, and the hypothesis ensures that it is not always $+\infty$, so fA is proper. Define a linear transformation D from $\mathbb{R}^{m+1}$ to $\mathbb{R}^{n+1}$ by $D(y,\eta) = (Ay,\eta)$. A point (y,η) belongs to $\operatorname{epi} fA$ exactly when the point $(Ay,\eta) = D(y,\eta)$ belongs to $\operatorname{epi} f$. Therefore $\operatorname{epi} fA = D^{-1}(\operatorname{epi} f)$.

The effective domain of fA consists of those y for which $Ay \in \operatorname{dom} f$: that is, $A^{-1}(\operatorname{dom} f)$. Under our hypothesis we have $\operatorname{ri}\operatorname{dom} fA = A^{-1}(\operatorname{ri}\operatorname{dom} f)$ by Proposition 1.2.11, and this set is nonempty. Let y' be any point in $\operatorname{ri}\operatorname{dom} fA$, so that $Ay' \in \operatorname{ri}\operatorname{dom} f$, and let y be any point of $\mathbb{R}^m$. By Corollary 6.2.16 we have

$$\begin{aligned}\operatorname{cl}(fA)(y) &= \lim_{\lambda\downarrow 0}(fA)[(1-\lambda)y + \lambda y']\\ &= \lim_{\lambda\downarrow 0} f[(1-\lambda)Ay + \lambda Ay']\\ &= (\operatorname{cl} f)(Ay)\\ &= [(\operatorname{cl} f)A](y),\end{aligned}$$

which proves the assertion about the closure. Applying our earlier analysis to $\operatorname{cl} f$ instead of to f we find that $\operatorname{epi}[(\operatorname{cl} f)A] = D^{-1}(\operatorname{epi}\operatorname{cl} f)$, so by Corollary 4.1.4 we

have

$$\begin{aligned}\operatorname{epi}\operatorname{rec}[(\operatorname{cl} f)A] &= \operatorname{rc}\operatorname{epi}[(\operatorname{cl} f)A]\\ &= \operatorname{rc}[D^{-1}(\operatorname{epi}\operatorname{cl} f)]\\ &= D^{-1}(\operatorname{rc}\operatorname{epi}\operatorname{cl} f)\\ &= D^{-1}\operatorname{epi}[\operatorname{rec}(\operatorname{cl} f)]\\ &= \operatorname{epi}\{[\operatorname{rec}(\operatorname{cl} f)]A\}.\end{aligned}$$

Accordingly, $\operatorname{rec}[(\operatorname{cl} f)A] = (\operatorname{rec}\operatorname{cl} f)A$. □

Theorem 9.1.4. *Let f be a proper convex function on $\mathbb{R}^n$, and let A be a linear transformation from $\mathbb{R}^m$ to $\mathbb{R}^n$. Assume that* $\operatorname{im} A$ *meets* $\operatorname{ri}\operatorname{dom} f$. *Then:*

a. fA and A^*f^* are proper convex functions and $(fA)^* = A^*f^*$; in particular, A^*f^* is closed;
b. $\operatorname{rec} A^*f^* = A^*(\operatorname{rec} f^*)$, *and this function is closed proper convex;*
c. For each $y^* \in \mathbb{R}^m$ the infimum in the definition of $A^*f^*(y^*)$ is attained, but may be $+\infty$.
d. For each $y \in \mathbb{R}^m$,

$$\partial(fA)(y) = A^*\partial f(Ay); \tag{9.1}$$

e. For each y^* in $\mathbb{R}^m$,

$$\partial(A^*f^*)(y^*) = \{y \mid \text{for some } (x^*,x) \in \partial f^*,\ A^*x^* = y^* \text{ and } Ay = x\}. \tag{9.2}$$

Proof. Theorem 9.1.3 says that fA is a proper convex function whose closure is $(\operatorname{cl} f)A$, which is therefore also a proper function. We have already seen that A^*f^* is convex, and Proposition 9.1.2 says that its conjugate is $f^{**}A^{**} = (\operatorname{cl} f)A$. As the conjugate is proper, A^*f^* is proper too. Suppose now that we could prove that A^*f^* is closed. Then we would have shown that

$$\begin{aligned}(fA)^* = (fA)^{***} = [(fA)^{**}]^* = [\operatorname{cl}(fA)]^* = [(\operatorname{cl} f)A]^*\\ = [(A^*f^*)^*]^* = \operatorname{cl}(A^*f^*) = A^*f^*,\end{aligned}$$

where the last equality would follow from the proof that A^*f^* is closed. This would be enough to prove (a).

We know A^*f^* is proper, so to show it is closed we need only show that $\operatorname{epi} A^*f^*$ is closed. Define a linear transformation E^* from $\mathbb{R}^{n+1}$ to $\mathbb{R}^{m+1}$ by $E^*(x^*,\xi) = (A^*x^*,\xi)$.

If $(y^*,\xi^*) \in E^*(\operatorname{epi} f^*)$ then there is some $(x^*,\xi^*) \in \operatorname{epi} f^*$ such that $y^* = A^*x^*$. Then as $\xi^* \geq f(x^*)$ we have $\xi^* \geq (A^*f^*)(y^*)$ and so $(y^*,\xi^*) \in \operatorname{epi} A^*f^*$.

Also, if $(z^*,\zeta^*) \in \operatorname{epi} A^*f^*$ then for each $\varepsilon > 0$ there is some v^* with $A^*v^* = z^*$ and $f^*(v^*) \leq \zeta + \varepsilon$. Then $(v^*,\zeta+\varepsilon) \in \operatorname{epi} f^*$ so $(z^*,\zeta) + (0,\varepsilon) \in E^*(\operatorname{epi} f^*)$. As ε was arbitrary we have

$$E^*(\operatorname{epi} f^*) \subset \operatorname{epi} A^*f^* \subset \operatorname{cl} E^*(\operatorname{epi} f^*). \tag{9.3}$$

We will prove that $E^*(\operatorname{epi} f^*)$ closed, which with (9.3) will show that

$$\operatorname{epi}(A^* f^*) = E^*(\operatorname{epi} f^*). \tag{9.4}$$

We can use Corollary 4.1.17 for that proof if we can show that $\operatorname{epi} f^*$ satisfies the recession condition for E^*. The first requirement of that condition is that $\operatorname{epi} f^*$ be closed, which we already have. The second is that $(\ker E^*) \cap (\operatorname{rc} \operatorname{epi} f^*)$ be a subspace. To show this, suppose $z^* \in (\ker E^*) \cap (\operatorname{rc} \operatorname{epi} f^*)$. Then $z^* = (k^*, 0)$ with $k^* \in \ker A$. Also

$$(k^*, 0) \in \operatorname{rc} \operatorname{epi} f^* = \operatorname{epi} \operatorname{rec} f^*,$$

so that $(\operatorname{rec} f^*)(k^*) \leq 0$. Therefore $k^* \in \operatorname{rc} f^*$, so that $k^* \in (\ker A^*) \cap (\operatorname{rc} f^*)$. The hypothesis says that $(\operatorname{im} A) \cap (\operatorname{ri} \operatorname{dom} f)$ is nonempty, which by Corollary 7.2.14 is equivalent to saying that $(\ker A^*) \cap (\operatorname{rc} f^*)$ is a subspace. The latter statement implies that $-k^* \in (\ker A^*) \cap (\operatorname{rc} f^*)$, which in particular lets us conclude that $\operatorname{rec} f^*(-k^*) \leq 0$, or equivalently

$$-z^* = (-k^*, 0) \in \operatorname{epi} \operatorname{rec} f^* = \operatorname{rc} \operatorname{epi} f^*.$$

We certainly have $-z^* \in \ker E^*$, so $-z^*$ belongs to $(\ker E^*) \cap (\operatorname{rc} \operatorname{epi} f^*)$, which must therefore be a subspace. Therefore $\operatorname{epi} f^*$ satisfies the recession condition for E^*, so Corollary 4.1.17 shows that $E^*(\operatorname{epi} f^*)$ is closed, and this proves (9.4) as well as showing that $\operatorname{epi} A^* f^*$ is closed. The latter tells us that $A^* f^*$ is lower semicontinuous; however, we already know it is proper, so it is in fact closed. This proves (a).

For (b), we begin by invoking Theorem 7.2.10 to show that, as f^* is proper,

$$\operatorname{rec} f^* = I^*_{\operatorname{dom} f} = I^*_{\operatorname{cl} \operatorname{dom} f},$$

and therefore $(\operatorname{rec} f^*)^* = I_{\operatorname{cl} \operatorname{dom} f}$. Then by Proposition 9.1.2, $[A^*(\operatorname{rec} f^*)]^* = (I_{\operatorname{cl} \operatorname{dom} f})A$. The hypothesis says that $\operatorname{im} A$ meets the set

$$\operatorname{ri} \operatorname{dom} f = \operatorname{ri}(\operatorname{cl} \operatorname{dom} f) = \operatorname{ri} \operatorname{dom}(I_{\operatorname{cl} \operatorname{dom} f}).$$

Theorem 9.1.3 then shows that $(I_{\operatorname{cl} \operatorname{dom} f})A$ is proper, and by Proposition 7.1.2 so then is $A^*(\operatorname{rec} f^*)$.

Applying the method of proof used for part (a) above to $\operatorname{rec} f^*$ instead of to f^*, we obtain

$$E^*(\operatorname{epi} \operatorname{rec} f^*) \subset \operatorname{epi} A^*(\operatorname{rec} f^*) \subset \operatorname{cl} E^*(\operatorname{epi} \operatorname{rec} f^*). \tag{9.5}$$

We also established there the hypotheses of Corollary 4.1.18, which says that $\operatorname{rc} E^*(\operatorname{epi} f^*)$ is closed and that

$$E^*(\operatorname{rc} \operatorname{epi} f^*) = \operatorname{rc} E^*(\operatorname{epi} f^*). \tag{9.6}$$

The left side of (9.6) is $E^*(\operatorname{epi} \operatorname{rec} f^*)$, which is then closed; therefore from (9.5) we obtain

$$\operatorname{epi} A^*(\operatorname{rec} f^*) = E^*(\operatorname{epi} \operatorname{rec} f^*). \tag{9.7}$$

This shows that $A^*(\operatorname{rec} f^*)$ is lower semicontinuous, but as it is proper it is also closed. Now by using successively (9.7), (9.6), and (9.4), we obtain

$$\begin{aligned} \operatorname{epi} A^*(\operatorname{rec} f^*) &= E^*(\operatorname{epi} \operatorname{rec} f^*) \\ &= E^*(\operatorname{rc} \operatorname{epi} f^*) \\ &= \operatorname{rc} E^*(\operatorname{epi} f^*) \\ &= \operatorname{rc} \operatorname{epi} A^* f^* \\ &= \operatorname{epi} \operatorname{rec}(A^* f^*). \end{aligned} \tag{9.8}$$

As we already know that $A^*(\operatorname{rec} f^*)$ is a closed proper convex function, this completes the proof of (b).

For (c), we find from (9.4) that a point of the form $(y^*, (A^* f^*)(y^*))$ in $\operatorname{epi} A^* f^*$ must be the image under E^* of some point $(x^*, \xi^*) \in \operatorname{epi} f^*$. The finite value ξ^* cannot be greater than $f^*(x^*)$ because of the definition of $A^* f^*$, so it must be equal to $f^*(x^*)$. The other possibility is that $(A^* f^*)(y^*) = +\infty$, in which case there are no points $(x^*, \xi^*) \in \operatorname{epi} f$ with $A^* x^* = y^*$. Then the set over which the infimum is taken is empty, and the infimum is $+\infty$ by convention.

To prove (d) let $y \in \mathbb{R}^m$ and $x = Ay$, and suppose that $(x, x^*) \in \partial f$. Let $y^* = A^* x^*$ and suppose that z is any point of $\mathbb{R}^m$. Then

$$fA(z) = f(Az) \geq f(Ay) + \langle x^*, Az - Ay \rangle = (fA)(y) + \langle y^*, z - y \rangle,$$

so that $y^* \in \partial(fA)(y)$. It follows that $A^* \partial f(Ay) \subset \partial(fA)(y)$. We did not need the regularity condition for this conclusion.

When the regularity condition does hold, we know from the part of the proof already completed that the conjugate of fA is $A^* f^*$. Suppose that $y^* \in \partial(fA)(y)$; as y belongs to $\operatorname{dom} \partial(fA)(\,\cdot\,) \subset \operatorname{dom}(fA)$, $(fA(y)$ must be finite. We then have $(A^* f^*)(y^*) = \langle y^*, y \rangle - f(Ay)$, so $(A^* f^*)(y^*)$ is finite too. From (c) we have the existence of some $x^* \in \mathbb{R}^n$ with $A^* x^* = y^*$ and $f^*(x^*) = A^* f^*(y^*)$. Then $f^*(x^*) = \langle A^* x^*, y \rangle - f(x)$, and if we write $x = Ay$ we see that $(x, x^*) \in \partial f$. Therefore $y^* \in A^* \partial f(Ay)$, so $\partial(fA)(y) \subset A^* \partial f(Ay)$. This proves (d).

For (e) we observe that the regularity condition in the hypothesis is satisfied for f if and only if it is satisfied for $\operatorname{cl} f$. Apply (9.1) to the function $\operatorname{cl} f$ to obtain

$$(y, y^*) \in \partial(\operatorname{cl} f) A \text{ if and only if } (x, x^*) \in \partial(\operatorname{cl} f),\ A^* x^* = y^*,\ x = Ay. \tag{9.9}$$

Now use Proposition 8.1.6 to rewrite (9.9) as

$$(y^*, y) \in \partial(A^* f^*) \text{ if and only if } (x^*, x) \in \partial f^*,\ A^* x^* = y^*,\ x = Ay. \tag{9.10}$$

The content of (9.10) is that $\partial(A^* f^*)(y^*)$ consists of those y for which there is some $(x^*, x) \in \partial f^*$ such that $A^* x^* = y^*$ and $x = Ay$. That is (9.2), so it completes the proof of (e). □

Theorem 9.1.4 has many applications, but one of the most useful is to sums of convex functions. We demonstrate this application after defining an additional functional operation that it requires.

Definition 9.1.5. Let $f_1, \dots, f_k$ be convex functions from $\mathbb{R}^n$ to $(-\infty, +\infty]$. The *infimal convolution* of these functions is the function defined by

$$\biguplus_{i=1}^{k} f_i(x) = \inf \left\{ \sum_{i=1}^{k} f_i(x_i) \mid \sum_{i=1}^{k} x_i = x \right\}.$$

A computation using Proposition 6.1.3 shows that this new function is convex. However, it might not be well behaved: for example, the infimum might not be attained. Moreover, even the function $\sum_{i=1}^{k} f_i$ can behave badly: it will be identically $+\infty$ if there is no point common to all of the sets $\operatorname{dom} f_i$. However, under a simple regularity condition these functions are well behaved and, in fact, closely related to each other.

Theorem 9.1.6. *Let $f_1, \dots, f_k$ be proper convex functions on $\mathbb{R}^n$. Assume that*

$$\cap_{i=1}^{k} \operatorname{ri} \operatorname{dom} f_i \neq \emptyset. \tag{9.11}$$

Then

a. $\sum_{i=1}^{k} f_i$ is a proper convex function whose closure is $\sum_{i=1}^{k} \operatorname{cl} f_i$ and whose conjugate is $\biguplus_{i=1}^{k} f_i^*$;
b. $\biguplus_{i=1}^{k} f_i^*$ is closed, proper, and convex, and the infimum in its definition is always attained;
c. The recession function of $\sum_{i=1}^{k} \operatorname{cl} f_i$ is $\sum_{i=1}^{k} \operatorname{rec} \operatorname{cl} f_i$;
d. $\partial(\sum_{i=1}^{k} f_i) = \sum_{i=1}^{k} \partial f_i$;
e. For each $x^* \in \mathbb{R}^n$, $\partial(\biguplus_{i=1}^{k} f_i^*)(x^*)$ is the set of $x \in \mathbb{R}^n$ such that there exist $x_1^*, \dots, x_k^*$ with $(x_i^*, x) \in \partial f_i^*$ for each i and with $\sum_{i=1}^{k} x_i^* = x^*$.

Proof. Define a function F on $\mathbb{R}^{nk}$ by $F(x_1, \dots, x_k) = \sum_{i=1}^{k} f_i(x_i)$. Then F is convex (Proposition 6.1.3) and proper, and $(\operatorname{cl} F)(x_1, \dots, x_n) = \sum_{i=1}^{k} \operatorname{cl} f_i(x_i)$ (Corollary 6.2.16). Define a linear transformation A from $\mathbb{R}^n$ to $\mathbb{R}^{nk}$ by $Ax = (x, \dots, x)$. Then $\sum_{i=1}^{k} f_i = FA$, and the condition (9.11) is the statement that $\operatorname{im} A$ meets $\operatorname{ri} \operatorname{dom} F$. By Theorem 9.1.3 we then have $\operatorname{cl}(FA) = (\operatorname{cl} F)A$: that is, $\operatorname{cl} \sum_{i=1}^{k} f_i = \sum_{i=1}^{k} \operatorname{cl} f_i$. The assertion about the recession function also follows from Theorem 9.1.3.

Routine computation shows that

$$F^*(x_1^*, \dots, x_n^*) = \sum_{i=1}^{k} f_i^*(x_i^*), \qquad A^*(x_1^*, \dots, x_k^*) = \sum_{i=1}^{k} x_i^*, \qquad A^* F^* = \biguplus_{i=1}^{k} f_i^*,$$

and that

$$\partial F(x_1 \dots, x_n) = \times_{i=1}^{k} \partial f_i(x_i).$$

Applying Theorem 9.1.4 we find that the conjugate of FA is $\biguplus_{i=1}^{k} f_i^*$, that the latter function is closed proper convex and the infimum in its definition is always attained,

and that the subdifferential formulas asserted in the statement of the theorem are valid. □

In the proof of Theorem 9.1.4 we were able to obtain the formula $\partial(fA)(y) \supset A^*\partial f(Ay)$ without any regularity assumptions. Therefore in the situation of Theorem 9.1.6 we always have

$$\partial(\sum_{i=1}^{k} f_i)(x) \supset \sum_{i=1}^{k} \partial f_i(x),$$

but we need the intersection condition (9.11) to conclude that equality holds.

With the results of Theorems 9.1.4 and 9.1.6 one can calculate the subdifferentials of many composite functions of practical interest for optimization. A typical result is the following characterization of optimality in the problem of minimizing a convex function on a convex set.

Proposition 9.1.7. *Let f be a proper convex function on $\mathbb{R}^n$ and let C be a convex subset of $\mathbb{R}^n$ containing a point x_0. For x_0 to minimize f on C it suffices that*

$$0 \in \partial f(x_0) + N_C(x_0), \tag{9.12}$$

and if $\operatorname{ri} C$ *meets* $\operatorname{ri}\operatorname{dom} f$ *then* (9.12) *is also necessary.*

Proof. If (9.12) holds then there is some $x^* \in \partial f(x_0)$ with $-x^* \in N_C(x_0)$. Suppose $x \in C$. Then

$$f(x) \geq f(x_0) + \langle x^*, x - x_0 \rangle.$$

However, $\langle x^*, x - x_0 \rangle \geq 0$ because $-x^* \in N_C(x_0)$. Therefore $f(x) \geq f(x_0)$, so x_0 minimizes f on C.

Now assume that $\operatorname{ri}\operatorname{dom} f$ meets $\operatorname{ri} C$. The latter set is the relative interior of the effective domain of I_C, and so by Theorem 9.1.6 we have for each x

$$\partial(f + I_C)(x) = \partial f(x) + \partial I_C(x) = \partial f(x) + N_C(x). \tag{9.13}$$

Saying that x_0 minimizes f on C is the same as saying that it minimizes $f + I_C$, and then by the definition of subdifferential we have $0 \in \partial(f + I_C)(x_0)$. Referring to (9.13), we see that we can rewrite this in the form (9.12), as claimed. □

To obtain a sufficient optimality condition we needed no regularity condition, while the added hypothesis was required for a necessary condition. This situation is typical for local minimization. Here, however, because of convexity it held for global minimization.

9.2 Moreau envelopes and applications

Given a convex function f and a positive parameter λ, one can create a new convex function e_λ, called the *Moreau envelope* of f. The envelope function minorizes f

and is always continuously differentiable. Moreover, the construction of e_λ provides additional useful information about the subdifferential ∂f.

Definition 9.2.1. Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a closed proper convex function, and let $\lambda > 0$. The *Moreau envelope* e_λ is a function defined from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$ by

$$e_\lambda(x) = \inf_y \{f(y) + (2\lambda)^{-1}\|y-x\|^2\}.$$

□

Theorem 9.2.2. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a closed proper convex function, and let $\lambda > 0$. Then*

a. e_λ is a closed proper convex function, and for each $x \in \mathbb{R}^n$ the infimum in its definition is attained at a unique point $y = p_\lambda(x)$ characterized by

$$x = y + \lambda y^*, \qquad (y, y^*) \in \partial f. \tag{9.14}$$

b. For each $x \in \operatorname{dom} f$ one has $e_\lambda(x) \le f(x)$, and if $x \in \operatorname{dom} \partial f$ then also

$$f(x) \le e_\lambda(x) + (1/2)\lambda d[0, \partial f(x)]^2. \tag{9.15}$$

c. For fixed $\lambda > 0$ the functions $p_\lambda(x)$ and

$$d_\lambda(x) := \lambda^{-1}[x - p_\lambda(x)]$$

are Lipschitz continuous on $\mathbb{R}^n$ with moduli 1 and λ^{-1} respectively.

d. e_λ is Fréchet differentiable on $\mathbb{R}^n$ with derivative d_λ.

Proof. Fix $x \in \mathbb{R}^n$. As f is proper convex, it is finite on its effective domain, which is nonempty. Let $x_0 \in \operatorname{ri} \operatorname{dom} f$, so that f is subdifferentiable at x_0, and let $x_0^* \in \partial f(x_0)$. Then

$$f(y) + (2\lambda)^{-1}\|y-x\|^2 \ge f(x_0) + \langle x_0^*, y - x_0\rangle + (2\lambda)^{-1}\|y-x\|^2 =: m(y).$$

The quadratic function m is closed proper convex, and for any $v \in \mathbb{R}^n$ and $\tau > 0$ we have

$$\tau^{-1}[m(x+\tau v) - m(x)] = \langle x_0^*, v\rangle + \tau(2\lambda)^{-1}\|v\|;$$

if $v \ne 0$ this difference quotient converges to $+\infty$ as τ does. Therefore $\operatorname{rc} m = \{0\}$, and by Theorem 7.2.9 the level sets of m are compact.

If we write $g(y) = f(y) + (2\lambda)^{-1}\|y-x\|^2$ then g is closed proper convex as the sum of f and $(2\lambda)^{-1}\|(\cdot) - x\|^2$, by Theorem 9.1.6. As $m \le g$, for each real ϕ we have $\operatorname{lev}_\phi g \subset \operatorname{lev}_\phi m$. The sets $\operatorname{lev}_\phi g$ are closed because g is closed, so they are compact. Applying Proposition 6.2.8 to any nonempty level set we conclude that g attains its minimum on that level set, and this is the global minimum of g.

However, $g(y)$ inherits the strong convexity of $(2\lambda)^{-1}\|y-x\|^2$, and this implies that any minimizer y of g is unique. For this y we must have

$$0 \in \partial g(y) = \partial f(y) + \lambda^{-1}(y-x),$$

where the equality comes from Theorem 9.1.6. Therefore there is some y^* with $(y,y^*) \in \partial f$ and with $x = y + \lambda y^*$. If we suppose that $(z,z^*) \in \partial f$ with $x = z + \lambda z^*$, then

$$0 = (y-z) + \lambda(y^* - z^*),$$

and as ∂f is monotone, we have

$$0 = \langle y-z, y-z\rangle) + \lambda\langle y^* - z^*, y-z\rangle \geq \|y-z\|^2,$$

so that $z = y$ and therefore $z^* = y^*$. Accordingly, for each $x \in \mathbb{R}^n$ (9.14) determines a unique pair $(y,y^*) \in \partial f$.

If we observe that the conjugate of $q(v) := (2\lambda)^{-1}\|v\|^2$ is $q^*(v^*) = (1/2)\lambda\|v^*\|^2$, whose effective domain is all of $\mathbb{R}^n$, that f is closed, and that $f^* + q^*$ is proper, then we can apply Theorem 9.1.6 to obtain

$$(f^* + q^*)^*(x) = \inf_y\{f(y) + (2\lambda)^{-1}\|y-x\|^2\} = e_\lambda(x), \tag{9.16}$$

which shows that e_λ is a closed proper convex function. The proves (a).

For (b), we have $e_\lambda(x) \leq g(x) = f(x)$ for each $x \in \mathbb{R}^n$. Suppose $x \in \operatorname{dom}\partial f$, and use (9.16) to write

$$f(x) - e_\lambda(x) = f(x) - (f^* + q^*)(x).$$

It follows that for any $x^* \in \mathbb{R}^n$ we have

$$f(x) - e_\lambda(x) \leq f(x) - [\langle x^*, x\rangle - f^*(x^*)] + q^*(x^*).$$

We can make the difference $f(x) - [\langle x^*, x\rangle - f^*(x^*)]$ as small as possible by choosing an x^* in $\partial f(x)$, for which this difference will be zero. As $\partial f(x)$ is closed, we may further restrict x^* by choosing it to be the smallest element of $\partial f(x)$ so that $\|x^*\| = d[0, \partial f(x)]$. Then

$$f(x) - e_\lambda(x) \leq q^*(x^*) = (1/2)\lambda d[0, \partial f(x)]^2,$$

which establishes (b).

To prove (c), fix $\lambda > 0$ and consider two points x_1 and x_2. For $i = 1,2$ the points $y_i = p_\lambda(x_i)$ satisfy $(y_i, y_i^*) \in \partial f$ and $x_i = y_i + \lambda y_i^*$. Therefore

$$\begin{aligned} \|x_1 - x_2\|^2 &= \|y_1 - y_2\|^2 + 2\lambda\langle y_1^* - y_2^*, y_1 - y_2\rangle + \lambda^2\|y_1^* - y_2^*\|^2 \\ &\geq \|y_1 - y_2\|^2 + \lambda^2\|y_1^* - y_2^*\|^2. \end{aligned} \tag{9.17}$$

As we may discard either of the terms on the right side of (9.17), this shows in particular that $p_\lambda(x)$ is Lipschitzian on $\mathbb{R}^n$ with modulus 1, and that if we define

$$d_\lambda(x) = \lambda^{-1}[x - p_\lambda(x)],$$

then d_λ is Lipschitzian on $\mathbb{R}^n$ with modulus λ^{-1}, because $d_\lambda(x)$ is just y^* in the equation $x = y + \lambda y^*$.

For (d), fix $x \in \mathbb{R}^n$ and $\lambda > 0$ and define for $h \in \mathbb{R}^n$

$$a(h) = e_\lambda(x+h) - e_\lambda(x) - \langle d_\lambda(x), h\rangle.$$

We will show that $\|h\|^{-1}|a(h)|$ approaches zero as h does, which will establish that $d_\lambda(x)$ is the F-derivative of $e_\lambda f$ at x.

Let $y = p_\lambda(x)$; then we have

$$e_\lambda(x) = f(y) + (2\lambda)^{-1}\|y-x\|^2, \qquad e_\lambda(x+h) \le f(y) + (2\lambda)^{-1}\|y-(x+h)\|^2.$$

Subtracting, we obtain

$$e_\lambda(x+h) - e_\lambda(x) \le (2\lambda)^{-1}\{\|y-(x+h)\|^2 - \|y-x\|^2\} = \langle d_\lambda(x), h\rangle + (2\lambda)^{-1}\|h\|^2,$$

so that $a(h) \le (2\lambda)^{-1}\|h\|^2$. However, $a(h)$ is a convex function that takes the value 0 at the origin. Accordingly,

$$-a(h) \le a(-h) \le (2\lambda)^{-1}\|h\|^2,$$

so that $|a(h)| \le (2\lambda)^{-1}\|h\|^2$. Therefore

$$\|h\|^{-1}|a(h)| \le (2\lambda)^{-1}\|h\|,$$

which approaches zero as $\|h\|$ does. □

The conclusions of (c) and (d) together show that e_λ actually has an F-derivative that is everywhere Lipschitz continuous with a uniform modulus.

We can use part (a) of Theorem 9.2.2 to establish some very significant facts about subdifferentials of closed proper convex functions. We defined monotone operators at the beginning of Section 8.1 to be multifunctions F from $\mathbb{R}^n$ to $\mathbb{R}^n$ such that for each pair of elements (x,y) and (x',y') in the graph of F one has $\langle x-x', y-y'\rangle \ge 0$. We also defined cyclically monotone operators in Definition 8.1.5, and showed in Proposition 8.1.6 that subdifferentials of closed proper convex functions are cyclically monotone. However, any monotone operator retains its monotonicity if we arbitrarily delete points of its graph. Thus, one might wonder whether one could augment the graphs of subdifferential operators to obtain larger operators—in the sense of graph inclusion—that were still monotone or cyclically monotone. We now show that this cannot be done.

