Caltech Division of the Humanities and Social Sciences

Miscellaneous Notes on Optimization Theory and Related Topics

or

A Dessert Course in Optimization*

KC Border

Continually under revision.
Last Seriously Revised: November 20, 2015
This version was printed:
November 20, 2015

*These notes have been compiled from my lecture notes and handouts in various courses over thirtysomething years. Consequently, they are neither complete nor well organized, and the notation is not consistent throughout. I am of two minds about the role of these notes. In some sections, the notes are a kind of tutorial, designed to teach students. In other places they are a reference, or crib sheet, where there is not much in the way of motivation or elaboration. I figure that if you are reading those parts, you already have something in mind that you want to know. I have concentrated on including and elucidating as many obscure proofs as I have had time to type up, and omitted what I think is widely known. More results will be included in future versions, and some may disappear. It is a sad fact that as we learn more, we must omit more topics from the curriculum to make way for the new ideas. As a result, we forget what past generations knew. Outside of typographical errors, I believe the results here to be correct, except for those marked as conjectures, which I also believe to be true. On the other hand, I keep finding errors. I have tried to indicate whence I have "borrowed" ideas. This is difficult because "there is nothing new under the sun" (Ecclesiastes 1.9), and as Pete Seeger once attributed to his father, a musicologist, "Plagiarism is basic to all culture." Unless otherwise noted, I learned every argument or example from Jim Quirk, Joel Franklin, Ket Richter, Leo Hurwicz, or Taesung Kim, or else I read it in one of the references and forgot it, or on occasion it is an original reinvention.

Contents

Section 1

Introduction

1.1 Miscellaneous notation

The set of real numbers is denoted $\boldsymbol{R}$, while n-dimensional Euclidean space is denoted $\boldsymbol{R}^{\mathrm{n}}$. I adopt the usual convention of denoting the i^{th} **unit coordinate vector** by e^i, regardless of the dimension of the underlying space. That is, the i^{th} coordinate of e^i is one and all the others are zero. Similarly, the **unit vector**, which has all its components equal to one, is denoted $\mathbf{1}$, regardless of the dimension of the space. (This includes infinite-dimensional sequence spaces.)

For vectors $x, y \in \boldsymbol{R}^{\mathrm{n}}$, define

$$x \cdot y = \sum_{i=1}^{n} x_i y_i,$$

the **Euclidean inner product**. The **Euclidean norm** of a vector in $\boldsymbol{R}^{\mathrm{n}}$ is defined by

$$\|x\| = (x \cdot x)^{\frac{1}{2}} = \left(\sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}}.$$

We may occasionally wish to think of coordinates of a vector as being numbered from 0 through n rather than 1 through $n+1$.

The usual ordering on $\boldsymbol{R}$ is denoted $\geqslant$ or $\leqslant$. On $\boldsymbol{R}^{\mathrm{n}}$, the ordering $x \geqq y$ means $x_i \geqslant y_i$, $i = 1, \ldots, n$, while $x \gg y$ means $x_i > y_i$, $i = 1, \ldots, n$. We may occasionally write $x > y$ to mean $x \geqq y$ and $x \neq y$. A vector x is **nonnegative** if $x \geqq 0$, **strictly positive** if $x \gg 0$, and **semipositive** if $x > 0$. I shall try to avoid using the adjective "positive" by itself, since to most mathematicians it means "nonnegative," but to many nonmathematicians it means "strictly positive." Define $\boldsymbol{R}^{\mathrm{n}}_{+} = \{x \in \boldsymbol{R}^{\mathrm{n}} : x \geqq 0\}$ and $\boldsymbol{R}^{\mathrm{n}}_{++} = \{x \in \boldsymbol{R}^{\mathrm{n}} : x \gg 0\}$, the **nonnegative orthant** and **strictly positive orthant** of $\boldsymbol{R}^{\mathrm{n}}$ respectively.

$$\begin{array}{rcl} x \geqq y & \iff & x_i \geqslant y_i,\ i = 1, \ldots, n \\ x > y & \iff & x_i \geqslant y_i,\ i = 1, \ldots, n \text{ and } x \neq y \\ x \gg y & \iff & x_i > y_i,\ i = 1, \ldots, n \end{array}$$

Figure 1.1. Orderings on $\boldsymbol{R}^{\mathrm{n}}$.

1.1.1 Extended real numbers

The **extended real number system** $\boldsymbol{R}^{\sharp}$ consists of the real numbers plus two additional entities ∞ (sometimes known as $+\infty$) and $-\infty$. The ordering of the real numbers is extended

so that for any real number α, we have $-\infty < \alpha < \infty$. Furthermore we extend the definitions of addition and multiplication as follows.

$$\alpha + \infty = \infty \quad \text{and} \quad \alpha - \infty = -\infty;$$

$$\infty \cdot \alpha = \infty, \text{ if } \alpha > 0 \quad \text{and} \quad \infty \cdot \alpha = -\infty, \text{ if } \alpha < 0;$$

$$\infty \cdot 0 = 0;$$

for any real number α. The combination $\infty - \infty$ of symbols has no meaning.

1.1.2 Infimum and supremum

A set A of real numbers is **bounded above** if there exists some real number α, satisfying $\alpha \geqslant x$ for all $x \in A$. In this case we say that x is an **upper bound** for A. Similar definitions apply for lower bounds.

A number is the **greatest element** of A if it belongs to A and is an upper bound for A. A lower bound for A that belongs to A is the least element of A. Note that greatest and least elements are unique, for if x and y are both upper bounds that belong to A, then $x \geqslant y$ and $y \geqslant x$, so $x = y$.

The **infimum** of a set A of real numbers, denoted $\inf A$, is the greatest lower bound of A in the set of extended real numbers. That is,

$$\forall \alpha \in A \quad \inf A \leqslant \alpha,$$

and for any other extended real β,

$$(\forall \alpha \in A \quad \beta \leqslant \alpha) \implies \beta \leqslant \inf A.$$

The **supremum** of A is the greatest lower bound of A in the extended real numbers. Note that the definitions imply the following.

$$\inf \varnothing = \infty \quad \text{and} \quad \sup \varnothing = -\infty.$$

The real numbers are constructed so that they have the following properties:

1 Fact (The real numbers are complete.) *If a nonempty set of real numbers is bounded above, then it has a supremum. If a nonempty set of real numbers is bounded below, then it has an infimum.*

1.1.3 Contours of a function

Given a real function $f\colon X \to \boldsymbol{R}$, we may use the statistician's convention where

$$[f > \alpha] \quad \text{means} \quad \{x \in X : f(x) > \alpha\}, \text{ etc.}$$

A set of the form $[f = \alpha]$ is a **level set** of f, $[f \geqslant \alpha]$ is an **upper contour set**, and $[f > \alpha]$ is a **strict upper contour set** of f.

The **graph** of a function $f\colon X \to \boldsymbol{R}^\sharp$ is just $\{(x, \alpha) \in X \times \boldsymbol{R} : \alpha = f(x)\}$. The **epigraph** is $\{(x, \alpha) \in X \times \boldsymbol{R} : \alpha \geqslant f(x)\}$, and the **hypograph**[1] is $\{(x, \alpha) \in X \times \boldsymbol{R} : \alpha \leqslant f(x)\}$.

[1] Some authors use the term **subgraph** instead of hypograph, but epi- and hypo- are Greek prefixes, while super- and sub- come from Latin. I will stick with Greek here, since no one says "supergraph."

1.2 Maxima and minima

A function $f\colon X \to \boldsymbol{R}$ attains a **global maximum** or **absolute maximum** over X at $x^* \in X$ if $f(x^*) = \sup_{x\in X} f(x)$. That is, if $f(x^*) \geqslant f(x)$ for every $x \in X$. We may also say that x^* is a **global maximizer** or **global maximum point** of f on X, or that x^* **maximizes** f over X. The value $f(x^*)$ is called the **global maximum** of f on (or over) X.[2]

The function f achieves a **minimum** at x^* if $f(x^*) \leqslant f(x)$ for every $x \in X$. An **extremum** of f is a point where f attains either a maximum or a minimum. Notice that f achieves a maximum at x^* if and only if $-f$ achieves a minimum there. Consequently, anything we say about maxima can be converted into a statement about minima by appropriately changing signs. In particular, all the definitions listed here regarding maxima have corresponding definitions for minima. A function f has a **strict maximum** on X at $x^* \in X$ if $f(x^*) > f(x)$ for every $x \in X$ satisfying $x \neq x^*$. We may sometimes use the term **weak maximum** instead of maximum to emphasize that we do not necessarily mean a strict maximum. When $X \subset \boldsymbol{R}^{\mathrm{n}}$, we say that f has a **local maximum** or **relative maximum** at x^* or that x^* is a **local maximizer** of f on X if there is a neighborhood U of x^* in X such that x^* maximizes f on U.

Introduce general notation for vector-valued functions.

Often we are interested in maxima and minima of f on a subset of its domain. A common way to define such a subset is in terms of inequality constraints of the form $g_j(x) \geqslant \alpha_j$, where $g_j\colon X \to \boldsymbol{R}$, $j = 1, \dots, m$. (Or, we may write $g(x) \geqq a$, where $g = (g_1, \dots, g_m)\colon X \to \boldsymbol{R}^{\mathrm{m}}$, and $a = (\alpha_1, \dots, \alpha_m)$.) We say that a point x **satisfies the constraints** if it belongs to X and $g(x) \geqq a$. The set of points satisfying the constraints, $[g \geqq a] = \{x \in X : g(x) \geqq a\}$, is called the **constraint set**. The function f itself may be referred to as the **objective function** to distinguish it from the **constraint functions** $g_1, \dots, g_m$.

We say that x^* is a **constrained maximizer** of f if x^* satisfies the constraints, $g(x^*) \geqq a$, and $f(x^*) \geqslant f(x)$ for every x satisfying the constraints. The point x is a **local constrained maximizer** if it satisfies the constraints and there is a neighborhood U of x such that $f(x^*) \geqslant f(x)$ for every $x \in U$ satisfying the constraints. The point x^* is an **interior** maximizer of f if its lies in the relative interior of the constraint set.

[2] Over the years I have slipped into the practice of referring to the point x^* as a maximum of f. Roko Aliprantis has convinced me that this is potentially confusing at best and wrong at worst. For instance, what is the maximum of the cosine function, 1 or 0? Most people will answer 1, but this value is attained at 0.

I shall endeavor to avoid this perhaps erroneous practice, but I may backslide. Incidentally, I am not alone. A quick random sample of my bookshelves reveals that Luenberger [103] and Varian [156] also refer to x^* as a maximum of f, while Apostol [7], Debreu [41], and Mas-Colell, Whinston, and Green [108] do not. Some authors equivocate, e.g., Marsden [107]. The venerable Hancock [71] says that "f is a maximum for x^*."

Section 2

Basic topologcal concepts

This section is to briefly summarize some of the theory of topological spaces, especially metric spaces, but is not intended to be comprehensive. An excellent treatment of metric spaces may be found in Rudin [134, Chapter 2] or Aliprantis and Border [3, Chapters 2 and 3]. My favorite books on general topology include Aliprantis and Border [3], Wilanksy [160], and Willard [161].

I recently combined several disjoint sets of notes here. I need to make sure that things are in order, and that I make it clear which results apply only to metric spaces.

2.1 Metric spaces

A **metric** (or **distance**) on a nonempty set X is a function $d\colon X \times X \to \boldsymbol{R}$ satisfying the following four properties:

1. Positivity: $d(x, y) \geqslant 0$ and $d(x, x) = 0$ for all $x, y \in X$.
2. Discrimination: $d(x, y) = 0$ implies $x = y$.
3. Symmetry: $d(x, y) = d(y, x)$ for all $x, y \in X$.
4. Triangle Inequality: $d(x, y) \leqslant d(x, z) + d(z, y)$ for all $x, y, z \in X$. (See Figure 2.1.)

If d is a metric on a set X, then the pair (X, d) is called a **metric space**. A **semimetric** satisfies properties (1), (3), and (4).

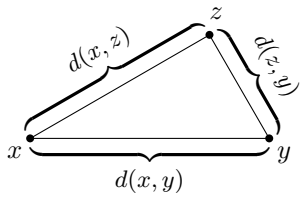

Figure 2.1. The triangle inequality.

2 Example (Metrics) Let X be any nonempty set. The function d defined by

$$d(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}$$

is a metric, called the **discrete metric**.

The natural metric on $\boldsymbol{R}$ is

$$d(x,y) = |x-y|.$$

There are several natural metrics on $\boldsymbol{R}^{\mathrm{n}}$.

The **Euclidean metric** is defined by

$$d(x,y) = \Big(\sum_{i=1}^{n} |x_i - y_i|^2\Big)^{\frac{1}{2}}.$$

Here is a roundabout but instructive proof of the triangle inequality for the Euclidean metric:

We start by proving the **Cauch–Schwartz inequality**,

$$|x \cdot y| \leqslant \|x\| \cdot \|y\|.$$

First observe that if either $x = 0$ or $y = 0$, this inequality holds an equality, so we need only consider the case where $x \neq 0$ and $y \neq 0$. Now observe that

$$\|\alpha x + \beta y\|^2 = (\alpha x + \beta y) \cdot (\alpha x + \beta y) = \alpha^2\|x\|^2 + 2\alpha\beta x \cdot y + \beta^2\|y\|^2 \tag{2.1}$$

so that

$$-2\alpha\beta x \cdot y = \alpha^2\|x\|^2 + \beta^2\|y\|^2 - \|\alpha x + \beta y\|^2 \leqslant \alpha^2\|x\|^2 + \beta^2\|y\|^2.$$

Considering α and β of various signs, we conclude that

$$2|\alpha\beta|\,|x \cdot y| \leqslant \alpha^2\|x\|^2 + \beta^2\|y\|^2.$$

Setting $\alpha = 1/\|x\|$ and $\beta = 1/\|y\|$ and rearranging gives the desired inequality.

Setting $\alpha = \beta = 1$ in (2.1) and applying the Cauchy–Schwartz yields

$$\|x+y\|^2 = \|x\|^2 + 2x \cdot y + \|y\|^2 \leqslant \|x\|^2 + 2|x \cdot y| + \|y\|^2 \leqslant \|x\|^2 + 2\,\|x\|\,\|y\| + \|y\|^2 = \big(\|x\| + \|y\|\big)^2$$

Taking square roots yields the following, which is also called the triangle inequality,

$$\|x+y\| \leqslant \|x\| + \|y\|.$$

Thus

$$d(x,z) = \|x-z\| = \|x-y+y-z\| \leqslant \|x-y\| + \|y-z\| = d(x,y) + d(y,z),$$

which proves the triangle inequality for the Euclidean metric.

The $\boldsymbol{\ell_1}$ **metric** is defined by

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|.$$

The **sup metric** or **uniform metric** is defined by

$$d(x,y) = \max_{i=1,\dots,n} |x_i - y_i|.$$

These definitions can be extended to spaces of infinite sequences, by replacing the finite sum with an infinite series. The infinite dimensional spaces of sequences where every point has a finite distance from zero under these metrics (that is, when the infinite series is absolutely convergent) are called ℓ_2, ℓ_1, and ℓ_∞, respectively. (Although for the metric on ℓ_∞, the maximum must replaced by a supremum: $d(x,y) = \sup_n |x_n - y_n|$.)

In general, a norm $\|\cdot\|$ on a vector space defines a metric by

$$d(x,y) = \|x-y\|.$$

□

Given a subset A of a metric space (X,d), the **distance function for** $\boldsymbol{A}$, $x \mapsto d(x,A)$, is defined by

$$d(x,A) = \inf\{d(x,y) : y \in A\}.$$

Clearly such a function is nonnegative.

2.2 Open sets in metric spaces

Let (X, d) be a metric space. A point x is an **interior point** of a subset A of X if there is some $\varepsilon > 0$ such that the **ε-neighborhood**, or **ε-ball**,

$$B_\varepsilon(x) = \{y \in X : d(y, x) < \varepsilon\}$$

is included in A. In that case we also say that A is a **neighborhood** of x.[1] The set of interior points is called the **interior** of X, denoted $\operatorname{int} X$ or sometimes X°. A set G is **open** if every point in G is an interior point, that is, $G = \operatorname{int} G$.

Let E be a subset of A. A point x is a **relative interior point** of E with respect to A if there is some $\varepsilon > 0$ such that $B_\varepsilon(x) \cap A = \{y \in A : d(y, x) < \varepsilon\} \subset E$. The set of relative interior points is called the **relative interior** of E. A set E is **relatively open in A** if every point in G is a relative interior point. Note that A is always relatively open in A.

- The interior $\operatorname{int} X$ of a set X is open. Indeed it is the largest open set included in X.
- The union of a family of open sets is open.
- The intersection of a finite family of open sets is open.
- The empty set and X are both open.
- Every open ball $B_r(x)$ is an open set.

  To see this, let $B_r(x)$ be an open ball and let $y \in B_r(x)$. Put $\varepsilon = r - d(x, y) > 0$. Now if $z \in B_\varepsilon(y)$, then the triangle inequality implies $d(x, z) \leqslant d(x, y) + d(y, z) < d(x, y) + \varepsilon = r$. So $B_\varepsilon(y) \subset B_r(x)$, which means that $B_r(x)$ is a d-open set.

The collection of open sets in a metric space is called the **topology** of the metric space. Two metrics generating the same topology are **equivalent**. The Euclidean, ℓ_1, and sup metrics on $\boldsymbol{R}^{\mathrm{n}}$ are equivalent. A property of a metric space that can be expressed in terms of open sets without mentioning a specific metric is called a **topological property**.

2.3 Topological spaces

The collection of open subsets of a metric space is closed under finite intersections is closed under finite intersections and arbitrary unions. Use that as the motivation for the following definition.

3 Definition *A **topology** τ on a nonempty set X is a family of subsets of X, called **open sets** satisfying*

1. *$\varnothing \in \tau$ and $X \in \tau$.*
2. *The family τ is closed under finite intersections. That is, if $U_1, \ldots, U_n$ belong to τ, then $\bigcap_{i=1}^n U_i$ belongs to τ.*
3. *The family τ is closed under arbitrary unions. That is, if U_α, $\alpha \in A$ belong to τ, then $\bigcup_{\alpha \in A} U_\alpha$ belongs to τ.*

*The pair (X, τ) is a **topological space**.*

*The topology τ is a **Hausdorff** topology if for every two distinct points x, y in X there are disjoint open sets U, V with $x \in U$ and $y \in V$.*

*The set A is a **neighborhood** of x if there is an open set U satisfying $x \in U \subset A$.*

[1] Be aware that Rudin [134] defines a neighborhood to be what I call an **ε-neighborhood**. Under my definition a neighborhood need not be a ball, nor need it be open.

Every metric topology is a Hausdorff topology. A property of X that can be expressed in terms of its topology is called a topological property.

4 Lemma *A set is open if and only it is a neighborhood of each of it points.*

Proof: Clearly an open set is neighborhood of its points. So assume the set G is a neighborhood of each of it points. That is, for each $x \in G$ there is an open set U_x satisfying $x \in U_x \subset G$. Then $G = \bigcup_{x \in G} U_x$ is open, being a union open sets. ∎

2.4 Closed sets

A set in a topological space is **closed** if its complement is open. Thus:

- The intersection of any family of closed sets is closed.
- The union of a finite family of closed sets is closed.
- The empty set and X are both closed.
- The smallest closed set including a set A is called the closure of A, denoted $\overline{A}$ or $\operatorname{cl} A$. It is the intersection of all closed sets that include A.
- It follows that A is closed if and only if $A = \overline{A}$. Also, if $A \subset B$, then $\overline{A} \subset \overline{B}$.
- In a metric space, $\overline{\{y : d(y,x) < \varepsilon\}} \subset \{y : d(y,x) \leqslant \varepsilon\}$.

  For the Euclidean metric on $\boldsymbol{R}^{\mathrm{n}}$, there is equality, but for arbitrary metric spaces the inclusion may be proper. (Consider the metric space of integers with its usual metric. Then $\overline{\{n : d(n,0) < 1\}} = \{0\} \subsetneq \{n : d(n,0) \leqslant 1\} = \{-1, 0, 1\}$.)

