1

Some Basic Convex Analysis

1.1 The Problem Domain

The names *Optimization* and *Mathematical Programming* are used more or less interchangeably to describe problems possessing the generic form: Find $\mathbf{x}^*$ such that

$$\mathbf{x}^* = \arg\left(\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})\right), \tag{1.1}$$

where we call the function $f : R^p \to R$ the *objective function* and the set $\mathcal{X}$ the *feasible region*. The point $\mathbf{x} \in R^p$ is feasible if $\mathbf{x} \in \mathcal{X}$. If $\mathcal{X} = R^p$ then the problem is *unconstrained*, while if $\mathcal{X} \subset R^p$ then the problem is *constrained* by the conditions defining $\mathcal{X}$. If $\mathcal{X}$ is defined by an explicit system of functional relations having the form

$$g_i(\mathbf{x}) \geq 0, \qquad i \in \sigma, \tag{1.2}$$

or

$$h_j(\mathbf{x}) = 0, \qquad j \in \nu, \tag{1.3}$$

where σ, ν are index sets, then these relations are called *constraints* (*inequality* and *equality* constraints respectively). Two aspects of Problem (1.1) are of principal concern:

1. theoretical: questions relating to the existence and characterization of solutions;
2. computational: questions relating to the development of algorithms for computing solutions.

What are the nature of these theoretical questions? They are not trivial even if $\mathcal{X}$ is a finite set of elements and f is a simple function (even linear) on $\mathcal{X}$.

In principal such a problem can be solved by direct enumeration, but the computational complexity of such a process can be enormous. Thus the interest lies in classifying problems for which fast algorithms are possible, and in proving their computational complexity. If f is linear then it cannot assume an extreme value at a finite point so the structure of $\mathcal{X}$ becomes of crucial importance, and the questions of interest are those which seek to characterize $\mathcal{X}$ in terms of its intrinsic geometry. The simplest class of functions generalizing linear functions, at least from our point of view, is the class of convex functions. This contains the functions of principal interest here. Also, it forms a convenient framework to use to illustrate aspects of more general theory because the passage from local scale to global scale introduces no surprises so that, for example, a local minimum is also a global minimum. In general, of course, this is not so, and global questions are both significant and difficult. One point of departure from [45] is the discussion of some naturally occurring polyhedral but nonconvex problems. One context in which these arise is in the modification of statistical estimation procedures to make them more robust in the presence of outlying observations. A curious point reported here is the tendency for these procedures to become more convex as the number of data points is increased.

The starting point used here notes that intuitively $\mathbf{x}^*$ is a local minimum if there are no points in the immediate neighborhood of $\mathbf{x}^*$ that are both contained in $\mathcal{X}$ and give a lower value of f. This property of separation can be exploited very effectively in the convex case. This is shown in Figure 1.1. There the

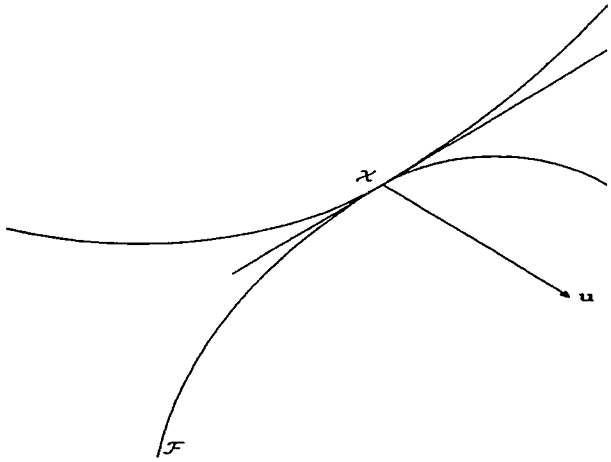

Figure 1.1. Separation is a basic paradigm.

parabolic curves could represent respectively the set $\mathcal{F} = \{\mathbf{x}; f(\mathbf{x}) < f(\mathbf{x}^*)\}$ and the constraint set $\mathcal{X}$.

Separation has the important advantage that it does not require differentiability to characterize the minimum. This is important for applications of polyhedral convex functions as the points at which the function fails to be differentiable are often the points of greatest interest. An alternative way to look at these ideas is also important. This notes that if $\mathbf{x}^*$ is not the minimum then $\mathcal{X} \cap \mathcal{F} \neq \emptyset$. It follows that there is a feasible direction at $\mathbf{x}^*$ in which it is possible to decrease f. Results that describe this situation, in which either $\mathbf{x}^*$ is optimal or there is a direction in which f can be decreased, are called "Theorems of the Alternative".