Definition 9.2.3. Let F be a monotone operator from $\mathbb{R}^n$ to $\mathbb{R}^n$. F is *maximal monotone* if its graph is not properly contained in the graph of any monotone operator. It is *maximal cyclically monotone* if its graph is not properly contained in the graph of any cyclically monotone operator. □

Thus, the maximal monotone operators are those whose graphs cannot be enlarged without their losing the property of monotonicity, and a similar statement

applies to maximal cyclically monotone operators and the property of cyclic monotonicity.

Part (a) of Theorem 9.2.2 established the *Moreau decomposition:* given a closed proper convex function f on $\mathbb{R}^n$ and a positive λ, we can express each point x of $\mathbb{R}^n$ uniquely in the form given by (9.14), namely

$$x = y + \lambda y^*, \qquad (y, y^*) \in \partial f.$$

This is a far-reaching generalization of the decompositions that we have already seen, which express a point of $\mathbb{R}^n$ uniquely as a sum $m + m^*$, with m belonging to a subspace M and m^* to $M^\perp$, or as $k + k^*$, where k belongs to a closed convex cone K and k^* to its polar cone K°, with $\langle k, k^* \rangle = 0$. Each of these is a special case of another decomposition, which uses a closed convex set C in $\mathbb{R}^n$ and expresses any point x of $\mathbb{R}^n$ uniquely as the projection c of x on C plus an element c^* of $N_C(x)$. As $N_C = \partial I_C$, we can now see that this is the special case of the Moreau decomposition that uses $f = I_C$.

One of the consequences of the Moreau decomposition is the following theorem on maximal monotonicity.

Theorem 9.2.4. *If f is a closed proper convex function on $\mathbb{R}^n$, then ∂f is a maximal monotone operator.*

Proof. Suppose $\partial f \subset T$, where T is a monotone operator. Let (t, t^*) be any point of T and write $z = t + t^*$. Apply the Moreau decomposition to z with $\lambda = 1$ to find $(y, y^*) \in \partial f$ with $y + y^* = z$. Then

$$0 = (t + t^*) - (y + y^*) = (t - y) + (t^* - y^*), \tag{9.18}$$

and so

$$0 = \|t - y\|^2 + \langle t^* - y^*, t - y \rangle. \tag{9.19}$$

As both (t, t^*) and (y, y^*) belong to T, the second term on the right in (9.19) is nonnegative, which implies $t = y$. Putting that result in (9.18) we obtain $t^* = y^*$, so that $(t, t^*) = (y, y^*)$ and therefore $(t, t^*) \in \partial f$. As (t, t^*) was an arbitrary point of T, we have $T = \partial f$ so that ∂f is maximal monotone. □

Corollary 9.2.5. *If f is a closed proper convex function on $\mathbb{R}^n$, then ∂f is a maximal cyclically monotone operator.*

Proof. The cyclically monotone operators are a subclass of the monotone operators. As ∂f is maximal monotone by Theorem 9.2.4 and is cyclically monotone, it is also maximal within the more restricted class of cyclically monotone operators. □

Here is another consequence of the Moreau decomposition. We use the term *Lipschitz homeomorphism* to denote a homeomorphism h of sets contained in normed linear spaces, such that both h and h^{-1} are Lipschitz continuous. In the following theorem we need to put a norm on the graph space $\mathbb{R}^n \times \mathbb{R}^n$; we can use $\|(a, a^*)\| = \max\{\|a\|, \|a^*\|\}$.

Theorem 9.2.6. *If f is a closed proper convex function on $\mathbb{R}^n$, then the graph of ∂f is Lipschitz homeomorphic to $\mathbb{R}^n$.*

Proof. Write G for the graph of ∂f and fix any $\lambda > 0$. The map $\rho(g,g^*) = g + \lambda g^*$ takes G into $\mathbb{R}^n$, and the map $\sigma(z) = [p_\lambda(z), d_\lambda(z)]$ takes $\mathbb{R}^n$ into G because $d_\lambda(z) \in \partial f(p_\lambda(z))$. In fact each of these maps is onto, because for any $z \in \mathbb{R}^n$ there is a unique $(g,g^*) \in G$ with $g + \lambda g^* = z$ so that $z = \rho(g,g^*)$, and for each $(g,g^*) \in G$ if we define $z = g + \lambda g^*$ then $g = p_\lambda(z)$ and $g^* = d_\lambda(z)$ so that $\sigma(z) = (g,g^*)$. The map ρ is evidently Lipschitz continuous, and part (c) of Theorem 9.2.2 shows that σ is Lipschitz continuous. □

9.3 Systems of convex inequalities

This section proves two basic results about finite or infinite systems of convex inequalities. We will use them in the next chapter to prove the von Neumann min-max theorem, but they have other uses as well as being of interest in themselves.

The first theorem deals with a finite set of inequalities, and one can regard it as an extension of the Gordan theorem of the alternative (Proposition 3.1.11). We write Λ_m for the *unit* $(m-1)$*-simplex:* that is, the set of points $(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m_+$ whose components sum to 1.

Theorem 9.3.1. *Let $f_1, \ldots, f_I$ be proper convex functions on $\mathbb{R}^n$, and let C be a convex subset of $\mathbb{R}^n$ such that* $\operatorname{ri} C \subset \cap_{i=1}^I \operatorname{dom} f_i$. *Then*

$$\text{for each } x \in C, \sup_{i=1}^{I} f_i(x) \geq 0 \tag{9.20}$$

if and only if there exists $p^* \in \Lambda_I$ such that

$$\text{for each } x \in \operatorname{cl} C, \sum_{i=1}^{I} p_i^* f_i(x) \geq 0. \tag{9.21}$$

The reason for regarding this as an extension of the Gordan theorem is that the supremum of the f_i will be nonnegative on C if and only if there is no $x \in C$ such that $f_i(x) < 0$ for $i = 1, \ldots, I$. This is an extension of the matrix inequality $Ax < 0$ in the Gordan theorem, if one regards the rows A_i of A as generating convex functions $f_i(x) = \langle A_i, x \rangle$.

Proof. (If) The supremum of the f_i majorizes $\sum_{i=1^I} p_i^* f_i$ whenever $p^* \in \Lambda_I$, so if (9.21) holds, then (9.20) must also hold.

(Only if) Assume (9.20). If C is empty then any $p^* \in \Lambda_I$ will work, so assume that C is nonempty. Define

$$D = \{w \in \mathbb{R}^I \mid \text{for some } x \in C \text{ and each } i = 1, \ldots, I,\ f_i(x) < w_i\}.$$

(9.20) implies that D is disjoint from $\mathbb{R}^I_-$. Moreover, D is convex: take two points in it, choose $\lambda \in (0,1)$, and apply Proposition 6.1.3 to each f_i. Therefore we can separate D and $\mathbb{R}^I_-$ properly by a hyperplane with normal p^*. Lemma 3.1.6 says that we can choose $p^* \in \mathbb{R}^I_+$ (the polar of $\mathbb{R}^I_-$), and that we can suppose that the hyperplane passes through the origin. Moreover, we can then scale p^* so that its components sum to 1. Then $p^* \in \Lambda_I$. As $\mathbb{R}^I_-$ lies in the lower closed halfspace of $H(p^*,0)$, D must lie in the upper closed halfspace. Then for each $w \in D$, $\langle p^*, w\rangle \geq 0$.

Let $X = C \cap [\cap_{i=1}^I \operatorname{dom} f_i]$. This X is a convex set on which every f_i is finite. Moreover, by hypothesis $\operatorname{ri} C \subset \operatorname{dom} f_i$ for each i, so we have

$$\operatorname{ri} C \subset X \subset C. \tag{9.22}$$

By taking closures through this inclusion and using Proposition 1.2.6 we obtain

$$\operatorname{cl} C = \operatorname{cl} X, \quad \operatorname{ri} C = \operatorname{ri} X.$$

Choose an arbitrary $x \in X$. For an arbitrary positive ε set $w_i = f_i(x) + \varepsilon$ for $i = 1,\ldots,I$, and let w be the element of $\mathbb{R}^I$ whose components are the w_i. The definition of w shows that it is in D, so that

$$0 \leq \langle p^*, w\rangle = \sum_{i=1}^I p_i^* f_i(x) + \varepsilon.$$

But ε was any positive number, so $\sum_{i=1}^I p_i^* f_i(x) \geq 0$. As x was arbitrary in X, the function $f := \sum_{i=1}^I p_i^* f_i$ is nonnegative on X.

As f is proper convex, if x is any point of $\operatorname{cl} X$ then $f(x) \geq (\operatorname{cl} f)(x)$, and $(\operatorname{cl} f)(x)$ is the limit of the values of f along a line segment from a point in $\operatorname{ri} X$ to x. These values are all nonnegative, so $f(x)$ is nonnegative. Accordingly, f is nonnegative on $\operatorname{cl} X$. As $\operatorname{cl} X = \operatorname{cl} C$, f is nonnegative on $\operatorname{cl} C$ and therefore (9.21) holds. □

The second theorem deals with a possibly infinite set of inequalities, but strengthens the requirement on the functions and on the set C.

Theorem 9.3.2. *Let $\{f_\alpha \mid \alpha \in A\}$ be a nonempty collection of closed proper convex functions on $\mathbb{R}^n$, and let C be a compact convex subset of $\mathbb{R}^n$ such that $C \subset \operatorname{dom} f_\alpha$ for each $\alpha \in A$. Then*

$$\text{for each } x \in C, \ \sup_{\alpha \in A} f_\alpha(x) > 0 \tag{9.23}$$

if and only if there exist a finite positive integer K, an element p^* of Λ_K, a positive ε, and a set of elements $\{\alpha_1,\ldots,\alpha_K\} \subset A$, such that

$$\text{for each } x \in C, \ \sum_{k=1}^K p_k^* f_{\alpha_k}(x) \geq \varepsilon. \tag{9.24}$$

Proof. (If) Assume (9.24) holds. If $p^* \in \Lambda_K$, then $\sup_{\alpha \in A} f_i(x)$ majorizes any finite sum of the form $\sum_{k=1}^K p_k^* f_{\alpha_k}(x)$ and therefore (9.23) holds.

(Only if) As before, if C is empty then any integer $K > 0$, element $p^* \in \Lambda_K$, and elements $\alpha_k \in A$ will work; therefore assume that C is nonempty. For each $x \in C$ we are given that the supremum of the $f_\alpha(x)$ is positive, so at least one of these values, say $f_{\alpha_x}(x)$, is greater than a positive number ε_x. Further, as f_{α_x} is lower semicontinuous there is an open neighborhood N_x of x on which f_{α_x} remains greater than ε_x.

The neighborhoods $\{N_x \mid x \in C\}$ form an open cover of the compact set C, so there is a finite subcover. For some positive integer K, let this subcover be the union of the sets N_{x_k} for $k = 1, \ldots, K$. For each such k write α_k, ε_k, f_k, and N_k for α_{x_k}, ε_{x_k}, $f_{\alpha_{x_k}}$, and N_{x_k} respectively. Define $\varepsilon := \min_{k=1}^K \varepsilon_k$; then $\varepsilon > 0$.

For each $x \in C$ there is then a k' such that $x \in N_{k'}$, and therefore

$$\sup_{k=1}^K [f_k(x) - \varepsilon] \geq f_{k'}(x) - \varepsilon > \varepsilon_{k'} - \varepsilon \geq 0.$$

Therefore the supremum of the functions $\{f_k(x) - \varepsilon \mid k = 1, \ldots, K\}$ is positive on all of C.

Apply Theorem 9.3.1 to this finite collection of functions to produce $p^* \in \Lambda_K$ such that the function $\sum_{k=1}^K p_k^* [f_k(x) - \varepsilon]$ is nonnegative for each $x \in C$. Then because the p_k^* sum to 1, for each such x

$$\sum_{k=1}^K p_k^* f_k(x) = \sum_{k=1}^K p_k^* [f_k(x) - \varepsilon] + \varepsilon \geq \varepsilon.$$

Accordingly, $\sum_{k=1}^K p_k^* f_k$ is never less than ε on C, so that (9.24) holds. □

Chapter 10
Duality

A particularly useful application area for the techniques of convexity has been *duality,* in which one associates with an optimization problem called the *primal problem* a *dual problem,* also involving optimization but typically in a different space, and then exploits information that each of the two problems provides about the other. Duality actually is a very simple concept that does not require any convexity at all. However, the best known and probably most effective applications have been in areas where some convexity appears, because the methods of convexity facilitate the functional operations required, and they also provide symmetry in the relationship between the primal and dual problems.

The initial section of this chapter introduces the underlying conceptual model and the fundamental concept of a saddle point, and it establishes some basic definitions and notation. These require no convexity. The next section proves a *min-max theorem* asserting that under certain hypotheses saddle points will exist.

We then show how, given a convex optimization problem, one may embed that problem in a larger problem to which the conceptual model applies. This embedding is the essence of duality. We illustrate it with a duality analysis of a general linear programming problem, which shows the procedure and the key objects involved with minimal technical complications. We then employ the same embedding method to treat a more general convex optimization problem, and show how elements of each of the primal and dual problems provide insight into the behavior of the other problem.

Finally we develop criteria for the existence of optimal solutions. These involve subgradients, and therefore the information that we developed in Chapters 8 and 9 about the behavior of subgradients and their connection with directional differentiation give additional information about the behavior of these solutions.

10.1 Conceptual model

To introduce the concept of duality, suppose that S and T are two sets and that ϕ is an extended real-valued function from $S \times T$ to $\bar{\mathbb{R}}$. One can think of function values $\phi(x,y)$ as the outcomes of actions by two players, Xavier and Yvette, who control the variables $x \in S$ and $y \in T$ respectively. Xavier wishes to make ϕ small, whereas Yvette wishes to make it large. Thus, one way to interpret ϕ would be as a money payment from Xavier to Yvette that can take on either positive or negative values, with negative values indicating a transfer of money from Yvette to Xavier. The fundamental model that we are using here is thus a two-person zero-sum game [10, 21, 27, 40], but with a general payoff function rather than the bilinear function resulting from the well known *finite* (or *matrix*) games.

Define two subsets of S and T respectively by

$$\begin{aligned} \operatorname{dom}_1 \phi &= \{x \in S \mid \text{for each } y \in T,\ \phi(x,y) < +\infty\}, \\ \operatorname{dom}_2 \phi &= \{y \in T \mid \text{for each } x \in S,\ \phi(x,y) > -\infty\}, \end{aligned} \tag{10.1}$$

and define $\operatorname{dom}\phi = \operatorname{dom}_1 \phi \times \operatorname{dom}_2 \phi$. Thus, if $(x,y) \in \operatorname{dom}\phi$ then $\phi(x,y)$ is finite, though it may also be finite at points outside $\operatorname{dom}\phi$. We say that ϕ is *proper* if $\operatorname{dom}\phi \neq \emptyset$.

Now use ϕ to define two functions $\xi : S \to \bar{\mathbb{R}}$ and $\eta : T \to \bar{\mathbb{R}}$ by

$$\xi(x) = \sup_{y \in T} \phi(x,y), \qquad \eta(y) = \inf_{x \in S} \phi(x,y). \tag{10.2}$$

In terms of the game model, ξ provides for each strategy x available to Xavier an upper bound on what he will have to pay if he uses x, given that Yvette observes his choice and then does the best she can against that strategy, and η provides for each y available to Yvette a lower bound on what she can get if she employs y.

If $x \notin \operatorname{dom}_1 \phi$ then there is a $y' \in T$ such that $\phi(x,y') = +\infty$, and therefore $\xi(x) = +\infty$. For that reason, Xavier would never want to choose x outside $\operatorname{dom}_1 \phi$. Similarly, if $y \notin \operatorname{dom}_2 \phi$ then $\eta(y) = -\infty$ and Yvette would never want to choose y outside $\operatorname{dom}_2 \phi$. We then have

$$\inf_{x \in S} \xi(x) = \inf_{x \in \operatorname{dom}_1 \phi} \xi(x), \qquad \sup_{y \in T} \eta(y) = \sup_{y \in \operatorname{dom}_2 \phi} \eta(y).$$

The definitions of ξ and η show that for each $x_0 \in S$ and each $y_0 \in T$,

$$\xi(x_0) = \sup_y \phi(x_0,y) \geq \phi(x_0,y_0) \geq \inf_x \phi(x,y_0) = \eta(y_0). \tag{10.3}$$

Thus $\xi(x) \geq \eta(y)$ whenever $(x,y) \in S \times T$. This is the *weak duality inequality.*

By taking the infimum over $x \in S$ and the supremum over $y \in T$ in (10.3) we find that

$$\inf_{x \in S} \sup_{y \in T} \phi(x,y) \geq \sup_{y \in T} \inf_{x \in S} \phi(x,y). \tag{10.4}$$

The two sides of (10.4) may very well be unequal: for example, the problem in which S and T are each $\{0,1\}$ and

$$\phi(x,y) = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{if } x \neq y, \end{cases} \tag{10.5}$$

has $\xi(x) \equiv 1$ and $\eta(y) \equiv 0$, so that $\inf_x \sup_y \phi(x,y) = 1$ and $\sup_y \inf_x \phi(x,y) = 0$.

Definition 10.1.1. A point $(x_0,y_0) \in \operatorname{dom}\phi$ is a *saddle point* of ϕ if for each $(x,y) \in \operatorname{dom}\phi$ one has

$$\phi(x,y_0) \geq \phi(x_0,y_0) \geq \phi(x_0,y). \tag{10.6}$$

Inspection of (10.6) shows that a saddle point satisfies

$$\eta(y_0) = \min_{x \in \operatorname{dom}_1 \phi} \phi(x,y_0) = \phi(x_0,y_0) = \max_{y \in \operatorname{dom}_2 \phi} \phi(x_0,y)\xi(x_0). \tag{10.7}$$

Thus, by choosing x_0 Xavier can guarantee that the payoff will be no larger than $\phi(x_0,y_0)$, though it may be smaller, and by choosing y_0 Yvette can guarantee that the payoff will be no smaller than $\phi(x_0,y_0)$, though it may be larger. In defining a saddle point we restrict the variables to $\operatorname{dom}\phi$ because, as observed above, players would never want to choose points outside $\operatorname{dom}\phi$.

A saddle point need not exist: e.g., the problem described in (10.5) has none. Moreover, if a saddle point exists it need not be unique, as one can see by amending (10.5) to make $\phi(x,y) \equiv 1$, so that every pair (x,y) is a saddle point.

We saw in (10.3) that we always had $\xi(x) \geq \phi(x,y) \geq \eta(y)$, and the example in (10.5) showed that we might never have $\xi(x) = \eta(y)$. The following theorem shows that a pair (x,y) achieves this equality precisely when it is a saddle point. It also develops some properties of the set of saddle points.

Theorem 10.1.2. *Let S and T be sets, with $\phi : S \times T \to \bar{\mathbb{R}}$, and define $\xi : S \to \bar{\mathbb{R}}$ and $\eta : T \to \bar{\mathbb{R}}$ by* (10.2)*. If ϕ is proper, then the following are equivalent:*

a. $(x_0,y_0) \in \operatorname{dom}\phi$ *and* $\xi(x_0) = \eta(y_0)$.
b. (x_0,y_0) *is a saddle point of* ϕ.

If these two conditions hold, then

$$\xi(x_0) = \phi(x_0,y_0) = \eta(y_0). \tag{10.8}$$

Further, there are subsets Φ_S of S and Φ_T of T such that the set Φ of saddle points has the form $\Phi = \Phi_S \times \Phi_T$, and every saddle point has the same value of ϕ.

Proof. (a) implies (b). If (a) holds then $(x_0,y_0) \in \operatorname{dom}\phi$. Moreover, by applying the weak duality inequality (10.3) to (x_0,y_0) and using (a), we find that for each $(x,y) \in \operatorname{dom}\phi$,

$$\phi(x,y_0) \geq \eta(y_0) = \xi(x_0) \geq \phi(x_0,y_0) \geq \eta(y_0) = \xi(x_0) \geq \phi(x_0,y),$$

which shows that (x_0, y_0) is a saddle point and also shows that (a) implies (10.8).

(b) implies (a). If (b) holds then $(x_0, y_0) \in \operatorname{dom}\phi$, and (10.7) implies that $\eta(y_0 = \phi(x_0, y_0) = \xi(x_0)$.

The final claims are true if there are no saddle points, so we can suppose that Φ is nonempty. Let

$$\Phi_S = \{x \in S \mid \text{for some } y' \in T,\ (x, y') \in \Phi\},$$
$$\Phi_T = \{y \in T \mid \text{for some } x' \in S,\ (x', y) \in \Phi\}.$$

We have $\Phi_S \subset \operatorname{dom}_1 \phi$, $\Phi_T \subset \operatorname{dom}_2 \phi$, and $\Phi \subset \Phi_S \times \Phi_T$.

Now let $(x, y) \in \Phi_S \times \Phi_T$; we will show that $(x, y) \in \Phi$. By construction, there are points $x' \in \operatorname{dom}_1 \phi$ and $y' \in \operatorname{dom}_2 \phi$ such that (x, y') and (x', y) belong to Φ. Then

$$\phi(x', y) \geq \phi(x', y') \geq \phi(x, y') \geq \phi(x, y) \geq \phi(x', y), \tag{10.9}$$

so that all of these function values are the same.

Recalling that (x', y) is a saddle point and using (10.7) we find that

$$\xi(x') = \phi(x', y) = \eta(y), \tag{10.10}$$

and using the same reasoning on (x, y') yields

$$\xi(x) = \phi(x, y') = \eta(y'). \tag{10.11}$$

Using the equality of the values in (10.9) we can rearrange (10.10) and (10.11) to obtain

$$\xi(x) = \phi(x, y) = \eta(y), \qquad \xi(x') = \phi(x', y') = \eta(y').$$

Now from the equivalence of (a) and (b) we conclude that both (x, y) and (x', y') are saddle points, so that in particular $(x, y) \in \Phi$ and therefore $\Phi = \Phi_S \times \Phi_T$. To see that any two saddle points have the same value of ϕ, let (u, v) and (u', v') be saddle points. Applying the argument that we just finished to these pairs instead of to (x, y') and (x', y) shows that both (u, v') and (u', v) are saddle points. These four pairs will then satisfy an inequality of the form (10.9), which shows in particular that $\phi(u, v) = \phi(u', v')$. □

The unique value that ϕ takes at each saddle point is called the *saddle value* of ϕ. The saddle value therefore represents the payoff resulting from a certain type of equilibrium in which neither player can do any better by unilaterally departing from the equilibrium choices represented by the saddle point. It is a very special case of the *Nash equilibrium* [10, Section 1.2]. However, although ϕ takes the saddle value at every saddle point, a point of $S \times T$ at which ϕ takes the saddle value may or may not be a saddle point.

One might have reservations about this model on the grounds that it begins with a function ϕ that takes extended real values, which might seem artificial. Could we start with a model involving a real-valued function ϕ_r and constraints $x \in X \subset S$ and $y \in Y \subset T$, then recover ϕ in a natural way by using extended real values to remove

the constraints on x and y? We can do so, but in the process we will discover an indeterminacy about ϕ that might not have been evident from the initial discussion above.

Let X and Y be nonempty subsets of S and T respectively, and let $\phi_r : X \times Y \to \mathbb{R}$. Define $\xi_r : X \to \bar{\mathbb{R}}$ and $\eta_r : Y \to \bar{\mathbb{R}}$ by

$$\xi_r(x) = \sup_{y \in Y} \phi_r(x,y), \qquad \eta_r(y) = \sup_{x \in X} \phi_r(x,y). \tag{10.12}$$

We can interpret these quantities in terms of a game just as we did in the extended-real-valued case.

Now define $\phi : S \times T \to \bar{\mathbb{R}}$ as follows:

$$\phi(x,y) = \begin{cases} \phi_r(x,y) & \text{if } (x,y) \in X \times Y, \\ +\infty & \text{if } x \notin X \text{ and } y \in Y, \\ -\infty & \text{if } x \in X \text{ and } y \notin Y, \\ \text{arbitrary values} & \text{if } x \notin X \text{ and } y \notin Y. \end{cases} \tag{10.13}$$

If we construct from this ϕ the various elements of the extended-real-valued model above, we recover exactly the real-valued model, as we now see.

Proposition 10.1.3. *Let X and Y be nonempty subsets of S and T respectively, and let $\phi_r : X \times Y \to \mathbb{R}$. Define $\phi : S \times T \to \bar{\mathbb{R}}$ by* (10.13)*. Then ϕ is proper, with*

$$\operatorname{dom}_1 \phi = X, \qquad \operatorname{dom}_2 \phi = Y, \tag{10.14}$$

and with the functions ξ and η defined in (10.2) *given by*

$$\xi(x) = \begin{cases} \xi_r(x) = \sup_{y \in Y} \phi_r(x,y) & \text{if } x \in X, \\ +\infty & \text{if } x \notin X, \end{cases} \tag{10.15}$$

and

$$\eta(y) = \begin{cases} \eta_r(y) = \inf_{x \in X} \phi_r(x,y) & \text{if } y \in Y, \\ -\infty & \text{if } y \notin Y. \end{cases} \tag{10.16}$$

Proof. From (10.13) we see that if $x \in X$ then the value of $\phi(x,y)$ will be $\phi_r(x,y)$ if $y \in Y$ and $-\infty$ if $y \notin Y$. Therefore $\phi(x,y) < +\infty$ for every y, so $x \in \operatorname{dom}_1 \phi$. On the other hand, if $x \notin X$ then if we choose $y \in Y$, which is nonempty by hypothesis, we have $\phi(x,y) = +\infty$ so that $x \notin \operatorname{dom}_1 \phi$. Therefore $\operatorname{dom}_1 \phi = X$. A similar argument shows that $\operatorname{dom}_2 \phi = Y$, which establishes (10.14).

If $x \in X$, then as $\phi(x,y) = -\infty$ for $y \notin Y$ we have

$$\xi(x) = \sup_{y} \phi(x,y) = \sup_{y \in Y} \phi(x,y) = \sup_{y \in Y} \phi_r(x,y) = \xi_r(x),$$

because ϕ agrees with ϕ_r on $X \times Y$. If $x \notin X$, then as Y is nonempty there is a point y for which $\phi(x,y) = +\infty$, so that $\xi(x) = +\infty$. This establishes (10.15). A parallel argument for η establishes (10.16). □

The indeterminacy mentioned just before (10.12) is in the assignment in (10.13) of arbitrary values to $\phi(x,y)$ when $x \notin X$ and $y \notin Y$. Evidently these values do not matter for any of the work we have done up to this point. They do matter, however, for operations to regularize saddle functions by imposing semicontinuity requirements. However, for the purpose of this section we are free to use any values we want to, and two especially natural choices are to set $\phi(x,y) = -\infty$ for $x \notin X$ and $y \notin Y$, or to use $+\infty$ instead of $-\infty$ in assigning these values. These two choices produce, respectively, the *lower simple extension* and the *upper simple extension* of ϕ_r.

The mention of semicontinuity might call to mind how little we have actually assumed about ϕ, X, Y, S, and T up to this point. We did not impose any structure on the spaces S and T, nor on X and Y except that they were nonempty subsets of S and T, and we assumed nothing about ϕ except that it was a function. Although even without such assumptions we were able to establish some relationships among various concepts and to give a game interpretation, it would be nice to have some idea of when saddle points can be expected to exist. In the next section we give an existence proof for saddle points in a setting with additional structure.

10.2 Existence of saddle points

Theorems asserting the existence of saddle points are usually called min-max theorems. One such theorem, proved by John von Neumann (1928), involves two compact convex sets $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ and a function $\phi : X \times Y \to \mathbb{R}$ such that for each $x \in X$ $\phi(x,\cdot)$ is concave and upper semicontinuous, and for each $y \in Y$ $\phi(\cdot,y)$ is convex and lower semicontinuous. It shows that any such ϕ must have a saddle point in $X \times Y$.

More recently there have been proofs of various min-max theorems that require weaker assumptions than does the von Neumann theorem. We will prove one, due to Sion (1958), that has been very useful. Versions of this theorem are available for general spaces, but we will work with nonempty subsets X of $\mathbb{R}^n$ and Y of $\mathbb{R}^m$. We will use a function $\phi : X \times Y \to \mathbb{R}$ that is lower semicontinuous in x for each $y \in Y$ and upper semicontinuous in y for each $x \in X$.

Before proving the von Neumann theorem we examine properties of sets and functions that we need for both theorems but that we did not use in Section 10.1. First, the semicontinuity assumptions on ϕ allow us to say a little more about the expressions $\inf_{x\in X} \sup_{y\in Y} \phi(x,y)$ and $\sup_{y\in Y} \inf_{x\in X} \phi(x,y)$. The function $\xi : X \to \bar{\mathbb{R}}$ given by $\xi(x) := \sup_{y\in Y} \phi(x,y)$ is the supremum of a collection of functions $\{\phi(\cdot,y) \mid y \in Y\}$, each of which is lower semicontinuous by hypothesis. By Proposition 6.2.5 ξ is then lower semicontinuous on X. A similar argument shows that $\eta(y) := \inf_{x\in X} \phi(x,y)$ is upper semicontinuous on Y.