5 Lemma *x belongs to $\overline{A}$ if and only if every neighborhood of x contains a point in A.*

Proof: ($\Longleftarrow$) If every neighborhood if x contains a point of A, then the open set $\overline{A}^c$ does not contain x, so $x \in \overline{A}$.

($\Longrightarrow$) Assume U is a neighborhood of x (that is, $x \in \operatorname{int} U$) such that $U \cap A = \varnothing$ (so that $A \subset U^c \subset (\operatorname{int} U)^c$). Then x does not belong to the closed set $(\operatorname{int} U)^c$ that includes A, so $x \notin \overline{A}$. ∎

6 Lemma *In a metric space, $x \in \overline{A}$ if and only if $d(x, A) = 0$.*

Proof: Let $D = \{x : d(x, A) = 0\}$. Then D is closed. To see this, observe that if $x \in D^c$, say $d(x, A) = \alpha > 0$, then $B_\alpha(x)$ is an open ball included in D^c, so D^c is open. Now clearly we have $A \subset D$, so $\overline{A} \subset \overline{D} = D$.

By definition,

$$\overline{A} = \bigcap \{F : A \subset F,\ F^c \text{ is open}\},$$

so if $x \notin \overline{A}$, then x belongs to some open set F^c with $A \subset F$. Therefore, there is some $\varepsilon > 0$ so that $B_\varepsilon(x) \subset F^c$. Thus $d(x, A) \geqslant \varepsilon$, so $x \notin D$.

That is, $x \in \overline{A}$ if and only if $d(x, A) = 0$. ∎

2.5 Compactness

Let K be a subset of a topological space. A family $\mathcal{A}$ of sets is a **cover** of K if

$$K \subset \bigcup_{A\in\mathcal{A}} A.$$

If each set in the cover $\mathcal{A}$ is open, then $\mathcal{A}$ is an **open cover** of K. A family $\mathcal{B}$ of sets is a **subcover** of $\mathcal{A}$ if $\mathcal{B} \subset \mathcal{A}$ and $K \subset \cup_{A\in\mathcal{B}} A$.

For example, let K be a subset of $\boldsymbol{R}$, and for each $x \in K$, let $\varepsilon_x > 0$. Then the family $\mathcal{A} = \{(x - \varepsilon_x, x + \varepsilon_x) : x \in K\}$ of open intervals is a open cover of K.

7 Definition *A subset K of a topological space is **compact** if every open cover of K has a finite subcover.*

There is an equivalent characterization of compact sets that is perhaps more useful. A family $\mathcal{A}$ of sets has the **finite intersection property** if every finite subset $\{A_1, \dots, A_n\}$ of $\mathcal{A}$ has a nonempty intersection, $\bigcap_{i=1}^n A_i \neq \varnothing$.

For example, the family $\{[x, \infty) : x \in \boldsymbol{R}\}$ of closed intervals has the finite intersection property. (Why?)

8 Theorem *A set K is compact if and only if every family of closed subsets of K having the finite intersection property has a nonempty intersection.*

Proof: Suppose that K is compact, and let $\mathcal{A}$ be a family of closed subsets of K. If $\bigcap_{A\in\mathcal{A}} A = \varnothing$, then $K \subset \bigcup_{A\in\mathcal{A}} A^c$, so $\{A^c : A \in \mathcal{A}\}$ is an open cover of K. Thus there are $A_1, \dots, A_n \in \mathcal{A}$ satisfying $K = \bigcup_{i=1}^n A_i^c$. This implies $\bigcap_{i=1}^n A_i = \varnothing$, so $\mathcal{A}$ does not have the finite intersection property. Thus, if $\mathcal{A}$ possesses the finite intersection property, then $\bigcap_{A\in\mathcal{A}} A \neq \varnothing$.

For the converse, assume that every family of closed subsets of K with the finite intersection property has nonempty intersection, and that $\mathcal{V}$ is an open cover of K. Then $\bigcap_{V\in\mathcal{V}} V^c = \varnothing$, so the finite intersection property must be violated. That is, there exist $V_1, \dots, V_n \in \mathcal{V}$ satisfying $\bigcap_{j=1}^n V_j^c = \varnothing$, or $K \subset \bigcup_{j=1}^n V_j$, which proves that K is compact. ∎

9 Lemma *A compact subset of a Hausdorff space is closed.*

Proof: Let K be compact, and let $x \notin K$. Then by the Hausdorff property, for each $y \in K$ there are disjoint open sets U_y and V_y with $y \in U_y$ and $x \in V_y$. Thus $\{U_y : y \in K\}$ is an open cover of K. Since K is compact, it has a finite subcover, that is, there are $y_1, \dots, y_k$ with $K \subset \bigcup_{i=1}^k U_{y_i} = U$. Then $V = \bigcap_{i=1}^k V_{y_i}$ is an open set satisfying $x \in V \subset U^c \subset K^c$. That is, K^c is a neighborhood of x. Since x is an arbitrary member of K^c, we see that K^c is open, so K is closed. ∎

10 Lemma *A closed subset of a compact set is compact.*

Proof: Let K be compact and $F \subset K$ be closed. Let $\mathcal{G}$ be an open cover of F. Then $\mathcal{G} \cup \{F^c\}$ is an open cover of K. Let $\{G_1, \dots, G_k, F^c\}$ be a finite subcover of K. Then $\{G_1, \dots, G_k\}$ is a finite subcover of F. ∎

It is easy to see the following.

11 Lemma *Finite unions of compact sets are compact.*

For a nonempty subset A of a metric space (X, d) its **diameter** is $\sup\{d(x, y) : x, y \in A\}$. A set is **bounded** if its diameter is finite. A subset of a metric space is **totally bounded** if for every $\varepsilon > 0$, it can be covered by finitely many ε-balls. Boundedness and total boundedness are not topological properties, they may depend on the particular metric.

The next result follows easily from the definitions.

12 Lemma *Every compact subset of a metric space is totally bounded.*

The following lemma is easy to prove.

13 Lemma *If subset of $\boldsymbol{R}^{\mathrm{n}}$ is bounded under the Euclidean metric (or the ℓ_1, or the ℓ_∞ metric), then it is totally bounded.*

The Heine–Borel–Lebesgue Theorem characterizes compactness for subsets of $\boldsymbol{R}^{\mathrm{n}}$. It follows from Lemma 13 and Theorem 23 below.

14 Heine–Borel–Lebesgue Theorem *A subset of $\boldsymbol{R}^{\mathrm{n}}$ is compact if and only if it is both closed and bounded in the Euclidean metric.*

This result is special. In general, a set may be closed and bounded without being totally bounded or compact, but see Theorem 23 below.

15 Example (A noncompact, closed bounded set in $\boldsymbol{\ell_\infty}$) Consider ℓ_∞, the vector space of bounded sequences, with the norm $\|x\| = \sup_n x_n$. The set C of unit coordinate vectors e^i, $i = 1, 2, \ldots$, is closed and bounded, but not compact. The collection of open balls of radius $\frac{1}{2}$ centered at each coordinate vector covers C but has no finite subcover. □

2.6 Convergence and continuity

Let (X, d) be a metric space. A sequence $x_1, x_2, \ldots$ in X **converges** to a point x in X, written

$$x_n \xrightarrow[n\to\infty]{} x$$

if $d(x_n, x) \to 0$ as a sequence of real numbers. In other words, if

$$\forall \varepsilon > 0 \; \exists N \; \forall n \geqslant N \quad d(x_n, x) < \varepsilon.$$

Or in yet other words, if the sequence **eventually** lies in any neighborhood of x.

For metric spaces (but not necessarily for more general topological spaces), limits of sequences determine the closure of a set.

16 Lemma *Let A be a subset of a metric space X. The closure of A consists of all the limits of sequences in A.*

Proof: By Lemma 6, $x \in \overline{A}$ if and only if $d(x, A) = 0$. If $d(x, A) = 0$, for each n there is x_n with $d(x, x_n) < 1/n$. Then $x_n \to x$. Conversely, if $x_n \to x$ and each $x_n \in A$, then $d(x, A) = 0$. ∎

17 Definition *Let X and Y be topological spaces and let $f\colon X \to Y$. Then f is **continuous** if the inverse image of open sets are open. That is, if U is an open subset of Y, then $f^{-1}(U)$ is an open subset of X.*

*The function f is **continuous at x** if the inverse image of every open neighborhood of $f(x)$ is a neighborhood of x.*

This corresponds to the usual ε-δ definition of continuity that you are familiar with.

18 Lemma *A function is continuous if and only if it is continuous at each point.*

Proof: Let $f\colon X \to Y$. Assume f is continuous, and let V be an open neighborhood of $f(x)$. Then $x \in f^{-1}(V)$, and since f is continuous, $f^{-1}(V)$ is a neighborhood of x.

Conversely, assume f is continuous at each point, and let V be an open set in Y. Let x belong to $f^{-1}(V)$, or $f(x) \in V$. Since V is an open neighborhood of $f(x)$, and f is continuous at x, the inverse image $f^{-1}(V)$ is a neighborhood of x. Since x is an arbitrary member of $f^{-1}(V)$, it is open (Lemma 4). ∎

In a metric space, f is continuous if and only if

$$x_n \to x \quad \Longrightarrow \quad f(x_n) \to f(x).$$

Or equivalently,

$$\forall x \in X \ \forall \varepsilon > 0 \ \exists \delta > 0 \ \forall z \in X \quad d(x,z) < \delta \implies \rho\big(f(x), f(z)\big) < \varepsilon.$$

19 Definition *Let X and Y be topological spaces. A function $f\colon X \to Y$ is a* ***homeomorphism*** *if it is one-to-one and onto, is continuous, and its inverse is continuous.*

If f is homeomorphism $U \leftrightarrow f(U)$ is a one-to-one correspondence between the topologies of X and Y. Thus X and Y have the same topological properties. They can in effect be viewed as the same topological space, where f simply renames the points.

20 Lemma *Let $f\colon X \to Y$ be continuous. If K is a compact subset of X, then $f(K)$ is a compact subset of Y.*

Proof: Let $\{G_i\}_{i\in I}$ be an open cover of $f(K)$. Then $\{f^{-1}(G_i)\}_{i\in I}$ is an open cover of K. Let $\{f^{-1}(G_1), \dots, f^{-1}(G_k)\}$ be a finite subcover of K. Then $\{G_1, \dots, G_k\}$ is a finite subcover of $f(K)$. ∎

21 Lemma *Let $f\colon X \to Y$ be one-to-one and continuous, where Y is a Hausdorff space and X is compact. Then $f\colon X \to f(X)$ is a homeomorphism.*

Proof: We need to show that f^{-1} is continuous on $f(X)$. So let G be any open subset of X. We must show that $(f^{-1})^{-1}(G) = f(G)$ is open. Now G^c is a closed subset of X, and thus compact. Therefore $f(G^c)$ is compact, and since Y is Hausdorff, $f(G^c)$ is a closed subset of Y. But then $f(G^c)^c = f(G)$ is open. ∎

2.7 Lipschitz continuity

Let (X,d) and (Y,ρ) be metric spaces. A function $f\colon X \to Y$ satisfies a **Lipschitz condition at x_0** or is **Lipschitz continuous at x_0** if there is some $M > 0$ and $\varepsilon > 0$ such that for every x

$$d(x_0,x) < \varepsilon \quad \Longrightarrow \quad \rho\big(f(x_0), f(x)\big) < Md(x_0,x).$$

The number M is called a **Lipschitz constant** for f at x_0. Apostol [6] makes the following more general definition. The function f satisfies a **Lipschitz condition of order α at x_0** if there is some $M > 0$ and $\varepsilon > 0$ such that for every x

$$d(x_0,x) < \varepsilon \quad \Longrightarrow \quad \rho\big(f(x_0), f(x)\big) < Md(x_0,x)^\alpha.$$

When α is not mentioned, it is tacitly assumed to be 1. Note that if f satisfies a Lipschitz condition of order $\alpha > 0$ at x_0, then it is indeed continuous at x_0.

We say that f is **uniformly Lipschitz continuous on the set A** if there is some $M > 0$ such that for all x, z

$$x, z \in A \quad \Longrightarrow \quad \rho\big(f(x), f(z)\big) < Md(x,z).$$

We also say that f is **locally uniformly Lipschitz continuous at x_0** if there is some $M > 0$ and a neighborhood U of x_0 such that f is uniformly Lipschitz continuous on U. Equivalently, there is some $\varepsilon > 0$ such that for all x, z

$$x, z \in B_\varepsilon(x_0) \quad \Longrightarrow \quad \rho\big(f(x), f(z)\big) < Md(x,z).$$

Note that a function can be continuous without being Lipschitz continuous. For example $f(x) = x^{\frac{1}{3}}$ is continuous at 0 but not Lipschitz continuous at 0. (As $x^{\frac{1}{3}} < Mx$ if and only if $x > M^{-\frac{2}{3}}$ for $x > 0$, but it does satisfy a Lipschitz condition of order $\frac{1}{3}$.) We shall see later on (Proposition 64, Lemma 91) that Lipschitz continuity at a point is implied by differentiability at that point. The converse is not true, as the function f given by $f(x) = x \sin \frac{1}{x}$ for $x \neq 0$ and $f(0) = 0$ is Lipschitz continuous at 0 with constant 1, but not differentiable at 0.

2.8 Complete metric spaces

A sequence $x_1, x_2, \ldots$ in the metric space (X, d) is a **Cauchy sequence** if

$$\forall \varepsilon > 0 \; \exists N \; \forall n, m \geqslant N \quad d(x_n, x_m) < \varepsilon.$$

A metric space (X, d) is **complete** if every Cauchy sequence converges to a point in X. It is easy to see that any closed subset of a complete metric space is itself complete under the restriction of the metric.

The next result is a profoundly useful fact about complete metric spaces. Let us say that a sequence $\{A_n\}$ of sets has **v**anishing diameter if

$$\lim_{n\to\infty} \text{diameter}\, A_n = 0.$$

22 Cantor Intersection Theorem *In a complete metric space, if a decreasing sequence of nonempty closed subsets has vanishing diameter, then the intersection of the sequence is a singleton.*

Proof: Let $\{F_n\}$ be a decreasing sequence of nonempty closed subsets of the complete metric space (X, d), and assume $\lim_{n\to\infty} \text{diameter}\, F_n = 0$. The intersection $F = \bigcap_{n=1}^{\infty} F_n$ cannot have more that one point, for if $a, b \in F$, then $d(a, b) \leqslant \text{diameter}\, F_n$ for each n, so $d(a, b) = 0$, which implies $a = b$.

To see that F is nonempty, for each n pick some $x_n \in F_n$. Since $d(x_n, x_m) \leqslant \text{diameter}\, F_n$ for $m \geqslant n$, the sequence $\{x_n\}$ is Cauchy. Since X is complete there is some $x \in X$ with $x_n \to x$. But x_n belongs to F_m for $m \geqslant n$, and each F_n is closed, so x belongs to F_n for each n. ∎

Completeness is not a topological property. That is, there can be two metrics d and d' on X that generate the same topology, and one can be complete, while the other isn't. For instance, under the usual metric $d(n, m) = |n - m|$, the set of natural numbers is complete, as the only Cauchy sequences are eventually constant. But the function $d'(n, m) = \left|\frac{1}{n} - \frac{1}{m}\right|$ is also a metric. Under d' the sequence $1, 2, 3, \ldots$ is Cauchy, but there is no natural number that is the limit of the sequence.

Similarly, total boundedness is not a topological property. Using the metrics d and d' of the previous paragraph, the natural numbers are totally bounded under d' but not under d.

However, there is a topological consequence to having a metric that is both complete and totally bounded.

23 Theorem (Compactness of metric spaces) *For a metric space the following are equivalent:*

1. *The space is compact.*

2. *The space is complete and totally bounded.*

3. *The space is sequentially compact. That is, every sequence has a convergent subsequence.*

Proof: Let (X, d) be a metric space.

(1) $\implies$ (2) Since $X = \bigcup_{x \in X} B_\varepsilon(x)$, there exist $x_1, \dots, x_k$ in X such that $X = \bigcup_{i=1}^k B_\varepsilon(x_i)$. That is, X is totally bounded. To see that X is also complete, let $\{x_n\}$ be a Cauchy sequence in X, and let $\varepsilon > 0$ be given. Pick n_0 so that $d(x_n, x_m) < \varepsilon$ whenever $n, m \geqslant n_0$. The sequence $\{x_n\}$ has a limit point, say x. I claim that $x_n \to x$. Indeed, if we choose $k \geqslant n_0$ such that $d(x_k, x) < \varepsilon$, then for each $n \geqslant n_0$, we have Write as a theorem.

$$d(x_n, x) \leqslant d(x_n, x_k) + d(x_k, x) < \varepsilon + \varepsilon = 2\varepsilon,$$

proving $x_n \to x$. That is, X is also complete.

(2) $\implies$ (3) Fix a sequence $\{x_n\}$ in X. Since X is totally bounded, there must be infinitely many terms of the sequence in a closed ball of radius $\frac{1}{2}$. (Why?) This ball is totally bounded too, so it must also include a closed set of diameter less than $\frac{1}{4}$ that contains infinitely many terms of the sequence. By induction, construct a decreasing sequence of closed sets with vanishing diameter, each of which contains infinitely many terms of the sequence. Use this and the Cantor Intersection Theorem 22 to construct a convergent subsequence.

(3) $\implies$ (1) Let $\{V_i\}_{i \in I}$ be an open cover of X. First, we claim that there exists some $\delta > 0$ such that for each $x \in X$ we have $B_\delta(x) \subset V_i$ for at least one i.[2] Indeed, if this is not the case, then for each n there exists some $x_n \in X$ satisfying $B_{1/n}(x_n) \cap V_i^c \neq \varnothing$ for each $i \in I$. If x is the limit point of some subsequence of $\{x_n\}$, then it is easy to see that $x \in \bigcap_{i \in I} V_i^c = \left(\bigcup_{i \in I} V_i\right)^c = \varnothing$, a contradiction.

Now fix some $\delta > 0$ such that for each $x \in X$ we have $B_\delta(x) \subset V_i$ for at least one i. We claim that there exist $x_1, \dots, x_k \in X$ such that $X = \bigcup_{i=1}^k B_\delta(x_i)$. To see this, assume by way of contradiction that this is not the case. Fix $y_1 \in X$. Since the claim is false, there exists some $y_2 \in X$ such that $d(y_1, y_2) \geq \delta$. Similarly, since $X \neq B_\delta(y_1) \cup B_\delta(y_2)$, there exists some $y_3 \in X$ such that $d(y_1, y_3) \geq \delta$ and $d(y_2, y_3) \geq \delta$. So by an inductive argument, there exists a sequence $\{y_n\}$ in X satisfying $d(y_n, y_m) \geq \delta$ for $n \neq m$. However, any such sequence $\{y_n\}$ cannot have any convergent subsequence, contrary to our hypothesis. Hence there exist $x_1, \dots, x_k \in X$ such that $X = \bigcup_{i=1}^k B_\delta(x_i)$.

Finally, for each $1 \leqslant j \leqslant k$ choose an index i_j such that $B_\delta(x_j) \subset V_{i_j}$. Then $X = \bigcup_{j=1}^k V_{i_j}$, proving that X is compact. ∎

2.9 Product topology

24 Definition *If X and Y are topological spaces, the* ***product topology*** *on $X \times Y$ consists of arbitrary unions of sets of the form $U \times V$, where U is an open set in X and V is an open set in Y.*

The product topology is indeed a topology.

The Cartesian product of two metric spaces (X, d_1) and (Y, d_2) is a metric space under the metric $d\big((x, y), (x', y')\big) = d_1(x, x') + d_2(y, y')$. The topology it defines is called the **product topology**. If (Z, d_3) is a metric space, and $f\colon X \times Y \to Z$ is continuous with respect to the product topology, we say that f is **jointly continuous** in (x, y). We say that f is **separately continuous** in (x, y) if for each x, the function $y \mapsto f(x, y)$ is continuous, and for each y, the function $x \mapsto f(x, y)$ is continuous. Joint continuity implies separate continuity, but the converse is not true.