Certain themes are important in developing computational algorithms. Here the idea of a descent method is central. In this there are two main ingredients:

1. a recipe for generating a direction contained in $\mathcal{X}$ (a feasible direction) in which f can be decreased; such a direction is called a *descent* direction;
2. a method for stepping in this direction to achieve an effective decrease in f; this calculation is called the *linesearch* step.

A further important question concerns the stability of the computational algorithm. This has to do with the manner in which perturbations of the problem data affect the results of the computation. Stability under perturbation is an important property in constructing convergence proofs. Again it is convenient to distinguish two aspects:

1. Local stability, which has to do with small perturbations about the current point. Verification of this property is frequently not too difficult.
2. Global stability, which has to do with behavior of the method along the possible class of solution trajectories and is frequently much harder to categorize.

A fast rate of convergence is a further desirable property of an algorithm. Other ideas include penalization as a means for treating the constraint set. It involves making compromises. Also, homotopy ideas can be relevant to developing algorithms as the idea of following a trajectory has much in common with the idea of a descent method.

Example 1.1.1 *The linear programming problem. This problem considers the minimum of a linear function subject to linear inequality constraints: The key point in this example is that the constraint geometry determines the solution.*

$$\min_{\mathbf{x}\in\mathcal{X}} \mathbf{c}^T\mathbf{x}; \qquad \mathcal{X} = \{\mathbf{x}; A\mathbf{x} \geq \mathbf{b}\}.$$

Cambridge University Press
978-0-521-79133-5: Simplicial Algorithms for Minimizing Polyhedral Functions
M. R. Osborne
Excerpt
More information

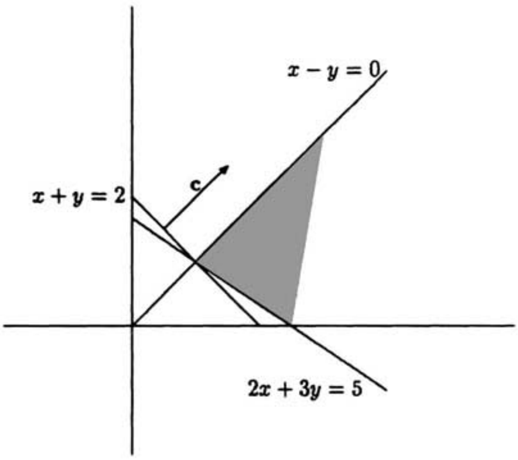

Figure 1.2. Simple linear programming example.

Let

$$\mathbf{c} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & -1 \\ 2 & 3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ 5 \end{bmatrix}.$$

Then it follows from Figure 1.2 that the minimum value is 2 and is obtained at the vertex $\mathbf{x} = \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$. *Also,*

$$\begin{bmatrix} 1 & 2 \\ -1 & 3 \end{bmatrix} \mathbf{u} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \mathbf{u} = \frac{1}{5} \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

This expresses the gradient of the objective function as a linear combination of the active constraint gradients and so has the form of a multiplier relation. A point to note is that the multiplier vector components are positive.

Example 1.1.2 *A simple l_1 fitting problem. This illustrates a simple nonsmooth optimization problem. Consider the problem*

$$\min_x \; |x| + |x-1| + |x-2| + |x-3| + |x-4|.$$

This has the graph shown in Figure 1.3. Here the function is linear except at the points where an absolute value term vanishes and these prove to be the only interesting points. The minimum occurs at $x = 2$, and this is also the median of the data points.

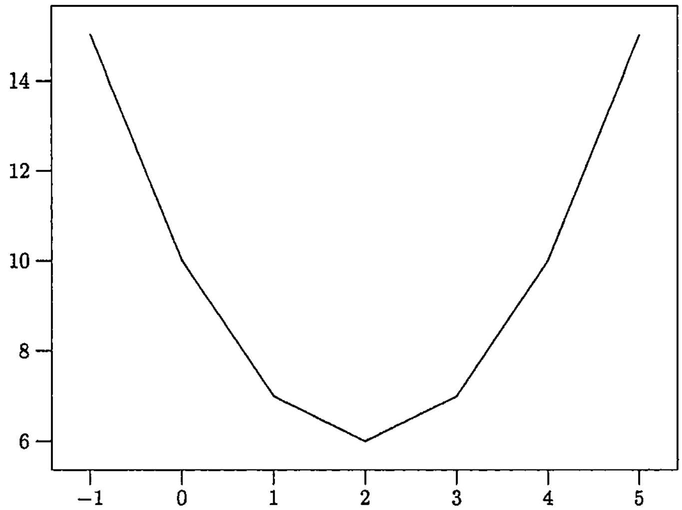

Figure 1.3. Simple l_1 example.