If either or both of the sets X and Y are compact, we can do more. First, if X is compact then for each $y \in Y$ we have by Proposition 6.2.8 $\inf_{x\in X} \phi(x,y) = \min_{x\in X} \phi(x,y)$ and therefore

$$\sup_{y\in Y} \inf_{x\in X} \phi(x,y) = \sup_{y\in Y} \min_{x\in X} \phi(x,y).$$

Also, Proposition 6.2.8 then shows that for some $x_0 \in X$,

$$\inf_{x\in X} \sup_{y\in Y} \phi(x,y) = \inf_{x\in X} \xi(x) = \xi(x_0) = \min_{x\in X} \xi(x) = \min_{x\in X} \sup_{y\in Y} \phi(x,y).$$

Similarly, if Y is compact then

$$\inf_{x\in X} \sup_{y\in Y} \phi(x,y) = \inf_{x\in X} \max_{y\in Y} \phi(x,y),$$

and for some $y_0 \in Y$,

$$\sup_{y\in Y} \inf_{x\in X} \phi(x,y) = \sup_{y\in Y} \eta(y) = \eta(y_0) = \max_{y\in Y} \eta(y) = \max_{y\in Y)} \inf_{x\in X} \phi(x,y).$$

If both X and Y are compact then we can replace inf by min and sup by max everywhere, and write

$$\min_{x\in X} \max_{y\in Y} \phi(x,y), \qquad \max_{y\in Y)} \min_{x\in X} \phi(x,y).$$

The next result is the von Neumann theorem. Its original statement referred only to a real-valued function on the product of two sets. The following version uses the extension of such a function developed in Section 10.1.

Theorem 10.2.1 (von Neumann, 1928). *Let X and Y be nonempty compact convex subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Let $\phi_r : X \times Y \to \mathbb{R}$, and suppose that for each $y \in Y$ $\phi_r(\cdot, y)$ is a lower semicontinuous convex function on X, and for each $x \in X$ $\phi_r(x, \cdot)$ is an upper semicontinuous concave function on Y. Let $\phi : \mathbb{R}^n \times \mathbb{R}^m \to \bar{\mathbb{R}}$ be an extension of ϕ_r satisfying* (10.13)*. Then ϕ has a saddle point in $X \times Y$.*

Proof. We first prove that the functions $\xi(x)$ and $\eta(y)$ are closed proper convex and closed proper concave respectively, and that the extrema in their definitions are attained. For $x \in X$ $\xi(x)$ is the supremum of $\phi_r(x,y)$ for $y \in Y$ (Proposition 10.1.3), and by hypothesis $\phi_r(x, \cdot)$ is upper semicontinuous, real-valued and concave and Y is compact. Therefore the supremum is attained at a finite value, and in fact

$$\xi(x) = \xi_r(x) = \max_{y\in Y} \phi_r(x,y) = \max_{y\in Y} \phi(x,y).$$

For $x \notin X$ and $y \in Y$ we have $\phi(x,y) = +\infty$, so that $\operatorname{dom} \xi = X$. A similar argument shows that the infimum in the definition of $\eta(y)$ is attained for each $y \in Y$,

$$\eta(y) = \eta_r(y) = \min_{x\in X} \phi_r(x,y) = \min_{x\in X} \phi(x,y),$$

and $\operatorname{dom}\eta = Y$.

The epigraph of ξ is the intersection, over all $y \in Y$, of the sets $\{(x,\alpha) \mid \phi(x,y) \leq \alpha\}$, which are the epigraphs of the functions $\phi(\cdot\,,y)$. For $y \in Y$ the functions $\phi_r(\cdot\,,y)$ are lower semicontinuous and convex by hypothesis, and as $\phi(\cdot\,,y)$ is the $+\infty$ extension of $\phi_r(\cdot\,,y)$ and X is closed, Proposition 6.2.6 shows that $\phi(\cdot\,,y)$ is lower semicontinuous. The epigraph of ξ is then an intersection of closed convex sets, so ξ is lower semicontinuous and convex. It never takes $-\infty$ because no $\phi(\cdot\,,y)$ does, and therefore ξ is a closed proper convex function. A similar proof shows that η is closed proper concave.

It now follows from the nonemptiness and compactness of X and Y that

$$\inf_{x\in\operatorname{dom}_1\phi} \xi(x) = \inf_{x\in X} \xi_r(x) = \min_{x\in X} \xi_r(x) = \min_{x\in X} \max_{y\in Y} \phi(x,y),$$

attained at some $x_0 \in \operatorname{dom}_1\phi$, and that

$$\sup_{y\in\operatorname{dom}_2\phi} \eta(y) = \sup_{y\in Y} \eta_r(y) = \max_{y\in Y} \eta_r(y) = \max_{y\in Y} \min_{x\in X} \phi(x,y),$$

attained at some $y_0 \in \operatorname{dom}_2\phi$.

We know from (10.3) that

$$\max_{y\in Y} \eta(y) = \max_{y\in Y} \min_{x\in X} \phi(x,y) \leq \min_{x\in X} \max_{y\in Y} \phi(x,y) = \min_{x\in X} \xi(x), \tag{10.17}$$

and by our analysis above, the two sides of this inequality are finite. Write β for the value of the right-hand side. Choose a positive ε; we will show that the left-hand side is greater than $\beta - \varepsilon$, and this will show that the two sides are actually equal.

For each $y \in Y$ the function $\phi(\cdot\,,y)$ is closed proper convex. As β is the right-hand side of (10.17), for each $x \in X$ we have $\max_{y\in Y} \phi(x,y) \geq \beta > \beta - \varepsilon$. Apply Theorem 9.3.2 to the collection $\{\theta(\cdot\,,y) = \phi(\cdot\,,y) - (\beta-\varepsilon) \mid y \in Y\}$ with $C = X$ to conclude that there are a finite positive integer K, an element $p^* \in \Lambda_K$, a positive δ, and a collection of points $\{y_1,\ldots,y_K\} \subset Y$ such that for each $x \in X$, $\sum_{k=1}^K p_k^* \theta(x,y_k) \geq \delta$. If we set $y_\varepsilon = \sum_{k=1}^K p_k^* y_k$, then for each $x \in X$ the concavity of the function $\theta(x,\cdot)$ yields

$$\theta(x,y_\varepsilon) \geq \sum_{k=1}^K p_k^* \theta(x,y_k) \geq \delta,$$

and therefore $\eta(y_\varepsilon) - (\beta-\varepsilon) \geq \delta$. Then

$$\max_{y\in Y} \min_{x\in X} \phi(x,y) = \max_{y\in Y} \eta(y) \geq \eta(y_\varepsilon) \geq \beta - \varepsilon + \delta > \beta - \varepsilon.$$

Therefore the two sides of (10.17) are equal, so that

$$\min_{x\in X} \xi(x) = \xi(x_0) = \eta(y_0) = \max_{y\in Y} \eta(y).$$

Applying Theorem 10.1.2, we find that (x_0,y_0) is a saddle point of ϕ. □

The Sion theorem is next. Instead of convex and concave functions it uses functions that are *quasiconvex* or *quasiconcave*. Here is a definition.

Definition 10.2.2. Let C be a convex subset of $\mathbb{R}^n$. A function $q : C \to \mathbb{R}$ is *quasiconvex on C* if for each real μ the set $\{x \in C \mid q(x) \leq \mu\}$ is convex. q is *quasiconcave on C* if $-q$ is quasicovex on C.

Thus quasiconvexity amounts to convexity of the lower level sets of a function. It is a much weaker assumption than convexity. An equivalent property is that for each x and x' in C and each $\lambda \in [0,1]$, one has $q[(1-\lambda)x+\lambda x'] \leq \max\{q(x), q(x')\}$. For more on quasiconvexity and related topics, see e.g. [12, Section 2.4] or [22, Chapter 9].

We will establish the Sion theorem by first proving a technical lemma from which the theorem follows immediately. To simplify the statements, we list first the following hypotheses that will apply to both results.

$$\begin{aligned}&\textit{Let } X \textit{ and } Y \textit{ be nonempty convex subsets of } \mathbb{R}^n \textit{ and } \mathbb{R}^m\\ &\textit{respectively, with } X \textit{ compact. Let } \phi : X \times Y \to \mathbb{R} \textit{ be a function}\\ &\textit{that is lower semicontinuous and quasiconvex in } x \textit{ for each } y \in Y,\\ &\textit{and upper semicontinuous and quasiconcave in } y \textit{ for each } x \in X.\end{aligned} \tag{10.18}$$

Here is the lemma.

Lemma 10.2.3. *Assume* (10.18) *and suppose in addition that $\{y_1, \ldots, y_k\}$ is a finite subset of Y and that*

$$0 < \min_{x \in X} \max_{i=1}^{k} \phi(x, y_i). \tag{10.19}$$

Then there exists $y_0 \in Y$ such that

$$0 < \min_{x \in X} \phi(x, y_0). \tag{10.20}$$

Proof. For $k = 1$ there is nothing to prove. We will show that the lemma is true in the case $k = 2$, which requires most of the work, then establish it for general k by induction.

For the case $k = 2$, if the lemma is not true then

$$\text{For each } y \in Y, \ \min_{x \in X} f(x, y) \leq 0. \tag{10.21}$$

Use (10.19) to find a real μ with

$$0 < \mu < \min_{x \in X} \max_{i=1}^{2} f(x, y_i). \tag{10.22}$$

Write $[y_1, y_2]$ for the closed line segment between y_1 and y_2, and for each $y \in [y_1, y_2]$ define

$$C_0(y) := \{x \in X \mid \phi(x,y) \le 0\}, \qquad C_\mu(y) := \{x \in X \mid \phi(x,y) \le \mu\}. \tag{10.23}$$

These two sets are nonempty (by (10.21)), closed, and convex. Write C_1 for $C_\mu(y_1)$ and C_2 for $C_\mu(y_2)$. We have $C_1 \cap C_2 = \emptyset$ because otherwise for some $x' \in X$ we would have $\max\{\phi(x',y_1), \phi(x',y_2)\} \le \mu$, contradicting (10.22).

For each $x \in X$ and each $y \in [y_1, y_2]$ we have

$$\phi(x,y) \ge \min\{\phi(x,y_1), \phi(x,y_2)\} \tag{10.24}$$

by quasiconcavity of $\phi(x,\cdot)$. Fix $y' \in [y_1,y_2]$. If $C_\mu(y') \not\subset C_1 \cup C_2$ then there is an $x' \in X$ with $\phi(x',y_i) > \mu$ for $i = 1,2$ but $\phi(x',y') \le \mu$, contradicting (10.24). Therefore $C_\mu(y') \subset C_1 \cup C_2$.

Now define $D_i = C\mu(y') \cap C_i$ for $i = 1,2$. These two sets are closed and convex; we have $D_1 \cap D_2 = \emptyset$ and $D_1 \cup D_2 = C\mu(y')$. If both sets were nonempty it would contradict the fact that by Proposition C.2.2 $C_\mu(y')$ is connected. Therefore one is empty, so we have

$$C_0(y') \subset C_\mu(y') \subset C_1 \text{ or } C_0(y') \subset C_\mu(y') \subset C_2. \tag{10.25}$$

Now define Y_1 to be the set of those $y' \in [y_1,y_2]$ for which the first alternative in (10.25) holds, and Y_2 to be the set of those for which the second holds. These are nonempty, because $y_i \in Y_i$ for $i = 1,2$, and we have

$$Y_1 \cap Y_2 = \emptyset, \qquad Y_1 \cup Y_2 = [y_1,y_2]. \tag{10.26}$$

We will show next that each Y_i is closed in $[y_1,y_2]$. The proof is similar for each, so we prove only that Y_1 is closed. Choose a sequence $\{w_m\} \subset Y_1$ that converges to some w_0 belonging to the closed set $[y_1,y_2]$. To show that $w_0 \in Y_1$ we need to show that $C_\mu(w_0) \subset C_1$. Choose any $x_0 \in C_0(w_0)$. We have $\phi(x_0,w_0) \le 0 < \mu$, so by upper semicontinuity of $\phi(x_0,\cdot)$ there is a neighborhood N of w_0 such that for each $w \in N$, $\phi(x_0,w) < \mu$. Take m large enough so that $y_m \in N$: then $\phi(x_0,y_m) < \mu$ so $x_0 \in C_\mu(y_m)$. We took y_m to be in Y_1, so $x_0 \in C_\mu(y_m) \subset C_1$. But x_0 was any point of $C_0(w_0)$, so $C_0(w_0) \subset C_1$ and then by (10.25) also $C_\mu(w_0) \subset C_1$. This shows that $w_0 \in Y_1$ and therefore that Y_1 is closed in $[y_1,y_2]$.

As Y_1 and Y_2 are nonempty, closed in $[y_1,y_2]$, and satisfy (10.26) we have a contradiction to the connectedness of $[y_1,y_2]$. This implies that (10.21) does not hold, and therefore that there exists $y_0 \in Y$ such that $0 < \min_{x\in X} \phi(x,y_0)$, which proves the lemma in the case $k = 2$.

Now suppose that we have established the lemma for all values of k less than some $k > 2$. Suppose that (10.19) holds. We will show that (10.20) holds.

Define Z to be $\{x \in X \mid \phi(x,y_k) \le 0\}$. This Z is compact and convex. If it is empty, take $y_0 = y_k$ and we have (10.20). Suppose it is nonempty. From (10.19) we have

$$0 < \min_{x\in X} \max_{i=1}^{k} \phi(x,y_i). \tag{10.27}$$

For $x \in Z$, $\phi(x, y_k) \leq 0$ and therefore the index k cannot contribute to the maximum in (10.27). Accordingly, we have

$$0 < \min_{x \in Z} \max_{i=1}^{k-1} \phi(x, y_i). \tag{10.28}$$

Apply the induction hypothesis to (10.28) to conclude that there is some $y' \in Y$ such that

$$0 < \min_{x \in Z} \phi(x, y').$$

Now we have

$$\begin{aligned} &\text{If } x \in Z, \text{ then } 0 < \phi(x, y') \leq \max\{\phi(x, y'), \phi(x, y_k)\}, \\ &\text{If } x \in X \backslash Z, \text{ then } 0 < \phi(x, y_k) \leq \max\{\phi(x, y'), \phi(x, y_k)\}, \end{aligned}$$

from which it follows that

$$0 < \min_{x \in X} \max\{\phi(x, y'), \phi(x, y_k)\},$$

and by the first part of the proof there is y_0 such that (10.20) holds. □

Next, the theorem:

Theorem 10.2.4 (Sion, 1958). *Assume* (10.18). *Then*

$$\min_{x \in X} \sup_{y \in Y} \phi(x, y) = \sup_{y \in Y} \min_{x \in X} \phi(x, y). \tag{10.29}$$

Proof. We already know from (10.4) that $\geq$ holds in (10.29), so we need only prove $\leq$. Let

$$\alpha < \min_{x \in X} \sup_{y \in Y} \phi(x, y); \tag{10.30}$$

we will show that

$$\alpha < \sup_{y \in Y} \min_{x \in X} \phi(x, y), \tag{10.31}$$

and that will complete the proof.

For $y \in Y$ define $S(y) := \{x \in X \mid \phi(x, y) \leq \alpha\}$. These sets are closed by lower semicontinuity of $\phi(\,\cdot\,, y)$ and compact because they are subsets of X. If their intersection is nonempty then there is some $x_0 \in X$ such that for each $y \in Y$, $\phi(x_0, y) \leq \alpha$. Then

$$\min_{x \in X} \sup_{y \in Y} \phi(x, y) \leq \sup_{y \in Y} \phi(x_0, y) \leq \alpha,$$

which contradicts (10.30). Therefore $\cap_{y \in Y} S(y) = \emptyset$, and by the finite intersection principle there is a finite subset $\{y_1, \ldots, y_k\}$ of Y such that $\cap_{i=1}^{k} S(y_i) = \emptyset$. This means that for each $x \in X$ we have $\max_{i=1}^{k} \phi(x, y_i) > \alpha$. The function on the left side of this inequality is lower semicontinuous in x by Proposition 6.2.5, and it attains its minimum for $x \in X$ by Proposition 6.2.8. Therefore the minimum must be strictly greater than α, so that

$$\alpha < \min_{x\in X} \max_{i=1}^{k} \phi(x, y_i). \tag{10.32}$$

We can rewrite (10.32) as

$$0 < \min_{x\in X} \max_{i=1}^{k} [\phi(x, y_i) - \alpha],$$

and if we apply Lemma 10.2.3 we find that there exists $y_0 \in Y$ such that

$$0 < \min_{x\in X} [\phi(x, y_0) - \alpha],$$

so that

$$\sup_{y\in Y} \min_{x\in X} \phi(x, y) \geq \min_{x\in X} \phi(x, y_0) > \alpha,$$

and this proves (10.31). □

With an additional assumption we can get a saddle point.

Corollary 10.2.5. *Assume* (10.18)*, and in addition suppose that Y is compact. Then ϕ has a saddle point.*

Proof. If Y is compact, then as shown in the discussion just before (10.18), the conclusion of the Sion theorem becomes

$$\min_{x\in X} \max_{y\in Y} \phi(x, y) = \max_{y\in Y} \min_{x\in X} \phi(x, y).$$

Using the notation of Section 10.1 we can rewrite this as

$$\min_{x\in X} \xi(x) = \max_{y\in Y} \eta(y).$$

There are then points $x_0 \in X$ and $y_0 \in Y$ such that

$$\min_{x\in X} \xi(x) = \xi(x_0), \qquad \max_{y\in Y} \eta(y) = \eta(y_0),$$

and therefore $\xi(x_0) = \eta(y_0)$. Now Theorem 10.1.2 shows that (x_0, y_0) is a saddle point of ϕ. □

10.3 Conjugate duality: formulation

The fundamental idea of duality is to start with a given optimization problem, called the *primal* problem, of finding the infimum of some given extended real-valued function f depending on variables $x \in \mathbb{R}^n$. We then embed this problem as one of a family of problems indexed by a parameter $p \in \mathbb{R}^m$. Specifically, we create a *perturbation function* $F : \mathbb{R}^n \times \mathbb{R}^m \to \bar{\mathbb{R}}$ such that $F(\cdot, 0) = f$ and for each x

the function $F(x,\cdot)$ is closed and convex. Then for each $p \in \mathbb{R}^m$ we consider the optimization problem of finding

$$v(p) := \inf_x F(x,p). \tag{10.33}$$

The function v is called the *value function,* or *marginal function,* of the problem. For $p = 0$ we have the original problem, as $F(x,0) = f(x)$ so that $v(0)$ is the infimum of f, which may be either finite or infinite.

The next step is to construct from F a saddle function in such a way that one half of the resulting saddle problem is an optimization problem that reduces to the original problem of minimizing f. The other half of the saddle problem turns out to be a new problem, invisible until now, the *dual* problem.

Section 10.3.1 shows how to construct such a saddle function. Section 10.3.2 extends the basic construction to produce a system that is symmetric in the sense that one can start from either the primal or the dual problem and then proceed through the duality procedure to obtain the other problem. Section 10.3.3 applies this system to some relatively simple problems, both to illustrate the calculations and to show how this structure unifies particular formulations that otherwise might not seem to be closely related.

10.3.1 Constructing the Lagrangian

A saddle function that fits the requirement outlined above is the *Lagrangian* $L : \mathbb{R}^n \times \mathbb{R}^m \to \bar{\mathbb{R}}$ defined by

$$L(x,p^*) = [-F(x,\cdot)]^*(p^*) = \inf_p\{\langle p^*,p\rangle + F(x,p)\}, \tag{10.34}$$

which we obtain by taking the conjugate in the concave sense of $-F$ in the second variable only. As under our conditions

$$\sup_{p^*} L(x,p^*) = \sup_{p^*}[-F(x,\cdot)]^*(p^*) = -[-F(x,\cdot)]^{**}(0) = F(x,0) = f(x), \tag{10.35}$$

the problem of finding the infimum of f becomes the problem of computing

$$\inf_x f(x) = \inf_x \sup_{p^*} L(x,p^*). \tag{10.36}$$

In (10.36) we have one half of a saddle point problem. For the other half, we can define a function of p^* by

$$g(p^*) := \inf_x L(x,p^*), \tag{10.37}$$

so that

$$\sup_{p^*} g(p^*) = \sup_{p^*} \inf_x L(x,p^*) \le \inf_x \sup_{p^*} L(x,p^*) = \inf_x f(x). \tag{10.38}$$

This g is the *dual objective function* associated with the particular duality structure induced by our choice of F. If we had chosen a different F satisfying the requirements above, we could have obtained a different dual objective. Thus, with the problem of finding the infimum of f we can associate as many different dual problems as we wish, and these may differ greatly in form.

For each x, (10.34) shows that $L(x,p^*) = [-F(x,\cdot)]^*(p^*)$, so that $L(x,p^*)$ is a closed concave function of p^*. The function g is the infimum of the collection of such functions, indexed by x, and therefore it is closed and concave no matter what f was.

In the above formulation (10.35) shows that $f(x)$ corresponds to the function $\xi(x)$ of Sections 10.1 and 10.2, and (10.37) shows that $g(p^*)$ corresponds to $\eta(y)$. The general analysis of Section 10.1 then shows that

$$\sup_{p^*} g(p^*) \le \inf_x f(x), \tag{10.39}$$

but that equality does not necessarily hold. When it fails to hold, the difference $\inf_x f(x) - \sup_{p^*} g(p^*)$ is called a *duality gap*. On the other hand, Theorem 10.1.2 shows that equality does hold for points $x_0 \in \mathrm{dom}_1 L$ and $p_0^* \in \mathrm{dom}_2 L$ exactly when (x_0, p_0^*) is a saddle point of L. It also shows that the set of saddle points has a product structure. Accordingly, if there are multiple solutions any primal optimal solution x_0 can be paired with any dual optimal solution p_0^* to form a saddle point of the Lagrangian L.

10.3.2 Symmetric duality

We would like to extend the basic construction to a duality structure that is symmetric in the sense that we can pass from a primal to a dual problem, then apply the same or an analogous technique to the dual problem and thereby recover the primal. We can do so by requiring that F be, not just closed and convex in p for fixed values of x, but closed and convex in (x,p).

Assuming that F is a closed convex function, define the *dual perturbation function* G by $G = F^A$, where the superscript A denotes the adjoint operation introduced in Section 7.1.2. As F is convex, we take the adjoint in the convex sense to obtain

$$\begin{aligned} G(p^*,x^*) &= F^A(p^*,x^*) \\ &= \inf_{x,p}\{-\langle x^*,x\rangle + \langle p^*,p\rangle + F(x,p)\} \\ &= \inf_x\{-\langle x^*,x\rangle + L(x,p^*)\}. \end{aligned} \tag{10.40}$$

This G is a closed concave function of (p^*, x^*), and by substituting $x^* = 0$ in (10.40) we find that

$$G(p^*, 0) = \inf_x L(x, p^*) = g(p^*),$$

so that $G(\cdot, 0)$ is the dual objective function g. Therefore we can regard G as a perturbation function playing the same role for the dual problem as F did for the primal. The last line of (10.40) shows that we could also have obtained G directly from L, because

$$-G(p^*, x^*) = \sup_x \{\langle x^*, x\rangle - L(x, p^*)\} = L(\cdot, p^*)^*(x^*).$$

We can also define a Lagrangian for the dual problem; by analogy with the primal construction we take

$$\begin{aligned} L'(x, p^*) &= [-G(p^*, \cdot)]^*(x) \\ &= \sup_{x^*} \{\langle x^*, x\rangle + G(p^*, x^*)\}. \end{aligned} \tag{10.41}$$

Now suppose that we want to dualize the dual problem. We can try to do so by taking the adjoint of the dual perturbation function G, and as G is concave we take the adjoint in the concave sense. However, $G^A = F^{AA} = F$, so that dualizing the dual problem allows us to recover the original primal perturbation function and hence, if we wish, the original primal objective function $f(x) = F(x, 0)$.

10.3.3 Example: linear programming

To put in perspective the various constructions that we just introduced, here is an example of what the duality transformations do when applied to a general linear programming problem. The formulation is general enough so that one can recover the dual of any linear programming problem by making particular choices of the cones involved. Later we will illustrate calculations for other problems.

Let J and K be polyhedral convex cones in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, and consider the linear programming problem

$$\inf_x \{\langle c^*, x\rangle \mid a - Ax \in K^\circ,\ x \in J\}. \tag{10.42}$$

By replacing K and J by particular cones, one can express any finite-dimensional linear programming problem in this form. We will illustrate how to obtain the conventional dual problem associated with (10.42) from one case of the general duality structure discussed above. The condensed derivation (through the Lagrangian) comes first, then the longer version that yields the complete dual perturbation structure.

We need first to embed (10.42) in a perturbation structure. For that purpose we introduce a variable $p \in \mathbb{R}^m$ and define

$$F(x,p) = \begin{cases} \langle c^*, x\rangle & \text{if } a - p - Ax \in K^\circ,\ x \in J, \\ +\infty & \text{otherwise.} \end{cases} \tag{10.43}$$

Then

$$L(x,p^*) = \inf_p \{\langle p^*, p\rangle + F(x,p)\}. \tag{10.44}$$

To calculate the infimum in (10.44), as x is fixed we have only p to work with. We can write the requirement $a - p - Ax \in K^\circ$ as

$$p = a - Ax - k^*, \qquad k^* \in K^\circ, \tag{10.45}$$

and then reason that if $a - p - Ax = k^*$ and $x \in J$ the quantity being minimized is $\langle p^*, p\rangle + \langle c^*, x\rangle$, so that as we can do nothing for the moment with x, we should try to make $\langle p^*, p\rangle$ as small as possible. We can see from (10.45) that if $p^* \in K$ then by taking $k^* = 0$ we can make $\langle p^*, p\rangle$ equal to $\langle p^*, a - Ax\rangle$, but no smaller. Therefore in this case the infimum is $\langle c^*, x\rangle + \langle p^*, a\rangle - \langle p^*, Ax\rangle$. On the other hand, if $p^* \notin K$ then we can find some $k^* \in K^\circ$ with $\langle p^*, k^*\rangle > 0$, and then by taking large multiples of this k^* we can make $\langle p^*, p\rangle + \langle c^*, x\rangle$ as small as we wish. In this case, the infimum is $-\infty$. Finally, if $x \notin J$ then $F(x,p) = +\infty$ for each p, so the infimum is $+\infty$. Collecting the results of these three cases, we have

$$L(x,p^*) = \begin{cases} \langle c^*, x\rangle + \langle p^*, a\rangle - \langle p^*, Ax\rangle & \text{if } p^* \in K \text{ and } x \in J, \\ -\infty & \text{if } p^* \notin K \text{ and } x \in J, \\ +\infty & \text{if } x \notin J. \end{cases} \tag{10.46}$$

To obtain the dual objective function $g(p^*)$ we calculate the infimum in x of $L(\cdot, p)$ by considering first the case when $p^* \in K$ and then that of $p^* \notin K$. In the first case we are taking the infimum of

$$\langle c^*, x\rangle + \langle p^*, a\rangle - \langle p^*, Ax\rangle = \langle p^*, a\rangle + \langle c^* - A^* p^*, x\rangle,$$

over $x \in J$ (we do not want to leave J because the function value is then $+\infty$, which will not help the infimum). This shows us that if $A^* p^* - c^* \in J^\circ$ then the best infimum we can get is $\langle p^*, a\rangle$ at $x = 0$, whereas if $A^* p^* - c^* \notin J^\circ$ then we will get $-\infty$. In the second case, when $p^* \notin K$, adjusting x makes no difference and the infimum is $-\infty$. We therefore have

$$g(p^*) = \inf_x L(x,p^*) = \begin{cases} \langle p^*, a\rangle & \text{if } A^* p^* - c^* \in J^\circ \text{ and } p^* \in K, \\ -\infty & \text{otherwise.} \end{cases} \tag{10.47}$$

The dual of (10.42) is then the problem of finding the supremum of g. In notation similar to that of (10.42), this is

$$\sup_{p^*} \{\langle p^*, a\rangle \mid A^* p^* - c^* \in J^\circ,\ p^* \in K\}. \tag{10.48}$$

For a particular instance of this general formulation, if we choose $K = \mathbb{R}^m$ and $J = \mathbb{R}^n_+$ then (10.42) and (10.48) become, respectively,

$$\inf_x \{\langle c^*, x\rangle \mid Ax = a,\ x \geq 0\} \tag{10.49}$$

and

$$\sup_{p^*} \{\langle p^*, a\rangle \mid A^* p^* \leq c,\ p^* \text{ free}\}. \tag{10.50}$$

Next we use the same problems (10.42) and (10.48) to illustrate the full duality structure with both primal and dual perturbation functions. To do so we replace the calculation of the Lagrangian in (10.46) by calculation of the adjoint of F as in (10.40):

$$\begin{aligned} G(p^*, x^*) &= F^A(p^*, x^*) \\ &= \inf_{x,p} \{-\langle x^*, x\rangle + \langle p^*, p\rangle + F(x, p)\} \\ &= \inf_x \inf_p \{-\langle x^*, x\rangle + \langle p^*, p\rangle + \langle c^*, x\rangle \mid a - p - Ax \in K^\circ,\ x \in J\} \\ &= \inf_x \begin{cases} \{\langle p^*, a\rangle + \langle c^* - x^* - A^* p^*, x\rangle \mid x \in J\} & \text{if } p^* \in K, \\ -\infty & \text{if } p^* \notin K, \end{cases} \\ &= \begin{cases} \langle p^*, a\rangle & \text{if } A^* p^* + x^* - c^* \in J^\circ,\ p^* \in K, \\ -\infty & \text{otherwise.} \end{cases} \end{aligned} \tag{10.51}$$

If we set the dual perturbation variable x^* to 0 then $G(p^*, x^*)$ reduces to $g(p^*)$, as we should expect.