25 Example (Separate continuity does not imply joint continuity)
****

□

[2] Such a number δ is known as a **L**ebesgue number of the cover.

It is possible to define a natural product topology for an arbitrary product of metric spaces, but unless the set of factors is countable, the result will not be a metric space. Nonetheless, the notion of compactness is topological, so the next result makes sense. A proof may be found, for instance, in Aliprantis and Border [3, Theorem 2.57, p. 52].

26 Tychonoff Product Theorem *The Cartesian product of an arbitrary family of compact sets is compact.*

2.10 Semicontinuous functions

The real-valued function $f\colon X \to \boldsymbol{R}$ is **upper semicontinuous on $\boldsymbol{X}$** if for each $\alpha \in \boldsymbol{R}$, the upper contour set $[f \geqslant \alpha]$ is closed, or equivalently, the strict lower contour set $[f < \alpha]$ is open. It is **lower semicontinuous** if every lower contour set $[f \leqslant \alpha]$ is closed, or equivalently, the strict upper contour set $[f > \alpha]$ is open.

The extended real valued function f is **upper semicontinuous at the point $\boldsymbol{x}$** if $f(x) < \infty$ and

$$\forall \varepsilon > 0\ \exists \delta > 0 \quad d(y,x) < \delta \implies f(y) < f(x) + \varepsilon.$$

Similarly, f is **lower semicontinuous at the point $\boldsymbol{x}$** if $f(x) > -\infty$ and

$$\forall \varepsilon > 0\ \exists \delta > 0 \quad d(y,x) < \delta \implies f(y) > f(x) - \varepsilon.$$

Equivalently, f is upper semicontinuous at x if $f(x) < \infty$ and

$$f(x) \geqslant \limsup_{y \to x} f(y) = \inf_{\varepsilon > 0} \sup_{0 < d(y,x) < \varepsilon} f(y).$$

Similarly, f is lower semicontinuous at x if $f(x) > -\infty$ and

$$f(x) \leqslant \liminf_{y \to x} f(y) = \sup \varepsilon > 0 \inf_{0 < d(y,x) < \varepsilon} f(y).$$

Note that f is upper semicontinuous if and only if $-f$ is lower semicontinuous.

27 Lemma *A real valued function $f\colon X \to \boldsymbol{R}$ is upper semicontinuous on X if and only if it is upper semicontinuous at each point of X. It is lower semicontinuous on X if and only if it is lower semicontinuous at each point of X.*

Proof: I'll prove the result for upper semicontinuity. Assume that f is upper semicontinuous on X. For any real number α, if $f(x) < \beta < \alpha$, then $\{y \in X : f(y) < \beta\}$ is an open neighborhood of x. Thus for $\varepsilon > 0$ small enough $d(y,x) < \varepsilon$ implies $f(y) < \beta$. Therefore $\limsup_{y \to x} f(y) \leqslant \beta < \alpha$. Setting $\alpha = \limsup_{y \to x} f(y)$, we see that it cannot be the case that $f(x) < \limsup_{y \to x} f(y)$, for then $f(x) < \alpha = \limsup_{y \to x} f(y) < \alpha$, a contradiction. That is, f is upper semicontinuous at x.

For the converse, assume that f is upper semicontinuous at each x. Fix a real number α, and let $f(x) < \alpha$. Since $f(x) \geqslant \limsup_{y \to x} f(y)$, there is $\varepsilon > 0$ small enough so that $\sup_{y : 0 < d(y,x) < \varepsilon} f(y) < \alpha$, but this implies $\{x \in X : f(x) < \alpha\}$ is open, so f is upper semicontinuous on X.

∎

28 Corollary *A real-valued function is continuous if and only if it is both upper and lower semicontinuous.*

29 Theorem *An extended real-valued function f is upper semicontinuous on X if and only if its hypograph is closed. It is lower semicontinuous on X if and only if its epigraph is closed.*

Proof: Assume f is upper semicontinuous, and let (x_n, α_n) be a sequence in its hypograph, that is, $f(x_n) \geqslant \alpha_n$ for all n. Therefore $\limsup_n f(x_n) \geqslant \limsup_n \alpha_n$. If $(x_n, \alpha_n) \to (x, \alpha)$, since f is upper semicontinuous at x, we have $\alpha = \lim_n \alpha_n \leqslant \limsup_n f(x_n) \leqslant f(x)$. Thus $\alpha \leqslant f(x)$, or (x, α) belong to the hypograph of f. Therefore the hypograph is closed.

Assume now that the hypograph is closed. Pick x and let $\alpha = \limsup_{y \to x} f(y)$. Then there is a sequence $x_n \to x$ with $f(x_n) \uparrow \alpha$. Since $\big(x_n, f(x_n)\big)$ belongs to the hypograph for each n, so does its limit (x, α). That is, $\limsup_{y \to x} f(y) = \alpha \leqslant f(x)$, so f is upper semicontinuous at x. ∎

30 Exercise Prove that if both the epigraph and hypograph of a function are closed, then the graph is closed. Give an example to show that the converse is not true. □

31 Proposition *The infimum of a family of upper semicontinuous functions is upper semicontinuous. The supremum of a family of lower semicontinuous functions is lower semicontinuous.*

Proof: Let $\{f_\nu\}_\nu$ be a family of upper semicontinuous functions, and let $f(x) = \inf_\nu f_\nu(x)$. Then $[f \geqslant \alpha] = \cap_\nu [f_\nu \geqslant \alpha]$, which is closed. Lower semicontinuity is dealt with *mutatis mutandis*. ∎

32 Definition *Given an extended real-valued function f on the metric space X, we define the* ***upper envelope*** $\overline{f}$ *of f by*

$$\overline{f}(x) = \max\{f(x), \limsup_{y \to x} f(y)\} = \inf_{\varepsilon > 0} \sup_{d(y,x) < \varepsilon} f(y),$$

and the ***lower envelope*** $\underline{f}$ *of f by*

$$\underline{f}(x) = \min\{f(y), \liminf_{y \to x} f(y)\}, = \sup_{\varepsilon > 0} \inf_{d(y,x) < \varepsilon} f(y).$$

Clearly if f is upper semicontinuous at x, then $f(x) = \overline{f}(x)$, and if f is lower semicontinuous at x, then $f(x) = \underline{f}(x)$. Consequently, f is upper semicontinuous if and only $f = \overline{f}$, and f is lower semicontinuous if and only $f = \underline{f}$.

We say that the real-valued function g **dominates** the real-valued function f on X if for every $x \in X$ we have $g(x) \geqslant f(x)$.

33 Theorem *The upper envelope $\overline{f}$ is the smallest upper semicontinuous function that dominates f and the lower envelope $\underline{f}$ is the greatest lower semicontinuous function that f dominates.*

Moreover,

$$\mathit{hypograph}\, \overline{f} = \overline{\mathit{hypograph}\, f}.$$

and

$$\mathit{epi}\, \underline{f} = \overline{\mathit{epi}\, f}.$$

Proof: Clearly, $\overline{f}$ dominates f and f dominates $\underline{f}$.

Now suppose g is upper semicontinuous and dominates f. Then for any x, we have $g(x) \geqslant \limsup_{y \to x} g(y) \geqslant \limsup_{y \to x}$, so $g(x) \geqslant \overline{f}(x)$. That is, g dominates $\overline{f}$.

Similarly if g is lower semicontinuous and f dominates g, then $\underline{f}$ dominates g.

It remains to show that $\overline{f}$ is upper semicontinuous. It suffices to prove that the hypograph of $\overline{f}$ is closed. We prove the stronger result that

$$\text{hypograph}\, \overline{f} = \overline{\text{hypograph}\, f}.$$

Let (x_n, α_n) be a sequence in the hypograph of f, and assume it converges to a point (x, α). Since $\alpha_n \leqslant f(x_n)$, we must have $\alpha \leqslant \limsup_{y \to x} f(y)$, so $\alpha \leqslant \overline{f}$. That is, $\overline{\text{hypograph}\, f} \subset \text{hypograph}\, \overline{f}$. For the opposite inclusion, suppose by way of contradiction that (x, α) belongs to the hypograph of $\overline{f}$, but not to $\overline{\text{hypograph}\, f}$. Then there is a neighborhood $B_\varepsilon(x) \times B_\varepsilon(\alpha)$ disjoint from $\overline{\text{hypograph}\, f}$. In particular, if $d(y, x) < \varepsilon$, then $f(y) < \alpha \leqslant \overline{f}(x)$, which implies $\overline{f}(x) > \max\{f(x), \limsup -y \to x f(y)\}$, a contradiction. Therefore $\overline{\text{hypograph}\, f} \supset \text{hypograph}\, \overline{f}$.

The case of $\underline{f}$ is similar. ∎

2.11 Existence of extrema

A fundamental theorem on continuity and compactness is:

34 Theorem *The continuous image of a compact set is compact.*

Proof: Let $f\colon X \to Y$ be continuous. Let K be a compact subset of X, and $\{U_\alpha\}$ be an open cover of $f(K)$. Then $\{f^{-1}(U_\alpha)\}$ is an open cover of K, and so has a finite subcover $f^{-1}(U_1), \dots, f^{-1}(U_k)$. Then $U_1, \dots, U_k$ is an open cover of $f(K)$. ∎

This theorem and the Heine–Borel–Lebesgue Theorem imply the next result, which is the fundamental result on the existence of maxima. But a more direct proof is instructive.

But first I will give a tediously detailed proof of an "obvious" fact.

35 Lemma *Every nonempty finite set of real numbers has a greatest and least element.*

Proof: I will prove the statement for greatest elements. The proof proceeds by induction. Let $\mathbb{P}(n)$ denote the proposition: Every set of n elements has a greatest element. By reflexivity of $\geqslant$, the statement $P(1)$ is true. We now prove that $P(n) \implies P(n+1)$. Assume $P(n)$ is true, and let A be a set of cardinality $n+1$. Pick some $x \in A$. Then $A \setminus \{x\}$ has n elements, and so has a greatest element y. That is, $y \geqslant z$ for all $z \in A \setminus \{x\}$. Since $\geqslant$ is a total order on $\boldsymbol{R}$ $x \geqslant y$ or $y \geqslant x$ (or both). If $y \geqslant x$, then $y \geqslant z$ for all $z \in (A \setminus \{x\}) \cup \{x\} = A$. On the other hand, if $x \geqslant y$, then by transitivity $x \geqslant z$ for all $z \in A \setminus \{x\}$, and by reflexivity $x \geqslant x$, so x is the greatest element of A. ∎

36 Weierstrass' Theorem *Let K be a compact set, and let $f\colon K \to \boldsymbol{R}$ be continuous. Then f achieves both a maximum and a minimum over K.*

Proof: I will prove that f achieves a maximum, the proof for a minimum is similar. Observe that the family $\big\{[f < \alpha] : \alpha \in \boldsymbol{R}\big\}$ is an open cover of K. Since K is compact, it has a finite subcover, say $\{[f < \alpha_1], \dots, [f < \alpha_m]\}$. Let $\alpha^* = \max_{i=1,\dots,m} \alpha_i$. The f is bounded above by α^* on K. The real numbers are complete, so the set of upper bounds has least upper bound $\beta = \sup\{f(x) : x \in K\}$. Assume by way of contradiction that there is no x in K with $f(x) = \beta$. Then $\{[f < \alpha] : \alpha < \beta\}$ is an open cover of K without a finite subcover, which contradicts the compactness of K. ∎

Alternate proof: I will prove that f achieves a maximum, the proof for a minimum is similar. Observe that the family $\big\{[f \geqslant \alpha] : \alpha \in f(K)\big\}$ has the finite intersection property. Since f is continuous, each set $[f \geqslant \alpha]$ is closed. Since K is compact, Theorem 8 implies that this family has nonempty intersection. Thus $\bigcap_{\alpha \in f(K)} \big\{x \in K : f(x) \geqslant \alpha\big\}$ is the nonempty set of maximizers of f. ∎

The same proofs demonstrates the following theorem.

37 Theorem *Let $f\colon X \to \boldsymbol{R}$ be upper semicontinuous. Suppose that for some $\alpha \in \boldsymbol{R}$ the upper contour set $[f \geqslant \alpha]$ is nonempty and compact. Then f achieves a maximum on X.*

If $f\colon X \to \boldsymbol{R}$ is lower semicontinuous and $[f \leqslant \alpha]$ is nonempty and compact for some α, then f achieves a minimum on X.

38 Example (Functions without maxima) To see what can go wrong, consider the following examples.

1. Set $X = (0, 1)$, which is bounded but not closed, and set $f(x) = x$, which is a well behaved continuous function, but there is no point in $(0, 1)$ that maximizes f. Thus f is bounded above but does not achieve a maximum.

2. Set $X = \boldsymbol{R}$, which is closed but not bounded, and set $f(x) = x$. Again there is no point in $\boldsymbol{R}$ that maximizes f. In this case f is not bounded above.

3. Set $X = [0, 1]$, which is closed and bounded, and define f by

$$f(x) = \begin{cases} x & 0 \leqslant x < 1 \\ 0 & x = 1. \end{cases}$$

Then f is not continuous, and no point in $[0, 1]$ maximizes f. Again f is bounded above but does not achieve a maximum.

□

2.12 Topological vector spaces

39 Definition *A (real)* ***topological vector space*** *is a vector space X together with a topology τ where τ has the property that the mappings scalar multiplication and vector addition are continuous functions. That is, the mappings*

$$(\alpha, x) \mapsto \alpha x$$

from $\boldsymbol{R} \times X$ to X and

$$(x, y) \mapsto x + y$$

from $X \times X$ to X are continuous. (Where, of course, $\boldsymbol{R}$ has its usual topology, and $\boldsymbol{R} \times X$ and $X \times X$ have their product topologies.)

For a detailed discussion of topological vector spaces, see chapter five of the Hitchhiker's Guide [3]. But here are some of the results we will use.

40 Fact *Every normed vector space (including $\boldsymbol{R}^{\mathrm{n}}$) is a topological vector space.*

41 Lemma *If V is open, then $V + y$ is open.*

Proof: Since $f\colon x \mapsto x - y$ is continuous, $V + y = f^{-1}(V)$ is open. ∎

42 Corollary *If V is open, then $V + A$ is open for any set A.*

Proof: $V + A = \cup_{y \in A} V + y$ is open as a union of open sets. ∎

43 Lemma *If V is open, and $\alpha \neq 0$, then αV is open.*

Proof: Since $f\colon x \mapsto (1/\alpha)x$ is continuous, $\alpha V = f^{-1}(V)$ is open. ∎

44 Corollary *If V is open, then $\alpha V + \beta A$ is open for any set A and scalars α, β with $\alpha \neq 0$.*

45 Definition *A set C in a vector space is* ***circled*** *or* ***radial*** *if $\alpha C \subset C$ whenever $|\alpha| \leqslant 1$.*

46 Lemma *Let V be a neighborhood of zero. Then there is an open circled neighborhood U of zero included in V.*

Proof: The mapping $f\colon (\alpha, x) \mapsto \alpha x$ is continuous, and $f(0,0) = 0$, the inverse image $f^{-1}(V)$ is a neighborhood of 0. Thus there is a $\delta > 0$ and an open neighborhood W of 0 such that $(-\delta, \delta) \times W \subset f^{-1}(V)$. This implies that for any α with $|\alpha| < \delta$ and $x \in W$, we have $\alpha x \in V$. In other words $\alpha W \subset V$. Set

$$U = \bigcup_{\alpha : 0 < |\alpha| < \delta} \alpha W$$

Then $U \subset V$, U is circled, and U is open, being the union of the open sets αW. ∎

47 Lemma *Let $T\colon X \to Y$ be a linear transformation between topological vector spaces. Then T is continuous on X if it is continuous at 0.*

Proof: Let V be an open set in Y. It suffices to prove that T is continuous at each point x. So let V be an open neighborhood of $T(x)$. Then $V - T(x)$ is an open neighborhood of 0. Since T is continuous at 0, the inverse image $T^{-1}\big(V - T(x)\big)$, is a neighborhood of 0, so $T^{-1}\big(V - T(x)\big) + x$ is a neighborhood of x. But by linearity, $T^{-1}\big(V - T(x)\big) + x = T^{-1}(V)$, and we are done. ∎

2.13 Continuity of the coordinate mapping

This section proves the following theorem, which everyone seems to take for granted.

48 Theorem *Let X be a Hausdorff topological vector space of dimension n, and let $x_1, \dots, x_n$ be an ordered basis for X. The* ***coordinate mapping*** *$T\colon X \to \boldsymbol{R}^{\mathrm{n}}$ defined by*

$$T\left(\sum_{i=1}^{n} \alpha_i x_i\right) = (\alpha_1, \dots, \alpha_n)$$

is a linear homeomorphism. That is, T is well-defined, linear, one-to-one, maps X onto $\boldsymbol{R}^{\mathrm{n}}$, is continuous, and T^{-1} is continuous.

Proof: Let X be an n-dimensional Hausdorff tvs, and let $v_1, \dots, v_n$ be an ordered basis for X. The coordinate mapping $T\colon X \to \boldsymbol{R}^{\mathrm{n}}$ is defined by

$$T\left(\sum_{i=1}^{n} \lambda_i v_i\right) = (\lambda_1, \dots, \lambda_n).$$

From basic linear algebra, T is a linear bijection from X onto $\boldsymbol{R}^{\mathrm{n}}$. Also $T^{-1}\colon (\lambda_1, \dots, \lambda_n) \mapsto \sum_{i=1}^{n} \lambda_i x_i$ is a linear bijection. Moreover, T^{-1} is continuous, as scalar multiplication and vector addition are continuous. It remains to prove that T is continuous. It suffices to prove that T is continuous at zero.

Let B be the open unit ball and let S be the unit sphere in $\boldsymbol{R}^{\mathrm{n}}$. Since S is compact and T^{-1} is continuous, $T^{-1}(S)$ is compact. Since X is Hausdorff, $T^{-1}(S)$ is closed. Now $0_X \notin T^{-1}(S)$, as $0_{\boldsymbol{R}^{\mathrm{n}}} \notin S$, so there exists a circled neighborhood V of zero such that $V \cap T^{-1}(S) = \varnothing$. Since V is circled, we have $V \subset T^{-1}(B)$: For if there exists some $x \in V$ such that $x \notin T^{-1}(B)$ (that is, $\|T(x)\| \geq 1$), then $\frac{x}{\|T(x)\|} \in V \cap T^{-1}(S)$, a contradiction.

Thus, $T^{-1}(B)$ is a neighborhood of zero. Since scalar multiples of B form a neighborhood base at zero in $\boldsymbol{R}^{\mathrm{n}}$, we see that T is continuous at zero, and therefore continuous. ∎

Informally this says that $\boldsymbol{R}^{\mathrm{n}}$ is the *only* n-dimensional Hausdorff topological vector space.

A useful corollary of this is.

49 Corollary *Let X be a Hausdorff topological vector space, and let $\{x_1, \dots, x_n\}$ be a linearly independent subset of X. Let α_m be a sequence in $\boldsymbol{R}^{\mathrm{n}}$. If $\sum_{i=1}^{n} \alpha_{mi} x_i \xrightarrow[m\to\infty]{} \sum_{i=1}^{n} \alpha_i x_i$, then for each $i = 1, \dots, n$, we have $\alpha_{mi} \xrightarrow[m\to\infty]{} \alpha_i$.*

Proof: The linear subspace X spanned by $\{x_1, \dots, x_n\}$ is a Hausdorff topological space in the relative topology, and Theorem 48 gives the conclusion. ∎

When X is already some $\boldsymbol{R}^{\mathrm{p}}$, there is a more familiar proof of the corollary.

Proof of Corollary for $\boldsymbol{R}^{\mathrm{p}}$: Let X be the $p \times n$ matrix whose j^{th} column is x_j. By the theory of ordinary least squares estimation if $x = X\alpha = \sum_{j=1}^{n} \alpha_j x_j$ is a linear combination of $\{x_1, \dots, x_n\}$, then the coordinate mapping $T(x)$ is given by

$$T(x) = (X'X)^{-1}X'x,$$

which is clearly continuous. ∎

The corollary is rather delicate—it can fail if either X is not Hausdorff or $\{x_1, \dots, x_n\}$ is dependent.