1.2 Properties of Convex Sets

In this section necessary material on the representation of convex sets and on separation theorems presented in [45] is summarized. As indicated in the previous section, separation properties are very important in characterizing optima in convex problems. Here the basic results governing the separation of disjoint convex sets are surveyed. An essential preliminary is the provision of an appropriate representation of convex sets.

Definition 1.1 *The set $\mathcal{X} \subseteq R^p$ is convex if*

$$\mathbf{x}, \mathbf{y} \in \mathcal{X} \Rightarrow \theta\mathbf{x} + (1-\theta)\mathbf{y} \in \mathcal{X}, \quad \textit{for } 0 \le \theta \le 1. \tag{1.4}$$

Equivalently, $\mathcal{X}$ is convex if all finite convex combinations of points in $\mathcal{X}$ are again in $\mathcal{X}$:

$$\mathbf{x}_i \in \mathcal{X}, \quad i = 1, 2, \ldots, n, \quad \sum_{i=1}^{n} \lambda_i = 1, \quad \lambda_i \ge 0, \quad 1 \le n < \infty,$$
$$\Rightarrow \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \in \mathcal{X}.$$

This follows by induction. For example, if the result holds for $k \le n$ then the following rearrangement

$$\sum_{i=1}^{n+1} \lambda_i \mathbf{x}_i = \frac{\sum_{i=1}^{n} \lambda_i}{\sum_{i=1}^{n+1} \lambda_i} \sum_{i=1}^{n} \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i} \mathbf{x}_i + \lambda_{n+1} \mathbf{x}_{n+1}$$

shows that it holds also for $n+1$.

Remark 1.2.1 *Given any set $\mathcal{X}$, the set obtained by taking all possible convex combinations of points in $\mathcal{X}$ is called the convex hull of $\mathcal{X}$ and is written conv($\mathcal{X}$). It is the smallest convex set containing $\mathcal{X}$.*

Example 1.2.1 *A function $\|*\| : R^p \to R$ is a norm provided it satisfies the conditions.*

N1 $\|\mathbf{x}\| > 0, \mathbf{x} \neq 0$;
N2 $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, *the triangle inequality*;
N3 $\|a\mathbf{x}\| = |a| \|\mathbf{x}\|, a \in R$.

Consider the set $\mathcal{X}$ defined by $\mathcal{X} = \{\mathbf{x}; \|\mathbf{x}\| \leq 1\}$. Then $\mathcal{X}$ satisfies the following conditions:

S1 $\mathcal{X}$ is convex – this follows directly from the triangle inequality;
S2 $\mathcal{X}$ is balanced so that $\mathbf{x} \in \mathcal{X} \Leftrightarrow -\mathbf{x} \in \mathcal{X}$; and
S3 $\mathcal{X}$ has a proper interior.

Alternatively, given $\mathcal{X}$ satisfying conditions S1–S3, then the function on $R^n \to R$ defined by

$$\|\mathbf{x}\| = \inf_{\lambda > 0} \lambda, \qquad \mathbf{x} \in \lambda \mathcal{X} \tag{1.5}$$

satisfies conditions N1–N3. For example, to demonstrate the triangle inequality N2, note that both $\frac{\mathbf{x}}{\|\mathbf{x}\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|} \in \mathcal{X}$. It follows from the convexity of $\mathcal{X}$ that

$$\begin{aligned} \frac{\|\mathbf{x}\|}{\|\mathbf{x}\| + \|\mathbf{y}\|} \frac{\mathbf{x}}{\|\mathbf{x}\|} + \frac{\|\mathbf{y}\|}{\|\mathbf{x}\| + \|\mathbf{y}\|} \frac{\mathbf{y}}{\|\mathbf{y}\|} \in \mathcal{X} \\ \Rightarrow \mathbf{x} + \mathbf{y} \in (\|\mathbf{x}\| + \|\mathbf{y}\|)\mathcal{X} \\ \Rightarrow \|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|. \end{aligned}$$

If $\mathcal{X}$ is not balanced then condition N3 is not satisfied.