We have now derived a dual perturbation structure $G(p^*, x^*)$ from the original primal $F(x, p)$ and, as we can check that F is closed proper convex, we know from the analysis of Section 10.3.2 that if we were to start with G then we could recover F. However, the symmetry in this general structure is not quite complete, and in fact it cannot be made complete.

To see this, use (10.41) to compute the dual Lagrangian:

$$\begin{aligned} L'(p^*, x) &= [-G(p^*, \cdot)] * (x) \\ &= \sup_{x^*} \{\langle x^*, x\rangle + G(p^*, x^*)\} \\ &= \sup_x \begin{cases} \{\langle x^*, x\rangle + \langle p^*, a\rangle \mid A^* p^* + x^* - c^* \in J^\circ\} & \text{if } p^* \in K, \\ -\infty & \text{if } p^* \notin K, \end{cases} \\ &= \sup_x \begin{cases} \{\langle c^* - A^* p^* + l^*, x\rangle + \langle p^*, a\rangle \mid l^* \in J^\circ\} & \text{if } p^* \in K, \\ -\infty & \text{if } p^* \notin K, \end{cases} \\ &= \begin{cases} \langle c^*, x\rangle + \langle p^*, a\rangle - \langle p^*, Ax\rangle & \text{if } x \in J,\ p^* \in K, \\ +\infty & \text{if } x \notin J,\ p^* \in K \\ -\infty & \text{if } p^* \notin K. \end{cases} \end{aligned} \tag{10.52}$$

Now if we compare (10.46) with (10.52) we see that there is a small but noticeable difference: L and L' disagree in the region where $x \notin J$ and $p^* \notin K$, the primal Lagrangian L taking the value $+\infty$ there while the dual Lagrangian L' takes the value $-\infty$. We saw a similar disagreement at the end of Section 10.1 when we looked at the lower and upper simple extensions of a saddle function.

In fact this disagreement expresses a general property of saddle functions like L and L': namely, a full treatment of such functions has to deal not with individual functions but with equivalence classes. The construction of L and L' has produced two members of such an equivalence class.

10.3.4 An economic interpretation

The dual linear programming problem in (10.48) may appear to have come out of nowhere as a result of the manipulations that we applied to the primal. However, dual problems arising in practice often have interpretations that help to illuminate aspects of the systems being studied. This is certainly the case with linear programming, where we can often assign a definite economic meaning to the dual.

To demonstrate how this happens, take the primal problem of (10.42) and make the particular choices $K = \mathbb{R}^I_+$ and $J = \mathbb{R}^H_+$, to obtain the problem

$$\inf_x \{\langle c^*, x\rangle \mid Ax \geq a,\ x \geq 0\}. \tag{10.53}$$

The problem in (10.53) is an abstraction of the decision problem faced by a manager with H production activities (e.g., plants or departments within plants), each represented by a column of A: for the jth activity and for goods numbered $1, \ldots, I$, the element a_{ij} of A is the number of units of good i produced by activity j in one time period if the activity is operated at a reference level. In this interpretation we assume constant returns to scale, so for a nonnegative real number x_j the activity if operated at level x_j will produce $a_{ij}x_j$ units of good i. The entries a_{ij} may be positive, zero, or negative, with a negative entry indicating net consumption of that good. On the other hand, the levels x_j cannot be negative: these processes are not reversible. Therefore, the net output of the entire production process, given a set of levels $x \in \mathbb{R}^H_+$, will be a vector $Ax \in \mathbb{R}^I$ of goods.

The cost of operating production activity j for one time unit at the reference level is c^*_j, and we assume that the cost of operating at level x_j is $c^*_j x_j$. We will assume that the manager's tasks is to produce at least a_i units of good i in one time period, where the a_i also may be negative: producing at least -5 units means not consuming more than 5 units.

This is the simplest formulation; it assumes free disposal in the sense that one can produce more than is needed of any good, then give away or discard the excess. If there is no free disposal then we can add disposal activities, the most elementary of which will be columns with a -1 in some entry and zeros in all others, and with a positive unit operating cost. Operating such an activity disposes of excess, but at a

cost. Then we take $K = \mathbb{R}^I$ instead of $K = \mathbb{R}^I_+$, so that the constraint $Ax \geq a$ becomes $Ax = a$.

Faced with the above requirement, in the version with free disposal the manager will consider only production plans $x \geq 0$ satisfying $Ax \geq a$, and among those will prefer an x_0 that achieves the least value of the total cost $\langle c^*, x \rangle$. This is (10.53).

Now consider the problem that the manager has in allocating cost to products. There is a total production cost $\langle c^*, x_0 \rangle$, and there is a vector $(a_1, \ldots, a_I)$ of products available for sale to bring in income to cover the cost. But it is not at all clear how to allocate the cost to the products, because the production costs appear to come from activities and not from products. We may know that a particular activity cost \$128,000 to operate for one time unit, but that activity produced 560 units of one product, 11,600 of another, consumed 8,470 units of a third, and so on. It is very difficult to see how to make a realistic allocation of operating cost to the products rather than to the activities.

Yet we need such an allocation, because we need to know whether making 10,000 units of product number 5 will yield as much or more, when the product is sold in the market, as it cost us to make the 10,000 units. If the answer is that it will yield less, then we had better change our production plan, or perhaps discontinue product 5 altogether. Thus, we actually need to find a set of prices that not only allocates all of the cost to products, but does so in a way that lets us predict the effects of changing our production plan. This means that we need what economists call *marginal costs* of changes in the elements of a.

This is where the dual problem can help. With the particular choices of K and L that we made for (10.53), the dual becomes

$$\sup_{p^*} \{ \langle p^*, a \rangle \mid A^* p^* \leq c^*, \ p^* \geq 0 \}. \tag{10.54}$$

We will interpret this as a problem of assigning prices $p_1^*, \ldots, p_I^*$ to the I products. As we show next, the constraints in (10.54) amount to requiring these prices to meet certain requirements of realism derived from observed behavior of markets.

First, if there is free disposal then prices should not be negative. If, for example, $p_1^* < 0$ then this means that people will pay us to take good 1 away from them. Given free disposal, we can take large quantities from those people, collect this subsidy of $|p_1^*|$ per unit, and then throw away the goods. The people offering us these goods would be behaving stupidly, as under free disposal they could throw the goods away themselves at no cost instead of paying us $|p_1^*|$ per unit to do so. It is not reasonable to assume that a price system depending on such behavior would last long, so we should impose the constraint that $p^* \geq 0$.

The argument for the second constraint is very similar. Suppose that for some j we were to have $(A^* p^*)_j > c_j^*$. As

$$(A^* p^*)_j = \sum_{i=1}^{I} p_i^* a_{ij},$$

this is the value, at the prices p^*, of the package of goods produced by activity j operating for one time unit at the reference level. The total cost of that operation is c_j^*, so that we have an *arbitrage opportunity,* which is a chance to make a transaction leaving us with no change in product inventory but a guaranteed net monetary gain: something for nothing. To do so we operate activity j at some positive level ρ and produce a bundle $\rho(a_{1j},\dots,a_{Ij})$ of goods. We have to pay a cost ρc_j^*. Then we exchange the goods on the market for a sum of money equaling $\rho(A^*p^*)_j$. We are left with the same inventory of goods as before, but with our cash increased by the positive amount $\rho[(A^*p^*)_j - c_j^*]$. As there is no limit to the level ρ at which we are allowed to operate activity j, we can make as much free money as we want to by using a sufficiently large operating level ρ.

Experience with markets has shown that arbitrage opportunities tend not to persist for very long, because people try to exploit the opportunity, buying on one side and selling on the other. This activity drives up prices of the products being bought and drives down prices of the products being sold, so the arbitrage opportunity disappears. Thus in order for our prices p^* to pass the reality test, they should not offer any arbitrage opportunities. That means that we should enforce the constraint $A^*p^* \le c^*$.

In fact, if we look back at the argument by which we decided to constrain the prices to be nonnegative, we can see that the existence of a negative price in the presence of free disposal was actually another arbitrage opportunity. Therefore we can give a very simple verbal description of the dual constraints that we have derived: the prices should not offer any arbitrage opportunity, or in other words they should be *arbitrage-free.*

A price system p^* satisfies the two requirements just discussed exactly when it satisfies the constraints of the dual programming problem (10.54). We will assume that the primal problem (10.53) has been properly formulated in the sense that it has a feasible solution and its objective function is bounded below on the set of feasible solutions. Then we know from the theory of linear programming that both the primal and dual problems are actually solvable, and that their optimal objective values are equal. This means that we can find an optimal solution x_0 of (10.53) and an optimal solution p_0^* of (10.54) such that

$$\langle c^*, x_0\rangle = \langle p_0^*, a\rangle. \tag{10.55}$$

But this means that we have solved our allocation problem: we can find arbitrage-free prices p_0^* so that the total value $\langle p_0^*, a\rangle$ of the goods a that we have produced, at the prices p_0^*, exactly equals the total cost $\langle c^*, x_0\rangle$ of producing those goods.

In fact, we will show in Proposition 10.4.3 below that the negatives of these prices are actually subgradients of the primal value function $v(p)$ at $p = 0$. For our choices of K and L the perturbation function we used was

$$F(x,p) = \begin{cases} \langle c^*, x\rangle & \text{if } Ax + p \ge a,\ x \ge 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{10.56}$$

Thus if we want to investigate the effect of changing a to $a+h$ for some $h \in \mathbb{R}^I$, then we should set $p = -h$. The new optimal value will then be $v(-h)$. Then for any dual optimal solution p_0^*,

$$v(-h) \geq v(0) + \langle -p_0^*, (-h) - 0 \rangle,$$

and therefore the change in total cost from this production change will be $v(-h) - v(0)$, which is at least as large as $\langle p_0^*, h \rangle$.

If the dual optimal solution is unique, as is often the case, then that solution is actually a Fréchet derivative (Theorem 8.3.2), so that the change in optimal cost to replace a by $a+h$ will be $\langle p_0^*, h \rangle + o(h)$. In this case, the elements of p_0^* are true marginal costs for the production of the goods $1, \ldots, I$. It is possible to use the polyhedrality of epigraphs associated with the linear programming problem to extend this result somewhat in this particular case, but we do not do so here.

10.4 Conjugate duality: existence of solutions

In this section we look at four questions:

1. When does the Lagrangian L have a saddle point?
2. When is there no duality gap?
3. How can one interpret the dual variables?
4. What conditions ensure that the extrema of the primal and/or dual problems are attained?

The answers to the first two questions are closely related, and we will look first at those, then at the rest.

Theorem 10.4.1. *Suppose that for each x, $F(x, \cdot)$ is closed and convex. Then the following are equivalent:*

a. The Lagrangian L has a saddle point (x_0, p_0^*) with finite saddle value $L(x_0, p_0^*)$.
b. $x_0 \in \operatorname{dom} f$, $p_0^* \in \operatorname{dom} g$, and $g(p_0^*) = f(x_0)$.

When these equivalent conditions hold,

$$\sup g(p^*) = g(p_0^*) = L(x_0, p_0^*) = f(x_0) = \inf f, \tag{10.57}$$

so that $L(x_0, p_0^*)$ equals the common optimal value of the primal and dual problems.

Proof. From (10.35) and (10.37) we see that for each x and p,

$$g(p^*) = \inf_x L(x, p^*), \qquad \sup_{p^*} L(x, p^*) = f(x). \tag{10.58}$$

If $x \in \operatorname{dom} f$ then the second assertion in (10.58) says that for each p^* we have $L(x, p^*) \leq f(x) < +\infty$, so $x \in \operatorname{dom}_1 L$ and therefore $\operatorname{dom} f \subset \operatorname{dom}_1 L$. A similar argument using the first assertion of (10.58) shows that $\operatorname{dom} g \subset \operatorname{dom}_2 L$.

(a) implies (b). If (x_0, p_0^*) is a saddle point of L with finite saddle value, then

$$f(x_0) = \sup_{p^*} L(x_0, p^*) = L(x_0, p_0^*) = \inf_x L(x, p_0^*) = g(p_0^*). \tag{10.59}$$

This shows that $f(x_0)$ and $g(p_0^*)$ are both finite, so that $x_0 \in \operatorname{dom} f$ and $p_0^* \in \operatorname{dom} g$. Therefore (b) holds.

(b) implies (a). If (b) holds then $(x_0, p_0^*) \in \operatorname{dom} f \times \operatorname{dom} g \subset \operatorname{dom} L$. Using (10.58) together with the assumption that $f(x_0) = g(p_0^*)$, we have

$$f(x_0) = g(p_0^*) \le L(x_0, p_0^*) \le f(x_0), \tag{10.60}$$

so that all of these values are equal. As $f(x_0) < +\infty$ and $g(p_0^*) > -\infty$, in fact the values are all finite. Using (10.58) once more, we find that for each x and p^*,

$$L(x_0, p^*) \le f(x_0) = L(x_0, p_0^*) = g(p_0^*) \le L(x, p_0^*), \tag{10.61}$$

so that (x_0, p_0^*) is a saddle point with finite saddle value, which proves (a).

If (a) and (b) hold, then we know that $f(x_0) = g(p_0^*)$. For each x and p^*, (10.39) showed that $g(p^*) \le f(x)$, so we have

$$g(p^*) \le f(x_0) = g(p_0^*) \le f(x),$$

showing that x_0 solves the primal problem and p_0^* solves the dual problem. We can then see from (10.60) that the optimal value of each problem equals $L(x_0, p_0^*)$. □

Theorem 10.4.1 shows that if the Lagrangian has a saddle point with finite saddle value, then both primal and dual problems are solvable with a common optimal value, so in that case there certainly is no duality gap. But there may be no duality gap even if it is not the case that both problems have optimal solutions. To see this, it helps to look at the value functions.We have already seen the primal value function $v(p) = \inf_x F(x, p)$ in (10.33), and by analogy we can define the dual value function to be

$$w(x^*) = \sup_{p^*} G(p^*, x^*). \tag{10.62}$$

Proposition 10.4.2. *Suppose that the perturbation function F is closed proper convex. Then*

$$\sup g^* = \operatorname{cl} v(0) \le v(0) = \inf f. \tag{10.63}$$

Accordingly, there is no duality gap if and only if $v(0) = \operatorname{cl} v(0)$.

Proof. The infimum of f is $v(0)$ by definition of v. Moreover, as F is convex v is a convex function by Exercise 6.1.19, so that $-v$ is concave. Then

$$\begin{aligned}(-v)^*(p^*) &= \inf_p\{\langle p^*, p\rangle - (-v)(p)\}\\ &= \inf_{x,p}\{\langle p^*, p\rangle + F(x,p)\}\\ &= \inf_x L(x, p^*)\\ &= g(p^*),\end{aligned} \tag{10.64}$$

so that the dual objective g is $(-v)^*$, and therefore

$$g^* = (-v)^{**} = \operatorname{cl}(-v) = -\operatorname{cl} v.$$

This means that

$$\sup g = -g^*(0) = \operatorname{cl} v(0) \le v(0) = \inf f,$$

which proves (10.63). □

Proposition 10.4.2 gives valuable insight into the possible difference between $\sup g$ and $\inf f$ by showing that any such difference must be exactly the difference between the values of v and of $\operatorname{cl} v$ at the origin. Evidently, there will be no difference if $0 \in \operatorname{ri}\operatorname{dom} v$ (Theorem 6.2.14), and there *may* be no difference even if this condition is not met. However, (10.64) can give us even more information, as we show next.

Proposition 10.4.3. *Suppose that for each x, $F(x,\cdot)$ is closed and convex, and that* $\inf f$ *is finite. Then*

$$-\partial v(0) = \{p^* \mid g(p^*) = \max g\}. \tag{10.65}$$

If $0 \in \operatorname{ri}\operatorname{dom} v$ *then this set is nonempty.*

Proof. Let p be a point at which $v(p)$ is finite. A point p^* belongs to $-\partial v(p)$ exactly when $p^* \in \partial(-v)(p)$ (use Definition 8.1.1, remembering that the inequality goes in opposite directions for convex and for concave functions). In turn, this occurs exactly when

$$(-v)^*(p^*) = \langle p^*, p\rangle - (-v)(p).$$

Rewriting this using (10.64), we see that $p^* \in -\partial v(p)$ if and only if

$$g(p^*) = \langle p^*, p\rangle + v(p).$$

Setting $p = 0$, where v is finite by hypothesis, and recalling that $v(0) = \inf f$, we find that $p^* \in -\partial v(0)$ if and only if $g(p^*) = \inf f$. But as we know that $\sup g \le \inf f$, we see that $-\partial v(0)$ is exactly the set of maximizers of the dual objective function g.

For this set to be nonempty it suffices by Proposition 6.2.13 that the origin belong to the relative interior of $\operatorname{dom} v$. □

Proposition 10.4.3 shows that the condition $0 \in \operatorname{ri}\operatorname{dom} v$ ensures that the dual problem has at least one optimal solution. Moreover, as we saw earlier this condition also suffices to ensure that

$$\sup g = \operatorname{cl} v(0) = v(0) = \inf f,$$

so that there will be no duality gap.

10.4.1 Example: convex nonlinear programming

Let $f_i(x)$ for $i = 0, \ldots, m$ be closed proper convex functions on $\mathbb{R}^n$, with $\cap_{i=0}^m \operatorname{dom} f_i \neq \emptyset$. We consider the constrained optimization problem

$$\inf_x \{ f_0(x) \mid f_i(x) \leq 0,\, i = 1, \ldots, m. \} \tag{10.66}$$

This representation can accommodate a constraint of the form $x \in C$ by including the function I_C among the constraint functions f_i.

We use the perturbation structure

$$F(x,p) = \begin{cases} f_0(x) & \text{if } f_i(x) \leq p_i,\, i = 1, \ldots, m, \\ +\infty & \text{otherwise.} \end{cases} \tag{10.67}$$

One can check directly that the epigraph of F is a closed convex subset of $\mathbb{R}^{n+m+1}$. Further, F cannot take $-\infty$ and it does not always take $+\infty$ (let $x \in \cap_{i=0}^m \operatorname{dom} f_i$ and take the p_i to be large enough so that all constraints are satisfied), so it is a closed proper convex function.

Now a computation using the definition of the Lagrangian shows that

$$L(x,p^*) = \begin{cases} f_0(x) + \sum_{i=1}^m p_i^* f_i(x) & \text{if } p^* \geq 0, \\ +\infty & \text{otherwise,} \end{cases} \tag{10.68}$$

and therefore the corresponding dual problem is that of finding the supremum in p^* of

$$\begin{aligned} g(p^*) &= \inf_x L(x,p^*) \\ &= \begin{cases} \inf_x \{ f_0(x) + \sum_{i=1}^m p_i^* f_i(x) \} & \text{if } p^* \geq 0, \\ +\infty & \text{otherwise.} \end{cases} \end{aligned} \tag{10.69}$$

Depending on the form of the functions f_i, it may or may not be possible to find a simple representation of the dual problem (in particular, one in which the "inf" operator does not explicitly appear).

Here is a condition that will ensure that the dual problem has a solution and that there is no duality gap.

Definition 10.4.4. The problem (10.66) satisfies the *Slater condition* if there is a point $\hat{x} \in \operatorname{dom} f_0$ such that

$$f_i(\hat{x}) < 0, \quad i = 1, \ldots, m.$$

□

To see why the Slater condition works, observe that if P is a sufficiently small neighborhood of the origin in $\mathbb{R}^m$, then for each $p \in P$ we will have

$$f_1(\hat{x}) < p_1, \ldots, f_m(\hat{x}) < p_m,$$

and therefore

$$v(p) = \inf_x F(x,p) \leq F(\hat{x},p) = f_0(\hat{x}) < +\infty.$$

Accordingly, $P \subset \operatorname{dom} v$ so that in fact $0 \in \operatorname{int}\operatorname{dom} v$, which is stronger than the condition $0 \in \operatorname{ri}\operatorname{dom} v$ for applicability of Propositions 10.4.2 and 10.4.3.

One could set up conditions for the primal problem to have a minimizer by applying Propositions 10.4.2 and 10.4.3 to the dual, using the symmetry of the duality framework. However, the dual perturbation structure is frequently not easy to work with. For that reason it may be better simply to impose a condition directly on the primal problem: for example, compactness of a level set.

10.4.2 The augmented Lagrangian

The Lagrangian in (10.68) contains sign constraints on the dual variables p^*, and the dual objective function g inherits these constraints. In computing the Lagrangian one sees that these sign constraints arise, apparently inevitably, from the presence of inequality constraints in the original problem. One might conclude that whenever the problem contains inequality constraints, they will produce explicit sign restrictions on the dual variables. However this is not so, because the form of the Lagrangian (and so also the dual objective) depends not just on the form of the constraints but also on other aspects of the function F.

To illustrate this we take the same primal problem and produce quite a different Lagrangian by altering the form of F. In fact, the alteration will be apparently very slight: for some positive real number r we take

$$F_r(x,p) = \begin{cases} f_0(x) + (r/2)\|p\|^2 & \text{if } f_i(x) \leq p_i,\ i = 1,\ldots,m, \\ +\infty & \text{otherwise.} \end{cases} \tag{10.70}$$

Now in calculating the Lagrangian we have to solve m minimization problems, each of the form

$$\min\{p_i^* p_i + (r/2)p_i^2 \mid f_i(x) \leq p_i\}. \tag{10.71}$$

The unconstrained minimum of the quadratic objective function in (10.71) occurs at $p_i = -r^{-1}p_i^*$, and it has the value $-(2r)^{-1}(p_i^*)^2$. Clearly, we want to make this choice if it is feasible for (10.71): that is, if $p_i^* + rf_i(x) \leq 0$. Otherwise (when $f_i(x)$ is greater than the minimizer of (10.71)) we want to make the quadratic as small as possible, which means setting $p_i = f_i(x)$ so that the quadratic has the value $p_i^* f_i(x) + (r/2)f_i(x)^2$. Therefore we have

$$L(x,p^*) = f_0(x) + \sum_{i=1}^{m} \theta_i(f_i(x), p_i^*, r), \tag{10.72}$$

where

$$\theta_i(\phi_i, p_i^*, r) = \begin{cases} -(2r)^{-1}(p_i^*)^2 & \text{if } p_i^* + r\phi_i \le 0, \\ p_i^*\phi_i + (r/2)\phi_i^2 & \text{otherwise.} \end{cases} \tag{10.73}$$

The quantity defined by (10.72) and (10.73) is the so-called *augmented Lagrangian,* which has received much attention because of its usefulness in computational solution of optimization problems. As we can see from this development, it is just another of the infinitely many possibilities for constructing Lagrangians for the same primal problem by selecting different functional forms for F.

10.5 Notes and references

The general setting of the problems outlined in Section 10.1 follows [34, Part VII], especially Sections 33 and 36.

For the saddle point theorems of Section 10.2, the theorem of Sion is adapted from [41]. The proof given here follows [20]. The proof of Theorem 10.2.1 is adapted from [5, Chapter 8, Section 7].

The duality formulation in Section 10.3 is due to Rockafellar: see [35], which also deals with problems in infinite-dimensional spaces. Many previous authors had dealt with duality in more restricted forms.

A fuller treatment of saddle functions, including equivalence classes, is in [34].

Appendix A
Topics from Linear Algebra

This appendix presents some useful results that are often not taught in elementary linear algebra courses. We give results only for the real field, as it is simpler and we do not need the extra generality.

The initial sections cover linear spaces, the concept of a norm, and a few basic properties of norms, particularly on finite-dimensional linear spaces, including that of $m \times n$ matrices.

We then apply the basic results to derive some structural properties of matrices, including the extremely useful singular value decomposition. As an application we show how to compute the Moore-Penrose generalized inverse. The final section develops properties of affine sets, which are indispensable in the study of convexity. The Moore-Penrose inverse appears again here as a useful tool.

A.1 Linear spaces

A *linear space,* or vector space, is a pair $V = (X, F)$ where

- X is a group whose operation and identity element we denote by $+$ and 0 respectively,
- F is a field whose elements we call *scalars*, and whose additive identity and unit element are 0 and 1 respectively,
- For each $\alpha \in F$ and each $x \in X$ the product αx is defined and is an element of X,

and where the following four properties hold for each x and y in X and each α and β in F:

a. $\alpha(x+y) = \alpha x + \alpha y$;
b. $(\alpha+\beta)x = \alpha x + \beta x$;
c. $\alpha(\beta x) = (\alpha\beta)x$;
d. $1x = x$.

A *real vector space* is one in which $F = \mathbb{R}$. Every vector space in this book is real. We usually take V to be the familiar vector space $\mathbb{R}^n$ in which the elements of X are n-tuples of real numbers $(x_1, \ldots, x_n)$ and $F = \mathbb{R}$. Courses on basic linear algebra develop the most important properties of $\mathbb{R}^n$, and we will use these. We will occasionally take a different X, such as the $m \times n$ real matrices.

If $S = \{x_1, \ldots, x_k\}$ is a finite subset of V, then a *linear combination* of elements of S is an element of the form $\mu_1 x_1 + \ldots + \mu_k x_k$, where $\mu_1, \ldots, \mu_k$ are scalars. The *span* of S is the set

$$\operatorname{span} S = \begin{cases} \text{All linear combinations of elements of } S & \text{if } S \neq \emptyset, \\ \{0\} & \text{otherwise.} \end{cases}$$

V is *finite-dimensional* if there exists a finite subset S of V such that $V = \operatorname{span} S$. Every vector space used in this book is finite-dimensional.

When $F = \mathbb{R}$ the linear space V is an *inner product space* if there is an operation $\langle \cdot, \cdot \rangle : V \times V \to F$ obeying the following for each x, y, and z in X and each α and β in F:

1. $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0$ if and only if $x = 0$;
2. $\langle x, y \rangle = \langle y, x \rangle$;
3. $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$.

If F were to consist of the complex numbers $\mathbb{C}$, this definition would still work if the the right-hand side of the equation in item (2) above were replaced by its complex conjugate.

For elements $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ of $\mathbb{R}^n$ the standard inner product is

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i, \tag{A.1}$$

but there are infinitely many others: for example, $\sum_{i=1}^{n} (x_i)(Ay)_i$, where A is a symmetric positive definite matrix. We will use the notation in (A.1) for the standard inner product, and will introduce other notation if needed for different inner products.

A.1.1 Exercises for Section A.1

Exercise A.1.1. Let A be an $n \times n$ matrix. Prove the following

1. If A is symmetric and positive definite, then the expression

$$[x, y] = \langle x, Ay \rangle \tag{A.2}$$

   defines an inner product on $\mathbb{R}^n$.
2. Under either or both of the following conditions, (A.2) does not define an inner product on $\mathbb{R}^n$: (1) A is not symmetric; (2) A is not positive definite.

A.2 Normed linear spaces

As shown in Section A.1, points in a real vector space V satisfy axioms concerning addition as well as multiplication by scalars. However, without more structure we cannot speak about the size of an element $x \in V$, except to say whether it is or is not zero. One approach to the question of size is to define a norm on that space.

Definition A.2.1. Let V be a real linear space. A *norm* on V is a function $\phi(\cdot) : V \to \mathbb{R}$ having the following properties for each x and y in V and each $\alpha \in \mathbb{R}$:

1. $\phi(x) \geq 0$, and $\phi(x) = 0$ if and only if $x = 0$.
2. $\phi(\alpha x) = |\alpha|\phi(x)$.
3. (the *triangle inequality*) $\phi(x+y) \leq \phi(x) + \phi(y)$.

□

The pair $(V, \|\cdot\|)$ is a *normed linear space*.

Norms are not unique (e.g., multiply by a nonnegative scalar). If $|\cdot|$ is any norm for V, then $|\cdot|$ is a metric on V so $(V, |\cdot|)$ is a metric space.