50 Example Let $X = \boldsymbol{R}^2$ under the semi-metric $d\big((x,y),(x',y')\big) = |x - x'|$. (This topology is not Hausdorff.) Then X is a topological vector space. Let $x_1 = (1,0)$ and $x_2 = (0,1)$ be the unit coordinate vectors. Then $\frac{1}{m}x_1 + 0x_2 = (1/m, 0) \to (0,1) = 0x_1 + 1x_2$, (since $d\big((1/m,0),(0,1)\big) = 1/m$, but the second coordinates do not converge $(0 \not\to 1)$. □

51 Example Let $X = \boldsymbol{R}^2$ with the Euclidean topology and let $x_1 = (1,0)$ and $x_2 = (-1,0)$. Then $nx_1 + nx_2 = (0,0) \to (0,0) = 0x_1 + 0x_2$, but $n \not\to 0$. □

2.14 Continuous linear transformations

A word about linear transformations is appropriate here. Between finite dimensional Euclidean spaces every linear transformation is continuous and has a representation in terms of matrix products. Between infinite-dimensional normed spaces, there can be discontinuous linear transformations! Here is a simple example.

52 Example (Discontinuous linear transformation) Let X be the set of real sequences that have a limit, and let $L(x)$ denote the limit of the sequence $x = (x_1, x_2, \dots)$. Then standard properties of the limit show that X is a linear space under termwise addition and scalar multiplication, and $L\colon X \to \boldsymbol{R}$ is a linear transformation. Consider the norm on X defined by

$$\|x\| = \sum_{k=1}^{\infty} \frac{|x_k|}{2^k}.$$

(It is easily verified that this is indeed a norm.) The sequence $(x^1, x^2, \dots)$ in X given by

$$x^n = (\underbrace{0, \dots, 0}_{n \text{ zeroes}}, 1, 1, \dots)$$

satisfies $\|x^n\| = 2^{-n}$, so $x^n \xrightarrow[n\to\infty]{\|\cdot\|} 0$ (where 0 is the sequence of zeroes), but $L(x^n) = 1$ for all n. Since $L(0) = 0$, we see that L is *not* continuous. □

A linear transformation $T\colon X \to Y$ between normed vector spaces is **bounded** if it is bounded on the unit ball. That is, if

$$\sup\Big\{\big\|T(x)\big\| : \|x\| \leqslant 1\Big\} < \infty.$$

This supremum is called the **operator norm** of T, denoted $\|T\|$. This norm has the property that

$$\|T(x)\| \leqslant \|T\| \cdot \|x\| \qquad \text{for all } x.$$

The next result is standard.

53 Theorem (Bounded and continuous linear transformations) *Let $T\colon X \to Y$ be a linear transformation between normed spaces. The following are equivalent.*

1. *T is continuous at zero.*
2. *T is continuous.*
3. *T is bounded.*
4. *T is Lipschitz continuous.*

Proof: (1) $\implies$ (2) Exercise.

(2) $\implies$ (3) Taking $\varepsilon = 1$ in the definition of continuity there is some $\delta > 0$ such that $\|x - 0\| = \|x\| < \delta$ implies $\|T(x)\| = \|T(x) - T(0)\| < 1$. Set $M = \frac{1}{\delta}$. Then by homogeneity $\|x\| < 1$ implies $\|\delta x\| < \delta$, so $\delta\|T(x)\| = \|T(\delta x)\| < 1$, or $\|T(x)\| < \frac{1}{\delta} = M$. That is, T is bounded.

(3) $\implies$ (4) Clearly $\|T(x - z)\| \leqslant \|T\| \cdot \|x - z\|$, so the operator norm is a global Lipschitz constant.

(4) $\implies$ (1) Obvious. ∎

The space of continuous linear functions from the normed space X into the normed space Y is denoted $\mathrm{L}XY$. On this space, the operator norm is indeed a norm.

A function $B\colon X \times X \to Y$ is called **bilinear** if is linear in each variable separately. That is, if $B\big(\alpha x + (1-\alpha)z, v\big) = \alpha B(x,v) + (1-\alpha)B(z,v)$ and $B\big(x, \alpha v + (1-\alpha)w\big) = \alpha B(x,v) + (1-\alpha)B(x,w)$. Clearly every continuous linear transformation T from X into $\mathrm{L}XY$ with the operator norm, corresponds to a continuous bilinear transformation $B\colon X \times X \to Y$ via $B(x,v) = T(x)(v)$, and vice-versa. (Well, if it's not clear, see Dieudonné [43, Theorem 5.7.8, p. 108].)

2.15 Correspondences

This section covers just enough about correspondences to prove the maximum theorem in the next section. More results are available in Border [30].

54 Definition *A **correspondence** φ from X to Y associates to each point in X a subset of Y. We write this as $\varphi\colon X \twoheadrightarrow Y$. For a correspondence $\varphi\colon X \twoheadrightarrow Y$, let $\operatorname{gr}\varphi$ denote the **graph** of φ. That is,*

$$\operatorname{gr}\varphi = \{(x,y) \in X \times Y : y \in \varphi(x)\}.$$

*Let $\varphi\colon X \twoheadrightarrow Y$, $E \subset Y$ and $F \subset X$. The **image** of F under φ is defined by*

$$\varphi(F) = \cup_{x \in F} \varphi(x).$$

For correspondences there are two useful notions of inverse.

55 Definition *The **upper** (or **strong**) **inverse** of E under φ, denoted $\varphi^u[E]$, is defined by*

$$\varphi^u[E] = \{x \in X : \varphi(x) \subset E\}.$$

*The **lower** (or **weak**) **inverse** of E under φ, denoted $\varphi^\ell[E]$, is defined by*

$$\varphi^\ell[E] = \{x \in X : \varphi(x) \cap E \neq \varnothing\}.$$

For a single y in Y, define

$$\varphi^{-1}(y) = \{x \in X : y \in \varphi(x)\}.$$

Note that $\varphi^{-1}(y) = \varphi^\ell[\{y\}]$.

Corresponding to the two notions of inverse are two notions of continuity.

56 Definition *A correspondence $\varphi\colon X \twoheadrightarrow Y$ is* ***upper hemicontinuous (uhc) at x*** *if whenever x is in the upper inverse of an open set, so is a neighborhood of x; and φ is* ***lower hemicontinuous (lhc) at x*** *if whenever x is in the lower inverse of an open set so is a neighborhood of x.*

The correspondence $\varphi\colon X \twoheadrightarrow Y$ is ***upper hemicontinuous*** *(resp.* ***lower hemicontinuous****) if it is upper hemicontinuous (resp. lower hemicontinuous) at every $x \in X$. Thus φ is upper hemicontinuous (resp. lower hemicontinuous) if the upper (resp. lower) inverses of open sets are open.*

A correspondence is called ***continuous*** *if it is both upper and lower hemicontinuous.*

Warning. The definition of upper hemicontinuity is not standard. Berge [18] requires in addition that φ have compact values in order to be called upper hemicontinuous.

If $\varphi\colon X \twoheadrightarrow Y$ is singleton-valued it can be considered a function from X to Y and we may sometimes identify the two. In this case the upper and lower inverses of a set coincide and agree with the inverse regarded as a function. Either form of hemicontinuity is equivalent to continuity as a function. The term "semicontinuity" has been used to mean hemicontinuity, but this usage can lead to confusion when discussing real-valued singleton correspondences. A semicontinuous real-valued function is not a hemicontinuous correspondence unless it is also continuous.

57 Definition *The correspondence $\varphi\colon E \twoheadrightarrow F$ is* ***closed at x*** *if whenever $x^n \to x$, $y^n \in \varphi(x^n)$ and $y^n \to y$, then $y \in \varphi(x)$. A correspondence is* ***closed*** *if it is closed at every point of its domain, that is, if its graph is closed.*

58 Example (Closedness vs. Upper Hemicontinuity) In general, a correspondence may be closed without being upper hemicontinuous, and vice versa.

Define $\varphi\colon \boldsymbol{R} \twoheadrightarrow \boldsymbol{R}$ via

$$\varphi(x) = \begin{cases} \{\frac{1}{x}\} & x \neq 0 \\ \{0\} & x = 0. \end{cases}$$

Then φ is closed but not upper hemicontinuous.

Define $\mu\colon \boldsymbol{R} \twoheadrightarrow \boldsymbol{R}$ via $\mu(x) = (0,1)$. Then μ is upper hemicontinuous but not closed.[3] □

59 Proposition *Let $E \subset \boldsymbol{R}^{\mathrm{m}}$, $F \subset \boldsymbol{R}^{\mathrm{k}}$ and let $\varphi\colon E \twoheadrightarrow F$.*

1. *If φ is upper hemicontinuous and closed-valued, then φ is closed.*
2. *If F is compact and φ is closed, then φ is upper hemicontinuous.*
3. *If φ is singleton-valued at x and upper hemicontinuous at x, then φ is continuous at x.*

Proof: 1. Suppose $(x,y) \notin \operatorname{gr}\varphi$. Then since φ is closed-valued, there is a closed neighborhood U of y disjoint from $\varphi(x)$. Then $V = U^c$ is an open neighborhood of $\varphi(x)$. Since φ is upper hemicontinuous, $\varphi^{\mathrm{u}}[V]$ contains an open neighborhood W of x, i.e., $\varphi(z) \subset V$ for all $z \in W$. Thus $(W \times U) \cap \operatorname{gr}\varphi = \varnothing$ and $(x,y) \in W \times U$. Hence the complement of $\operatorname{gr}\varphi$ is open, so $\operatorname{gr}\varphi$ is closed.

2. Suppose not. Then there is some x and an open neighborhood U of $\varphi(x)$ such that for every neighborhood V of x, there is a $z \in V$ with $\varphi(z) \not\subset U$. Thus we can find $z^n \to x$, $y^n \in \varphi(z^n)$ with $y^n \notin U$. Since F is compact, $\{y^n\}$ has a convergent subsequence converging to $y \notin U$. But since φ is closed, $(x,y) \in \operatorname{gr}\varphi$, so $y \in \varphi(x) \subset U$, a contradiction.

∎

[3] Again, under Berge's definition, an upper hemicontinuous correspondence is automatically closed.

60 Proposition *Let $E \subset \boldsymbol{R}^{\mathrm{m}}$, $F \subset \boldsymbol{R}^{\mathrm{k}}$, $\varphi\colon E \twoheadrightarrow F$.*

1. *If φ is compact-valued, then φ is upper hemicontinuous at x if and only if for every sequence $x^n \to x$ and $y^n \in \varphi(x^n)$ there is a convergent subsequence of $\{y^n\}$ with limit in $\varphi(x)$.*

2. *Then φ is lower hemicontinuous if and only if $x^n \to x$ and $y \in \varphi(x)$ imply that there is a sequence $y^n \in \varphi(x^n)$ with $y^n \to y$.*

Proof: 1. Suppose φ is upper hemicontinuous at x, $x^n \to x$ and $y^n \in \varphi(x^n)$. Since φ is compact-valued, $\varphi(x)$ has a bounded neighborhood U. Since φ is upper hemicontinuous, there is a neighborhood V of x such that $\varphi(V) \subset U$. Thus $\{y^n\}$ is eventually in U, thus bounded, and so has a convergent subsequence. Since compact sets are closed, this limit belongs to $\varphi(x)$.

Now suppose that for every sequence $x^n \to x$, $y^n \in \varphi(x^n)$, there is a subsequence of $\{y^n\}$ with limit in $\varphi(x)$. Suppose φ is not upper hemicontinuous; then there is a neighborhood U of x and a sequence $z^n \to x$ with $y^n \in \varphi(z^n)$ and $y^n \notin U$. Such a sequence $\{y^n\}$ can have no subsequence with limit in $\varphi(x)$, a contradiction.

∎

61 Proposition *Let $E \subset \boldsymbol{R}^{\mathrm{m}}$, $F \subset \boldsymbol{R}^{\mathrm{k}}$ and $\varphi, \mu\colon E \twoheadrightarrow F$, and define $(\varphi \cap \mu)\colon E \twoheadrightarrow F$ by $(\varphi \cap \mu)(x) = \varphi(x) \cap \mu(x)$. Suppose $\varphi(x) \cap \mu(x) \neq \varnothing$.*

1. *If φ and μ are upper hemicontinuous at x and closed-valued, then $(\varphi \cap \mu)$ is upper hemicontinuous at x.*

2. *If μ is closed at x and φ is upper hemicontinuous at x and $\varphi(x)$ is compact, then $(\varphi \cap \mu)$ is upper hemicontinuous at x.*

Proof: Let U be an open neighborhood of $\varphi(x) \cap \mu(x)$. Put $C = \varphi(x) \cap U^c$.

1. Note that C is closed and $\mu(x) \cap C = \varnothing$. Thus there are disjoint open sets V_1 and V_2 with $\mu(x) \subset V_1$ and $C \subset V_2$. Since μ is upper hemicontinuous at x, there is a neighborhood W_1 of x with $\mu(W_1) \subset V_1 \subset V_2^c$. Now $\varphi(x) \subset U \cup V_2$, which is open and so x has a neighborhood W_2 with $\varphi(W_2) \subset U \cup V_2$, as φ is upper hemicontinuous at x. Put $W = W_1 \cap W_2$. Then for $z \in W$, $\varphi(z) \cap \mu(z) \subset V_2^c \cap (U \cup V_2) \subset U$. Thus $(\varphi \cap \mu)$ is upper hemicontinuous at x.

2. Note that in this case C is compact and $\mu(x) \cap C = \varnothing$. Since μ is closed at x, if $y \notin \mu(x)$ then we cannot have $y^n \to y$, where $y^n \in \mu(x^n)$ and $x^n \to x$. Thus there is a neighborhood U_y of y and W_y of x with $\mu(W_y) \subset U_y^c$. Since C is compact, we can write $C \subset V_2 = U_{y^1} \cup \cdots \cup U_{y^n}$; so setting $W_1 = W_{y^1} \cap \cdots \cap W_{y^n}$, we have $\mu(W_1) \subset V_2^c$. The rest of the proof is as in (1).

∎

2.16 The maximum theorem

One of the most useful and powerful theorems employed in mathematical economics and game theory is the "maximum theorem." It states that the set of solutions to a maximization problem varies upper hemicontinuously as the constraint set of the problem varies in a continuous way. Theorem 62 is due to Berge [18] and considers the case of maximizing a continuous real-valued function over a compact set which varies continuously with some parameter vector. The set of solutions is an upper hemicontinuous correspondence with compact values. Furthermore, the value of the maximized function varies continuously with the parameters.

62 Berge Maximum Theorem *Let P, X be metric spaces and let $\varphi\colon P \twoheadrightarrow X$ be a compact-valued correspondence. Let $f\colon X \times P \to \boldsymbol{R}$ be continuous. Define the "argmax" correspondence $\mu\colon P \twoheadrightarrow X$ by*

$$\mu(p) = \{x \in \varphi(p) : x \text{ maximizes } f(\cdot, p) \text{ on } \varphi(p)\},$$

and the value function $V\colon P \to \boldsymbol{R}$ by

$$V(p) = f(x, p) \qquad \text{for any } x \in \mu(p).$$

If φ is continuous at p, then μ is closed and upper hemicontinuous at p and V is continuous at p. Furthermore, μ is compact-valued.

Proof: First note that since φ is compact-valued, μ is nonempty and compact-valued. It suffices to show that μ is closed at p, for then $\mu = \varphi \cap \mu$ and Proposition 61 (2) implies that μ is upper hemicontinuous at p. Let $p^n \to p$, $x^n \in \mu(p^n)$, $x^n \to x$. We wish to show $x \in \mu(p)$ and $V(p^n) \to V(p)$. Since φ is upper hemicontinuous and compact-valued, Proposition 59 (1) implies that indeed $x \in \varphi(p)$. Suppose $x \notin \mu(p)$. Then there is $z \in \varphi(p)$ with $f(z, p) > f(x, p)$. Since φ is lower hemicontinuous at p, by Proposition 60 there is a sequence $z^n \to z$ with $z^n \in \varphi(p^n)$. Since $z^n \to z$, $x^n \to x$, and $f(z) > f(x)$, the continuity of f implies that eventually $f(z^n, p^n) > f(x^n, p^n)$, contradicting $x^n \in \mu(p^n)$.

To see that V is continuous at p, let G be an open neighborhood of $V(p)$. Then $f^{-1}(G)$ is open in $X \times P$. Thus for each $x \in \mu(p)$ there are open neighborhoods A_x of x and B_x of p such that

$$(x, p) \in A_x \times B_x \subset f^{-1}(G).$$

Clearly $\{A_x : x \in \mu(p)\}$ is an open cover of the compact set $\mu(p)$, so there is a finite subcollection $A_{x_1}, \dots, A_{x_n}$ with $\mu(p) \subset A_{x_1} \cup \cdots \cup A_{x_n}$. Set $A = \bigcup_{i=1}^n A_{x_i}$ and $B = \bigcap_{i=1}^n B_{x_i}$. Observe that $A \times B \subset f^{-1}(G)$. Since μ is upper hemicontinuous, $C = B \cap \mu^{\mathrm{u}}(A)$ is an open neighborhood of p. Furthermore, if $q \in C$ and $x \in \mu(q)$, then $(x, q) \in A \times B \subset f^{-1}(G)$. Therefore $V(q) = f(x, q) \in U$, so $C \subset V^{-1}(G)$, which shows that V is continuous at p. ∎

2.17 Lebesgue measure and null sets

In this section I describe just enough about Lebesgue measure to understand the meaning of "almost everywhere." Lebesgue measure is a generalization of the concept of length from intervals to more general subsets of the real line (or more generally $\boldsymbol{R}^{\mathrm{n}}$).

By definition, the Lebesgue measure λ of any interval is its length. This is true regardless of whether the interval is closed, open, or half-open. In particular, the Lebesgue measure of a point is zero and the measure of $\boldsymbol{R}$ is ∞. The measure of a disjoint union of intervals is the sum of the lengths of the intervals. For a countable family of pairwise disjoint intervals, the measure is the infinite series of the lengths, which may be finite or ∞.

A set A of real numbers has **Lebesgue measure zero**, or is a **null set**, if for every $\varepsilon > 0$ there is a countable collection of intervals whose union includes A and whose total length is no more than ε. For instance, every countable set has Lebesgue measure zero. To see this, enumerate the countable set A as $x_1, x_2, \dots$. Let I_n be the (open) interval $(x_n - \frac{1}{2^{n+1}}\varepsilon, x_n + \frac{1}{2^{n+1}}\varepsilon)$, which has length $\varepsilon 2^{-n}$ and contains x_n. Then the total length of these intervals is ε and the union includes A. There are uncountable sets of Lebesgue measure zero, too. One example is the Cantor set described in Section 3.4. The idea is that a set of measure zero cannot be very big in the sense of length. A property is said to hold **almost everywhere** if the set of points where it fails is of Lebesgue measure zero.

2.17.1 A little more on Lebesgue measure

A ***σ*-algebra** Σ of subsets of a set X is a family of subsets with the following properties.

1. X and $\varnothing$ belong to Σ.

2. If A belongs to Σ, then so does A^c. (That is, Σ is closed under complements.)

3. If $A_1, A_2, \dots$ is a countable family of elements of Σ, then $\bigcup_{n=1}^{\infty} A_n$ belongs to Σ, and $\bigcap_{n=1}^{\infty} A_n$ belongs to Σ.

A measure μ on a σ-algebra Σ is a function from Σ into $[0, \infty]$ satisfying the following **countable additivity** property. If $A_1, A_2, \dots$ is a countable sequence of pairwise disjoint elements of Σ, then

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

The power set of a set X is a σ-algebra of subsets of X, and every family of subset of X is included in a smallest σ-algebra. The smallest σ-algebra of subset of $\boldsymbol{R}$ that includes all the intervals is called the **Borel $\boldsymbol{\sigma}$-algebra**, and it members are **Borel sets**.

There is exactly one measure on the Borel σ-algebra that agrees with length on intervals. It is called **Lebesgue measure**. It is possible to extend Lebesgue measure to a measure on a largest σ-algebra of subsets of $\boldsymbol{R}$, whose members are called **Lebesgue measurable sets**. This extension is unique, and is also called Lebesgue measure. Every Lebesgue measurable set is of the form $B \cup N$, where B is a Borel set and N is a null set (has Lebesgue measure zero).