Definition 1.2 *The set of codimension* 1

$$H(\mathbf{u}, \upsilon) = \{\mathbf{x} \in R^p; \mathbf{u}^T \mathbf{x} = \upsilon\} \tag{1.6}$$

is called a hyperplane.

Remark 1.2.2 *It is important to note that $H(\mathbf{u}, \upsilon)$ separates R^p into two distinct halfspaces. Here the convention is followed that these are labeled*

$$H^+(\mathbf{u}, \upsilon) = \{\mathbf{x};\ \mathbf{u}^T\mathbf{x} > \upsilon\},$$
$$H^-(\mathbf{u}, \upsilon) = \{\mathbf{x};\ \mathbf{u}^T\mathbf{x} \le \upsilon\}.$$

Theorem 1.1 *(Separating hyperplane lemma – simplest case). Let $\mathcal{X} \neq \emptyset$, $\mathcal{X} \subset R^p$ be convex and closed, and let $\mathbf{x}_0 \notin \mathcal{X}$. Then there exists H such that $\mathcal{X} \subset H^-$, $\mathbf{x}_0 \in H^+$.*

Proof Let $\mathbf{x}_1 \in \mathcal{X}$. Then

$$\inf_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \mathbf{x}_0\|_2^2 \le \|\mathbf{x}_1 - \mathbf{x}_0\|_2^2 \le r^2.$$

It follows that $\|\mathbf{x} - \mathbf{x}_0\|_2^2$ is continuous on the closed bounded set $\mathcal{X} \cap \{\mathbf{x};\ \|\mathbf{x} - \mathbf{x}_0\|_2 \le r\}$, so that there exists $\mathbf{x}^* \in \mathcal{X}$ at which the minimum is attained. Choose $\mathbf{y} \in \mathcal{X}$, set $\mathbf{z} = \mathbf{y} - \mathbf{x}^*$, and let $0 \le \gamma \le 1$. Then $\mathbf{x} = \mathbf{x}^* + \gamma\mathbf{z} \in \mathcal{X}$, and

$$\begin{aligned}\|\mathbf{x} - \mathbf{x}_0\|_2^2 &= \|\mathbf{x}^* - \mathbf{x}_0\|_2^2 + 2\gamma\mathbf{z}^T(\mathbf{x}^* - \mathbf{x}_0) + \gamma^2\|\mathbf{z}\|_2^2 \\ &\Rightarrow 2\gamma\mathbf{z}^T(\mathbf{x}^* - \mathbf{x}_0) + \gamma^2\|\mathbf{z}\|_2^2 \ge 0 \\ &\Rightarrow \mathbf{z}^T(\mathbf{x}^* - \mathbf{x}_0) \ge 0,\ \forall \mathbf{y} \in \mathcal{X},\end{aligned}$$

by choosing γ small enough. It follows that

$$H(\mathbf{x}^* - \mathbf{x}_0, \mathbf{x}^{*T}(\mathbf{x}^* - \mathbf{x}_0) - \varepsilon)$$

is a suitable hyperplane for a small enough $\varepsilon > 0$. In this case there is a small ball about $\mathbf{x}_0$ not containing points of $\mathcal{X}$. This is called *strong separation.* ■

This result can be extended in several ways. The most important of these refer to cases where the quantities involved have no points in common, but this is not true of their closures. Such a result is the following:

Theorem 1.2 *(Separating hyperplane lemma – stronger case). Let $\mathcal{X}, \mathcal{Y}$ be convex sets, $\mathcal{X}, \mathcal{Y} \subset R^p$, $\mathcal{X} \cap \mathcal{Y} = \emptyset$. Then there exists a hyperplane H such that, after possible relabeling of $\mathcal{X}$ and $\mathcal{Y}$ so that the set in H^- is closed in the case that either or both are closed,*

$$\mathcal{X} \subseteq H^-, \quad \mathcal{Y} \subset H^+.$$

Equivalently, there exists $\mathbf{u} \in R^p$ such that

$$\sup_{\mathbf{x}\in\mathcal{X}} \mathbf{u}^T\mathbf{x} \le \inf_{\mathbf{x}\in\mathcal{Y}} \mathbf{u}^T\mathbf{x}. \tag{1.7}$$

This result can be proved by induction when p is finite [59, p. 96]. It requires the Hahn–Banach Theorem in general [37].