Definition A.2.2. A function f between two normed linear spaces $(X, |\cdot|)$ and $(Y, \|\cdot\|)$ is *Lipschitz continuous,* or *Lipschitzian,* if there is some nonnegative real number λ such that for each x and x' in X,

$$\|f(x) - f(x')\| \leq \lambda |x - x'|.$$

It is *nonexpansive* if it is Lipschitzian with $\lambda = 1$. □

Lemma A.2.3. *If $(V, \|\cdot\|)$ is a real normed linear space then $\|\cdot\|$ is a nonexpansive function from $(V, \|\cdot\|)$ to $\mathbb{R}$.*

Proof. Let v and v' be elements of V. Then the triangle inequality for $\|\cdot\|$ shows that $\|v\| \leq \|v - v'\| + \|v'\|$. Interchanging v and v' we find that $\|v'\| \leq \|v' - v\| + \|v\|$. Therefore

$$|\,\|v\| - \|v'\|\,| \leq \|v - v'\|,$$

which shows that $\|\cdot\|$ is nonexpansive. □

If an inner product $[\cdot, \cdot]$ is available on V, then it yields a very useful norm defined by

$$\|v\| = [v, v]^{1/2}. \tag{A.3}$$

In order to verify the properties of that norm, we first establish the Schwarz inequality.

Lemma A.2.4. *Let V be a real linear space having an inner product $[\cdot, \cdot]$, and for $v \in V$ let $\|v\|$ be defined by* (A.3)*. If x and y are any elements of V, then*

$$|[x, y]| \leq \|x\|\|y\|, \tag{A.4}$$

with equality if and only if one of x and y is a scalar multiple of the other.

Proof. Choose x and y in V; if $y = 0$ then each side of (A.4) is zero and y is a scalar multiple of x, so the result holds in that case. If $y \neq 0$ then for each $\alpha \in \mathbb{R}$ the square of the norm of $x + \alpha y$ is nonnegative:

$$0 \leq \|x+\alpha y\|^2 = [x+\alpha y, x+\alpha y] = \|x\|^2 + 2\alpha[x,y] + \alpha^2\|y\|^2. \tag{A.5}$$

If we take $\alpha = -[x,y]/\|y\|^2$ then (A.5) becomes

$$0 \leq \|x+\alpha y\|^2 = \|x\|^2 - 2[x,y]^2/\|y\|^2 + [x,y]^2/\|y\|^2 = \|y\|^{-2}(\|x\|^2\|y\|^2 - [x,y]^2), \tag{A.6}$$

so that (A.4) holds. Moreover, if one of x and y is a scalar multiple of the other then the two sides of (A.4) are equal, but otherwise the right-hand side of (A.6) must be positive because $\|x+\alpha y\| > 0$, and then strict inequality must hold in (A.4). □

Proposition A.2.5. *If V is a real linear space having an inner product $[\cdot,\cdot]$, then the function $\|\cdot\|$ defined in* (A.3) *is a norm on V.*

Proof. The properties of the inner product show that $\|u\|$ is nonnegative for each $u \in V$, and is zero if and only if $u = 0$. They also show that for any scalar α,

$$\|\alpha u\| = [\alpha u, \alpha u]^{1/2} = |\alpha|\|u\|.$$

To verify the triangle inequality, apply the Schwarz inequality to obtain

$$\|u+v\|^2 = \|u\|^2 + 2[u,v] + \|v\|^2 \leq \|u\|^2 + 2\|u\|\|v\| + \|v\|^2 = (\|u\| + \|v\|)^2,$$

and take the nonnegative square root on each side. This also shows that for the norm $\|\cdot\|$, the triangle inequality is strict unless one of u and v is a scalar multiple of the other. This does not hold for norms in general. □

If $V = \mathbb{R}^n$ and $[\cdot,\cdot]$ is the standard inner product $\langle\cdot,\cdot\rangle$, then the norm $\|\cdot\|$ constructed above is the *Euclidean norm,* given by $\|x\| = (\sum_{i=1}^n x_i^2)^{1/2}$. It is also called the *2-norm* or l_2 norm.

Two other norms commonly used on $\mathbb{R}^n$ are the *maximum norm* or l_∞ norm, given by

$$\|x\|_\infty = \max_{i=1}^n |x_i|, \tag{A.7}$$

and the l_1 norm, given by

$$\|x\|_1 = \sum_{i=1}^n |x_i|. \tag{A.8}$$

The presence of these other norms raises the question of whether one might define some inner products on $\mathbb{R}^n$ that would yield these norms via (A.3). The following theorem characterizes those real normed linear spaces in which an inner product generates the norm.

Theorem A.2.6. *Let $(V, \|\cdot\|)$ be a real normed linear space. The following are equivalent.*

a. An inner product $[\cdot,\cdot]$ *exists on* V *with the property that for each* $v \in V$

$$[v,v] = \|v\|^2. \tag{A.9}$$

b. For each u *and* v *in* V*,*

$$\|u+v\|^2 + \|u-v\|^2 = 2(\|u\|^2 + \|v\|^2). \tag{A.10}$$

If these equivalent properties hold, then the inner product in (a) is uniquely defined for each u *and* v *in* V *by*

$$[u,v] = (1/4)(\|u+v\|^2 - \|u-v\|^2). \tag{A.11}$$

Proof. (a) implies (b). Suppose $[\cdot,\cdot]$ exists and satisfies (A.9). Then for each u and v in V,

$$\|u+v\|^2 + \|u-v\|^2 = [u+v,u+v] + [u-v,u-v] = 2[u,u] + 2[v,v] = 2(\|u\|^2 + \|v\|^2). \tag{A.12}$$

Therefore (A.10) holds. Further,

$$\|u+v\|^2 - \|u-v\|^2 = [u+v,u+v] - [u-v,u-v] = 4[u,v], \tag{A.13}$$

so (A.11) uniquely defines the inner product.

(b) implies (a). Now suppose that (A.10) holds for each u and v in V. Define $[u,v]$ by (A.11); we show that it satisfies (A.9) and obeys the requirements for an inner product. By taking $u = v$ in (A.11) we find that $[u,u] = \|u\|^2$ for each $u \in V$, so that (A.9) holds. The properties of the norm then show that $[u,u]$ is always nonnegative and is zero if and only if $u = 0$, and that $[u,v] = [v,u]$ for each u and v.

Next, choose u, u', and v in V, and let

$$s = (u+u')/2, \qquad s' = (u-u')/2. \tag{A.14}$$

Then (A.10) yields

$$\begin{aligned} \|(s+v)+s'\|^2 + \|(s+v)-s'\|^2 &= 2(\|s+v\|^2 + \|s'\|^2) \\ \|(s-v)+s'\|^2 + \|(s-v)-s'\|^2 &= 2(\|s-v\|^2 + \|s'\|^2). \end{aligned} \tag{A.15}$$

Subtract the second equation from the first to get

$$\begin{aligned} (\|(s+s')+v\|^2 - \|(s+s')-v\|^2) + (\|(s-s')+v\|^2 - \|(s-s')-v\|^2) \\ = 2(\|s+v\|^2 - \|s-v\|^2), \end{aligned}$$

and then use (A.11) to rewrite this as

$$4([s+s',v] + [s-s',v]) = 8[s,v].$$

Finally, divide by 4 and use (A.14) to bring this to the form

$$[u,v]+[u',v]=[u+u',v], \tag{A.16}$$

which shows that part of the third requirement for an inner product holds. To complete the proof we show that for each real α and each u and v in V, $[\alpha u,v]=\alpha[u,v]$. Let

$$C:=\{\alpha\in\mathbb{R}\mid \text{for each } u \text{ and } v \text{ in } V,\ [\alpha u,v]=\alpha[u,v]\}.$$

We will show that $C=\mathbb{R}$.

First, the definition of C shows that it contains 1. From (A.11) we see that $0=[0,v]$ and then (A.16) shows that

$$0=[0,v]=[u+(-u),v]=[u,v]+[-u,v],$$

so that $[-u,v]=-[u,v]$ and therefore $-1\in C$. If α and β belong to C then

$$[(\alpha+\beta)u,v]=[\alpha u+\beta u,v]=[\alpha u,v]+[\beta u,v]=\alpha[u,v]+\beta[u,v]=(\alpha+\beta)[u,v],$$

so that $\alpha+\beta\in C$, and an induction shows that any finite sum of elements of C is in C. As any integer is a finite sum of the numbers $+1$ and -1, all of the integers belong to C. Now if p and q are two integers with $q\neq 0$, we have

$$[(p/q)u,v]=q^{-1}q[(p/q)u,v]=q^{-1}[pu,v]=(p/q)[u,v],$$

so that every rational number is in C.

Lemma A.2.3 shows that the norm is a Lipschitz continuous function from $(V,\|\cdot\|)$ to $\mathbb{R}$. Then (A.11) shows that the function $[\cdot,\cdot]$ is continuous from $V\times V$ to $\mathbb{R}$. Now let $\alpha\in\mathbb{R}$ and let $\{\rho_k\}$ be a sequence of rationals converging to α. Then for each k, $[\rho_k u,v]=\rho_k[u,v]$ by what we have already shown. Taking the limit and, on the left, using the continuity of $[\cdot,\cdot]$ we obtain $[\alpha u,v]=\alpha[u,v]$. This shows that in fact $C=\mathbb{R}$ as required. □

Although any norm on a linear space V is a metric, it is not obvious whether the metric topologies are all the same or whether they differ for different norms. The key point here is the dimensionality of the space: if the space is finite-dimensional then they are all the same, whereas if it is infinite-dimensional then they may differ. The following definition and theorem show why this is so.

Definition A.2.7. Let V be a linear space and let $|\cdot|_a$ and $|\cdot|_b$ be norms on V. These norms are *equivalent* if there exist positive constants α and β such that for each $v\in V$,

$$\alpha|v|_a\leq|v|_b\leq\beta|v|_a. \tag{A.17}$$

□

If the norms $|\cdot|_a$ and $|\cdot|_b$ on V are equivalent, let v be a point of V and X be a neighborhood of v in $(V,|\cdot|_a)$. Then by taking a small enough positive ε one can always arrange for the ε-ball about v in $(V,|\cdot|_b)$ to be contained in X. The equivalence of two norms thus implies that the metric topologies associated with those norms are the same.

The following theorem gives some fundamental properties of a finite-dimensional linear space; then a corollary shows that any two norms on such a space are equivalent. An *isometry* between two linear spaces is a map that preserves distance, and an *isomorphism* from one onto the other is a linear map with a single-valued linear inverse.

Theorem A.2.8. *If V is a real linear space having finite dimension n, then there is an isomorphism x of $\mathbb{R}^n$ onto V. If $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^n$, then the function*

$$|u| := \|x^{-1}(u)\|$$

is a norm on V, and x is then an isometry from $(\mathbb{R}^n, \|\cdot\|)$ to $(V, |\cdot|)$. Further, the unit sphere

$$S := \{u \in V \mid |u| = 1\}$$

of $(V, |\cdot|)$ is compact.

Proof. Let $\dim V = n$ and choose a basis $v_1, \ldots, v_n$. For $\xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n$ define $x(\xi) = \sum_{i=1}^n \xi_i v_i$. Then $x(\cdot)$ is a linear map from $\mathbb{R}^n$ onto V, because every $x \in V$ is a linear combination of the v_i. As the v_i are linearly independent the coefficients in such a linear combination are unique, so that the map x is injective and it has a single-valued linear inverse defined on V by

$$x^{-1}(\sum_{i=1}^n \xi_i v_i) = (\xi_1, \ldots, \xi_n).$$

Therefore x is an isomorphism of $\mathbb{R}^n$ onto V.

Now define a real-valued function $|\cdot|$ on V by $|v| = \|x^{-1}(v)\|$, where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^n$. This function satisfies the conditions for a norm on V, so V is a normed linear space. Moreover, if ξ and ξ' are points of $\mathbb{R}^n$ then

$$|x(\xi) - x(\xi')| = \|x^{-1}[x(\xi)] - x^{-1}[x(\xi')]\| = \|\xi - \xi'\|, \tag{A.18}$$

so that x and hence also x^{-1} are isometries between $(\mathbb{R}^n, \|\cdot\|)$ and $(V, |\cdot|)$, hence in particular Lipschitz continuous.

A point $u \in V$ has $|u| = 1$ if and only if $u = x(\xi)$ and $\|\xi\| = 1$. Therefore S is the image $x(S^{n-1})$ of the unit sphere

$$S^{n-1} = \{\xi \in \mathbb{R}^n \mid \|\xi\| = 1\}$$

of $\mathbb{R}^n$ under the map x. But S^{n-1} is a compact subset of $\mathbb{R}^n$ and x is continuous, so the image S is compact. □

Without the finite dimensionality requirement the unit sphere of a normed linear space may not be compact.

Corollary A.2.9. *If V is a finite-dimensional real linear space, then any two norms on V are equivalent.*

Proof. Let $\dim V = n$ and choose a basis $\{v_1, \ldots, v_n\}$ for V; let the function x be as defined in Theorem A.2.8. That theorem shows that V has a norm $|\cdot|$. Write $\|\cdot\|$ for the Euclidean norm on $\mathbb{R}^n$. We know from Lemma A.2.3 that $|\cdot|$ is Lipschitz continuous (in fact, nonexpansive) on V, We will now show that *any* norm on V is Lipschitz continuous in the metric topology of $(V, |\cdot|)$.

Let $|\cdot|_a$ be any norm on V and define $\mu_+ = (\sum_{i=1}^n |v_i|_a^2)^{1/2}$. Choose ξ and ξ' in $\mathbb{R}^n$ and define $u = x(\xi)$ and $u' = x(\xi')$. Then

$$\begin{aligned} |\,|u|_a - |u'|_a\,| \le |x(\xi) - x(\xi')|_a &= |\sum_{i=1}^n (\xi_i - \xi_i')v_i|_a \\ &\le \sum_{i=1}^n |\xi_i - \xi_i'|\,|v_i|_a \le \mu_+ \|\xi - \xi'\| = \mu_+ |u - u'|, \end{aligned} \tag{A.19}$$

where the first inequality follows from Lemma A.2.3, the second from the triangle inequality for $|\cdot|_a$, and the third from the Schwarz inequality applied to the n-tuples of real numbers involved in the sum. The last equality is from Theorem A.2.8. Then (A.19) shows that $|\cdot|_a$ is a (Lipschitz) continuous function on $(V, |\cdot|)$.

Theorem A.2.8 says that the unit sphere $S := \{u \in V \mid |u| = 1\}$ of $(V, |\cdot|)$ is compact. The function $|u|_a/|u|$ is continuous on S in the topology of $(V, |\cdot|)$, so it attains a maximum α_+ and a minimum α_- on S. Moreover, α_- must be positive because otherwise the point $u_- \in S$ at which it is attained would have $|u_-|_a = 0$ and would therefore be zero, contradicting $|u_-| = 1$. For any nonzero $u \in V$ we have $u/|u| \in S$, so

$$\alpha_- \le |\,u/|u|\,|_a \le \alpha_+,$$

and therefore

$$\alpha_- |u| \le |u|_a \le \alpha_+ |u|. \tag{A.20}$$

However, (A.20) holds also for $u = 0$ and therefore it holds for every $u \in V$.

If $|\cdot|_b$ is a norm on V then the same argument produces positive constants β_- and β_+ such that

$$\beta_- |u| \le |u|_b \le \beta_+ |u|. \tag{A.21}$$

Rearrangement of (A.20) and (A.21) yields

$$\alpha_+^{-1} \beta_- |u|_a \le |u|_b \le \alpha_-^{-1} \beta_+ |u|_a,$$

so that $|\cdot|_a$ and $|\cdot|_b$ are equivalent. □

Corollary A.2.9 shows that all of the norms on a finite-dimensional linear space define the same metric topology. Accordingly, we are free to use whichever norm is most convenient for a particular application, with no concern about the effect of the choice on the topology of the space.

A.2.1 Exercises for Section A.2

Exercise A.2.10. For each of the norms $\|\cdot\|_\infty$ and $\|\cdot\|_1$ determine the dimensions n, if any, for which there is an inner product $[\cdot,\cdot]$ on $\mathbb{R}^n$ such that for each x, $[x,x]$ is the square of the norm of x. Prove your assertions.

Exercise A.2.11. Given a positive dimension n, find constants μ_- and μ_+ which may depend on n, such that for each $x \in \mathbb{R}^n$

$$\mu_- \|x\|_\infty \le \|x\| \le \mu_+ \|x\|_\infty, \tag{A.22}$$

where $\|\cdot\|$ is the Euclidean norm. Show that your constants are sharp: that is, show that with your μ_- and μ_+, for each of the two inequalities in (A.22) there is a nonzero point of $\mathbb{R}^n$ at which that inequality becomes an equality.

A.2.2 Notes and references

The characterization of inner product spaces in Theorem A.2.6 is due to Jordan and von Neumann [18, Theorem I]. Their characterization is for complex spaces; the proof we give here for real spaces is slightly simpler.

A.3 Linear spaces of matrices

Write $\mathbb{R}^{m\times n}$ for the set of real $m \times n$ matrices, with each of m and n being at least 1. When equipped with the usual operations of matrix addition and of multiplication by scalars, this becomes a linear space in which the zero element is the zero matrix. The resulting space is of dimension mn because the matrices A_{ij} having 1 in position ij and zero elsewhere form a basis.

A useful inner product on $\mathbb{R}^{m\times n}$, defined below, employs the *trace* of a square matrix. The trace is the sum of the diagonal elements, but it is also the sum of the eigenvalues of the matrix. To see why, let the matrix be $F \in \mathbb{R}^{n\times n}$; then the eigenvalues $\lambda_1, \ldots, \lambda_n$ are the zeros of the characteristic polynomial

$$P(\lambda) = \det(\lambda I_n - F) = \Pi_{i=1}^n (\lambda - \lambda_i). \tag{A.23}$$

The product on the right in (A.23) shows that the coefficient of λ^{n-1} in $P(\lambda)$ must be

$$\alpha := -\sum_{i=1}^n \lambda_i. \tag{A.24}$$

However, the only way in which a product containing λ^{n-1} can occur in the expansion of the determinant in (A.23) is if $n-1$ of the elements come from diagonal

entries $(\lambda - F_{ii})$ in which λ is chosen for inclusion in the product, and the nth element comes from the remaining diagonal entry but with $-F_{ii}$ selected instead of λ. This element must come from the remaining diagonal entry because no two elements in such a product can come from the same row or the same column, and all of the other rows and columns have already been used for the first $n-1$ terms. Then the product is $-F_{ii}\lambda^{n-1}$, and as we can select any i for the element $-F_{ii}$ there are n of these products. Accordingly, we have $\alpha = -\sum_{i=1}^{n} F_{ii}$, which together with (A.24) implies

$$\operatorname{tr} F = \sum_{i=1}^{n} F_{ii} = \sum_{i=1}^{n} \lambda_i.$$

Definition A.3.1. The *standard inner product* on $\mathbb{R}^{m\times n}$ assigns to points A and B the number $\langle A,B\rangle = \operatorname{tr}(AB^*)$.

The trace of the $m \times m$ matrix AB^* has the form

$$\operatorname{tr}(AB^*) = \sum_{i=1}^{m} (AB^*)_{ii} = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}B_{ij}. \tag{A.25}$$

Equation (A.25) shows that $\langle A,B\rangle = \langle B,A\rangle$, and that the inner product is linear in the first argument with the second held constant (and vice versa). Setting $B = A$ we see that

$$\langle A,A\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2; \tag{A.26}$$

hence $\langle A,A\rangle$ is always nonnegative and is zero if and only if $A = 0$. Therefore the quantity in Definition A.3.1 really is an inner product on $\mathbb{R}^{m\times n}$.

As $\langle\cdot,\cdot\rangle$ is an inner product on $\mathbb{R}^{m\times n}$, Proposition A.2.5 says that it defines a norm on that space. From (A.26) we see that this norm must be

$$\|A\|_F := [A,A]^{1/2} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2\right)^{1/2}, \tag{A.27}$$

which is usually called the *Frobenius norm.* This definition shows that for any such A we have $\|A\|_F = \|A^*\|_F$.

The Frobenius norm is *submultiplicative:* that is, it satisfies the property in (A.28) below.

Proposition A.3.2. *Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times k}$. Then*

$$\|AB\|_F \le \|A\|_F \|B\|_F. \tag{A.28}$$

The three Frobenius norms in (A.28) are in three different spaces: $\mathbb{R}^{m\times k}$, $\mathbb{R}^{m\times n}$, and $\mathbb{R}^{n\times k}$.

Proof. Write $a_{i\cdot}$ for the ith row of A and $b_{\cdot j}$ for the jth column of B. The element $(AB)_{ij}$ of the product is $\langle a_{i\cdot}, b_{\cdot j}\rangle$, and by the Schwarz inequality (in $\mathbb{R}^n$) its absolute

value is not greater than the product $\|a_{i\cdot}\|\|b_{\cdot j}\|$ of the 2-norms of the two vectors. Therefore

$$\|AB\|_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} |(AB)_{ij}|^2 \le \sum_{i=1}^{m}\sum_{j=1}^{n} \|a_{i\cdot}\|^2\|b_{\cdot j}\|^2 = (\sum_{i=1}^{m}\|a_{i\cdot}\|^2)(\sum_{j=1}^{n}\|b_{\cdot j}\|^2)$$
$$= \|A\|_F^2\|B\|_F^2.$$

□

Definition A.3.3. Let $|\cdot|_M$ be a norm on $\mathbb{R}^{m\times n}$ and let $|\cdot|_m$ and $|\cdot|_n$ be vector norms on $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. The triple $(|\cdot|_m, |\cdot|_M, |\cdot|_n)$ is *consistent* if for each $A \in \mathbb{R}^{m\times n}$ and each $x \in \mathbb{R}^n$,

$$|Ax|_m \le |A|_M|x|_n. \tag{A.29}$$

If a norm $|\cdot|$ is defined on $\mathbb{R}^n$ for each n, as is the Euclidean norm, and a matrix norm $\|\cdot\|$ is defined on $\mathbb{R}^{m\times n}$ for each pair (m,n), then if for each such pair the triple $(|\cdot|, \|\cdot\|, |\cdot|)$ is consistent we say the matrix norm $\|\cdot\|$ is *subordinate* to the vector norm $|\cdot|$.

Corollary A.3.4. *The Frobenius norm is subordinate to the Euclidean vector norm.*

Proof. Take $k = 1$ in Proposition A.3.2 and observe that the Frobenius norm of an $n \times 1$ matrix is the same as the Euclidean norm of the n-vector having the same components. □

The Frobenius norm is important because it comes from the standard inner product of Definition A.3.1, but it sometimes produces surprising results. For example, if we take $m = n$ and $A = I_n$ in (A.29) we obtain the bound $\|x\| \le n^{1/2}\|x\|$, which is surely true but not particularly useful. For that reason we will investigate some other possible norms for $\mathbb{R}^{m\times n}$. The first of these is the very important 2-norm of A, also called the *spectral norm.* We usually write $\|A\|$ for this norm, but if multiple norms are in use we write $\|A\|_2$.

Definition A.3.5. The 2-norm of $A \in \mathbb{R}^{m\times n}$ is

$$\|A\| := \max_{x\in S^{n-1}} \|Ax\|.$$

□

We can write max instead of sup because S^{n-1} is compact and the Euclidean vector norm $\|\cdot\|$ is continuous.

Definition A.3.5 shows that $\|A\|$ is nonnegative, that it equals zero if and only if $A = 0$, and that $\|\alpha A\| = |\alpha|\|A\|$ for any scalar α. A calculation using the definition along with the triangle inequality for the vector norm also shows that if $C \in \mathbb{R}^{m\times n}$ then

$$\|A + C\| \le \|A\| + \|C\|.$$

Therefore $(\mathbb{R}^{m\times n}, \|\cdot\|)$ is a normed linear space.

Definition A.3.5 also shows that for each $x \in \mathbb{R}^n$ one has $\|Ax\| \le \|A\|\|x\|$, so the 2-norm on $\mathbb{R}^{m\times n}$ is subordinate to the Euclidean vector norm. Also, if $D \in \mathbb{R}^{m\times k}$ and $E \in \mathbb{R}^{k\times n}$, then for each $x \in S^{n-1}$,

$$\|(DE)x\| = \|D(Ex)\| \le \|D\|\|Ex\| \le \|D\|\|E\|\|x\| = \|D\|\|E\|,$$

so that $\|DE\| \le \|D\|\|E\|$ and therefore $\|\cdot\|$ is submultiplicative.

To develop an additional characterization of the 2-norm, recall that the elements of a set $v_1, \ldots, v_m$ of points in $\mathbb{R}^n$ are *orthonormal* if for each i and j, $\langle v_i, v_j\rangle = \delta_{ij}$, where the right-hand side is the *delta function* taking the values 0 if $i \ne j$ and 1 if $i = j$. If we let V be the element of $\mathbb{R}^{m\times n}$ having columns $v_1, \ldots, v_m$, then V^*V is the identity I_m of $\mathbb{R}^m$. The columns of V must be linearly independent, because if $0 = Vz$ then $0 = V^*Vz = z$.

If $m = n$ then we call such a V an *orthogonal matrix*. In that case the columns of V form a basis of $\mathbb{R}^n$. Given any $x \in \mathbb{R}^n$ we can then find $y \in \mathbb{R}^n$ with $Vy = x$; then

$$(VV^*)x = V(V^*V)y = Vy = x,$$

so that $VV^* = I_n = V^*V$. The first of these equalities shows that the rows of V are also orthogonal.

Matrices with orthonormal columns have the very convenient property that they do not change the Euclidean norm of a vector.

Lemma A.3.6. *Let $A \in \mathbb{R}^{k\times n}$ have orthonormal columns $a_1, \ldots, a_n$. Then $k \ge n$ and for each $x \in \mathbb{R}^n$,*

$$\|Ax\| = \|x\|. \tag{A.30}$$

Proof. For each $x \in \mathbb{R}^n$ we have

$$\|Ax\|^2 = \langle Ax, Ax\rangle = \langle x, A^*Ax\rangle. \tag{A.31}$$

However, the element $(A^*A)_{ij}$ is $\langle a_i^*, a_j\rangle = \delta_{ij}$, so that $A^*A = I_n$, the identity of $\mathbb{R}^n$. Then (A.31) yields $\|Ax\|^2 = \|x\|^2$ and therefore (A.30). But (A.30) also shows that A has full column rank, so we have $k \ge n$. □

Proposition A.3.7. *Let $A \in \mathbb{R}^{m\times n}$. Then $\|A\|^2$ is the largest eigenvalue of A^*A.*

Proof. If $x \in \mathbb{R}^n$ then $\langle x, A^*Ax\rangle = \|Ax\|^2$, so A^*A is positive semidefinite. It is also symmetric, so there exist an orthogonal matrix $Q \in \mathbb{R}^{n\times n}$ and a diagonal matrix $\Lambda \in \mathbb{R}^{n\times n}$ with nonnegative diagonal elements $\lambda_1 \ge \ldots, \ge \lambda_n$ (the eigenvalues of A^*A) such that $A^*A = Q\Lambda Q^*$. Then for each $x \in \mathbb{R}^n$,

$$\begin{aligned}\|Ax\|^2 = \langle Ax, Ax\rangle = \langle x, A^*Ax\rangle = \langle x, Q\Lambda Q^*x\rangle = \langle Q^*x, \Lambda(Q^*x)\rangle \\ = \sum_{i=1}^n \lambda_i (Q^*x)_i^2 \le \lambda_1 \sum_{i=1}^n (Q^*x)_i^2 = \lambda_1 \|Q^*x\|^2 = \lambda_1 \|x\|^2,\end{aligned} \tag{A.32}$$

where the last equality is from Lemma A.3.6. It follows that the maximum of $\|Ax\|^2$ for $x \in S^{n-1}$ is not greater than λ_1. However, if the orthonormal columns of Q are $q_1, \ldots, q_n$ then $q_1 \in S^{n-1}$ and

$$\|Aq_1\|^2 = \langle q_1, A^*Aq_1 \rangle = \langle q_1, \lambda_1 q_1 \rangle = \lambda_1,$$

showing that the maximum is exactly λ_1, the largest eigenvalue of A^*A. □

Corollary A.3.8. *If $Q \in \mathbb{R}^{n\times n}$ is an orthogonal matrix, then $\|Q\| = 1$.*

Proof. Proposition A.3.7 says that $\|Q\|^2$ is the largest eigenvalue of Q^*Q, which is the identity and therefore has all of its eigenvalues equal to 1. □

We have seen already that an orthogonal matrix does not change the length of a vector. A somewhat similar property holds for matrices.

Proposition A.3.9. *Let $A \in \mathbb{R}^{m\times n}$, and let $P \in \mathbb{R}^{m\times m}$ and $Q \in \mathbb{R}^{n\times n}$ be orthogonal matrices. Then $\|PAQ^*\| = \|A\|$.*

Proof. Let $x \in S^{n-1}$; then

$$\|Ax\| = \|P(Ax)\| = \|(PAQ^*)(Qx)\| \le \|PAQ^*\|\|Qx\| = \|PAQ^*\|,$$

where the first equality holds because P does not change the length of Ax and the last holds because Q does not change the length of x. Therefore $\|A\| \le \|PAQ^*\|$. But also

$$\|PAQ^*\| \le \|P\|\,\|A\|\,\|Q^*\| = \|A\|,$$

where the equality is from Corollary A.3.8. □

If we were to apply Proposition A.3.7 to A^* then we would find that $\|A^*\|^2$ is the largest eigenvalue of the matrix AA^*, which in general is not equal to A^*A and is not even of the same dimensions unless $m = n$. The following result shows that this difference does not matter.