The Lebesgue measure of an interval or collection of intervals is **translation invariant**, that is, the length of the interval $[\alpha, \beta]$ is the same as the interval $[\alpha + \gamma, \beta + \gamma]$ for every γ. Lebesgue measure is also translation invariant on σ-algebra of Lebesgue measurable sets. In fact, it is the only translation invariant measure that extends length.

The bad news is that there are subsets of $\boldsymbol{R}$ that are not Lebesgue measurable.

For details and proofs of all these assertions see Aliprantis and Border [3, Chapter 10], Halmos [70, Chapter 3], or Royden [132, Chapter 3].

In $\boldsymbol{R}^{\mathrm{n}}$, the same definitions are made where the role of the intervals is played by the "rectangles," which are Cartesian products of intervals, and length is replaced by n-dimensional volume.

Section 3

Calculus

3.1 A little calculus

This section is not intended to be a comprehensive review of calculus. Rather, I just mention some of the more salient results, and have written down some of the results for which I have a hard time remembering the exact statement. My favorite sources for basic calculus are Apostol [7, 8] and Hardy [72]. Good references for advanced calculus include Apostol [6, Chapter 5] (a bit old fashioned now, which is useful at times), Dieudonné [43] (elegant, but abstract), Loomis and Sternberg [100] (quite abstract with slightly eccentric notation and terminology, but with better versions of most theorems), Marsden [107, Chapter 6] (quite readable and chatty), Rudin [134] (quite readable), and Spivak [145] (also quite readable, but fast moving).

Let f be a real-valued function of a real variable. We start by recalling that the notation

$$\lim_{x\to\alpha} f(x) = \beta$$

means that for every $\varepsilon > 0$, there is some $\delta > 0$ such that if $0 < |x - \alpha| < \delta$, then $|f(x) - \beta| < \varepsilon$. The reason we restrict attention to x with $0 < |x - \alpha|$ is so that we can divide by $x - \alpha$, as in the next definition.

63 Definition (Derivative of a function) *Let $f\colon (\alpha, \beta) \to \boldsymbol{R}$ be a real function of one real variable. If the limit*

$$\lim_{v\to 0} \frac{f(x+v) - f(x)}{v}$$

*exists (as a finite real number), then we say that f is **differentiable** at x and that the limit is the **derivative** of f at x, denoted $f'(x)$ or $Df(x)$.*

Introduce one-sided derivates.

According to this definition, a function has a derivative only at interior points of its domain, and the derivative is always finite.[1]

64 Proposition (Differentiability implies local Lipschitz continuity) *If f is differentiable at a point x in (α, β), then there exist $\delta > 0$ and $M > 0$ for which $0 < |v| < \delta$ implies*

$$|f(x+v) - f(x)| < M|v|.$$

[1] There are occasions when we may wish to allow the derivative to assume the values $\pm\infty$, and we shall indicate this by explicitly. But most of the time when we say that a function is differentiable, we want the derivative to be finite. To give you an idea of why, consider the function $f\colon \boldsymbol{R} \to \boldsymbol{R}$ by $f(x) = x^{1/2}$ for $x \geqslant 0$, and $f(x) = (-x)^{1/2}$ for $x < 0$. This function is continuously differentiable by our definition everywhere except at 0, where we might be tempted to set $f'(0) = \infty$. Now consider the everywhere continuously differentiable function $g(x) = x^2$. Then $g \circ f(x) = |x|$, which is not even differentiable at 0.

Proof: Since f is differentiable at x, taking $\varepsilon = 1$, there is some $\delta > 0$ such that $0 < |v| < \delta$ implies

$$\left| \frac{f(x+v) - f(x)}{v} - f'(x) \right| < 1.$$

This in turn implies

$$\big|f(x+v) - f(x)\big| < \big(\big|f'(x)\big| + 1\big)|v|,$$

which proves the assertion. ∎

The conclusion of the theorem is that f satisfies a Lipschitz condition at x_0. It readily implies the following.

65 Corollary *If f is differentiable at a point x in (α, β), then f is continuous at x.*

66 Squeezing Lemma *Suppose $f \geqslant h \geqslant g$ everywhere and $f(x) = h(x) = g(x)$. If f and g are differentiable at x, then h is also differentiable at x, and $f'(x) = h'(x) = g'(x)$.*

Next we present the well known Mean Value Theorem, see e.g., Apostol [7, Theorem 4.5, p. 185], and an easy corollary, cf. [7, Theorems 4.6 and 4.7].

67 Mean Value Theorem *Suppose f is continuous on $[\alpha, \beta]$ and differentiable on (α, β). Then there exists $\gamma \in (\alpha, \beta)$ satisfying*

$$f(\beta) - f(\alpha) = f'(\gamma)(\beta - \alpha).$$

This result has some corollaries relating derivatives and monotonicity of functions.

68 Definition (Monotonicity) *Let $f\colon [\alpha, \beta] \to \boldsymbol{R}$. We say that f is*

strictly increasing *on (α, β) if $\alpha < x < y < \beta$ implies $f(x) < f(y)$.*

increasing *or* ***nondecreasing*** *on (α, β) if $\alpha < x < y < \beta$ implies $f(x) \leqslant f(y)$. The term* ***isotone*** *is occasionally used to mean this.*

strictly decreasing *on (α, β) if $\alpha < x < y < \beta$ implies $f(x) > f(y)$.*

decreasing *or* ***nonincreasing*** *on (α, β) if α, β implies $f(x) \geqslant f(y)$. The term* ***antitone*** *is occasionally used to mean this.*[2]

monotone *on (α, β) if it is either increasing on (α, β) or decreasing on (α, β).*

Some authors, notably Hardy [72] and Landau [97, Definition 27, p. 88], say that f is ***increasing at γ*** *if there exists some $\varepsilon > 0$ such that $\gamma - \varepsilon < x < \gamma < y < \gamma + \varepsilon$ implies $f(x) \leqslant f(\gamma) \leqslant f(y)$.*

f is ***decreasing at γ*** *if there exists some $\varepsilon > 0$ such that $\gamma - \varepsilon < x < \gamma < y < \gamma + \varepsilon$ implies $f(x) \geqslant f(\gamma) \geqslant f(y)$.*

69 Corollary (Derivatives and Monotonicity) *Suppose $f\colon [\alpha, \beta] \to \boldsymbol{R}$ is continuous on $[\alpha, \beta]$ and differentiable on (α, β). If $f'(x) \geqslant 0$ for all $x \in (\alpha, \beta)$, then f is nondecreasing on $[\alpha, \beta]$. If $f'(x) > 0$ for all $x \in (\alpha, \beta)$, then f is strictly increasing on $[\alpha, \beta]$.*

Similarly if $f'(x) \leqslant 0$ for all $x \in (\alpha, \beta)$, then f is nonincreasing on $[\alpha, \beta]$. If $f'(x) < 0$ for all $x \in (\alpha, \beta)$, then f is strictly decreasing on $[\alpha, \beta]$.

If $f'(x) = 0$ for all $x \in (\alpha, \beta)$, then f is constant on $[\alpha, \beta]$.

[2] Topkis [149] points out that the negation of the statement "f is increasing" is *not* "f is nonincreasing." For instance the sine function is not an increasing function, nor is it a nonincreasing function in my terminology. This does not seem to lead to much confusion however.

It is extremely important in the above theorem that $f'(x) > 0$ for all $x \in (\alpha, \beta)$ in order to conclude that f is strictly increasing. If we know only that $f'(x_0) > 0$, we cannot conclude even that f is monotone! The following example is well known, see e.g., Marsden [107, Exercise 7.1.3, p. 209], but was introduced to me by Ket Richter.

70 Example (Nonmonotonicity with $f'(x_0) > 0$) Consider the function on $\boldsymbol{R}$ given by

$$f(x) = x + 2x^2 \sin \frac{1}{x^2}.$$

Then f is differentiable everywhere on $\boldsymbol{R}$, and $f'(0) = 1$, but f is not monotone on any open interval around 0. To see that f is differentiable, the only difficulty is at zero. But observe that f is squeezed between $g(x) = x + 2x^2$ and $h(x) = x - 2x^2$, which have the property that $g \geqslant f \geqslant h$ everywhere, $g(0) = f(0) = h(0) = 0$, and $g'(0) = h'(0) = 1$, so by the Squeezing Lemma 66, f is differentiable at zero and $f'(0) = 1$. For nonzero x, $f'(x) = 1 + 4x \sin \frac{1}{x^2} - \frac{2}{x} \cos \frac{1}{x^2}$, which is continuous and attains arbitrarily large positive and negative values in every neighborhood of zero. Therefore f cannot be monotone on a neighborhood of zero. See Figure 3.1. Note that the derivative of f is discontinuous at zero.

The idea of the derivative as the slope of a tangent line is weird in this case.

Note that this function is increasing at zero in Landau's sense, but is not monotone on any open interval containing zero. □

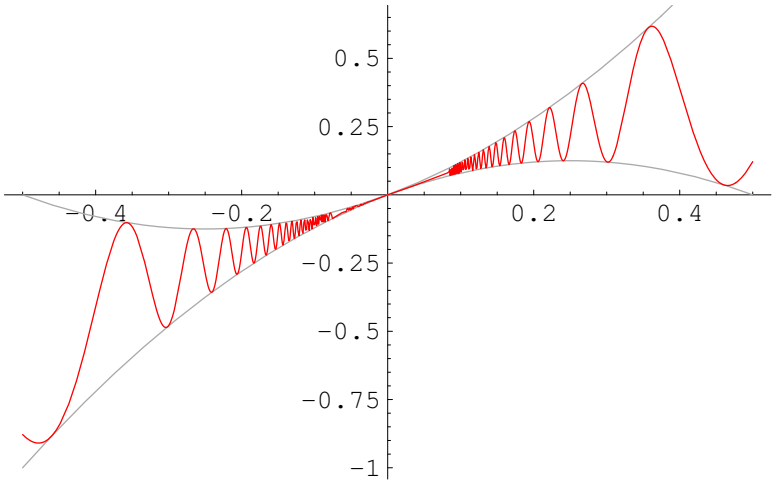

Figure 3.1. The nonmonotone function $x + 2x^2 \sin \frac{1}{x^2}$.

The next lemma is not hard to see. Compare it to Theorem 75 below.

71 Lemma *Let $f\colon (\alpha, \beta) \to \boldsymbol{R}$ be differentiable at x with $f'(x) > 0$. Then f is increasing at x (in Landau's sense). Likewise if $f'(x) < 0$. Then f is increasing at x.*

The next result is another consequence of the Mean Value Theorem, see Apostol [7, Exercise 10, p. 187].

72 Intermediate Value Theorem *Let $f\colon (\alpha, \beta) \to \boldsymbol{R}$ be everywhere differentiable on (α, β). Then f' assumes every value between $f'(a)$ and $f'(b)$ somewhere in (α, β).*

Are $f'(\alpha)$ and $f'(\beta)$ defined?

This implies among other things that the derivative of a function cannot have any jump discontinuities. *If a derivative is not continuous, the discontinuities must be of the second kind. That is, if f' is not continuous at some point γ it must be that $\lim_{x \to \gamma} f'(x)$ fails to exist.*

Many of the results on maximization can be derived from Taylor's Theorem, which has two useful forms. The first is a generalization of the Mean Value Theorem, and assumes $n-1$ continuous derivatives on an open set and the existence everywhere of the n^{th} derivative. The version here is taken from Apostol [6, Theorem 5–14, p. 96].

73 Taylor's Theorem *Let $f\colon (\alpha,\beta) \to \boldsymbol{R}$ be $n-1$ times continuously differentiable on (α,β) and assume that f has an n^{th} derivative at each point of (α,β). Fix a point x in (α,β). For any $v \neq 0$ such that $x+v$ belongs to (α,β), there is a point u strictly between 0 and v such that*

$$f(x+v) = f(x) + \sum_{k=1}^{n-1} \frac{f^{(k)}(x)}{k!}v^k + \frac{f^{(n)}(x+u)}{n!}v^n.$$

The other useful form of Taylor's Theorem is Young's form. It too assumes $n-1$ continuous derivatives on an open set, but assumes only that the n^{th} derivative exists at a point, and has a remainder term. This statement is a slight rewording of Serfling [138, Theorem C, p. 45], who cites Hardy [72, p. 278].

74 Young's Form of Taylor's Theorem *Let $f\colon (\alpha,\beta) \to \boldsymbol{R}$ be $n-1$ times continuously differentiable on (α,β) and assume that f has an n^{th} derivative at the point x in (α,β). For any v such that $x+v$ belongs to (α,β),*

$$f(x+v) = f(x) + \sum_{k=1}^{n} \frac{f^{(k)}(x)}{k!}v^k + \frac{r(v)}{n!}v^n,$$

where the remainder term $r(v)$ satisfies

$$\lim_{v\to 0} r(v) = 0.$$

I will prove the multivariable version of this for the case $n = 2$ in Theorem 106 below. That will demonstrate how the proof works in this case.

3.2 Extrema of a function of one variable

The main reference for these results is Apostol [7, pp. 181–195, 273–280].

3.2.1 Necessary first order conditions

We present the so-called **first order necessary conditions** for an interior extremum. They are called first order conditions because they involve first derivatives.

75 Theorem (Necessary First Order Conditions) *Let $f\colon [\alpha,\beta] \to \boldsymbol{R}$ and let x^* be a local maximizer of f. Assume f has a derivative at x^*.*

If $\alpha < x^* < \beta$ (x^* is an interior point), then

$$f'(x^*) = 0.$$

If $x^* = \alpha$, then $f'(x^*) \leqslant 0$. If $x^* = \beta$, then $f'(x^*) \geqslant 0$. In short, if x^* is a local maximizer, then

$$f'(x^*)(x^* - x) \geqslant 0 \quad \text{for all } x \in [\alpha,\beta].$$

If x^* is a local minimizer, then

$$f'(x^*)(x^* - x) \leqslant 0 \quad \text{for all } x \in [\alpha,\beta].$$

Proof: If f has a local maximum at x^*,

$$f(x^*) \geqslant f(x^* + v)$$

if v is small enough and $x^* + v \in [\alpha, \beta]$. Thus

$$\frac{f(x^* + v) - f(x^*)}{v} \quad \text{is} \quad \begin{cases} \leqslant 0 & \text{for} \quad v > 0 \\ \geqslant 0 & \text{for} \quad v < 0. \end{cases}$$

Now take limits as $v \to 0$ or $v \downarrow 0$ or $v \uparrow 0$, as appropriate.

The case of a minimum is similar, but the inequalities are reversed. ∎

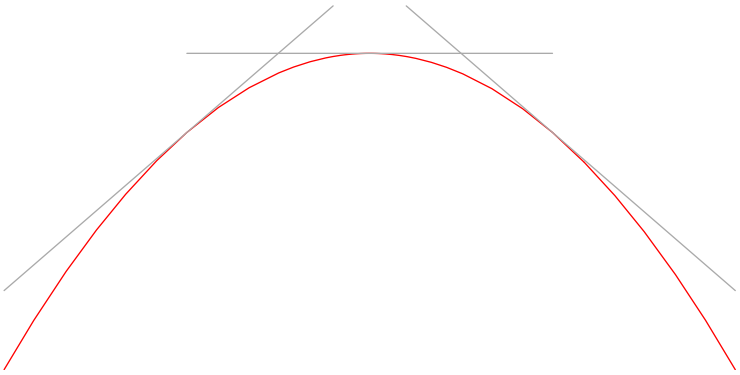

Figure 3.2. A nicely behaved maximum.

3.2.2 Sufficient first order conditions

The next result is a straightforward consequence of the Mean Value Theorem 67, cf. Apostol [7, Theorems 4.6 and 4.7].

76 Theorem (Sufficient First Order Conditions) *Suppose f is continuous on $[\alpha, \beta]$, and differentiable on (α, β) except perhaps at $\gamma \in (\alpha, \beta)$.*

- *If $f'(x) > 0$ for $x \in (\alpha, \gamma)$ and $f'(x) < 0$ for $x \in (\gamma, \beta)$, then γ is a strict maximizer of f on $[\alpha, \beta]$.*
- *If $f'(x) \geqslant 0$ for $x \in (\alpha, \gamma)$ and $f'(x) \leqslant 0$ for $x \in (\gamma, \beta)$, then γ maximizes f on $[\alpha, \beta]$.*
- *If $f'(x) < 0$ for $x \in (\alpha, \gamma)$ and $f'(x) > 0$ for $x \in (\gamma, \beta)$, then γ is a strict minimizer of f on $[\alpha, \beta]$.*
- *If $f'(x) \leqslant 0$ for $x \in (\alpha, \gamma)$ and $f'(x) \geqslant 0$ for $x \in (\gamma, \beta)$, then γ is a minimizer of f on $[\alpha, \beta]$.*

The conditions of Theorem 76 are sufficient conditions for the existence of a maximum at a point, but are hardly necessary.

77 Example (An unruly maximum) This is almost the same as Example 70. Consider the function

$$f(x) = \begin{cases} -x^2\big(2 + \sin\big(\frac{1}{x}\big)\big) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

(see Figure 3.3). Clearly f is differentiable for all nonzero x, and by the Squeezing Lemma 66, f is differentiable at zero since it is squeezed between $-x^2$ and $-3x^2$. In fact,

$$f'(x) = \begin{cases} -2x\big(2 + \sin\big(\frac{1}{x}\big)\big) + \cos\big(\frac{1}{x}\big) & x \neq 0 \\ 0 & x = 0. \end{cases}$$

Observe that f achieves a strict local maximum at zero, but its derivative switches sign infinitely often on both sides of zero (since for small x the $\cos\big(\frac{1}{x}\big)$ term determines the sign of $f'(x)$). In particular, the function is neither increasing on any interval $(\alpha, 0)$ nor decreasing on any interval $(0, \beta)$. This example appears in Sydsaeter [146]. □

Morse Theorem.

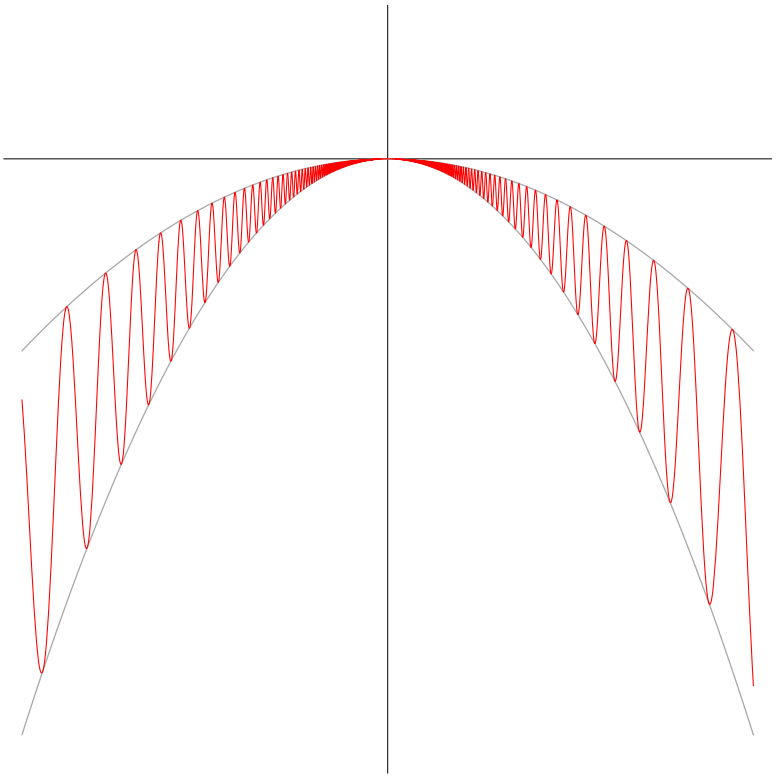

Figure 3.3. An unruly maximum at 0 for $-x^2\left(2 + \sin\big(\frac{1}{x}\big)\right)$.