Remark 1.2.3 *Let $\mathcal{X}$ be an open convex set, and $\mathbf{x}_0 \in \mathrm{cl}(\mathcal{X}) \setminus \mathcal{X}$. Then there exists a separating hyperplane H containing $\mathbf{x}_0$ and such that $\mathcal{X} \subset H^-$. Thus H contains only points of $\mathrm{cl}(\mathcal{X}) \setminus \mathcal{X}$, and H is said to be a supporting hyperplane to $\mathcal{X}$ at $\mathbf{x}_0$. This result extends to more general sets $\mathcal{Y}$, $\mathcal{X} \cap \mathcal{Y} = \emptyset$, that can be contained in a hyperplane. Then the result tells us that $\mathcal{Y}$ can be contained in a hyperplane supporting $\mathcal{X}$.*

Remark 1.2.4 *Let the* sup *in (1.7) be attained at $\mathbf{x}_0 \in \mathcal{X}$. Then $\mathbf{x}_0$ is a boundary point of $\mathcal{X}$, $H(\mathbf{u}, \mathbf{u}^T\mathbf{x}_0)$ supports $\mathcal{X}$ at $\mathbf{x}_0$, and $\mathrm{cl}(\mathcal{X}) \subseteq H^-$. The convention will be followed that if H supports $\mathcal{X}$ then $\mathcal{X} \subseteq H^-$.*

Definition 1.3 *Associated with the convex set $\mathcal{X}$ is the support function δ^* given by*

$$\delta^*(\mathbf{u} \mid \mathcal{X}) = \sup_{\mathbf{x} \in \mathcal{X}} \mathbf{u}^T\mathbf{x}. \tag{1.8}$$

Remark 1.2.5 *This is not the only definition of support function. A useful alternative in two dimensions is sketched in Figure 1.4. Here the support function, often denoted by p in this context, is defined as the perpendicular distance from the origin to the tangent to the set at the current point $\mathbf{r}(\theta) = x(\theta)\mathbf{i} + y(\theta)\mathbf{j}$. It is convenient to parametrize $p = p(\theta)$ by the angle θ that the tangent vector $\mathbf{t}$ to the set at $\mathbf{r}$ makes with the vertical axis that it intersects at the point $(0, \overline{y})$. If arc length on the set boundary is given by s then*

$$\mathbf{t} = \frac{d\mathbf{r}}{ds}, \qquad \|\mathbf{t}\| = 1,$$
$$\cos\theta = \frac{d\mathbf{r}}{ds} \cdot \mathbf{j} = \mathbf{t} \cdot \mathbf{j} = \frac{dy}{ds}.$$

Also, as $\mathbf{t}$ is a unit vector,

$$\sin\theta = -\frac{dx}{ds}, \quad \frac{d\mathbf{t}}{ds} = \frac{1}{\rho}\mathbf{n},$$
$$\mathbf{n} = -\frac{dy}{ds}\mathbf{i} + \frac{dx}{ds}\mathbf{j} = \frac{d\mathbf{t}}{d\theta},$$

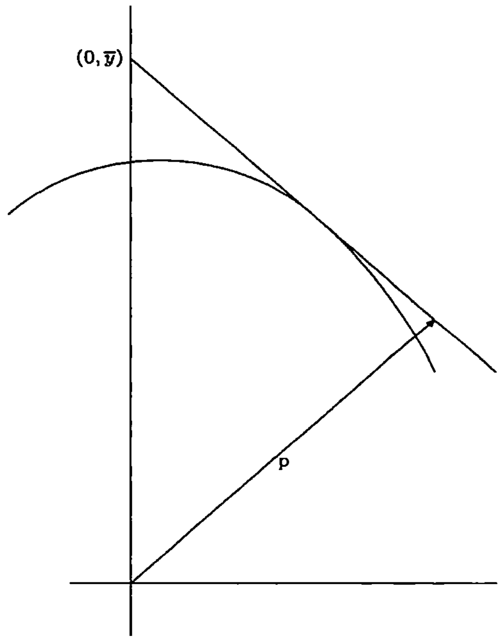

Figure 1.4. Support function illustrated.