Theorem A.3.10. *If $A \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{n\times m}$, then the nonzero eigenvalues of AC and of CA are the same in value and in multiplicity.*

Proof. Let I_k be the identity matrix of $\mathbb{R}^k$ and define matrices F and G in $\mathbb{R}^{(n+m)\times(n+m)}$ by

$$F = \begin{bmatrix} \tau I_m & A \\ C & I_n \end{bmatrix}, \qquad G = \begin{bmatrix} I_m & 0 \\ -C & \tau I_n \end{bmatrix}.$$

Then

$$FG = \begin{bmatrix} \tau I_m - AC & \tau A \\ 0 & \tau I_n \end{bmatrix}, \qquad GF = \begin{bmatrix} \tau I_m & A \\ 0 & \tau I_n - CA \end{bmatrix},$$

and taking determinants yields

$$\tau^n \det(\tau I_m - AC) = \det FG = \det GF = \tau^m \det(\tau I_n - CA). \tag{A.33}$$

If λ is a nonzero eigenvalue of one of AC and CA with multiplicity k then $(\tau-\lambda)^k$ is a factor of the characteristic polynomial of that matrix. We see from (A.33) that it must also be a factor of the characteristic polynomial of the other matrix, so any nonzero eigenvalue of one matrix is a nonzero eigenvalue of the other with the same multiplicity. □

Corollary A.3.11. *If $A \in \mathbb{R}^{m\times n}$ then $\|A^*\| = \|A\|$.*

Proof. Proposition A.3.7 says that $\|A^*\|^2$ is the largest eigenvalue of AA^*, and that $\|A\|^2$ is the largest eigenvalue of A^*A. Theorem A.3.10 says that these are the same. □

A.3.1 Exercises for Section A.3

Exercise A.3.12. Suppose we define a function on $\mathbb{R}^{m\times n}$ by $\|A\|_X := \max_{i=1}^m \max_{j=1}^n |A_{ij}|$.

a. Show that $\|\cdot\|_X$ is a matrix norm.
b. Exhibit matrices A and B conformable for multiplication, with $\|AB\|_X > \|A\|_X\|B\|_X$.

A.3.2 Notes and references

The proof of Theorem A.3.10 given here is due to Josef Schmid [38].

A.4 The singular value decomposition of a matrix

The singular value decomposition (SVD) of a general real matrix is a versatile and useful tool both for analysis and for computation. This section establishes the existence of the SVD and some of its properties.

Theorem A.4.1. *Let m and n be positive integers and $A \in \mathbb{R}^{m\times n}$. Then there are orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ such that*

$$A = U\Sigma V^*, \tag{A.34}$$

where the diagonal elements $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$ of Σ, called the singular values *of A, are nonnegative and nonincreasing, and all other elements of Σ are zero.*

Proof. If $A = 0$ we can take $U = I_m$, $V = I_n$, and $\Sigma = 0$. Therefore suppose that A has at least one nonzero element. There is also no loss of generality in assuming that $m \le n$, as otherwise we can work with A^*.

The matrix AA^* is symmetric and positive semidefinite, so there exist an orthogonal matrix $U \in \mathbb{R}^{m\times m}$ and a positive semidefinite diagonal matrix $\Delta \in \mathbb{R}^{m\times m}$ such

that $AA^* = U\Lambda U^*$, where the diagonal elements $\lambda_1, \ldots, \lambda_m$ of Λ are nonincreasing. Of these, assume that $\lambda_1, \ldots, \lambda_r$ are positive, and let those elements form the diagonal of a positive definite matrix $\Delta \in \mathbb{R}^{r\times r}$; then

$$\Lambda = \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix}.$$

If $r = m$ then some of the blocks in the following derivation disappear, but that does not affect the argument.

Let

$$T := A^*U = \begin{bmatrix} T_1 & T_2 \end{bmatrix},$$

where $T_1 \in \mathbb{R}^{n\times r}$ and $T_2 \in \mathbb{R}^{n\times(m-r)}$. Then

$$\begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} = \Lambda = U^*AA^*U = T^*T = \begin{bmatrix} T_1^*T_1 & T_1^*T_2 \\ T_2^*T_1 & T_2^*T_2 \end{bmatrix}. \tag{A.35}$$

This shows that $T_2^*T_2 = 0$, but as $\operatorname{tr} T_2^*T_2 = \|T_2\|_F^2$ this implies that $T_2 = 0$.

Also, (A.35) shows that $T_1^*T_1 = \Delta$, so that

$$I_r = (T_1\Delta^{-1/2})^*(T_1\Delta^{-1/2}),$$

and therefore $V_1 := T_1\Delta^{-1/2}$ has orthonormal columns. Here and later in the proof we always use the nonnegative square root.

Adjoin to V_1 a matrix $V_2 \in \mathbb{R}^{n\times(n-r)}$ such that the resulting matrix $V := [V_1\ V_2]$ is orthogonal. Then

$$U^*AV = T^*V = \begin{bmatrix} T_1^*T_1\Delta^{-1/2} & T_1^*V_2 \\ T_2^*T_1\Delta^{-1/2} & T_2^*V_2 \end{bmatrix} = \begin{bmatrix} \Delta^{1/2} & 0 \\ 0 & 0 \end{bmatrix}, \tag{A.36}$$

where we used the relations

$$T_1^*T_1 = \Delta, \qquad T_2 = 0, \qquad T_1^*V_2 = \Delta^{1/2}V_1^*V_2 = 0.$$

Defining Σ to be the matrix on the right in (A.36), we obtain $U^*AV = \Sigma$, so that (A.34) holds. The singular values in Σ are nonnegative, and are nonincreasing because their squares $\lambda_1, \ldots, \lambda_m$ were nonincreasing. □

The matrices U and V appearing in the SVD are generally not unique. For example, if $A = I_n$ then for any orthogonal $Q \in \mathbb{R}^{n\times n}$, $A = Q(I_n)Q^*$.

Although the method used in the proof of Theorem A.4.1 provides a fairly transparent derivation, in general one should not use it for computation because calculating AA^* and A^*A can introduce needless inaccuracy [13, Example 5.3.2]. For discussion of some SVD algorithms, see [13, Section 5.4.5]. Good software is available: one example is the routine `_GESVD` of LAPACK, available through Netlib [26].

Corollary A.4.2. *The singular values of A are unique, and are the nonnegative square roots of the* $\min\{m,n\}$ *largest eigenvalues of AA^* or equivalently of A^*A.*

The columns of U and of V are complete sets of eigenvectors for AA^* and for A^*A respectively.

Proof. Use (A.34) to obtain $AA^* = U\Lambda U^*$, where $\Lambda := \Sigma^2$. This shows that $\sigma_1^2, \ldots, \sigma_m^2$ are the eigenvalues of AA^* in nonincreasing order, and that the columns of U form a complete set of eigenvectors of AA^*. Uniqueness of the singular values follows, because they are nonnegative and because there is a unique list of eigenvalues in nonincreasing order. A similar calculation using A^*A shows that all of its nonzero eigenvalues are included in $\sigma_1^2, \ldots, \sigma_m^2$, and that the columns of V form a complete set of eigenvectors. □

In the proof of Theorem A.4.1 we used U to compute V. In situations where the nonzero eigenvalues of AA^* are not distinct it may not be possible to start with arbitrary choices of eigenvectors U of AA^* and V of A^*A and then to obtain (A.34) from these (Exercise A.4.4).

Corollary A.4.3. *Let r be the maximum index i for which $\sigma_i > 0$, or be zero if $\Sigma = 0$. This r is then the rank of A, and if $r \neq 0$ then* $\mathrm{im}A$ *is the span of the first r columns of U and* $\ker A$ *is the span of the last $n - r$ columns of V.*

Proof. If $r = 0$ then $A = 0$ and there is nothing to prove. If $r > 0$ use (A.34) to obtain $\mathrm{im}A \subset \mathrm{im}U\Sigma$. For the opposite inclusion, suppose $t \in \mathrm{im}U\Sigma$, so that for some y,

$$t = U\Sigma y = (U\Sigma V^*)(Vy) = A(Vy) \in \mathrm{im}A.$$

Therefore

$$\mathrm{im}A = \mathrm{im}U\Sigma = \mathrm{im}[u_1, \ldots, u_r],$$

where u_i denotes the ith column of U. Returning to (A.34) we see that $\ker A \supset \ker \Sigma V^*$. However, if $s \in \ker A$ then $0 = As = (U^*\Sigma V^*)s$ Multiplying by U we see that $s \in \ker \Sigma V^*$, so $\ker A = \ker \Sigma V^*$. The right-hand side is the set of elements orthogonal to the first r columns of V, which is the span of the last $n - r$ columns of V. The assertion that r is the rank of A follows from $\mathrm{im}A = \mathrm{im}U\Sigma$ and the fact that U is nonsingular. □

The SVD is an extremely useful tool that can clarify considerably the structure and properties of a matrix. For an example, let $A \in \mathbb{R}^{m\times n}$, and let

$$A = U\Sigma V^* \tag{A.37}$$

be an SVD of A. If the diagonal elements of Σ are $\sigma_1 \geq \ldots \geq \sigma_{\min\{m,n\}}$ and if of these $\sigma_1, \ldots, \sigma_r$ are positive, for some positive $r \leq \min\{m,n\}$, then we can partition U and V by taking the first r columns of each to create matrices U_+ and V_+, then writing

$$U = [U_+ \; U_0], \qquad V = [V_+ \; V_0]. \tag{A.38}$$

The matrices U_+ and V_+ have orthonormal columns, but they will not be orthogonal matrices unless they are square.

As $\sigma_k = 0$ if $k > r$, only columns of U_+ and columns of V_+ (i.e., rows of $(V_+)^*$) are involved in the product (A.37), so that

$$A = U_+ \Sigma_+ (V_+)^*, \tag{A.39}$$

where $\Sigma_+ \in \mathbb{R}^{r\times r}$ contains those elements Σ_{ij} with $1 \le i, j \le r$ and therefore has a strictly positive diagonal.

Corollary A.4.3 shows that

$$\mathrm{im} A = \mathrm{im} U_+ \tag{A.40}$$

and

$$\ker A = \mathrm{im} V_0. \tag{A.41}$$

If $A \in \mathbb{R}^{\mu\times\nu}$ and $B \in \mathbb{R}^{\nu\times\rho}$ and if we write $a_{(\cdot\, j)}$ for the jth column of A and $b_{(j\,\cdot)}$ for the jth row of B, then

$$AB = \sum_{j=1}^{\nu} a_{(\cdot\, j)} b_{(j\,\cdot)}, \tag{A.42}$$

where the right-hand side is a sum of ν matrices, each of rank not more than 1. This formula holds because if for fixed m and r we compute the element of the right side in position (m,r) we obtain $\sum_{j=1}^{\nu} a_{mj} b_{jr}$, and this is the same as $(AB)_{mr}$.

Applying this to (A.39) and writing u_k for the kth column of U_+ and v_k for the kth column of V_+, we obtain

$$A = \sum_{k=1}^{r} u_k \sigma_k (v_k)^*, \tag{A.43}$$

which expresses A as the sum of r rank-one matrices, where r is the rank of A.

A.4.1 Exercises for Section A.4

Exercise A.4.4. Give an example of a matrix $A \in \mathbb{R}^{m\times n}$, and orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ whose columns are eigenvectors of AA^* and A^*A respectively, such that U^*AV is not an SVD of A.

A.4.2 Notes and references

The singular value decomposition appears to be due originally to Beltrami and Jordan in the latter part of the nineteenth century, but much more information has appeared since then. For an excellent discussion of its early history, see [43].

The method of proving existence follows [1, Theorem A.2.3].

A.5 Generalized inverses of linear operators

In general, linear operators don't have single-valued inverses. But if we relax the requirements slightly then for any linear operator we can define a *generalized inverse* that acts somewhat like an inverse, in a sense made precise below. It becomes the inverse of the operator if the latter is nonsingular.

A.5.1 Existence of generalized inverses

Definition A.5.1. Two subspaces U and V of $\mathbb{R}^k$ are *independent* if $U \cap V = \{0\}$. If U and V are independent and their dimensions sum to k then they are *complementary* to each other. □

Independence of U and V is equivalent to the requirement that $U+V$ be the direct sum $U \oplus V$; complementarity is equivalent to having $\mathbb{R}^k = U \oplus V$.

Definition A.5.2. A *projector* on $\mathbb{R}^k$ is a function $f : \mathbb{R}^k \to \mathbb{R}^k$ that is idempotent: i.e., $f \circ f = f$. If a projector P is a linear transformation then we call it a *linear projector*. In the case of a linear projector P, if U and V are complementary subspaces of $\mathbb{R}^k$ such that $U = \operatorname{im} P$ and $V = \ker P$, then we call P the *linear projector on U along V*. □

In the last sentence of Definition A.5.2, P is "the" linear projector on U along V because it is unique. To see that, observe first that P fixes any element u of U, because as $U = \operatorname{im} P$ we must have $u = Py$ for some y, and then

$$Pu = P(Py) = Py = u.$$

Now let $x \in \mathbb{R}^k$ and let the unique expression of x as the sum of an element in U and an element in V be $x = x_U + x_V$. The linear operator P fixes U and annihilates V, so $Px = Px_U = x_U$. Thus the direct-sum decomposition completely determines the values of P, so P is unique. We will write $\pi_{U,V}$ for this projector.

A linear projector is *orthogonal* if its kernel is the orthogonal complement of its image. Such projectors have a special property.

Proposition A.5.3. *A linear projector P on $\mathbb{R}^n$ is orthogonal if and only if it is symmetric.*

Proof. If a linear projector P is symmetric then $\ker P = (\operatorname{im} P^*)^\perp = (\operatorname{im} P)^\perp$, so that P is an orthogonal projector. Conversely, if P is an orthogonal projector then $\ker P = (\operatorname{im} P)^\perp$, so that for each x and y in $\mathbb{R}^n$,

$$0 = \langle (I-P)x, Py \rangle = \langle P^*(I-P)x, y \rangle.$$

This implies that $P^*(I-P)$ is the zero operator, and therefore so also is its transpose $(I-P)^*P$. Expanding, we find that

$$P = P^*P = P^*,$$

so that P is symmetric. □

Let $A : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation of rank r. Its image $R \subset \mathbb{R}^m$ then has dimension r, and its kernel $K \subset \mathbb{R}^n$ has dimension $n-r$. Choose any subspaces $M \subset \mathbb{R}^m$ of dimension $m-r$ and $N \subset \mathbb{R}^n$ of dimension r such that $\mathbb{R}^m = M \oplus R$ and $\mathbb{R}^n = N \oplus K$. The proof of the following theorem shows how to construct a linear operator A^-, the generalized inverse of A having image N and kernel M.

Theorem A.5.4. *Let $A : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation having image R and kernel K, and let M and N be subspaces of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively such that $R \oplus M = \mathbb{R}^m$ and $K \oplus N = \mathbb{R}^n$. There is then a unique linear transformation $A^- : \mathbb{R}^m \to \mathbb{R}^n$ such that*

$$\operatorname{im} A^- = N, \quad \ker A^- = M, \qquad AA^- = \pi_{R,M}, \quad A^-A = \pi_{N,K}. \tag{A.44}$$

Proof. The restriction $\tau := A|_N$ of A to N is a linear operator from N into $\mathbb{R}^m$, and because $K = \ker A$ we have

$$A = \tau \circ \pi_{N,K}. \tag{A.45}$$

This shows that $R = \operatorname{im} A \subset \operatorname{im} \tau$, and the opposite inclusion also holds because τ is a restriction of A. Therefore τ takes N onto R. Moreover, its kernel is $\{0\}$ because N and K are independent, so that τ is actually an isomorphism of N onto R. It follows that τ has a single-valued inverse $\tau^{-1} : R \to \mathbb{R}^n$ with image N.

Let

$$A^- = \tau^{-1} \circ \pi_{R,M}. \tag{A.46}$$

Then A^- is a linear transformation from $\mathbb{R}^m$ to $\mathbb{R}^n$ with image N and kernel M. We have

$$AA^- = \tau \circ \pi_{N,K} \circ \tau^{-1} \circ \pi_{R,M} = \pi_{R,M}, \tag{A.47}$$

because the image of τ^{-1} is N, so that $\pi_{N,K} \circ \tau^{-1} = \tau^{-1}$. Similarly,

$$A^-A = \tau^{-1} \circ \pi_{R,M} \circ \tau \circ \pi_{N,K} = \pi_{N,K}, \tag{A.48}$$

because $\pi_{R,M} \circ \tau = \tau$.

To show uniqueness, let B and C be linear operators from $\mathbb{R}^m$ to $\mathbb{R}^n$, each having the properties claimed for A^- in (A.44). Then

$$B = B \circ \pi_{R,M} = BAC = \pi_{N,K} \circ C = C,$$

so that A^- is unique. □

For much more information on generalized inverses of various kinds, and many additional references, see [25].

A.5.2 The Moore-Penrose generalized inverse

If we specialize the construction of Section A.5.1 by requiring our two direct-sum decompositions to be orthogonal, then we have

$$N = K^{\perp} = \mathrm{im}A^*, \qquad M = R^{\perp} = \ker A^*. \tag{A.49}$$

In that case the projectors

$$AA^- = \pi_{R,M}, \qquad A^-A = \pi_{N,K} \tag{A.50}$$

will be orthogonal and so, by Proposition A.5.3, symmetric. Then it is customary to write A^+ in place of A^- for the generalized inverse; this is the *Moore-Penrose generalized inverse* of L. For more on this special case, including computational methods, see [17, Section 1.3] and [13, Section 5.5.4].

Some presentations of the Moore-Penrose inverse define it as a matrix $X \in \mathbb{R}^{m\times n}$ satisfying the conditions

$$\begin{aligned} &[a.]\ (AX)^* = AX, \\ &[b.]\ (XA)^* = XA, \\ &[c.]\ AXA = A, \\ &[d.]\ XAX = X. \end{aligned} \tag{A.51}$$

We can put this into the general framework already developed.

Proposition A.5.5. *Let $A \in \mathbb{R}^{n\times m}$ with $\mathrm{im}A = R$ and $\ker A = K$, and let $X \in \mathbb{R}^{m\times n}$. The following are equivalent.*

1. *X satisfies the conditions given for A^- in Theorem A.5.4 with the choices $N = K^{\perp}$ and $M = R^{\perp}$.*
2. *X satisfies* (A.51).

If these equivalent statements hold, then X is unique.

Proof. First suppose that assertion (1) holds. Then $AX = \pi_{RM} = \pi_{RR^{\perp}}$ so AX is an orthogonal projector and therefore is symmetric. This establishes (a) in (A.51), and a similar argument proves (b). The proof of Theorem A.5.4 shows that $\ker X = M$, so let $y \in \mathbb{R}^m$ and write it uniquely as $y = y_R + y_M$. Then

$$(XAX)y = X(AX)y = X\pi_{RM}y = Xy_R = Xy,$$

where the last equality holds because $y_M \in \ker X$. Therefore (d) in (A.51) holds, and a similar argument establishes (c) and therefore assertion (2).

Next, suppose assertion (2) holds. Using (c) we obtain $(AX)^2 = (AXA)X = AX$, so AX is a projector; then (a) together with Proposition A.5.3 shows that it is an orthogonal projector. We have $\mathrm{im}AX \subset \mathrm{im}A$, while if $y \in \mathrm{im}A$ then for some $x \in \mathbb{R}^m$ we have

$$y = Ax = (AXA)x = (AX)Ax = (AX)y,$$

which shows that $\mathrm{im}AX = \mathrm{im}A$. Also $\ker X \subset \ker AX$, while if $(AX)z = 0$ then

$$0 = X(AX)z = Xz,$$

showing that $\ker AX = \ker X$. We have thus shown that

$$AX = \pi_{\mathrm{im}A,\ker X}, \qquad \ker X = (\mathrm{im}A)^{\perp}, \tag{A.52}$$

where the second inequality comes from the fact that AX is an orthogonal projector with kernel $\ker X$ and image $\mathrm{im}A$. A similar argument shows that

$$XA = \pi_{\mathrm{im}X,\ker A}, \qquad \ker A = (\mathrm{im}X)^{\perp}. \tag{A.53}$$

Now if we let $R = \mathrm{im}A$ and $K = \ker A$, then (A.52) shows that $AX = \pi_{R,R^{\perp}}$ and $\ker X = R^{\perp}$, while (A.53) shows that $XA = \pi_{K^{\perp},K}$ and $\mathrm{im}X = K^{\perp}$. Therefore assertion (1) holds. The uniqueness claim follows because if assertion (1) holds then Theorem A.5.4 guarantees that A^{-} is unique. □

Let the rank of A be $r \geq 1$ and write the special form of the SVD given in (A.39):

$$A = U_{+}\Sigma_{+}(V_{+})^{*}.$$

Then one can verify directly that the Moore-Penrose generalized inverse of the linear operator represented by A is the operator $A^{+} : \mathbb{R}^{n} \to \mathbb{R}^{m}$ represented by

$$A^{+} = V_{+}(\Sigma_{+})^{-1}(U_{+})^{*}, \tag{A.54}$$

by simply checking that with the choices in (A.49) the operator given in (A.54) satisfies the requirements in (A.44).

A.5.3 Notes and references

E. H. Moore originally developed what is now called the Moore-Penrose generalized inverse [24]. Some interesting historical background is in [4]. The idea was rediscovered and extended by Penrose [28].

A.6 Affine sets

Affine sets are very simple sets whose properties follow from basic facts of linear algebra. They can be points, lines, or flat surfaces (like subspaces but not necessarily passing through the origin). In this section we explain their basic properties and

establish some ways of representing them. We also introduce three important concepts: the *affine transformation,* the *affine hull* of a set, and the property of *general position,* and develop some relationships among these objects.

A.6.1 Basic ideas

Definition A.6.1. A subset A of $\mathbb{R}^n$ is *affine* if for any two points x and y in A, and for any real μ, $(1-\mu)x+\mu y \in A$. □

This means that A contains the whole line through any pair of its points. In particular, the empty set is affine, as is the whole space $\mathbb{R}^n$. Some of the exercises for this section ask you to use the definition to show that the class of affine sets is closed with respect to the operations of translation, of addition or intersection (of finitely many sets), and of taking forward or inverse images under linear transformations.

It turns out that one is not restricted to forming combinations of just two points of such a set. Suppose that we use the term *affine combination* to refer to any linear combination of points of $\mathbb{R}^n$ of the form $\sum_{i=1}^k \mu_i x_i$ with $\sum_{i=1}^k \mu_i = 1$. Then the class of affine sets is closed under the operation of taking affine combinations, as the following proposition shows.

Proposition A.6.2. *A subset A of $\mathbb{R}^n$ is affine if and only if A contains all affine combinations of finitely many points of A.*

Proof. (If): This follows directly from the definition of affine set.

(Only if): Suppose A is affine. The definition shows that the claim is true if such a combination consists of 1 or 2 points. For an induction, assume that $k > 2$ and that A contains all affine combinations of no more than $k-1$ of its points. Let $x = \sum_{i=1}^k \mu_i x_i$, where the x_i belong to A and the μ_i sum to 1. As $k > 2$ not all of the μ_i can be 1, so by renumbering if necessary we can assume that $\mu_k \neq 1$. Then

$$x = (1-\mu_k)[\sum_{i=1}^{k-1}(1-\mu_k)^{-1}\mu_i x_i] + \mu_k x_k.$$

The quantity in square brackets, say s, is an affine combination of $k-1$ points of A, so it belongs to A by the induction hypothesis. Then x is an affine combination of s and x_k, so it belongs to A. □

The *Minkowski sum* of two subsets A and B of $\mathbb{R}^n$ is the set $A+B = \{a+b \mid a \in A,\ b \in B\}$. If one of the sets is a singleton, say $\{a\}$, then we usually write $a+B$ rather than $\{a\}+B$. This operation is associative, so we can extend this definition to any finite number of summands by setting $A+B+C = (A+B)+C$, etc.

Definition A.6.3. Two affine subsets A and B of $\mathbb{R}^n$ are *parallel* if there is some $x \in \mathbb{R}^n$ with $A = B+x$. In that case we say that A is a *translate* of B. □

Suppose that A is an affine subset of $\mathbb{R}^n$ and $x \in \mathbb{R}^n$. Choose any two points of the translate $A+x$ and write them as $a+x$ and $a'+x$, where a and a' belong to A. For any $\mu \in \mathbb{R}$ we have

$$(1-\mu)(a+x)+\mu(a'+x) = [(1-\mu)a+\mu a']+x.$$

The quantity in square brackets belongs to A, so the right-hand side belongs to $A+x$. Accordingly, any translate of an affine set is affine.

It turns out that affine sets are just translates of subspaces. This should not be surprising, as a subset L of $\mathbb{R}^n$ is a nonempty subspace if and only if L is affine and contains the origin. Indeed, if L is a nonempty subspace then it certainly contains the origin. It is also closed under linear combinations, hence certainly under affine combinations, so it is affine. On the other hand, if A is affine and contains the origin then let x and x' belong to A and μ and μ' be real numbers. Then

$$\mu x+\mu' x' = \mu x+\mu' x'+(1-\mu-\mu')0 \in A,$$

so that A is closed under linear combinations of its elements and therefore it is a nonempty subspace. The following proposition shows that any affine set has a unique subspace that is parallel to it.

Proposition A.6.4. *Given any affine subset A of $\mathbb{R}^n$, the set $L := A-A$ is the unique subspace of $\mathbb{R}^n$ that is parallel to A. For any point $a \in A$ one has $A = L+a$.*

Proof. If A is empty then so is L, and there is nothing else to prove. Therefore let A be nonempty and let $z_i \in L$ for $i = 1,2$. Then $z_i = a_i - a_i'$ for points a_i and a_i' of A. For any real numbers α_1 and α_2,

$$\alpha_1 z_1+\alpha_2 z_2 = [(1+\alpha_1)a_1-\alpha_1 a_1'+\alpha_2 a_2-\alpha_2 a_2'] - a_1.$$

The quantity in square brackets belongs to A, so the left-hand side belongs to $A-A = L$ and therefore L is a subspace.

Next, choose any $a \in A$; then $L = A-A \supset A-a$ so that $A \subset L+a$. But if $z \in L$ then there are elements a' and a'' of A with $z = a'-a''$; then $z+a = a'-a''+a \in A$, so that $L+a \subset A$ and therefore $A = L+a$.

Finally, suppose L_1 and L_2 are subspaces of $\mathbb{R}^n$ that are parallel to A. Then there are points x_1 and x_2 of $\mathbb{R}^n$ with $L_i+x_i = A$ for $i = 1,2$. As the subspaces L_i contain the origin, the x_i belong to A. Suppose $z_1 \in L_1$; then the point $a_1 := z_1+x_1$ belongs to A, and then so does $a_1-x_1+x_2$ because A is affine. But as $A = L_2+x_2$ there is some $z_2 \in L_2$ with $z_2+x_2 = a_1-x_1+x_2$, so that $z_2 = a_1-x_1 = z_1$ As z_1 was an arbitrary point of L_1 we have $L_1 \subset L_2$, and reversing the roles of L_1 and L_2 leads to the opposite inclusion. Therefore $L_1 = L_2$, so the subspace parallel to A is unique. □

We write $\mathrm{par}A$ for the unique subspace parallel to an affine set A.

Definition A.6.5. The *dimension* of an affine subset A of $\mathbb{R}^n$ is the dimension of its parallel subspace (-1 if $A = \emptyset$). □

Corollary A.6.6. *Suppose A and A' are two affine subsets of $\mathbb{R}^n$, each having dimension k. If $A \subset A'$, then $A = A'$.*

Proof. If $k = -1$ there is nothing to prove, so assume that A is nonempty. Let L_A and $L_{A'}$ be the subspaces parallel to A and A' respectively. If $a \in A$, then as $a \in A'$ also, Proposition A.6.4 yields

$$a + L_A = A \subset A' = a + L_{A'}, \tag{A.55}$$

so that $L_A \subset L_{A'}$. Any basis B for L_A has k independent elements that also belong to $L_{A'}$, but as $\dim L_{A'} = k$ by hypothesis, by linear algebra B also generates $L_{A'}$. Then $L_A = L_{A'}$, so by (A.55) $A = A'$. □

Another corollary of Proposition A.6.4 extends the standard direct-sum decomposition of $\mathbb{R}^n$.

Corollary A.6.7. *If A is a nonempty affine subset of $\mathbb{R}^n$ with parallel subspace L, then*

$$\mathbb{R}^n = A \oplus L^\perp. \tag{A.56}$$

Proof. Evidently $A + L^\perp \subset \mathbb{R}^n$. To show that the reverse inclusion holds, let $x \in A$ and use the standard decomposition

$$\mathbb{R}^n = L \oplus L^\perp \tag{A.57}$$

to write $x = x_L + x_P$ where $x_L \in L$ and $x_P \in L^\perp$. Proposition A.6.4 shows that

$$A = L + x = L + x_P,$$

where the second equality holds because $L + x_L = L$. Now we can write any $y \in \mathbb{R}^n$ as

$$y = y_L + y_P = (y_L + x_P) + (y_P - x_P), \tag{A.58}$$

where $y_L \in L$ and $y_P \in L^\perp$. The first quantity in parentheses belongs to A and the second to $L^\perp$, so that $y \in A + L^\perp$ and hence $\mathbb{R}^n \subset A + L^\perp$.