3.2.3 Second order (and higher order) sufficient conditions

The next theorem has slightly weaker hypotheses than the standard statement, which assumes that f is n times continuously differentiable on an interval, see, e.g., Apostol [6, Theorem 7–7, p. 148].

78 Theorem (Higher order sufficient conditions) *Let $f\colon (\alpha, \beta) \to \boldsymbol{R}$ be $n-1$ times continuously differentiable on (α, β) and assume that it has an n^{th} derivative at the interior point x^*. Suppose in addition that*

$$f'(x^*) = f''(x^*) = \cdots = f^{(n-1)}(x^*) = 0 \quad \text{and} \quad f^{(n)}(x^*) \neq 0.$$

- *If n is even, and $f^{(n)}(x^*) < 0$, then f has a strict local maximum at x^*.*

- *If n is even, and $f^{(n)}(x^*) > 0$, then f has a strict local minimum at x^*.*
- *If n is odd, then f has neither a local maximum nor a local minimum at x^*.*

Proof: From Young's form of Taylor's Theorem 74, for any v with $x^* + v \in (\alpha, \beta)$ we have

$$f(x^* + v) - f(x^*) = \frac{f^{(n)}(x^*) + r(v)}{n!} v^n,$$

where $\lim_{v \to 0} r(v) = 0$. So for $|v|$ small enough we have $|r(v)| < |f^{(n)}(x^*)|$. Then the sign of $f(x^* + v) - f(x^*)$ is the same as the sign of $f^{(n)}(x^*)v^n$. When n is even $v^n > 0$ and when n is odd v^n switches signs, and the conclusion follows. ∎

79 Corollary (Necessary Second Order Conditions) *Let $f\colon (\alpha, \beta) \to \boldsymbol{R}$ is continuously differentiable on a neighborhood of x^* and suppose $f''(x^*)$ exists. If x^* is a local maximizer of f, then $f''(x^*) \leqslant 0$. If x^* is a local minimizer of f, then $f''(x^*) \geqslant 0$.*

Proof: Assume first that x^* is a local maximizer. Then by Theorem 75 we must have $f'(x^*) = 0$, so by Theorem 78, if $f''(x^*) > 0$, then it is a *strict* local minimizer, so it cannot be a local maximizer. By contraposition then, $f''(x^*) \leqslant 0$. For the case of a local minimizer reverse the signs. ∎

80 Example (All derivatives vanishing at a strict maximum) It is possible for f to have derivatives of all orders that all vanish at a strict local maximizer. E.g., define

$$f(x) = \begin{cases} -e^{-\frac{1}{x^2}} & x \neq 0 \\ 0 & x = 0. \end{cases}$$

Then 0 is a strict global maximizer of f. Furthermore f has derivatives of all orders everywhere. (Why?) Nevertheless $f^{(n)}(0) = 0$ for all n. This example appears in Apostol [6, Exercise 5–4, p. 98] and in Sydsaeter [146]. It is closely related to the function

Elaborate. $f^{(n)}(x) = \frac{e^{-x^{-2}}}{p(x)}$, where p is a polynomial.

$$f(x) = \begin{cases} e^{\frac{1}{x^2 - 1}} & |x| \leqslant 1 \\ 0 & |x| \geqslant 1, \end{cases}$$

which was shown by Cauchy to have continuous derivatives of all orders.[3] □

3.3 The Classical Fundamental Theorems

This section is a review of the Fundamental Theorems of Calculus, as presented in Apostol [7]. The notion of integration employed is the Riemann integral. Recall that a bounded function is Riemann integrable on an interval $[\alpha, \beta]$ if and only it is continuous except on a set of Lebesgue measure zero. In this case its Lebesgue integral and its Riemann integral are the same.

Recall that an *indefinite integral* of f over the interval I is a function F satisfying $F(x) = \int_\alpha^x f(s)\,ds$ for every x in I, for some fixed choice of a in I. Different values of a give rise to different indefinite integrals. By the remarks just made, every function that is continuous almost everywhere has an indefinite integral on a bounded interval I.

A function P is a *primitive* or *antiderivative* of a function f on an open interval I if $P'(x) = f(x)$ for every x in I. Leibniz' notation for this is $\int f(x)\,dx = P(x) + C$. Note that if P is an antiderivative of f, then so is $P + C$ for any constant function C.

[3] At least Horváth [77, p. 166] attributes this function to Cauchy. See Aliprantis and Burkinshaw [4, Problem 21.2, p. 152] for a proof of the properties of Cauchy's function.

Despite the similarity in notation, we see that an antiderivative is conceptually distinct from an indefinite integral. The statement that P is an antiderivative of f is a statement about the derivative of P, namely that $P'(x) = f(x)$ for all x in I; whereas the statement that F is an indefinite integral of f is a statement about the integral of f, namely that there exists some α in I with $\int_\alpha^x f(s)\,ds = F(x)$ for all x in I. Nonetheless there is a close connection between the concepts, which justifies the similar notation. The connection is laid out in the two Fundamental Theorems of Calculus.

81 First Fundamental Theorem of Calculus [7, Theorem 5.1, p. 202] *Let f be integrable on $[\alpha, x]$ for each x in $I = [\alpha, \beta]$. Let $\alpha \leqslant \gamma \leqslant \beta$, and define F by*

$$F(x) = \int_\gamma^x f(s)\,ds.$$

Then F is differentiable at every x in (α, β) where f is continuous, and at such points $F'(x) = f(x)$.

That is, an indefinite integral of a continuous integrable function is also an antiderivative of the function.

This result is often loosely stated as, "the integrand is the derivative of its (indefinite) integral," which is not strictly true unless the integrand is continuous.

82 Second Fundamental Theorem of Calculus [7, Theorem 5.3, p. 205] *Let f be continuous on (α, β) and suppose that f possesses an antiderivative P. That is, $P'(x) = f(x)$ for every x in (α, β). Then for each x and γ in (α, β), we have*

$$P(x) = P(\gamma) + \int_\gamma^x f(s)\,ds = P(\gamma) + \int_c^x P'(s)\,ds.$$

That is, an antiderivative of a continuous function is also an indefinite integral.

This result is often loosely stated as, "a function is the (indefinite) integral of its derivative," which is not true. What is true is that "a function that happens to be an indefinite integral of something continuous, is an (indefinite) integral of its derivative." To see this, suppose that F is an indefinite integral of f. That is, for some a the Riemann integral $\int_\alpha^x f(s)\,ds$ is equal to $F(x)$ for every x in the interval I. In particular, f is Riemann integrable over $[\alpha, x]$, so it is continuous everywhere in I except possibly for a set N of Lebesgue measure zero. Consequently, by the First Fundamental Theorem, except possibly for a set N of measure zero, F' exists and $F'(x) = f(x)$. Thus the Lebesgue integral of F' over $[\alpha, x]$ exists for every x and is equal to the Riemann integral of f over $[\alpha, x]$, which is equal to $F(x)$. In that sense, F is the integral of its derivative. Thus we see that a necessary condition for a function to be an indefinite integral is that it be differentiable almost everywhere.

Is this condition sufficient as well? It turns out that the answer is no. There exist continuous functions that are differentiable almost everywhere that are not an indefinite integral of their derivative. Indeed such a function is not an indefinite integral of any function. The commonly given example is the Cantor ternary function.

3.4 The Cantor ternary function

Given any number x with $0 \leqslant x \leqslant 1$ there is an infinite sequence $(a_1, a_2, \ldots)$, where each a_n belongs to $\{0, 1, 2\}$, such that $x = \sum_{n=1}^\infty \frac{a_n}{3^n}$. This sequence is called a **ternary expansion** of x and the infinite series is the **ternary representation** of x. If x is of the form $\frac{N}{3^m}$ (in lowest terms), then it has two ternary representations and expansions: a terminating representation of the form $x = \sum_{n=1}^\infty \frac{a_n}{3^n}$, where $a_m > 0$ and $a_n = 0$ for $n > m$, and a repeating representation of

the form $x = \sum_{n=1}^{m-1} \frac{a_n}{3^n} + \frac{a_m - 1}{3^m} + \sum_{n=m+1}^{\infty} \frac{2}{3^n}$. But these are the only cases of a nonunique ternary representation, and there are only countably many such numbers. (See, e.g., [33, Theorem 1.23, p. 20].)

Given $x \in [0,1]$, let $N(x)$ be the first n such that $a_n = 1$ in the ternary expansion of x. If x has two ternary expansions use the one that gives the larger value for $N(x)$. For this discussion we shall call this the "preferred" expansion. If x has a ternary expansion with no $a_n = 1$, then $N(x) = \infty$. The **Cantor set** $\mathcal{C}$ consists of all numbers x in $[0,1]$ for which $N(x) = \infty$. That is, those that have a preferred ternary expansion where no $a_n = 1$. That is, all numbers x of the form $x = \sum_{n=1}^{\infty} \frac{2b_n}{3^n}$, where each b_n belongs to $\{0,1\}$. Each distinct sequence of 0s and 1s gives rise to a distinct element of $\mathcal{C}$. Indeed some authors identify the Cantor set with $\{0,1\}^{\mathbb{N}}$ endowed with its product topology, since the mapping $(b_1, b_2, \ldots) \mapsto \sum_{n=1}^{\infty} \frac{2b_n}{3^n}$ is a homeomorphism. Also note that a sequence $(b_1, b_2, \ldots)$ of 0s and 1s also corresponds to a unique subset of $\mathbb{N}$, namely $\{n \in \mathbb{N} : b_n = 1\}$. Thus there are as many elements $\mathcal{C}$ as there are subset of $\mathbb{N}$, so the Cantor set is uncountable. (This follows from the Cantor diagonal procedure.) Yet the Cantor set includes no interval.

It is perhaps easier to visualize the complement of the Cantor set. Let

$$\mathcal{A}_n = \{x \in [0,1] : N(x) = n\}.$$

The complement of the Cantor set is $\bigcup_{n=1}^{\infty} \mathcal{A}_n$. Define

$$\mathcal{C}_n = [0,1] \setminus \bigcup_{k=1}^{n} \mathcal{A}_k,$$

so that $\mathcal{C} = \bigcap_{n=0}^{\infty} \mathcal{C}_n$. Now $\mathcal{A}_1$ consists of those x for which $a_1 = 1$ in its preferred ternary expansion. This means that

$$\mathcal{A}_1 = \left(\tfrac{1}{3}, \tfrac{2}{3}\right) \qquad \text{and} \qquad \mathcal{C}_1 = \left[0, \tfrac{1}{3}\right] \cup \left[\tfrac{2}{3}, 1\right].$$

Note that $N(\frac{1}{3}) = \infty$ since $\frac{1}{3}$ has as its preferred representation $\sum_{n=2}^{\infty} \frac{2}{3^n}$ ($a_1 = 0$, $a_n = 2$ for $n > 1$). Now $\mathcal{A}_2$ consists of those x for which $a_1 = 0$ or $a_1 = 2$ and $a_2 = 1$ in its preferred ternary expansion. That last sentence is ambiguous. To be precise,

$$\mathcal{A}_2 = \left(\tfrac{1}{9}, \tfrac{2}{9}\right) \cup \left(\tfrac{7}{9}, \tfrac{8}{9}\right) \qquad \text{and} \qquad C_2 = \left[0, \tfrac{1}{9}\right] \cup \left[\tfrac{2}{9}, \tfrac{1}{3}\right] \cup \left[\tfrac{2}{3}, \tfrac{7}{9}\right] \cup \left[\tfrac{8}{9}, 1\right].$$

Each $\mathcal{C}_n$ is the union of 2^n closed intervals, each of length $\frac{1}{3^{n-1}}$, and $\mathcal{A}_{n+1}$ consists of the open middle third of each of the intervals in $\mathcal{C}_n$. The total length of the removed open segments is

$$\frac{1}{3} + 2 \cdot \frac{1}{9} + 4 \cdot \frac{1}{27} + \cdots = \sum_{n=0}^{\infty} \frac{2^n}{3^{n+1}} = \frac{1}{3} \sum_{n=0}^{\infty} \left(\frac{2}{3}\right)^n = \frac{1}{3} \cdot \frac{1}{1 - \frac{2}{3}} = 1.$$

Thus the total length of the Cantor set is $1 - 1 = 0$.

The Cantor ternary function f is defined as follows. On the open middle third $(\frac{1}{3}, \frac{2}{3})$ its value is $\frac{1}{2}$. On the open interval $(\frac{1}{9}, \frac{2}{9})$ its value is $\frac{1}{4}$ and on $(\frac{7}{9}, \frac{8}{9})$ its value is $\frac{3}{4}$. Continuing in this fashion, the function is defined on the complement of the Cantor set. The definition is extended to the entire interval by continuity. See Figure 3.4. A more precise but more opaque definition is this:

$$f(x) = \sum_{n=1}^{N(x)-1} \frac{\frac{1}{2}a_n}{2^n} + \frac{a_{N(x)}}{2^{N(x)}},$$

where $(a_1, a_2, \ldots)$ is the preferred ternary expansion of x. (If $N(x) = \infty$ we interpret this as the infinite series without the last term.)

Note that the range of f is all of $[0,1]$ since we can obtain the binary expansion of any real in $[0,1]$ as a value for f. (Since for $n < N(x)$ we have $a_n = 0$ or $a_n = 2$, so $\frac{1}{2}a_n$ is either zero

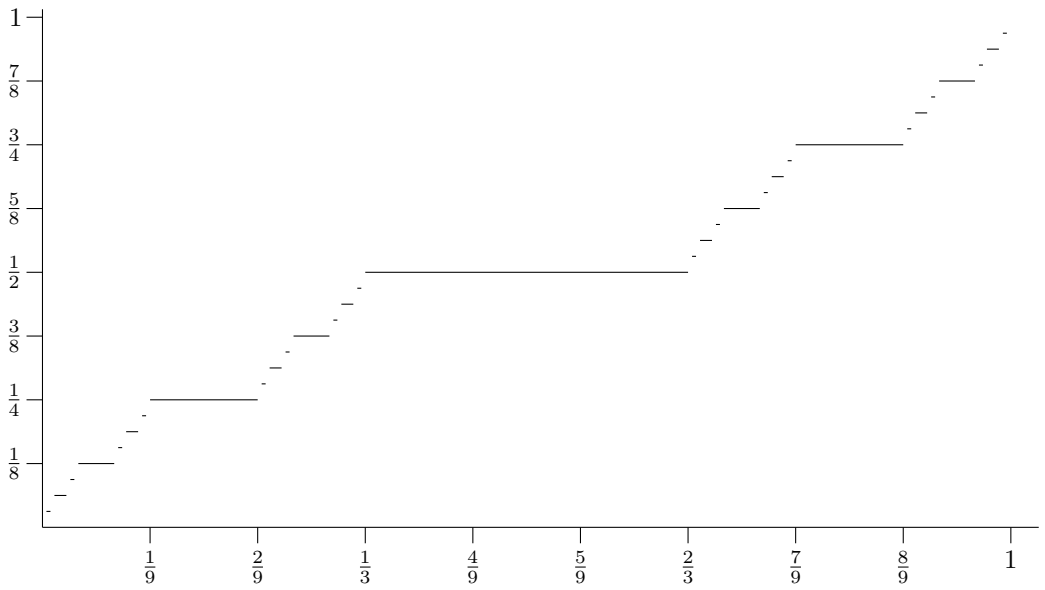

Figure 3.4. Partial graph of the Cantor ternary function.

or one.) Finally, I claim that f is continuous. (To see this make two observations. If $N(x)$ is finite, then f is constant on some neighborhood of x, and thus continuous at x. If $N(x) = \infty$, so that $x \in \mathcal{C}$, then for every $\varepsilon > 0$, there is some M such that $\sum_{n=M}^{\infty} \frac{1}{2^n} < \varepsilon$ and $\delta > 0$ such that $|x - y| < \delta$ implies $N(y) > M$, so $|f(x) - f(y)| < \varepsilon$.)

In any event, notice that f is constant on each open interval in some $\mathcal{A}_n$, so it is differentiable there and $f' = 0$. Thus f is differentiable almost everywhere, and $f' = 0$ wherever it exists, but

$$f(1) - f(0) = 1 \neq 0 = \int_0^1 f'(x)\,dx.$$

The Cantor function is also an example of a continuous function whose derivative is zero almost everywhere, but manages to increase from $f(0) = 0$ to $f(1) = 1$. There are more perverse functions on $[0, 1]$ that are continuous and strictly increasing, yet have a derivative that is zero almost everywhere. I won't go into that here, but see, for instance, Kannan and Krueger [88, § 8.6, pp. 208ff.] or Royden [132, Exercise 16e, p. 111].

3.5 Differentiation on normed spaces

In this section I want to introduce progressively more restrictive notions of derivatives and differentiability for functions between real vector spaces. The real line is considered to be the one-dimensional vector space. Recall that a derivative is some kind of limit of line segments joining points on the graph of a function. The simplest way to take such a limit is along a line segment containing x.

83 Definition (One-sided directional derivative) *Let A be a subset of the vector space X, let Y be a topological vector space, and let $f\colon A \to Y$.*

We say that f has the ***one-sided directional derivative*** *$f'(x;v)$ at x in the direction v, if $f'(x;v)$ is a vector in Y satisfying*

$$f'(x;v) = \lim_{\lambda \downarrow 0} \frac{f(x+\lambda v) - f(x)}{\lambda}.$$

In order for this definition to make sense, we implicitly require that there is some $\varepsilon > 0$ such that $0 \leqslant \lambda \leqslant \varepsilon$ implies that $x + \lambda v$ belongs to A, so that $f(x + \lambda v)$ is defined.

For the case $Y = \boldsymbol{R}$, we also permit f to assume one of the extended values $\pm\infty$, and also permit $f'(x;v)$ to assume one of the values $\pm\infty$.

Note that in the definition of $f'(x,v)$, the limit is taken in Y, so a topology is needed on Y, but none is necessary on X. Also note that $x + \lambda v$ need not belong to A for $\lambda < 0$. Considering $\lambda = 0$ implies $x \in A$. The next lemma shows that the set of v for which a one-sided directional derivative exists is a cone, and that $f'(x;v)$ is positively homogeneous in v on this cone.

84 Lemma *The one-sided directional derivative is positively homogeneous of degree one. That is, if $f'(x;v)$ exists, then*

$$f'(x;\alpha v) = \alpha f'(x;v) \qquad \text{for } \alpha \geqslant 0.$$

Proof: This follows from $\frac{f(x+\lambda\alpha v)-f(x)}{\lambda} = \alpha\frac{f(x+\beta v)-f(x)}{\beta}$, where $\beta = \lambda\alpha$, and letting $\lambda, \beta \downarrow 0$. ∎

85 Definition *If $f'(x;v) = -f'(x;-v)$, then we denote the common value by $D_v f(x)$, that is,*

$$D_v f(x) = \lim_{\lambda \to 0} \frac{f(x+\lambda v) - f(x)}{\lambda},$$

and we say that f has ***directional derivative*** *$D_v f(x)$ at x in the direction v.*

It follows from Lemma 84 that if $D_v(x)$ exists, then $D_{\alpha v}(x) = \alpha D_v(x)$ for all α. In $\boldsymbol{R}^{\mathrm{n}}$, the i^{th} **partial derivative** of f at x, if it exists, is the directional derivative in the direction e^i, the i^{th} unit coordinate vector.

Note that this definition still uses no topology on X. This generality may seem like a good thing, but it has the side effect that since it does not depend on the topology of X, it cannot guarantee the continuity of f at x in the normed case. That is, f may have directional derivatives in all nonzero directions at x, yet not be continuous at x. Moreover, we may not be able to express directional derivatives as a linear combination of partial derivatives.

86 Example (Directional derivatives without continuity or linearity) Let $f\colon \boldsymbol{R}^2 \to \boldsymbol{R}$ via

$$f(x,y) = \begin{cases} \dfrac{xy}{x^2+y} & y \neq -x^2 \\ 0 & y = -x^2. \end{cases}$$

Observe that f has directional derivatives at $(0,0)$ in every direction, for

$$\frac{f(\lambda x, \lambda y) - f(0,0)}{\lambda} = \frac{\left(\dfrac{\lambda^2 xy}{\lambda^2 x^2 + \lambda y}\right)}{\lambda} = \frac{xy}{\lambda x^2 + y}.$$

If $y \neq 0$, then the limit of this expression is x as $\lambda \to 0$, and if $y = 0$, the limit is 0. Thus the directional derivative exists for every direction (x, y), but it is not continuous at the x-axis.