where ρ *is the radius of curvature* ($\rho \geq 0$ *for convex sets), and* $\mathbf{n}$, $\mathbf{n} \cdot \mathbf{t} = 0$, *is the inward pointing unit normal at* $\mathbf{r}$. *Note that*

$$\frac{d\mathbf{t}}{ds} = \frac{d\mathbf{t}}{d\theta}\frac{d\theta}{ds} \Rightarrow \rho = \frac{ds}{d\theta}.$$

Elementary geometry gives

$$p = -\mathbf{n} \cdot \mathbf{r} = x(\theta)\cos\theta + y(\theta)\sin\theta,$$

whence

$$\frac{dp}{d\theta} = -\frac{d\mathbf{n}}{d\theta} \cdot \mathbf{r} = \mathbf{t} \cdot \mathbf{r} = -x(\theta)\sin\theta + y(\theta)\cos\theta.$$

Differentiating again gives

$$\frac{d^2 p}{d\theta^2} = \mathbf{n} \cdot \mathbf{r} + \frac{ds}{d\theta}$$

as $\mathbf{t} \cdot \mathbf{t} = 1$ *and* $\frac{d\mathbf{t}}{d\theta} = \mathbf{n}$. *It follows that*

$$\frac{d^2 p}{d\theta^2} + p = \rho \geq 0,$$

and this last condition gives a criterion for the convexity of the set. The connection with the previous definition of support function follows readily. We have

$$\delta^*(-\mathbf{n} \mid \mathcal{X}) = -\mathbf{n} \cdot \mathbf{r} = p$$

as the supremum (1.8) must be taken on the boundary of the set.

Definition 1.4 *The point $\mathbf{x}_0 \in \mathcal{X}$ is an extreme point of $\mathcal{X}$ if it does not lie on the line joining of any two distinct points in $\mathrm{cl}(\mathcal{X})$, the closure of $\mathcal{X}$.*

Definition 1.5 *The point $\mathbf{x}_0 \in \mathcal{X}$ is an exposed point of $\mathcal{X}$ if $\exists H$, $H \cap \mathrm{cl}(\mathcal{X}) = \{\mathbf{x}_0\}$.*

Definitions 1.4 and 1.5 are equivalent for convex sets if the set of extreme points of $\mathcal{X}$ is finite.

Lemma 1.1 *Let the hyperplane $H = H(\mathbf{u}, v)$ support $\mathcal{X}$ at $\mathbf{x}$. If $\mathbf{y}$ is an extreme point of $H \cap \mathcal{X}$, then $\mathbf{y}$ is an extreme point of $\mathcal{X}$.*

Proof If $H \cap \mathcal{X} = \{\mathbf{y}\}$ there is nothing to prove. Thus assume $H \cap \mathcal{X} = \mathcal{Y}$, where $\mathcal{Y}$ is not a singleton, and that $\mathbf{y}$ is an extreme point of $\mathcal{Y}$ but not of $\mathcal{X}$. Then

$$\exists \mathbf{w}, \quad \mathbf{z} \in \mathcal{X}, \quad \notin \mathcal{Y}, \quad 0 \le \theta \le 1, \quad \ni \mathbf{y} = \theta \mathbf{w} + (1-\theta)\mathbf{z}.$$

But $\mathbf{y} \in H$, $\mathbf{w}, \mathbf{z} \in (H^-)^o$. It follows that

$$0 = \mathbf{u}^T\mathbf{y} - v = \theta(\mathbf{u}^T\mathbf{w} - v) + (1-\theta)(\mathbf{u}^T\mathbf{z} - v) < 0.$$

This gives a contradiction. ■

Lemma 1.2 *Let $\mathcal{X}$ be a closed bounded convex set. Then $\mathcal{X}$ has extreme points.*

Proof This is another induction argument using induction with respect to dimension. The result is clear if $\mathcal{X}$ is a singleton. Let H support $\mathcal{X}$ at $\mathbf{x}_0$, a boundary point of $\mathcal{X}$. Then $H \cap \mathcal{X}$ is closed, bounded, and has dimension less than p. The intersection has extreme points by the induction hypothesis. By Lemma 1.1 these are extreme points of $\mathcal{X}$. ■

Theorem 1.3 *A closed bounded convex set $\mathcal{X} \subset R^p$ is the closed convex hull of its extreme points.*