For uniqueness, suppose that $y = y'_A + y'_P$ with $y'_A \in A$ and $y'_P \in L^\perp$. By subtracting this equation from (A.58) we obtain

$$0 = [(y_L + x_P) - y'_A] + [(y_P - x_P) - y'_P].$$

But the first quantity in square brackets belongs to $A - A = L$ and the second to $L^\perp$, so (A.57) shows that each is zero. Therefore $y'_A = y_L + x_P$ and $y'_P = y_P - x_P$, so the decomposition in (A.58) is unique and therefore $\mathbb{R}^n = A \oplus L^\perp$. □

The next proposition shows that affine sets are the solution sets of systems of linear equations.

Proposition A.6.8. *A subset D of $\mathbb{R}^n$ is an affine set of dimension $k \geq 0$ if and only if there exist an $(n-k) \times n$ matrix A of rank $n-k$ and an element $a \in \mathbb{R}^{n-k}$ such that $D = \{\, x \mid Ax = a \,\}$.*

Proof. Suppose that D is affine and has dimension k. Let $d \in D$ and let $L = D - d$. We know that L has dimension k, so its orthogonal complement $L^\perp$ has dimension $n-k$. Choose any basis for $L^\perp$ and form a $(n-k) \times n$ matrix A whose rows are the chosen basis vectors. We have

$$L = L^{\perp\perp} = \{\, x \mid Ax = 0 \,\}.$$

Writing $a = Ad$ we have $D = L + d = \{\, x \mid Ax = a \,\}$, as required.

Conversely, if D is the solution set of a system of equations $Ax = a$ with A having rank $n-k$, it follows immediately from the definition that D is affine. If $d \in D$ then $a = Ad$, so that $D - d = \{\, x \mid Ax = 0 \,\}$. As A has rank $n-k$, we see that the subspace $D - d$ has dimension k, and therefore so does D. □

The names *point*, *line*, and *hyperplane* denote affine sets in $\mathbb{R}^n$ of dimensions 0, 1, and $n-1$ respectively, the last two requiring that $n \geq 1$. Since hyperplanes have dimension $n-1$, their parallel subspaces are the orthogonal complements of one-dimensional subspaces, each of which is determined by a nonzero vector that is unique up to nonzero scalar multiplication. This vector is called the *normal* to the hyperplane in question. The last proposition then shows that a hyperplane H can always be represented in the form $H = \{\, z \mid \langle y^*, z \rangle = \eta \,\}$ with $y^* \neq 0$, and since the subspace complementary to par$H$ has dimension 1 it is easy to see that the pair (y^*, η) is unique up to multiplication by a nonzero scalar.

For this hyperplane H we define the *lower closed halfspace* of H to be $\{x \mid \langle y^*, x \rangle \leq \eta\}$, and the *upper closed halfspace* of H to be the corresponding set with $\leq$ replaced by $\geq$. There are also lower and upper *open halfspaces* associated with H, defined in the same way but using strict instead of weak inequalities. These halfspaces depend only on H and not on the particular y^* and η used to represent it, though which halfspace is lower and which is upper will depend on y^*.

In the rest of this book we use standard notation for hyperplanes and for their closed halfspaces as follows: for $y^* \in \mathbb{R}^n \backslash \{0\}$ and $\eta \in \mathbb{R}$, we let

$$\begin{aligned} H_0(y^*, \eta) &= \{x \in \mathbb{R}^n \mid \langle y^*, x \rangle = \eta\}, \\ H_-(y^*, \eta) &= \{x \in \mathbb{R}^n \mid \langle y^*, x \rangle \leq \eta\}, \\ H_+(y^*, \eta) &= \{x \in \mathbb{R}^n \mid \langle y^*, x \rangle \geq \eta\}. \end{aligned} \tag{A.59}$$

A.6.2 The affine hull

Definition A.6.9. Let S be a subset of $\mathbb{R}^n$. The *affine hull* of S, written aff$S$, is the intersection of all affine sets containing S. □

The set aff$S$ is itself affine. Definition A.6.9 gives what one can think of as an outer representation of aff$S$, but we can also develop an inner representation using affine combinations.

Proposition A.6.10. *Let S be a subset of $\mathbb{R}^n$. Then* $\operatorname{aff} S$ *is the set of all affine combinations of finite subsets of S.*

Proof. Let C be the collection of all affine combinations of finite subsets of S. If u and v are two elements of C, then any affine combination w of u and v is itself an affine combination of the two finite subsets of S that produced u and v, so C is an affine set. Each element t of S is the affine combination $(1)t$ of the finite subset $\{t\}$ of S, and it is therefore in C, so that $S \subset C$. As C is affine, we have $C \supset \operatorname{aff} S$.

Now suppose that A is an affine set that contains S. Proposition A.6.2 shows that then A contains all affine combinations of finite subsets of S, so $A \supset C$. As $\operatorname{aff} S$ is is the intersection of all such A, it contains C. Then $\operatorname{aff} S \supset C \supset \operatorname{aff} S$, so $\operatorname{aff} S = C$. □

The affine-hull operator is the first of several special operators for dealing with sets, which we'll introduce in succeeding sections. Much of the technical content of convexity deals with using these operators, and one of the important questions about them is when and if they commute with other operators.

We'll establish various results about such commutativity as we proceed, but to begin we consider the affine-hull operator and the closure operator. As an affine set is closed, for any subset S of $\mathbb{R}^n$ the set $\operatorname{aff} S$ is closed and contains S; hence it contains $\operatorname{cl} S$. The definition of affine hull then shows that in fact $\operatorname{aff} S \supset \operatorname{aff}\operatorname{cl} S$. The opposite inclusion is trivial because $S \subset \operatorname{cl} S$. So we have $\operatorname{aff} S = \operatorname{aff}\operatorname{cl} S$, and the fact that affine sets are closed means that $\operatorname{aff} S = \operatorname{cl}\operatorname{aff} S$. Therefore the operators aff and cl commute.

A.6.3 General position

If we consider three points of $\mathbb{R}^2$ arranged in a triangle, then it is not hard to see that we can represent any point of $\mathbb{R}^2$ as an affine combination of these three points, and it turns out that the coefficients in such an affine combination are unique. However, if we move one of the points so that it lies on the line through the other two, then we can represent only points on that line as affine combinations of the three points, and moreover the coefficients are no longer unique. The following definition gives, for points in $\mathbb{R}^n$, the crucial distinction between these two situations.

Definition A.6.11. Let $S = \{x_0, \ldots, x_k\}$ be a set of $k+1$ points in $\mathbb{R}^n$. The points of S are said to be in *general position* if $\operatorname{aff} S$ has dimension k. □

Points in general position are sometimes called *affinely independent*, but we do not use that term here. If the points of a finite subset of $\mathbb{R}^n$ are in general position, the set cannot contain more than $n+1$ points.

Because the concept of general position is very useful, we give three equivalent criteria, one or another of which is often easier to apply than is the definition.

Theorem A.6.12. *Let $S = \{x_0, \ldots, x_k\}$ be a subset of $\mathbb{R}^n$. The following are equivalent:*

a. The points of S are in general position.
b. The points $x_1 - x_0, \ldots, x_k - x_0$ are linearly independent in $\mathbb{R}^n$.
c. The points

$$\begin{pmatrix} x_0 \\ 1 \end{pmatrix}, \ldots, \begin{pmatrix} x_k \\ 1 \end{pmatrix}$$

are linearly independent in $\mathbb{R}^{n+1}$.
d. For any point x of aff$S$*, the coefficients μ_i in the representation*

$$x = \sum_{i=0}^{k} \mu_i x_i, \qquad \sum_{i=0}^{k} \mu_i = 1,$$

of x as an affine combination of $x_0, \ldots, x_k$, are unique.

When the points of S are in general position, the coefficients μ_i in (d) are called the *barycentric coordinates* of x with respect to $x_0, \ldots, x_k$.

Proof. First let L be the subspace parallel to aff$S$, and write $y_i = x_i - x_0$ for $i = 1, \ldots, k$. The points y_i belong to L. Further, if $x = \sum_{i=0}^{k} \mu_i x_i$ is any affine combination of the points of S, we must have $\mu_0 = 1 - \sum_{i=1}^{k} \mu_i$, and therefore $x = x_0 + \sum_{i=1}^{k} \mu_i y_i$. Since $\mu_1, \ldots, \mu_k$ are arbitrary, we find that aff$S = x_0 + \text{span}\{ y_1, \ldots, y_k \}$. Therefore L must be the span of the y_i, so the dimension of L (and hence that of aff$S$) is equal to the number of linearly independent y_i.

(a) implies (b). To say aff$S$ has dimension k is to say that L has dimension k, so the set $y_1, \ldots, y_k$ contains k linearly independent elements.

(b) implies (c). Write $z_0, \ldots, z_k$ for the points shown in (c). If the z_i are linearly dependent, then we have

$$\mu_0 x_0 + \ldots + \mu_k x_k = 0, \qquad \mu_0 + \ldots + \mu_k = 0,$$

for some $\mu_0, \ldots, \mu_k$ that are not all zero. As the μ_i sum to zero, at least one of $\mu_1, \ldots, \mu_k$ must be nonzero. Further, $\mu_0 = -(\mu_1 + \ldots + \mu_k)$, so that $\sum_{i=1}^{k} \mu_i y_i = 0$, and therefore the y_i are linearly dependent.

(c) implies (d). If we write a point x as an affine combination of the x_i then we have written $(x, 1)$ as a linear combination of the z_i with the same coefficients. If (c) holds, these coefficients must be unique.

(d) implies (a). The dimension of aff$S$ is, by definition, that of L. If this dimension is less than k then we can find scalars $\mu_1, \ldots, \mu_k$, not all zero, with $\sum_{i=1}^{k} \mu_i (x_i - x_0) = 0$. Then

$$x_0 = (1 - \sum_{i=1}^{k} \mu_i) x_0 + \sum_{i=1}^{k} \mu_i x_i,$$

and the two sides of this equation display x_0 written as an affine combination of the points of S in two different ways. □

Remark A.6.13. As the numbering of the points is at our disposal, any of the points could have played the role of x_0. This theorem therefore shows, for example, that

if the points of some finite set are in general position then the points of any of its subsets are also in general position.

Part (b) of the theorem also shows that if the points $\{f_0,\dots,f_k\}$ of a finite set $F \subset \mathbb{R}^n$ are in general position, then so are the points of $F+x$ for any $x \in \mathbb{R}^n$, because $f_i - f_0 = (f_i + x) - (f_0 + x)$.

Let A be a nonempty affine subset of $\mathbb{R}^n$, and suppose that the dimension of A is k. To find $k+1$ points in A that are in general position, first choose any $a_0 \in A$ and note that if $k = 0$ we are finished. If $k > 0$, write the parallel subspace L of A as $A - a_0$. As L has dimension k, it contains k linearly independent elements $v_1,\dots,v_k$. For $i = 1,\dots,k$ let $a_i = a_0 + v_i$. These points belong to A, and Theorem A.6.12 shows that $a_0,\dots,a_k$ are in general position. The following corollary shows that we can recover A from such a set of points.

Corollary A.6.14. *Let A be an affine subset of $\mathbb{R}^n$ of dimension $k \geq 0$, and let $S = \{a_0,\dots,a_m\}$ be a finite subset of A whose points are in general position. Then $m \leq k$, and if $m = k$ then $A = \operatorname{aff} S$.*

Proof. If the points of S are in general position, then by Theorem A.6.12 the points $a_1 - a_0,\dots,a_m - a_0$ are linearly independent. These points belong, by Proposition A.6.4, to the subspace $\operatorname{par} A$, which has dimension k. Therefore $m \leq k$.

If $m = k$ then the points just mentioned are in fact a basis for $\operatorname{par} A$, so for any point $x \in \operatorname{par} A$ there are scalars $\mu_1,\dots,\mu_k$ such that

$$x = \sum_{i=1}^{k} \mu_i (a_i - a_0). \tag{A.60}$$

Choose any $y \in A$. Proposition A.6.4 says that $A = a_0 + \operatorname{par} A$, so we can use (A.60) to write

$$y = a_0 + x = a_0 + \sum_{i=1}^{k} \mu_i (a_i - a_0) = (1 - \sum_{i=1}^{k} \mu_i) a_0 + \sum_{i=1}^{k} \mu_i a_i. \tag{A.61}$$

The right side of (A.61) is an affine combination of the points $a_0,\dots,a_k$ and is therefore in $\operatorname{aff} S$. Then $A \subset \operatorname{aff} S$, but as $S \subset A$ and A is affine we have $\operatorname{aff} S \subset A$, so that $A = \operatorname{aff} S$. □

Another consequence of Theorem A.6.12 is that if points in general position are slightly perturbed, they remain in general position. To prove this we need a lemma about perturbed matrices.

Lemma A.6.15. *Let $A \in \mathbb{R}^{n\times n}$ be nonsingular. If $\Delta \in \mathbb{R}^{n\times n}$ satisfies*

$$\|A^{-1}\Delta\| < 1, \tag{A.62}$$

then $A + \Delta$ is nonsingular and

$$\|(A+\Delta)^{-1}\| \leq (1 - \|A^{-1}\Delta\|)^{-1} \|A^{-1}\|. \tag{A.63}$$

Proof. Fix A and Δ, and suppose $x \in \mathbb{R}^n$ with $(A+\Delta)x = 0$. Then $-x = A^{-1}\Delta x$, so that

$$\|x\| = \|-x\| = \|A^{-1}\Delta x\| \le \|A^{-1}\Delta\|\|x\|.$$

As $\|A^{-1}\Delta\| < 1$, this inequality can hold only if $x = 0$. Therefore $A+\Delta$ is nonsingular.

Let y be any point of $\mathbb{R}^n$ such that $\|y\| = 1$, and write $x = (A+\Delta)^{-1}y$. Multiplying by $A^{-1}(A+\Delta)$, we obtain $x = A^{-1}y - A^{-1}\Delta x$. Then

$$\|x\| \le \|A^{-1}y\| + \|A^{-1}\Delta\|\|x\|,$$

so

$$\|x\| \le (1 - \|A^{-1}\Delta\|)^{-1}\|A^{-1}y\| \le [(1 - \|A^{-1}\Delta\|)^{-1}\|A^{-1}\|]\|y\|,$$

which proves (A.63). □

As $\|A^{-1}\Delta\| \le \|A^{-1}\|\|\Delta\|$, (A.62) always holds if $\|\Delta\| < \|A^{-1}\|^{-1}$, so the set of nonsingular matrices is open in $(\mathbb{R}^{n\times n}, \|\cdot\|)$.

Corollary A.6.16. *If the points $x_0, \dots, x_k$ are in general position in $\mathbb{R}^n$, then there is a positive ε such that whenever points $x'_0, \dots, x'_k$ of $\mathbb{R}^n$ satisfy $\|x'_i - x_i\| < \varepsilon$ for $i = 0, \dots, k$, the x'_i are in general position.*

Proof. Theorem A.6.12 says that the x_i are in general position exactly when the vectors $z_0, \dots, z_k$ in $\mathbb{R}^{n+1}$, formed by augmenting each x_i with a last component of 1, are linearly independent. If $k < n-1$ then choose additional points $z_{k+1}, \dots, z_n$ so that $z_0, \dots, z_n$ is a linearly independent set. The matrix A whose columns are $z_0, \dots, z_n$ is then nonsingular. Let

$$\Delta = \left[\begin{bmatrix} x'_0 - x_0 \\ 0 \end{bmatrix}, \dots \begin{bmatrix} x'_k - x_k \\ 0 \end{bmatrix}, 0, \dots 0 \right];$$

then the first $k+1$ columns of $A+\Delta$ are the vectors $z'_0, \dots, z'_k$ in $\mathbb{R}^{n+1}$ formed by augmenting each x'_i with a last component of 1. If $\|x'_i - x_i\| < \varepsilon$ for $i = 0, \dots, k$ then $\|\Delta\| \le (k+1)\varepsilon$. If we choose ε to be less than $(k+1)^{-1}\|A^{-1}\|^{-1}$ then $\|\Delta\|\|A^{-1}\| < 1$, and Lemma A.6.15 says that $A+\Delta$ is nonsingular. Then in particular its first $k+1$ columns, which are $z'_0, \dots, z'_k$, are linearly independent. Another application of Theorem A.6.12 then shows that $x'_0, \dots, x'_k$ are in general position. □

A.6.4 Affine transformations

Definition A.6.17. An *affine transformation* from an affine subset A of $\mathbb{R}^n$ to $\mathbb{R}^m$ is a function from A to $\mathbb{R}^m$ whose graph is an affine subset of $A \times \mathbb{R}^m$. □

This means that for each a and a' in A and each real μ one has $T[(1-\mu)a + \mu a'] = (1-\mu)T(a) + \mu T(a')$. For such a transformation the image $T(A) \subset \mathbb{R}^m$ is

an affine set, and if T happens to be one-to-one on A, so that it has an inverse function defined on $T(A)$, then the inverse function is also an affine transformation. Proposition A.6.2 shows that for an affine transformation T of an affine set A and for each set of points $a_0 \ldots, a_k$ of A and each set $\mu_0, \ldots, \mu_k$ of affine coefficients, one has $T(\sum_{i=0}^{k} \mu_i a_i) = \sum_{i=0}^{k} \mu_i T(a_i)$.

The following lemma shows the close association between affine transformations defined on an affine set A and linear transformations defined on the parallel subspace $\mathrm{par}A$.

Lemma A.6.18. *Let A be a nonempty affine subset of $\mathbb{R}^n$, and let $a_0 \in A$ and $v_0 \in \mathbb{R}^m$.*

a. If $T : A \to \mathbb{R}^m$ is an affine transformation such that $T(a_0) = v_0$, then the function $L : \mathrm{par}A \to \mathbb{R}^m$ defined by

$$L(u) := T(u + a_0) - v_0 \tag{A.64}$$

is a linear transformation.

b. If $L : \mathrm{par}A \to \mathbb{R}^m$ is a linear transformation, then the function $T : A \to \mathbb{R}^m$ defined by

$$T(a) := L(a - a_0) + v_0 \tag{A.65}$$

is an affine transformation with $T(a_0) = v_0$.

c. The linear transformation L defined in (a) is independent of the choice of (a_0, v_0) in the graph of T.

Proof. Suppose the hypotheses of (a) hold, and choose p and p' in $\mathrm{par}A$. For any α and β in $\mathbb{R}$ we have

$$\begin{aligned} L(\alpha p + \beta p') &= T[(\alpha p + \beta p') + a_0] - v_0 \\ &= T[\alpha(p + a_0) + \beta(p' + a_0) + (1 - \alpha - \beta)a_0] - v_0 \\ &= \alpha T(p + a_0) + \beta T(p' + a_0) + (1 - \alpha - \beta)T(a_0) - v_0 \\ &= \alpha[T(p + a_0) - v_0] + \beta[T(p' + a_0) - v_0] \\ &= \alpha L(p) + \beta L(p'), \end{aligned}$$

so that L is a linear transformation.

Now let the hypotheses of (b) hold. The definition of T in (A.65) implies that $T(a_0) = v_0$. Choose a and a' in A, and let $\mu \in \mathbb{R}$. Then

$$\begin{aligned} T[(1 - \mu)a + \mu a'] &= L[(1 - \mu)a + \mu a' - a_0] + v_0 \\ &= L[(1 - \mu)(a - a_0) + \mu(a' - a_0)] + v_0 \\ &= (1 - \mu)[L(a - a_0) + v_0] + \mu[L(a' - a_0) + v_0] \\ &= (1 - \mu)T(a) + \mu T(a'), \end{aligned}$$

so that T is an affine transformation.

For (c), choose two pairs (a_0, v_0) and (b_0, w_0) in the graph of T and for any $u \in \mathrm{par}A$ define $L(u)$ by (A.64) and $M(u) := T(u + b_0) - w_0$. Then

$$\begin{aligned} L(u) &= T(u+a_0) - v_0 \\ &= T(u+a_0) - T(a_0) + T(b_0) - w_0 \\ &= T[(u+a_0) - a_0 + b_0] - w_0 \\ &= M(u). \end{aligned}$$

□

Remark A.6.19. A consequence of (A.65) is that

$$\|T(a) - T(a')\| = \|L(a-a')\| \leq \|L\|\|a-a'\|,$$

so that in particular any affine transformation is Lipschitz continuous. We can apply this observation in the case of a k-dimensional affine set $A \subset \mathbb{R}^n$ by taking $T(a)$ to be the function sending the point $a \in A$ to the vector $[\mu_0(a), \ldots, \mu_k(a)]$ of barycentric coordinates of a with respect to some subset $\{a_0, \ldots, a_k\}$ of A that is in general position. The definition of barycentric coordinates shows that this T is an affine transformation, so that the barycentric coordinates of a are Lipschitz continuous functions of $a \in A$. □

If X and Y are topological spaces, we say they are *homeomorphic* if there is an invertible function f from X onto Y such that both f and f^{-1} are continuous. We then call f and f^{-1} *homeomorphisms.* A *Lipschitz homeomorphism* is a homeomorphism f with the added property that f and f^{-1} are Lipschitz continuous. Homeomorphisms are very useful because if two spaces are homeomorphic, then for topological purposes we can regard each as a copy of the other.

If T is an affine map that is also a homeomorphism, then T has an inverse which, by our previous observation, is also affine. In that case we call T an *affine homeomorphism.* We use the term *isometry* for a function between two metric spaces that preserves distances.

The next theorem shows how to construct an affine Lipschitz homeomorphism between any two nonempty affine sets having the same dimension. Moreover, this construction preserves distances.

Theorem A.6.20. *Let A and D be affine subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, each having dimension $k \geq 0$; let $a_0 \in A$ and $d_0 \in D$. Then there is an affine Lipschitz homeomorphism ϕ of A onto D having the property that $\phi(a_0) = d_0$ and, for each a_1, a_2, a_3, and a_4 in A,*

$$\langle \phi(a_1) - \phi(a_2), \phi(a_3) - \phi(a_4) \rangle = \langle a_1 - a_2, a_3 - a_4 \rangle. \tag{A.66}$$

In particular, ϕ and its inverse are isometries.

Proof. Let $A = a_0 + L$ and $D = d_0 + M$, where L and M are k-dimensional subspaces of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Let $V \in \mathbb{R}^{n\times k}$ and $W \in \mathbb{R}^{m\times k}$ with $L = \operatorname{im} V$ and $M = \operatorname{im} W$, and suppose that each of V and W has orthonormal columns. For $a \in A$ and $d \in D$ define

$$\phi(a) = d_0 + WV^*(a - a_0), \qquad \psi(d) = a_0 + VW^*(d - d_0). \tag{A.67}$$

Then $\phi : A \to D$ and $\psi : D \to A$ are affine, hence Lipschitz continuous by RemarkA.6.19, with $\phi(a_0) = d_0$ and $\psi(d_0) = a_0$. We have for $a \in A$

$$\psi \circ \phi(a) = a_0 + VW^*[WV^*(a - a_0)] = a_0 + VV^*(a - a_0).$$

But $VV^* = \pi_{L,L^\perp}$, the orthogonal projector on L along $L^\perp$, and $a - a_0 \in L$, so $\psi \circ \phi(a) = a$ and therefore $\psi \circ \phi = \mathrm{id}_A$. A similar argument shows that $\phi \circ \psi = \mathrm{id}_D$, so each is surjective and $\psi = \phi^{-1}$. They are thus affine Lipschitz homeomorphisms between A and D.

Let $a_1, \ldots a_4$ be elements of A. Then

$$\begin{aligned} \langle \phi(a_1) - \phi(a_2), \phi(a_3) - \phi(a_4) \rangle &= \langle WV^*(a_1 - a_2), WV^*(a_3 - a_4) \rangle \\ &= \langle a_1 - a_2, VW^*WV^*(a_3 - a_4) \rangle \\ &= \langle a_1 - a_2, VV^*(a_3 - a_4) \rangle \\ &= \langle a_1 - a_2, a_3 - a_4 \rangle, \end{aligned} \tag{A.68}$$

where the last equality follows because $VV^* = \pi_{L,L^\perp}$ and $a_3 - a_4 \in L$. In particular, ϕ is an isometry because if we take $a_3 = a_1$ and $a_4 = a_2$ then (A.68) says that

$$\|\phi(a_1) - \phi(a_2)\| = \|a_1 - a_2\|.$$

Using $\psi = \phi^{-1}$ in (A.68) shows that the same statements apply to ψ. □

If for some reason one does not want to require V and W to have orthonormal columns, then the homeomorphism result of Theorem A.6.20 still holds—without the isometry assertion—provided one substitutes V^+ and W^+ for V^* and W^* in the proof. Here V^+ and W^+ are the Moore-Penrose inverses discussed in Section A.5.2.

A.6.5 The lineality space of a general set

Sometimes one can simplify working with a subset of $\mathbb{R}^n$ by taking account of its *lineality space,*which one can think of as a subspace on which the set is uninteresting. As it is uninteresting, we can factor out that space and then deal with the rest of the set in a smaller space.

Proposition A.6.21. *For any nonempty subset S of $\mathbb{R}^n$ there is a unique largest subspace L (in the sense of inclusion) satisfying the equation $S + L = S$.*

Proof. Let Q be the collection of all subspaces L of $\mathbb{R}^n$ such that $S + L = S$. Q is nonempty because it contains $\{0\}$. If L and L' are elements of Q then

$$S + (L + L') = (S + L) + L' = S + L' = S,$$

so that $L + L' \in Q$. Now let L be an element of Q having maximal dimension. If K is any element of Q then by what we have already shown, $L + K \in Q$. If $K \not\subset L$

then $\dim(L+K) > \dim L$, which contradicts the choice of L. Therefore $K \subset L$, so the subspace L contains every element of Q and thus is the largest element of Q. It is unique because if elements L and L' of Q each contained every element of Q, then they would contain each other so that $L = L'$. □

Proposition A.6.21 shows that the following definition makes sense.

Definition A.6.22. Let S be a nonempty subset of $\mathbb{R}^n$. The *lineality space* $\operatorname{lin} S$ of S is the unique largest subspace L (in the sense of inclusion) satisfying the equation $S + L = S$. □

Proposition A.6.23. *Let S be a nonempty subset of $\mathbb{R}^n$. If L is a nonempty subspace of $\operatorname{lin} S$ and M is any subspace such that $\mathbb{R}^n = L \oplus M$, then*

$$S = L \oplus (S \cap M). \tag{A.69}$$

Proof. Choose any $s \in S$. There are unique points $s_L \in L$ and $s_M \in M$ with $s = s_L + s_M$. Then s_M belongs to M and, as $s_M = s - s_L \in S + \operatorname{lin} S = S$, also to S. As $s_L \in L$, we have shown that $s \in L + (S \cap M)$ and therefore that $S \subset L + (S \cap M)$. We also have $L + (S \cap M) \subset \operatorname{lin} S + S = S$, so $S = L + (S \cap M)$. The direct-sum decomposition in (A.69) follows because $\mathbb{R}^n = L \oplus M$. □

Proposition A.6.23 shows why a set S is uninteresting in directions contained in $\operatorname{lin} S$. We can see from (A.69) that S is a cylinder with cross-section $S \cap M$. We can then factor out L and work with $S \cap M$, which may have much smaller dimension. It is often convenient to take $M = L^{\perp}$.

A.6.6 Exercises for Section A.6

We make the convention that the intersection of an empty collection of subsets of $\mathbb{R}^n$ is $\mathbb{R}^n$.

Exercise A.6.24. The intersection of any collection of affine subsets of $\mathbb{R}^n$ is affine.

Exercise A.6.25. Let S be a subset of $\mathbb{R}^n$. Show that S and its closure have the same affine hull.

Exercise A.6.26. Use the definition of affine hull to show that for two sets $S \subset \mathbb{R}^n$ and $T \subset \mathbb{R}^m$ one has $\operatorname{aff}(S \times T) = (\operatorname{aff} S) \times (\operatorname{aff} T)$.

Exercise A.6.27. Show that if T is an affine transformation from an affine set A to an affine set D, then the image under T of an affine subset of A is an affine subset of D. Also show that the inverse image under T of an affine subset of D is an affine subset of A. Do not assume that T is one-to-one.

Exercise A.6.28. The sum of any finite collection of affine subsets of $\mathbb{R}^n$ is affine.

Exercise A.6.29. Give an example of a closed subset S of $\mathbb{R}^2$ and a linear transformation $L : \mathbb{R}^2 \to \mathbb{R}$ such that $L(S)$ is not closed.

Exercise A.6.30. Let S be a subset of $\mathbb{R}^n$ and suppose that f is a continuous function from $\mathbb{R}^n$ to $\mathbb{R}^m$. Show that if $f(\operatorname{cl} S)$ is closed, then $f(\operatorname{cl} S) = \operatorname{cl} f(S)$.