But f is not continuous at $(0,0)$. For instance, for $\varepsilon > 0$,

$$f(\varepsilon, -\varepsilon^2 - \varepsilon^4) = \frac{-\varepsilon(\varepsilon^2 + \varepsilon^4)}{\varepsilon^2 - \varepsilon^2 - \varepsilon^4} = \frac{1}{\varepsilon} + \varepsilon \to \infty \text{ as } \varepsilon \to 0.$$

Note too that the mapping $v \mapsto D_v f(0)$ is not linear. □

87 Definition (The Gâteaux derivative) *Let X and Y be normed vector spaces. If $D_v f(x)$ exists for all $v \in X$ and the mapping $T\colon v \mapsto D_v f(x)$ from X to Y is a continuous linear mapping, then T is called the* ***Gâteaux derivative*** *or* ***Gâteaux differential*** *of f at x,*[4] *and we say that f is* ***Gâteaux differentiable*** *at x.*

This notion uses the topology on X to define continuity of the linear mapping, but Gâteaux differentiability of f is still not strong enough to imply continuity of f, even in two dimensions. The next example may be found, for instance, in Aubin and Ekeland [15, p. 18].

88 Example (Gâteaux differentiability does not imply continuity) Define $f\colon \boldsymbol{R}^2 \to \boldsymbol{R}$ by

$$f(x,y) = \begin{cases} \dfrac{y}{x}(x^2 + y^2) & x \neq 0 \\ 0 & x = 0. \end{cases}$$

Then for $x \neq 0$,

$$\frac{f(\lambda x, \lambda y) - f(0,0)}{\lambda} = \frac{\left(\dfrac{\lambda y}{\lambda x}\lambda^2(x^2 + y^2)\right)}{\lambda} = \frac{\lambda y}{x}(x^2 + y^2) \to 0.$$

Thus $D_v f(0) = 0$ for any v, so f has a Gâteaux derivative at the origin, namely the zero linear map.

But f is not continuous at the origin. For consider $v(\varepsilon) = (\varepsilon^4, \varepsilon)$. The $v(\varepsilon) \to 0$ as $\varepsilon \to 0$, but

$$f\big(v(\varepsilon)\big) = \frac{\varepsilon}{\varepsilon^4}(\varepsilon^8 + \varepsilon^2) = \varepsilon^5 + 1/\varepsilon.$$

Thus $f\big(v(\varepsilon)\big) \to \infty$ as $\varepsilon \downarrow 0$, and $f\big(v(\varepsilon)\big) \to -\infty$ as $\varepsilon \uparrow 0$, so $\lim_{\varepsilon \to 0} f\big(v(\varepsilon)\big)$ does not exist. □

A stronger notion of derivative has proven useful. Gâteaux differentiability requires that chords have a limiting slope along straight lines approaching x. The stronger requirement is that chords have a limiting slope along arbitrary approaches to x. The definition quite naturally applies to functions between any normed vector spaces, not just Euclidean spaces, so we shall work as abstractly as possible. Dieudonné [43] claims that this makes everything clearer, but I know some who may disagree.

[4] This terminology disagrees with Luenberger [102, p. 171], who does not require linearity. It is however, the terminology used by Aubin [13, Definition 1, p. 111] , Aubin and Ekeland [15, p. 33], and Ekeland and Temam [47, Definition 5.2, p. 23].

89 Definition (The differential or Fréchet derivative) *Let X and Y be normed real vector spaces. Let U be an open set in X and let $f\colon U \to Y$. The Gâteaux derivative is called* ***differential*** *at x (also known as a* ***Fréchet derivative****, a* ***total derivative****, or simply a* ***derivative****) if it satisfies*

$$\lim_{v\to 0} \frac{\|f(x+v) - f(x) - D_v f(x)\|}{\|v\|} = 0. \tag{D}$$

The differential is usually denoted $Df(x)$, and it is a function from X into Y. Its value at a point v in X is denoted $Df(x)(v)$ rather than $D_v f(x)$. The double parentheses are only slightly awkward, and you will get used to them after a while.

When f has a differential at x, we say that f is ***differentiable*** *at x, or occasionally for emphasis that f is* ***Fréchet differentiable*** *at x.*

Actually my definition is a bit nonstandard. I started out with directional derivatives and said that if the mapping was linear and satisfied (D), then it was the differential. That is, I defined the differential in terms of directional derivatives. The usual approach is to say that f has a differential at x if there is some continuous linear mapping T that satisfies

$$\lim_{v\to 0} \frac{\|f(x+v) - f(x) - T(v)\|}{\|v\|} = 0. \tag{D$'$}$$

It is then customary to prove the following lemma.

90 Lemma *If T satisfies* (D$'$)*, then $T(v) = f'(x;v)$. Consequently, T is unique, so*

$$Df(x)(v) = D_v f(x) = f'(x;v) = \lim_{\lambda\downarrow 0} \frac{f(x+\lambda v) - f(x)}{\lambda}.$$

Proof: Fix $v \neq 0$ and replace v by λv in (D$'$), and conclude

$$\lim_{\lambda\downarrow 0} \frac{\|f(x+\lambda v) - f(x) - T(\lambda v)\|}{\lambda\|v\|} = \lim_{\lambda\downarrow 0} \frac{1}{\|v\|} \left\| \frac{f(x+\lambda v) - f(x)}{\lambda} - T(v) \right\| = 0.$$

That is, $T(v) = \lim_{\lambda\downarrow 0} \frac{f(x+\lambda v)-f(x)}{\lambda} = f'(x;v)$. ∎

The continuity (equivalently boundedness) of $Df(x)(\cdot)$ implies the continuity of f.

91 Lemma (Differentiability implies Lipschitz continuity) *If f is differentiable at x, then f is continuous at x. Indeed f is Lipschitz continuous at x. That is, there is $M \geqslant 0$ and $\delta > 0$ such that if $\|v\| < \delta$, then*

$$\Delta_v f(x) < M\ \|v\|.$$

Proof: Setting $\varepsilon = 1$ in the definition of differentiability, there is some $\delta > 0$ so that $\|v\| < \delta$ implies $\|\Delta_v f(x) - Df(x)(v)\| < \|v\|$, so by the triangle inequality,

$$\|\Delta_v f(x)\| < \|v\| + \|Df(x)(v)\| \leqslant \big(\|Df(x)\| + 1\big)\|v\|,$$

where of course $\|Df(x)\|$ is the operator norm of the linear transformation $Df(x)$. Thus f is continuous at x. ∎

Rewriting the definition

There are other useful ways to state this definition that I may use from time to time. Start by defining the **first difference function** $\Delta_v f$ of f at x by v, where $\Delta_v f\colon X \to Y$, by[5]

$$\Delta_v f(x) = f(x+v) - f(x).$$

We can rewrite the definition of differentiability in terms of the first difference as follows: f is (Fréchet) differentiable at x if

$$\forall \varepsilon > 0\ \exists \delta > 0\ \forall v \quad 0 < \|v\| < \delta \implies \|\Delta_v f(x) - Df(x)(v)\| < \varepsilon \|v\|.$$

Another interpretation of the definition is this. Fix x and define the difference quotient function d_λ by

$$d_\lambda(v) = \frac{f(x+\lambda v) - f(x)}{\lambda}.$$

If f is differentiable at X, then d_λ converges uniformly on norm-bounded sets to the linear function $Df(x)$ as $\lambda \to 0$.

Further notes on the definition

When $X = Y = \boldsymbol{R}$, the differential we just defined is closely related to the derivative defined earlier for functions of one variable. The differential is the linear function $Df(x)\colon v \mapsto f'(x)v$, where $f'(x)$ is the numerical derivative defined earlier. Despite this difference, some authors (including Dieudonné [43], Luenberger [102], Marsden [107], and Spivak [145]) call the differential a derivative, but with modest care no serious confusion results. Loomis and Sternberg [100, pp. 158–159] argue that the term differential ought to be reserved for the linear transformation and derivative for its *skeleton* or matrix representation. But these guys are rather extreme in their views on notation and terminology—for instance, on page 157 they refer to the "barbarism of the classical notation for partial derivatives."

Also note that my definition of differentiability does not require that f be continuous anywhere but at x. In this, I believe I am following Loomis and Sternberg [100, p. 142]. Be aware that some authors, such as Dieudonné [43, p. 149] only define differentiability for functions continuous on an open set. As a result the function $f\colon \boldsymbol{R}^2 \to \boldsymbol{R}$ defined by

$$f(x,y) = \begin{cases} x^2 + y^2 & x = y \\ 0 & x \neq y \end{cases}$$

is differentiable at $(x,y) = (0,0)$ under my definition, but not under Dieudonné's definition. By the way, Dieudonné does not require that the differential be a continuous linear transformation, he proves it using the continuity of f. Since we do not assume that f is continuous, we must make continuity of $Df(x)(\cdot)$ part of the definition (as do Loomis and Sternberg).

More variations on the definition

In Definition 89, I required that f be defined on an open set U in a normed space. Some authors, notably Graves [64], do not impose this restriction. Graves's definition runs like this.

Let X and Y be normed real vector spaces. Let A be a subset of X and let $f\colon A \to Y$. The **differential** at x is a linear transformation $T\colon X \to Y$ such that ********** This differential is also denoted $Df(x)$. Note that if A is small enough, the differential may not be uniquely defined.

[5] Loomis and Sternberg would write this as $\Delta f_x(v)$. Their notation makes it awkward to write second differences (see section 3.9).

3.6 Chain rule

92 Chain rule *Let X, Y, and Z be normed vector spaces, and let U be an open subset of X and V be an open subset of Y. Let $f\colon U \to V$, and $g\colon V \to Z$. If f is differentiable at x and g is differentiable at $f(x)$, then $g \circ f$ is differentiable at x and*

$$D(g \circ f)(x) = Dg\big(f(x)\big) \circ Df(x).$$

Proof: There are elegant proofs of this based on Landau's little-o notation, or what Loomis and Sternberg call infinitesimals. I've never been able to remember the rules for manipulating these for more than fifteen minutes, and I only need to do so every three or four years, so here is a somewhat tedious proof based only on the definitions.

Observe that

$$\begin{aligned} g\big(f(x+v)\big) - g\big(f(x)\big) &= g\big(f(x) + \Delta_v f(x)\big) - g\big(f(x)\big) \\ &= \Delta_{\Delta_v f(x)} g\big(f(x)\big). \end{aligned}$$

Fix $\varepsilon > 0$, and without loss of generality choose $\varepsilon < 1$. Set $M = \|Df(x)\| + 1$. To ease notation to come, define $u(v) = \Delta_v f(x)$ and $r(v) = \Delta_v f(x) - Df(x)(v)$. Since f is differentiable at x, there is $\delta_1 > 0$ such that $0 < \|v\| < \delta_1$ implies $\|r(v)\| < \varepsilon\|v\|$ and that

$$\|u(v)\| = \|Df(x)(v) + r(v)\| \leqslant \|Df(x)(v)\| + \|r(v)\| \leqslant \big(\|Df(x)\| + \varepsilon\big)\|v\| \leqslant M\|v\|.$$

(Note the weak inequalities. It is quite possible that $u(v) = 0$ even if $v \neq 0$. If $u(v) = 0$, then we actually have equality, otherwise we have strict inequality.)

Since g is differentiable at $f(x)$, there is some $\delta_2 > 0$ such that $\|u\| < \delta_2$ implies

$$\|\Delta_u g\big(f(x)\big) - Dg\big(f(x)\big)(u)\| \leqslant \varepsilon\|u\|.$$

Thus for $\|v\| < \delta = \min \delta_1, \frac{\delta_2}{M}$, we have $\|u(v)\| < M\|v\| < M\frac{\delta_2}{M} = \delta_2$, so

$$\|\Delta_{u(v)} g\big(f(x)\big) - Dg\big(f(x)\big)\big(u(v)\big)\| \leqslant \varepsilon\|u(v)\|.$$

In other words,

$$\|g\big(f(x+v)\big) - g\big(f(x)\big) - Dg\big(f(x)\big)\big(u(v)\big)\| \leqslant \varepsilon\|u(v)\|.$$

But

$$\begin{aligned} Dg\big(f(x)\big)\big(u(v)\big) &= Dg\big(f(x)\big)\big(Df(x)(v) + r(v)\big) \\ &= Dg\big(f(x)\big)\big(Df(x)(v)\big) + Dg\big(f(x)\big)\big(r(v)\big), \end{aligned}$$

so for $\|v\| < \delta$,

$$\begin{aligned} \big\|g\big(f(x+v)\big) - g\big(f(x)\big) - \big(Dg\big(f(x)\big) \circ Df(x)\big)(v)\big\| &\leqslant \varepsilon\big\|u(v)\big\| + \big\|Dg\big(f(x)\big)\big(r(v)\big)\big\| \\ &\leqslant \varepsilon\big\|u(v)\big\| + \big\|Dg\big(f(x)\big)\big\|\,\|r(v)\| \\ &\leqslant \varepsilon\big(\|Df(x)\| + 1\big)\|v\| + \varepsilon\big\|Dg\big(f(x)\big)\big\|\,\|v\| \\ &= \varepsilon\left(\|Df(x)\| + \big\|Dg\big(f(x)\big)\big\| + 1\right)\,\|v\|, \end{aligned}$$

which shows that $Dg\big(f(x)\big) \circ Df(x)$ satisfies the definition of $D(g \circ f)(x)$. ∎

3.7 Computing the differential

Let's compute a few simple differentials.

93 Lemma (Differential of a linear mapping) *If $T\colon X \to Y$ is a continuous linear function between normed spaces, then T is everywhere differentiable, and for each x,*

$$DT(x) = T.$$

Proof: To see this, just note that T satisfies the definition, as $T(x+v) - T(x) - T(v) = 0$. ∎

94 Example (Differential of the evaluation mapping) The **evaluation map** δ_x at x assigns to each function f its value at the point x. The evaluation map $\delta_x\colon \mathsf{L}(X,Y) \to Y$,

$$\delta_x\colon T \mapsto T(x)$$

is clearly linear. Moreover it is continuous (in the operator norm) and therefore differentiable, so by Lemma 93, for any $T \in \mathsf{L}(X,Y)$,

$$D\delta_x(T) = \delta_x$$

□

Do this for general bilinear functions, and composition of general functions. And evaluation of general functions.

95 Example (Differential of the composition mapping) The composition of a linear function from Y into Z with a linear function from X into Y is a linear function from X into Z. Consider composition as a mapping c from the vector space $\mathsf{L}(X,Y) \times \mathsf{L}(Y,Z)$ into $\mathsf{L}(X,Z)$ defined by

$$c(S,T) = T \circ S.$$

If X, Y, and Z are normed spaces, then the operator norm makes each of $\mathsf{L}(X,Y)$, $\mathsf{L}(Y,Z)$, and $\mathsf{L}(X,Z)$ into normed spaces. Is the composition map c differentiable? The answer is yes, and moreover,

$$Dc(S,T)(U,V) = T \circ U + V \circ S.$$

To see this observe that

$$\begin{aligned} c(S+\lambda U, T+\lambda V)(x) &= (T+\lambda V)\big(S(x)+\lambda U(x)\big) \\ &= T\big(S(x)\big) + \lambda T\big(U(x)\big) + \lambda U\big(S(x)\big) + \lambda^2 V\big(U(x)\big), \end{aligned}$$

Oops this is just the directional derivative.

so

$$\frac{c(S+\lambda U, T+\lambda V) - c(S,T)}{\lambda} = T \circ U + V \circ S + \lambda V \circ U \xrightarrow[\lambda \to 0]{} T \circ U + V \circ S.$$

That is, $Dc(S,T)(U,V) = T \circ U + V \circ S = c(U,T) + c(S,V)$. □

96 Example (Differential of a Cartesian product) Let $f\colon X \to Y$ and $g\colon X \to Z$ be differentiable, and define $p\colon X \to Y \times Z$ by

$$p(x) = \big(f(x), g(x)\big).$$

Then p is differentiable, $Dp(x) \in \mathsf{L}\big(X, \mathsf{L}(Y \times Z)\big)$, and

$$Dp(x)(v) = \big(Df(x)(v), Dg(x)(v)\big).$$

The only nontrivial part of this is to figure out the appropriate norm on $Y \times Z$. There are many. I like $\|(y,z)\| = \|y\| + \|z\|$. □

3.8 The mean value theorem

State a conventional version.

The mean value theorem is the workhorse of calculus. We give here a result that seems misnamed, but it is the version favored by Dieudonné [43, Theorem 8.5.1, p. 158]. (See also Loomis and Sternberg [100], Theorem 7.3 on p. 148 and Exercise 7.15 on p. 152.)

97 Mean Value Theorem *Let X be a normed space and $[a,b]$ be a nontrivial compact interval in $\boldsymbol{R}$. Let $f\colon [a,b] \to X$ and $g\colon [a,b] \to \boldsymbol{R}$ be continuous and differentiable everywhere on (a,b). Assume*

$$\|Df(t)\| \leqslant g'(t) \qquad t \in (a,b).$$

(Remember, $\|Df(t)\|$ is the operator norm of the linear transformation.) Then

$$\|f(b) - f(a)\| \leqslant g(b) - g(a).$$

Proof: Let $\varepsilon > 0$ be given and define A to be the set of points c in $[a,b]$ such that for all $a \leqslant t < c$, we have

$$\|f(t) - f(a)\| \leqslant g(t) - g(a) - \varepsilon(t-a).$$

It is easy to see that A is an interval, that $a \in A$, and that A is closed. Moreover if $A = [a,d]$, then by continuity of f and g we have $\|f(d)-f(a)\| \leqslant g(d)-g(a)-\varepsilon(d-a)$. Since ε is arbitrary, we conclude that in fact $\|f(d) - f(a)\| \leqslant g(d) - g(a)$. Therefore it suffices to show that $d = b$.

Suppose by way of contradiction that $d < b$. From the definition of differentiability, there is some $\delta > 0$ so that $|v| < \delta$ implies

$$\|f(d+v) - f(d) - Df(d)(v)\| \leqslant \frac{\varepsilon}{2}|v|$$

and

$$|g(d+v) - g(d) - g'(d)\cdot v| \leqslant \frac{\varepsilon}{2}|v|.$$

Thus for any $0 < v < \delta$, we have

$$\begin{aligned}
\|f(d+v) - f(a)\| &\leqslant \|f(d+v) - f(d)\| + \|f(d) - f(a)\| \\
&\leqslant \big(\|Df(d)(v)\| + \frac{\varepsilon}{2}|v|\big) + g(d) - g(a) - \varepsilon(d-a) \\
&\leqslant \big(g'(d)\cdot v + \frac{\varepsilon}{2}|v|\big) + g(d) - g(a) - \varepsilon(d-a) \\
&\leqslant \varepsilon|v| + g(d) - g(a) - \varepsilon(d-a) \\
&\leqslant g(d) - g(a) - \varepsilon(d+v-a),
\end{aligned}$$

which contradicts the maximality of d. Thus $d = b$, so for every $t < b$,

$$\|f(t) - f(a)\| \leqslant g(t) - g(a) - \varepsilon(t-a).$$

By the continuity of f and g and the fact that $\varepsilon > 0$ is arbitrary we see that

$$\|f(b) - f(a)\| \leqslant g(b) - g(a).$$

∎

Note that Dieudonné [43, Theorem 8.5.1, p. 158] proves a stronger result. He requires differentiability only at all but countably many points in (a,b).

A consequence of this result is the following uniform approximation theorem.

98 Theorem *Let X and Y be normed spaces and let C be a convex subset of X. Let $f\colon C \to Y$ be differentiable everywhere in C, and let $T \in \mathrm{L}(X,Y)$ satisfy*

$$\|Df(x) - T\| \leqslant \varepsilon$$

for all x in C. Then

$$\|\Delta_v f(x) - T(v)\| \leqslant \varepsilon\|v\|$$

whenever x and $x+v$ belong to C.