Exercise A.6.31. Suppose A is the affine hull of the points $(1,1,1,1)$, $(2,1,2,1)$, and $(0,3,1,2)$ in $\mathbb{R}^4$. Construct explicitly an affine isometry of A onto $\mathbb{R}^2$ using the method of Theorem A.6.20, and give an example to show that your isometry preserves the inner product.

Exercise A.6.32. Show that the operator aff commutes with affine functions: that is, for any set $S \subset \mathbb{R}^n$ and any affine function f from $\operatorname{aff} S$ to another affine set, one has $f(\operatorname{aff} S) = \operatorname{aff} f(S)$.

Appendix B
Topics from Analysis

This appendix develops techniques from analysis that we need in parts of this book. The initial section covers three basic inequalities.

B.1 Three inequalities

We start with a general form of the inequality between arithmetic and geometric mean, then prove the inequalities of Hölder and Minkowski as consequences.

B.1.1 The arithmetic-geometric mean inequality

Given real numbers $x_1, \ldots, x_I$ and positive weights $\mu_1, \ldots, \mu_I$, the (weighted) *arithmetic mean* of the x_i is $\sum_{i=1}^{I} \mu_i x_i$. If the x_i are all positive then their (weighted) *geometric mean* is $(\Pi_{i=1}^{I} x_i^{\mu_i}$. For positive x_i the arithmetic mean is never less than the geometric mean, and it is strictly greater unless the x_i are all equal. Particular cases of these means are the unweighted versions, in which each μ_i equals I^{-1}.

Theorem B.1.1. *Let $x_1, \ldots, x_I$ and $\mu_1, \ldots, \mu_I$ be positive real numbers such that $\sum_{i=1}^{I} \mu_i = 1$. Then*

$$\Pi_{i=1}^{I} x_i^{\mu_i} \leq \sum_{i=1}^{I} \mu_i x_i, \tag{B.1}$$

with equality if and only if the x_i are all equal.

Proof. The assertion is obvious if $I = 1$. For $I = 2$, if $x_1 = x_2$ then (B.1) holds as an equality. If $x_1 \neq x_2$ there is no loss of generality in assuming that $x_1 < x_2$; we prove that strict inequality then holds.

Define $y = x_2/x_1$. Then

$$(\mu_1 x_1 + \mu_2 x_2)/(x_1^{\mu_1} x_2^{\mu_2}) = \mu_1 y^{-\mu_2} + \mu_2 y^{\mu_1} =: \phi(y).$$

The mean value theorem says that for some $\xi \in (1, y)$,

$$\phi(y) = \phi(1) + \phi'(\xi)(y-1) = \phi(1) + [\mu_1 \mu_2 \xi^{-\mu_2}(1 - \xi^{-1})](y-1).$$

We have $\phi(1) = 1$, $y > 1$, and the quantity in square brackets is positive, so $\phi(y) > 1$. Then (B.1) holds with strict inequality for $I = 2$.

Suppose now that $I = k > 2$ and that we have proved the conclusion for each value of I less than k. Let $\mu_0 = \sum_{i=1}^{k-1} \mu_i$, and for $i = 1, \ldots, k-1$ define $\nu_i = \mu_i/\mu_0$; then the ν_i are positive and they sum to 1. Then $\mu_0 + \mu_k = 1$ and

$$\begin{aligned} \Pi_{i=1}^{k} x_i^{\mu_i} = \left(\Pi_{i=1}^{k-1} x_i^{\nu_i}\right)^{\mu_0} x_k^{\mu_k} &\le \left(\sum_{i=1}^{k-1} \nu_i x_i\right)^{\mu_0} x_k^{\mu_k} \\ &\le \mu_0 \left(\sum_{i=1}^{k-1} \nu_i x_i\right) + \mu_k x_k = \sum_{i=1}^{k} \mu_i x_i, \end{aligned} \tag{B.2}$$

where the first inequality comes from (B.1) for $I = k-1$ (the induction hypothesis) and the second from (B.1) for $I = 2$. The first inequality is strict unless $x_1 = \ldots = x_{k-1}$ and the second is strict unless $x_k = \sum_{i=1}^{k-1} \nu_i x_i$. Therefore strict inequality holds in (B.2) unless $x_1 = \ldots = x_k$, so the assertion of the theorem holds for $I = k$. □

The next two sections develop useful corollaries of Theorem B.1.1.

B.1.2 The Hölder inequality

Corollary B.1.2. *Suppose that $a_1, \ldots, a_I$ and $b_1, \ldots, b_I$ are positive numbers, and that $p > 1$ and $q > 1$ with $p^{-1} + q^{-1} = 1$. Then*

$$\sum_{i=1}^{I} a_i b_i \le \left(\sum_{i=1}^{I} a_i^p\right)^{1/p} \left(\sum_{i=1}^{I} b_i^q\right)^{1/q}. \tag{B.3}$$

The inequality in (B.3) *is strict unless there is some positive ρ such that for each i, $a_i^p = \rho b_i^q$.*

Proof. Let $\gamma = \sum_i a_i^p$ and $\delta = \sum_i b_i^q$, and for each i let $c_i = \gamma^{-1} a_i^p$ and $d_i = \delta^{-1} b_i^q$, so that the c_i and the d_i each sum to 1. To simplify the notation, write α for p^{-1} and β for q^{-1}, so that $\alpha + \beta = 1$. If we write $\sum_i$ for $\sum_{i=1}^{I}$, then

$$\begin{aligned}\sum_i a_i b_i &= \sum_i (\gamma c_i)^\alpha (\delta d_i)^\beta \\ &= \gamma^\alpha \delta^\beta \sum_i c_i^\alpha d_i^\beta \\ &\le \gamma^\alpha \delta^\beta \sum_i (\alpha c_i + \beta d_i) \\ &= \gamma^\alpha \delta^\beta \\ &= \left(\sum_i a_i^p\right)^{1/p} \left(\sum_i b_i^q\right)^{1/q},\end{aligned} \tag{B.4}$$

where the inequality comes from applying Theorem B.1.1 to each term $c_i^\alpha d_i^\beta$ in turn. This proves (B.3). Moreover, the inequality will be strict unless for each i we have $c_i^\alpha = d_i^\beta$; that is, unless $a_i = \rho b_i$ with $\rho = \gamma^{(p^{-1})} \delta^{(q^{-1})}$. □

B.1.3 Minkowski's inequality

Text to be written.

B.2 *Projectors and distance functions

In Section 1.1.2 we introduced the point-to-set distance function

$$d_S(x) = \inf\{\|x - s\| \mid s \in S\}.$$

This section introduces some additional properties of this distance function and of projectors.

The infimum in the definition of d_S need not be attained in general, but if S is closed then the argument used in the proof of the "if" part of Proposition 2.2.2 shows that for any point $x \in \mathbb{R}^n$ there is always a point s of S that is closest to x so that $d_S(x) = \|s - z\|$, although s may not be unique. If x and x' are points of $\mathbb{R}^n$ then for any positive ε there are points s and s' of S with

$$\|x - s\| \le d_S(x) + \varepsilon, \qquad \|x' - s'\| \le d_S(x') + \varepsilon,$$

so that

$$d_S(x) - d_S(x') \le \|x - s'\| - d_S(x') \le \|x - x'\| + \|x' - s'\| - d_S(x') \le \|x - x'\| + \varepsilon,$$

and by interchanging the roles of x and x' we find that $|d_S(x) - d_S(x')| \le \|x - x'\|$, so that the distance function is Lipschitzian with modulus 1.

If a set C is not only closed but also convex, then d_C has an additional useful property. Suppose that x and x' are points of $\mathbb{R}^n$, and let $\mu \in (0,1)$. Find points s and s' in S such that

$$\|x-s\| < d_C(x)+\varepsilon, \qquad \|x'-s'\| < d_C(x')+\varepsilon,$$

and let $s_\mu = (1-\mu)s+\mu s'$. If we define $x_\mu = (1-\mu)x+\mu x'$, then

$$\begin{aligned} d_C(x_\mu) \le \|x_\mu - s_\mu\| &= \|(1-\mu)(x-s)+\mu(x'-s')\| \\ &< (1-\mu)[d_C(x)+\varepsilon]+\mu[d_C(x')+\varepsilon] \\ &= (1-\mu)d_C(x)+\mu d_C(x')+\varepsilon, \end{aligned}$$

and therefore for each $\lambda \in [0,1]$ we have

$$d_C[(1-\lambda)x+\lambda x'] \le (1-\lambda)d_C(x)+\lambda d_C(x').$$

This means that d_C belongs to the class of *convex functions,* about which we will see much more in future chapters.

Again for closed convex S, Lemma 2.2.1 shows that the closest point must be unique. In that case we can define a function $\Pi_C : \mathbb{R}^n \to S$ taking x to the closest point in S. This function is the *Euclidean projector* on S. This projector has remarkable properties, two of which we will develop here; more properties will follow after we have the technical tools to prove them.

The first property concerns the relationship between Π_C and the function $I-\Pi_C$ that takes x to $x-\Pi_C(x)$, a residual quantity measuring the extent to which x cannot be approximated by points of S.

Proposition B.2.1. *Let S be a nonempty closed convex subset of $\mathbb{R}^n$. For any x and x' in $\mathbb{R}^n$ one has*

$$\|\Pi_C(x)-\Pi_C(x')\|^2 + \|(I-\Pi_C)(x)-(I-\Pi_C)(x')\|^2 \le \|x-x'\|^2. \tag{B.5}$$

Proof. For brevity write p for $\Pi_C(x)$ and r for $(I-\Pi_C)(x)$, with p' and r' being the analogous quantities for x'. Then

$$\begin{aligned} \|x-x'\|^2 &= \|(r-r')+(p-p')\|^2 \\ &= \|r-r'\|^2 + 2\langle -r, p'-p\rangle + 2\langle -r', p-p'\rangle + \|p-p'\|^2. \end{aligned} \tag{B.6}$$

The four quantities in the last line of (B.6) are all nonnegative, the middle two being so because of Lemma 2.1. This proves the assertion, but we will shortly return to (B.6) for additional proofs. □

An immediate consequence of (B.5) is that each of Π_C and $I-\Pi_C$ is *nonexpansive:* that is, Lipschitzian on $\mathbb{R}^n$ with Lipschitz constant 1. However, we can show considerably more. The next result uses information from the proof of Proposition B.2.1 to establish additional properties of Π_C and $I-\Pi_C$

Proposition B.2.2. *For a nonempty closed convex subset C of $\mathbb{R}^n$ and for any x and x' in $\mathbb{R}^n$ one has*

$$\langle x-x', \Pi_C(x)-\Pi_C(x')\rangle \geq \|\Pi_C(x)-\Pi_C(x')\|^2 \tag{B.7}$$

and

$$\langle x-x', (I-\Pi_C)(x)-(I-\Pi_C)(x')\rangle \geq \|(I-\Pi_C)(x)-(I-\Pi_C)(x')\|^2. \tag{B.8}$$

Proof. We use the same notation as in the proof of Proposition B.2.1. For (B.7), write

$$\begin{aligned}\langle x-x', p-p'\rangle &= \langle (r-r')+(p-p'), p-p'\rangle \\ &= \langle -r, p'-p\rangle + \langle -r', p-p'\rangle + \|p-p'\|^2 \\ &\geq \|p-p'\|^2,\end{aligned}$$

where we again used Lemma 2.2.1. To obtain (B.7), interchange the roles of p and r. □

Proposition B.2.2 says in particular that both Π_C and $I-\Pi_C$ are *monotone functions:* that is, functions $f(x)$ having the property that for any x and x' in the domain of f, $\langle x-x', f(x)-f(x')\rangle \geq 0$. However, it says more because of the squared terms on the right-hand sides of the inequalities.

Later in the book we will develop information about more general *monotone operators,* which may be multivalued (so that $f(x)$ can be an entire set of points instead of a single point as it is here). However, we can use the information we now have about the projector to find out more about differentiability properties of d_C. We write $\mathrm{bd}\,C$ for the boundary of C.

Proposition B.2.3. *Let C be a nonempty closed convex subset of $\mathbb{R}^n$. Then $(1/2)d_C(\cdot)^2$ is Fréchet differentiable on $\mathbb{R}^n$ with its derivative being $(I-\Pi_C)$, which has Lipschitz modulus 1 everywhere on $\mathbb{R}^n$. The function d_C is Fréchet differentiable on $\mathbb{R}^n \setminus \mathrm{bd}\,C$, with derivative at x given by $(I-\Pi_C)(x)/\|(I-\Pi_C)(x)\|$ on the complement of C and, if C has full dimension, by zero on the interior of C. This derivative is locally Lipschitzian wherever it exists.*

Proof. Again using the notation from the proof of Proposition B.2.1, we have for x and x' in $\mathbb{R}^n$

$$\begin{aligned}(1/2)d_C(x')^2-(1/2)d_C(x)^2-\langle r, x'-x\rangle &= \|r'\|^2/2-\|r\|^2/2-\langle r, x'-x\rangle \\ &= [\|r'\|^2/2-\langle r', r\rangle+\|r\|^2/2]-[\|r\|^2-\langle r', r\rangle+\langle r, x'-x\rangle] \\ &= \|r'-r\|^2/2+\langle -r, p'-p\rangle.\end{aligned} \tag{B.9}$$

The last line in (B.9) is one-half of the sum of the first two of the four terms in the last line of of (B.6). Recalling that each of those four terms was nonnegative and that their sum was $\|x'-x\|^2$, we see that

$$0 \leq (1/2)d_C(x')^2 - (1/2)d_C(x)^2 - \langle r, x' - x \rangle \leq (1/2)\|x' - x\|^2,$$

which shows that r is the derivative of $(1/2)d_C^2$ at x. The assertions about Lipschitz continuity follow from the nonexpansivity proved in Proposition B.2.1.

For the second claim, observe that everywhere on the complement of C $(I - \Pi_C)(x)$ is nonzero and $d_C(x) = \|(I - \Pi_C)(x)\|$, which is positive. The form of the derivative of d_C then follows from the chain rule applied to the function $(d_C^2)^{1/2}$, and local Lipschitz continuity holds because $z/\|z\|$ is locally Lipschitz continuous at each nonzero point and because the composition of Lipschitz continuous functions is Lipschitz continuous. If C has an interior, then d_C is identically zero there, and so then is its derivative. □

B.2.1 Notes and references

The method of proof in Theorem B.1.1 follows that of [15, Appendix I].

Appendix C
Topics from Topology

This appendix develops a few results from topology required elsewhere in this book. Sections C.1 and C.2 use the set $\mathbb{R}^n$ with the standard Euclidean norm topology, while Section C.3 defines the extended real line $\bar{\mathbb{R}}$ and exhibits a topology for it.

Two excellent general references for basic topology are [19] and [2].

C.1 Compactness

Definition C.1.1. A subset S of $\mathbb{R}^n$ is *compact* if whenever a family $\mathscr{G} := \{G_\alpha \mid \alpha \in A\}$ of open subsets of $\mathbb{R}^n$ *covers* S (i.e., $S \subset \cup_{\alpha\in A} G_\alpha$), there is a finite subfamily $\{G_{\alpha_1}, \ldots, G_{\alpha_k}\} \subset \mathscr{G}$ that also covers S.

One can show from this definition that subsets of $\mathbb{R}^n$ are compact if and only if they are closed and bounded.

Proposition C.1.2. *Let D be an open set in $\mathbb{R}^n$ and let K be a compact subset of D If $f : D \to \mathbb{R}^m$ is a continuous function, then $f(K)$ is compact.*

Proof. If K is empty there is nothing to prove. Suppose K is nonempty, and define $J := f(K)$. Let $\{Q_\alpha \mid \alpha \in A\}$ be an open cover of J. As f is continuous, for each $\alpha \in A$ the set $E_\alpha := f^{-1}(Q_\alpha)$ is open. For each $k \in K$, $f(k)$ belongs to some Q_α so that $k \in E_\alpha$, and thus the sets $\{E_\alpha \mid \alpha \in A\}$ constitute an open cover of K. As K is compact this open cover has a finite subcover $\{E_{\alpha_1}, \ldots, E_{\alpha_m}\}$. Then $\{Q_{\alpha_1}, \ldots, Q_{\alpha_m}\}$ is a finite subcover of J, so J is compact. □

The following proposition develops a useful fact that is often called the *finite intersection property.* Here, as in earlier chapters, we use cB to denote the complement of a set B.

Proposition C.1.3. *Let K be a compact subset of $\mathbb{R}^n$, and for some index set $\mathscr{A}$ let $\{D_\alpha \mid \alpha \in \mathscr{A}\}$ be a collection of closed subsets of K. If for each finite subset $\mathscr{F}$ of $\mathscr{A}$*

$$\cap_{\alpha\in\mathscr{F}} D_\alpha \neq \emptyset, \tag{C.1}$$

then

$$\cap_{\alpha\in\mathscr{A}} D_\alpha \neq \emptyset. \tag{C.2}$$

Proof. Suppose that (C.2) is not true; we show that then (C.1) cannot be true.

The assumption that (C.2) does not hold implies that

$$\cap_{\alpha\in\mathscr{A}} D_\alpha = \emptyset \text{ and therefore } \cup_{\alpha\in\mathscr{A}} cD_\alpha = \mathbb{R}^n \supset K.$$

As the sets cD_α for $\alpha \in \mathscr{A}$ are open, they form an open cover of K. As K is compact, there must then be a finite subcover. Accordingly, there is some finite $\mathscr{F} \subset \mathscr{A}$ such that

$$\cup_{\alpha\in\mathscr{F}} cD_\alpha \supset K, \text{ so that then } \cap_{\alpha\in\mathscr{F}} D_\alpha \subset cK. \tag{C.3}$$

However, the D_α are subsets of K, so the intersection in (C.3) is empty. This contradicts (C.1), so (C.2) must be true. □

The finite intersection property can be particularly useful if the sets D_α are not only nonempty but also *nested:* for example, when the α are real numbers and if α is greater than β then $D_\alpha \subset D_\beta$. In such a case, the intersection shown in (C.1) will consist of whichever D_α in the finite subset has the largest α, and as that set will be nonempty, so will be the intersection. The proposition then ensures that the intersection of all the sets will be nonempty. Proposition 6.2.8 provides an example of the use of this device.

C.2 Connectedness

Definition C.2.1. Let S be a subset of $\mathbb{R}^n$. S is *connected* if for any two nonempty subsets P and Q of S such that $P \cup Q = S$, at least one of $P \cap \mathrm{cl}\, Q$ and $(\mathrm{cl}\, P) \cap Q$ is nonempty.

Convex sets have the very useful property of being always connected.

Proposition C.2.2. *A convex subset C of $\mathbb{R}^n$ is connected.*

Proof. If C is empty or is a singleton, then it is connected. Otherwise, suppose that D and E are nonempty subsets of C with $D \cup E = C$. If D and E intersect then the connectedness condition is satisfied. If they do not, then as each is nonempty we can find two points $c_D \in D$ and $c_E \in E$ and these points must be distinct. For $\lambda \in [0,1]$ let $x_\lambda = (1-\lambda)c_D + \lambda c_E$, and let $\mu = \sup\{\lambda \in [0,1] \mid c_\lambda \in D\}$. If $c_\lambda \in D$ then $\lambda < 1$, as otherwise D and E would intersect. Then there are points c_μ with $\mu > \lambda$ but as close to λ as we wish. These points must be in E, so $c_\lambda \in D \cap \mathrm{cl}\, E$, which is then nonempty. If $c_\lambda \in E$ then λ cannot be zero. As it is a supremum, there are points $c_\mu \in D$ with $\mu < \lambda$ but as close to it as we wish. Then $c_\lambda \in (\mathrm{cl}\, D) \cap E$, showing that C is connected. □

C.3 The extended real line

We develop here a topology for the extended real line $\bar{\mathbb{R}}$ introduced at the beginning of Chapter 6.

Proposition C.3.1. *Let $\bar{\mathbb{R}}$ be $\mathbb{R}$ augmented by the two symbols $+\infty$ and $-\infty$. We make this collection an ordered set by retaining the standard order on $\mathbb{R}$ and stipulating that $-\infty$ and $+\infty$ are respectively less than and greater than any element of $\mathbb{R}$. Let T be the collection of subsets consisting of*

- *All open subsets of $\mathbb{R}$;*
- *All sets of the form $[-\infty,\alpha) := \{-\infty\} \cup (-\infty,\alpha)$ where α is any element of $\mathbb{R}$;*
- *All sets of the form $(\beta,+\infty] := (\beta,+\infty) \cup \{+\infty\}$ where β is any element of $\mathbb{R}$;*
- *The set $[-\infty,+\infty] := \{-\infty\} \cup \mathbb{R} \cup \{+\infty\} = \bar{\mathbb{R}}$.*

Then T is a base for a topology $\mathscr{T}$ and $(\bar{\mathbb{R}},\ \mathscr{T})$ is a topological space.

Proof. The empty set belongs to T because it is an open subset of $\mathbb{R}$. The union of all sets in T is in T because it is $[-\infty,+\infty]$. Let P and Q be any two sets in T and suppose $x \in P \cap Q$. We will show that there is an element W of T with $x \in W \subset P \cap Q$. Then [19, Theorem 11] shows that T is the base of a topology $\mathscr{T}$ for $\bar{\mathbb{R}}$.

As x exists, neither P nor Q can be empty. Here are the cases that can occur in the choice of P and Q.

- P and Q are open subsets of $\mathbb{R}$: then let $W := P \cap Q$, which is an open subset of $\mathbb{R}$ containing x and contained in T.
- One of P and Q is an open subset of $\mathbb{R}$, and the other has the form $[-\infty,\alpha)$ (case 1) or $(\beta,+\infty]$ (case 2). Suppose for convenience that P is the open subset, say (σ,τ). Then in case 1 we have $\alpha > \sigma$. Let $W := (\sigma,\alpha)$, which is in $P \cap Q$. In case 2 we have $\tau > \beta$ and we let $W := (\beta,\tau)$, which is in $P \cap Q$.
- Each of P and Q has one of the forms $[-\infty,\alpha)$ or $(\beta,+\infty]$. If both are of the first form or both are of the second form, we choose W to be whichever of P or Q is contained in the other. If they are of different forms, then $\alpha > \beta$ and we choose $W := (\beta,\alpha)$. In both cases $W \in P \cap Q$.
- One of P and Q is $[-\infty,+\infty]$. Then the other is a subset of $[-\infty,+\infty]$, so we choose W to be the subset, which equals $P \cap Q$.

□

References

1. Anderson, T.W.: Introduction to Multivariate Statistical Analysis, 2d edn. Wiley, New York (1984). ISBN-10: 0-471-88987-3
2. Armstrong, M.A.: Basic Topology. Undergraduate Texts in Mathematics. Springer, New York (1983). ISBN-13: 978-0-387-90839-7; originally published 1979 by McGraw-Hill (UK)
3. Ben-Israel, A.: Linear equations and inequalities on finite dimensional, real or complex, vector spaces: A unified theory. Journal of Mathematical Analysis and Applications **27**, 367–389 (1969)
4. Ben-Israel, A.: The Moore of the Moore-Penrose inverse. Electronic Journal of Linear Algebra **9**, 150–157 (2002)
5. Berge, C.: Topological Spaces. Macmillan, New York (1963)
6. Carathéodory, C.: Ueber den Variabilitätsbereich der Fourierschen Konstanten von positiven harmonischen Funktionen. Rend. Circ. Mat. Palermo **32**, 193–217 (1911)
7. Farkas, G.: A Fourier-féle mechanikai elv alkalmazásainak algebrai alapjáról [On the algebraic basis of the applications of the mechanical principle of Fourier]. Mathematikai és Physikai Lapok **5**, 49–54 (1896)
8. Farkas, G.: A Fourier-féle mechanikai elv alkalmazásának algebrai alapja [The algebraic basis of the application of the mechanical principle of Fourier]. Mathematikai és Természettudományi Értesitő **16**, 361–364 (1898)
9. Farkas, J.: Über die Theorie der einfachen Ungleichungen. Journal für die reine und angewandte Mathematik **124**, 1–27 (1902)
10. Fudenberg, D., Tirole, J.: Game Theory. MIT Press, Cambridge, MA (1991). ISBN-10 0-262-06141-4
11. Gale, D.: The Theory of Linear Economic Models. University of Chicago Press, Chicago, IL (1989). ISBN-13: 978-0226278841. Originally published 1960 by McGraw-Hill, New York.
12. Giannessi, F.: Constrained Optimization and Image Space Analysis, vol. 1. Springer, New York (2005). ISBN-10: 0-387-24770-X
13. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3d edn. The Johns Hopkins University Press, Baltimore, MD (1996)
14. Gordan, P.: Ueber die Auflösung linearer Gleichungen mit reellen Coefficienten. Mathematische Annalen **6**, 23–28 (1873)
15. Hardy, G.H.: A Course of Pure Mathematics, 10th edn. Cambridge University Press, Cambridge, UK (1967). (First edition 1908.)
16. Hoffman, A.J.: On approximate solutions of systems of linear inequalities. Journal of Research of the National Bureau of Standards **49**, 263–265 (1952)
17. Householder, A.S.: The Theory of Matrices in Numerical Analysis. Dover Publications, Inc., New York (1975). Originally published 1964 by Blaisdell Publishing Co.
18. Jordan, P., von Neumann, J.: On inner products in linear metric spaces. The Annals of Mathematics **36**, 719–723 (1935)
19. Kelley, J.L.: General Topology. No. 27 in Graduate Texts in Mathematics. Springer, New York (1975). Originally published 1955 by Van Nostrand
20. Komiya, H.: Elementary proof for Sion's minimax theorem. Kodai Mathematical Journal **11**, 5–7 (1988)
21. Luce, R.D., Raiffa, H.: Games and Decisions: Introduction and Critical Survey. Dover, New York (1989). ISBN-10 0-486-65943-7; originally published 1957 by John Wiley & Sons
22. Mangasarian, O.L.: Nonlinear Programming. No. 10 in Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (1994). ISBN-13: 978-0898713411. Originally published 1969 by McGraw-Hill, New York
23. Minty, G.J.: A "from-scratch" proof of a theorem of Rockafellar and Fulkerson. Mathematical Programming **7**, 368–375 (1974)
24. Moore, E.H.: On the reciprocal of the general algebraic matrix. Bulletin of the American Mathematical Society **26**, 385–396 (1920). [This is a report entitled "The fourteenth western meeting of the American Mathematical Society"; Moore's report appears on pp. 394–395.]

25. Nashed, M.Z. (ed.): Generalized Inverses and Applications. Academic Press, New York (1976)
26. Netlib: `http://www.netlib.org/index.html`
27. Neumann, J.v., Morgenstern, O.: Theory of Games and Economic Behavior. John Wiley & Sons, New York (1964). Originally published 1944 by Princeton University Press
28. Penrose, R.: A generalized inverse for matrices. Mathematical Proceedings of the Cambridge Philosophical Society **51**, 406–413 (1955)
29. Plambeck, E.L., Fu, B.R., Robinson, S.M., Suri, R.: Sample-path optimization of convex stochastic performance functions. Mathematical Programming **75**, 137–176 (1996)
30. Prékopa, A.: On the development of optimization theory. American Mathematical Monthly **87**, 527–542 (1980)
31. Robinson, S.M.: Some continuity properties of polyhedral multifunctions. Mathematical Programming Studies **14**, 206–214 (1981)
32. Robinson, S.M.: Analysis of sample-path optimization. Mathematics of Operations Research **21**, 513–528 (1996)
33. Rockafellar, R.T.: The elementary vectors of a subspace of $\mathbf{R}^n$. In: R.C. Bose, T.A. Dowling (eds.) Combinatorial Mathematics and Its Applications, pp. 104–127. University of North Carolina Press, Chapel Hill, NC (1969)
34. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton, NJ (1970)
35. Rockafellar, R.T.: Conjugate Duality and Optimization. No. 16 in CBMS Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, PA (1974)
36. Rockafellar, R.T., Wets, R.J.: Variational Analysis. No. 317 in Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin (1998). ISBN13: 978-3-540-62772-3
37. Royden, H.L.: Real Analysis, 2d edn. Macmillan, New York (1968)
38. Schmid, J.: A remark on characteristic polynomials. American Mathematical Monthly **77**, 998–999 (1970)
39. Shapiro, A.: On concepts of directional differentiability. Journal of Optimization Theory and Applications **66**, 477–487 (1990)
40. Shubik, M.: Game Theory in the Social Sciences: Concepts and Solutions. MIT Press, Cambridge, MA (1982). ISBN-10 0-262-19195-4
41. Sion, M.: On general minimax theorems. Pacific Journal of Mathematics **8**, 171–176 (1958)
42. Starr, R.M.: Quasi-equilibria in markets with non-convex preferences. Econometrica **37**, 25–38 (1969)
43. Stewart, G.W.: On the early history of the singular value decomposition. SIAM Review **35**, 551–566 (1993). Stable URL: http://www.jstor.org/stable/2132388
44. Walkup, D.W., Wets, R.J.B.: A Lipschitzian characterization of convex polyhedra. Proceedings of the American Mathematical Society **23**, 167–173 (1969)

Index