Proof: Fix x and $x+v$ in C, and consider the parameterized path $\varphi\colon [0,1] \to Y$ by $\varphi(\lambda) = f(x+\lambda v) - T(\lambda v)$. By the Chain Rule

$$D\varphi(\lambda) = Df(x+\lambda v)(v) - \lambda T(v).$$

Thus

$$\|D\varphi(\lambda)\| = \|Df(x+\lambda v)(v) - T(\lambda v)\| \leqslant \varepsilon\|\lambda v\| \leqslant \varepsilon\|v\|$$

for $0 < \lambda < 1$. Thus by the Mean Value Theorem 97

$$\|\varphi(1) - \varphi(0)\| = \|f(x+v) - T(v) - f(x) + T(0)\| = \|\Delta_v f(x) - T(v)\| \leqslant \varepsilon\|v\|.$$

∎

3.9 Second differentials

In this section, I want to discuss higher differentials in the abstract framework of linear transformations, including Taylor's Theorem. From a practical standpoint this may seem superfluous, since it is possible to use mixed partial derivatives, which are conceptually simpler. Indeed many authors do this. There are two reasons for the exercise. The first is to make sure I understand it. The second is that for optimization in infinite-dimensional function spaces, the partial derivative approach is not so simple.

If f is differentiable at each point in U, we say that f is **continuously differentiable** if the mapping $x \mapsto Df(x)$ from U into $\mathrm{L}(X,Y)$ is continuous, where $\mathrm{L}(X,Y)$ is given the operator norm. We may also ask if this mapping has a differential at x_0. If it does, then f is **twice differentiable** at x_0 and the differential is called the **second differential** of f at x_0, denoted $D^2f(x_0)$.

Let's examine the second differential $D^2f(x_0)$ at x_0 more carefully. It is a linear transformation from X into $\mathrm{L}(X,Y)$, so that $D^2f(x_0)(v)$, where $v \in X$, is a linear function from X into Y. Thus for w in X, $D^2f(x_0)(v)(w)$ belongs to Y. What is the interpretation of this vector? I first claim that if Df is differentiable at x_0, then each directional derivative $D_v f$ is differentiable at x_0. Moreover, the directional derivative of $D_v f$ in the direction w at x_0 is $D^2f(x_0)(v)(w)$.

L&S assume continuous differentiability, but I don't see where it is used.

99 Lemma (Second differential and directional derivatives) *If f is differentiable on the open set U and is twice differentiable at x_0, then the mapping $D_v f\colon x \mapsto D_v f(x)$ from $U \subset X$ into Y is differentiable at x_0 for each $v \in X$. Moreover, the directional derivative of $D_v f$ at x_0 in the direction w is given by*

$$D_w(D_v f)(x_0) = D^2f(x_0)(v)(w).$$

Proof: (Loomis and Sternberg [100, Theorem 16.1, p. 186]) Observe that for each x, the directional derivative in direction v satisfies

$$D_v f(x) = Df(x)(v) = \delta_v\big(Df(x)\big),$$

where δ_v is the evaluation mapping at v. That is,

$$x \mapsto D_v f(x) = \delta_v \circ Df \colon X \to Y.$$

Now we have already seen that δ_v is differentiable on $\mathrm{L}(X,Y)$ (Example 94), and Df is differentiable by hypothesis. Therefore by the chain rule, $x \mapsto D_v f(x)$ is differentiable at x_0 and its differential is a linear function from X into Y that satisfies

$$D(D_v f)(x_0) = (D\delta_v \circ D^2 f)(x) = D^2 f(x)(v) \in \mathrm{L}(X,Y),$$

as $D\delta_v = \delta_v$. Thus

$$D_w(D_v f)(x_0) = D^2 f(x)(v)(w) \in Y.$$

∎

Note that $(v,w) \mapsto D^2 f(x)(v)(w)$ is bilinear. Thus the second differential of f at x can be thought of as a bilinear function $D^2 f(x) \colon X \times X \to Y$, and henceforth we may write $D^2 f(x)(v,w)$ to economize on parentheses. In fact, $D^2 f(x)$ is a symmetric bilinear function. To see this, we start with the following lemma, which is of some independent interest.

Define the **second difference function** $\Delta^2_{v,w} f$ of f by (v,w) by

$$\Delta^2_{v,w} f(x) = \Delta_w\big(\Delta_v f\big)(x) = \Delta_v f(x+w) - \Delta_v f(x) = f(x+w+v) - f(x+w) - \big(f(x+v) - f(x)\big).$$

Notice that this is symmetric in v and w.

The second difference is the discrete analogue of the second differential.

100 Lemma (Second difference and second differential) *If f is differentiable at each point of an open set U, and twice differentiable at $x_0 \in U$, then the second differential $D^2 f(x_0)$ approximates the second difference $\Delta^2 f(x_0)$ in the following sense. For every $\varepsilon > 0$ there is some $\delta > 0$ such that $\|v\|, \|w\| < \delta$ implies*

$$\|\Delta^2_{v,w} f(x_0) - D^2 f(x_0)(v,w)\| \leqslant \varepsilon \|v\| \cdot \|w\|.$$

Note that this implies that we may fix either v or w and let the other tend to zero, and the difference above goes to zero.

Proof: (Loomis and Sternberg [100, p. 188]) Pick $\varepsilon > 0$. Since the mapping $x \mapsto Df(x)$ is differentiable at x_0, we see that there exists some $\delta > 0$, such that if $\|v\| < \delta$, then

$$\|Df(x_0+v)(\cdot) - Df(x_0)(\cdot) - D^2 f(x_0)(v,\cdot)\| \leqslant \varepsilon \|v\|.$$

Now fix u with $\|u\| < \frac{\delta}{3}$. Then for any v with $\|v\| < \frac{2\delta}{3}$, we $\|v+u\| < \delta$ have

$$\|Df(x_0+v+u)(\cdot) - Df(x_0)(\cdot) - D^2 f(x_0)(v+u,\cdot)\| \leqslant \varepsilon \|v+u\|.$$

Thus the linear transformations $Df(x_0+v) - D^2 f(x_0)(v)$ and $Df(x_0+v+u) - D^2 f(x_0)(v+u)$ are both close to $Df(x_0)$ in the operator norm, and hence close to each other. Indeed,

$$\big\|\big(Df(x_0+v) - D^2 f(x_0)(v)\big) - \big(Df(x_0+v+u) - D^2 f(x_0)(v+u)\big)\big\| \leqslant \varepsilon\big(\|v\| + \|v+u\|\big)$$

whenever $\|v\| < \frac{2\delta}{3}$ (since $\|u\| < \frac{\delta}{3}$). Since $D^2 f(x_0)$ is bilinear this reduces to

$$\big\|Df(x_0+v) - Df(x_0+v+u) - D^2 f(x_0)(-u)\big\| \leqslant \varepsilon\big(\|v\| + \|v+u\|\big) \leqslant 2\varepsilon\big(\|v\| + \|u\|\big). \quad (3.1)$$

Now the term $Df(x_0+v) - Df(x_0+v+u)$ in the expression above is the differential of some function g. Namely,

$$g(x) = f(x) - f(x+u) = -\Delta_u f(x).$$

We can thus write (3.1) as

$$\big\|Dg(x_0+v) - D^2f(x_0)(-u)\big\| \leqslant 2\varepsilon\big(\|v\| + \|u\|\big)$$

provided $\|v\| < \frac{2\delta}{3}$. The important thing to note is that this holds for all v sufficiently small for some fixed linear transformation $D^2f(x_0)(-u)\colon X \to Y$. This is exactly the situation dealt with by Theorem 98.

So consider the ball B of radius $\frac{2\delta}{3}$ around x_0. This is a convex set, so by Theorem 98, where the role of $T(\cdot)$ is played by $D^2f(x_0)(-u,\cdot)$, we see that (3.1) implies

$$\|\Delta_w g(x_0+v) - D^2f(x_0)(-u,w)\| \leqslant 2\varepsilon\big(\|v\| + \|u\|\big)\|w\| \tag{3.2}$$

whenever x_0+v and x_0+v+w belong to B. That is, whenever $\|v\| < \frac{2\delta}{3}$ and $\|v+w\| < \frac{2\delta}{3}$. This will certainly be satisfied if $\|v\| < \frac{\delta}{3}$ and $\|w\| < \frac{\delta}{3}$.

Since u was an arbitrary vector with $\|u\| < \frac{\delta}{3}$, (3.2) holds whenever $\|u\|, \|v\|, \|w\| < \frac{\delta}{3}$. In particular, it holds for $u = -v$. In this case,

$$\begin{aligned}
\Delta_w g(x_0+v) &= g(x_0+v+w) - g(x_0+v)\\
&= f(x_0+v+w) - f(x_0+v+w+u) - \big(f(x_0+v) - f(x_0+v+u)\big)\\
&= f(x_0+w+v) - f(x_0+w) - \big(f(x_0+v) - f(x_0)\big)\\
&= \Delta_v f(x_0+w) - \Delta_v f(x_0)\\
&= \Delta^2_{v,w} f(x_0).
\end{aligned}$$

Thus for all v, w with $\|v\|, \|w\| < \frac{\delta}{3}$, (3.2) (with $u = -v$) implies

$$\|\Delta^2_{v,w} f(x_0) - D^2f(x_0)(v,w)\| \leqslant 4\varepsilon\|v\| \cdot \|w\|.$$

This completes the proof. ∎

101 Corollary (Symmetry of the second differential) *If f is differentiable at each point of an open set U, and twice differentiable at $x \in U$, then the second differential $D^2f(x)$ is* ***symmetric****. That is, for all v, w*

$$D^2f(x)(v,w) = D^2f(x)(w,v).$$

Consequently the mixed directional derivatives satisfy

$$D_w(D_v f)(x) = D_v(D_w f)(x).$$

Proof: This follows readily from Lemma 100 and the fact that $\Delta_{v,w} f(x) = \Delta_{w,v} f(x)$, but here are the details.

Let $\varepsilon > 0$ be given. Applying Lemma 100 twice we see that there exist $\delta_1 > 0$ and $\delta_2 > 0$ such that for all v, w with $\|v\|, \|w\| < \delta_1$, we have

$$\|\Delta^2_{v,w} f(x_0) - D^2f(x_0)(v,w)\| \leqslant \frac{\varepsilon}{2}\|v\| \cdot \|w\|,$$

and for all v, w with $\|v\|, \|w\| < \delta_2$, we have

$$\|\Delta^2_{w,v} f(x_0) - D^2f(x_0)(w,v)\| \leqslant \frac{\varepsilon}{2}\|v\| \cdot \|w\|.$$

Setting $\delta = \min \delta_1, \delta_2$, and using the symmetry of the second difference function we see that for all v, w with $\|v\|, \|w\| < \delta$, we have

$$\|D^2f(x_0)(v,w) - D^2f(x_0)(w,v)\| \leqslant \varepsilon\|v\| \cdot \|w\|. \tag{3.3}$$

But homogeneity of the norm and the second differential imply that this must hold for all v and w regardless of norm. (For given any v and w, let the (possibly small) numbers $\alpha > 0$ and $\lambda > 0$ satisfy $\|\alpha v\|, \|\lambda w\| < \delta$. Then (3.3) implies

$$\|D^2 f(x_0)(\alpha v, \lambda w) - D^2 f(x_0)(\lambda w, \alpha v)\| \leqslant \frac{\varepsilon}{2}\|\alpha v\| \cdot \|\lambda w\|.$$

But by bilinearity $D^2 f(x_0)(\alpha v, \lambda w) = \alpha\lambda D^2 f(x_0)(v, w)$, so we have

$$\alpha\lambda\|D^2 f(x_0)(v, w) - D^2 f(x_0)(w, v)\| \leqslant \frac{\varepsilon}{2}\alpha\lambda\|v\| \cdot \|w\|,$$

and dividing by $\alpha\lambda > 0$ gives the desired result.)

Since ε is an arbitrary positive number, we must have $D^2 f(x_0)(v, w) = D^2 f(x_0)(w, v)$. ∎

3.10 Higher yet differentials

We have already introduced the first difference $\Delta_v f(x)$ and the second difference $\Delta^2_{v,w} f(x)$ defined by

$$\Delta_v f(x) = f(x + v) - f(x)$$

and

$$\begin{aligned}
\Delta^2_{v,w} f(x) &= \Delta_w \Delta_v f(x) \\
&= \Delta_v f(x + w) - \Delta_v f(x) \\
&= \big(f(x + v + w) - f(x + w)\big) - \big(f(x + v) - f(x)\big) \\
&= f(x + v + w) - f(x + w) - f(x + v) + f(x).
\end{aligned}$$

More generally we can define higher order differences inductively by

$$\Delta^n_{v_1,\dots,v_n} f(x) = \Delta_{v_n} \Delta^{n-1}_{v_1,\dots,v_{n-1}} f(x) = \Delta^{n-1}_{v_1,\dots,v_{n-1}} f(x + v_n) - \Delta^{n-1}_{v_1,\dots,v_{n-1}} f(x).$$

Out of pure laziness, I leave it to you as an exercise in induction to show that $\Delta^n_{v_1,\dots,v_n} f(x)$ is a symmetric function of $v_1, \dots, v_n$, and that

$$\Delta^n_{v,\dots,v} f(x) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} f(x + kv).$$

Finish this.

Higher differentials are also defined inductively. Suppose U is an open subset of X and $f\colon U \to Y$ is $n-1$ times differentiable at each point of U. More generally, the n^{th}-order differential of f at x, if it exists, can be identified with a multilinear function $D^n f(x)\colon \underbrace{X \times \cdots \times X}_{n \text{ copies}} \to Y$.

3.11 Matrix representations and partial derivatives

For Euclidean spaces the differentials have representations in terms of partial derivatives. I will follow Spivak [145] and denote the matrix representation of $f(x)$ by $f'(x)$. I will also follow the relatively common usage of denoting the j^{th} partial derivative of the real function f at the point x by $D_j f(x)$. Also following Spivak, I denote the second order **mixed partial derivative** $D_j\big(D_i f(x)\big)$ by $D_{i,j} f(x)$. Note the reversal of the order of i and j. (If f is differentiable, then $D_{i,j} f(x) = D_{j,i} f(x)$, so it hardly matters.)

We have the following special cases. If $f\colon \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}$, the differential of f corresponds to the **gradient vector** $f'(x) = \big(D_1 f(x), \dots, D_n f(x)\big)$, via

$$Df(x)(v) = f'(x) \cdot v = \sum_{j=1}^{n} D_j f(x) v_j.$$

The second differential of f corresponds to the $n \times n$ **Hessian matrix**[6]

$$f''(x) = \begin{bmatrix} D_{1,1} f(x) & \cdots & D_{1,n} f(x) \\ \vdots & & \vdots \\ D_{n,1} f(x) & \cdots & D_{n,n} f(x) \end{bmatrix}$$

via

$$D^2 f(x)(v, w) = w \cdot f''(x) v = \sum_{i=1}^{n} \sum_{j=1}^{n} D_{i,j} f(x) w_i v_j.$$

If $f\colon \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}^{\mathrm{m}}$, the differential of $Df(x)$ corresponds to the $m \times n$ **Jacobian matrix**

$$f'(x) = \begin{bmatrix} D_1 f^1(x) & \cdots & D_n f^1(x) \\ \vdots & & \vdots \\ D_1 f^m(x) & \cdots & D_n f^m(x) \end{bmatrix},$$

via

$$Df(x)(v) = f'(x)v = \begin{bmatrix} D_1 f^1(x) & \cdots & D_n f^1(x) \\ \vdots & & \vdots \\ D_1 f^m(x) & \cdots & D_n f^m(x) \end{bmatrix} v = \begin{bmatrix} \sum_{j=1}^{n} D_j f^1(x) v_j \\ \vdots \\ \sum_{j=1}^{n} D_j f^m(x) v_j \end{bmatrix} = \begin{bmatrix} f^{1\prime}(x) \cdot v \\ \vdots \\ f^{n\prime}(x) \cdot v \end{bmatrix}.$$

3.11.1 A word on notation

In the three hundred or so years since the calculus was invented in the late seventeenth century, many different notations have been developed for differentials and partial derivatives. Apostol [7, § 4.8, pp. 171–172] describes some of the history, and Spivak [145, pp. 44-45] discusses the advantages and disadvantages of each. I have never been very consistent in my notation, but I will try to be better in these notes, unless there are good reasons for using a variant notation. Table 3.1 is a guide to the various usages in some of my sources.

While we're on the subject of notation, if you try to read Loomis and Sternberg [100], they use $\mathrm{Hom}\,(X, Y)$ for $\mathsf{L}(X, Y)$, the vector space of all continuous linear transformations from X into Y when X and Y are normed spaces [100, p. 129], and also to denote the vector space of all linear transformations from X into Y when X and Y are not normed spaces [100, p. 45].

3.12 Chain rule revisited

102 Proposition (Chain rule for second differentials) *Let X, Y, and Z be normed vector spaces, and let U be an open subset of X and V be an open subset of Y. Let $f\colon U \to V$, and $g\colon V \to Z$. If f is twice differentiable at x and g is twice differentiable at $f(x)$, then $g \circ f$ is twice differentiable at x and*

$$D^2(g \circ f)(x)(v, w) = D^2 g\big(f(x)\big)\big(Df(x)(v), Df(x)(w)\big) + Dg(x)\big(D^2 f(x)(v, w)\big).$$

[6] Note well that Marsden [107, p. 184] defines the Hessian matrix to be the *negative* of my definition. He asserts that "the minus sign is purely conventional and is not of essential importance. The reader who is so inclined can make the appropriate changes in the text…" One place where such a change needs to be made is in the definition of the index of a critcal point [107, p. 223]. My usage follows Apostol [8].

Author	Differential	Directional derivative	Partial derivative	Gradient vector	Jacobian matrix	Second derivative	Mixed partials	Hessian Matrix
KCB	$Df(x)(v)$	$D_v f(x)$	$D_j f(x)$	$f'(x)$	$f'(x)$	$D^2 f(x)$	$D_{i,j} f(x)$	$f''(x)$
Apostol [6]	$df(x;v)$	$D_v f(x)$	$D_j f(x)$	$\nabla f(x)$	$J_f(x)$		$D_{j,i} f(x)$	
Apostol [7, 8]	$f'(x)(v)$	$f'(x,v)$	$D_j f(x)$	$\nabla f(x)$	$Df(x)$		$D_{ji} f(x)$	$H(x)$
			$\frac{\partial f}{\partial x_j}(x)$					
Dieudonné [43]	$f'(x) \cdot v$		$D_j f(x)$		$(D_j f_i)$	$D^2 f(x)$	$D_j\big(D_i f(x)\big)$	
	$Df(x)$					$f''(x)$		
Loomis and Sternberg [100]	$df_x(v)$	$D_v(x)$	df_x^j			$d^2 f_x$		
			$\frac{\partial f}{\partial x_j}(x)$					
Marsden [107]	$Df(x)$		$\frac{\partial f}{\partial x_j}(x)$	grad f		$D^2 f(x)$	$\frac{\partial^2 f}{\partial x_j \partial x_i}(x)$	
				$\nabla f(x)$				
Rudin [134]	$f'(x)v$	$(D_v f)(x)$	$(D_j f)(x)$	$\nabla f(x)$			$D_{ji} f(x)$	
			$D_j f(x)$					
Spivak [145]	$Df(x)(v)$	$D_v f(x)$	$D_j f(x)$	$f'(x)$	$f'(x)$		$D_{i,j} f(x)$	

Table 3.1. A guide to notation.

Proof: Define $\psi\colon X \to \mathsf{L}(Y,Z) \times \mathsf{L}(X,Y)$ by

$$\psi(x) = \big(Dg\big(f(x)\big), Df(x)\big).$$

Then the Chain Rule 92 can be restated as

$$D(g \circ f) = c \circ \psi,$$

where c is the composition operator. Applying the chain rule to this equation, we get

$$D^2(g \circ f)(x) = Dc\big(\psi(x)\big) \circ D\psi(x).$$

Thus we may write

$$\begin{aligned}
D^2(g \circ f)(x)(v) &= Dc\big(Dg\big(f(x)\big), Df(x)\big)\Big(\big(D^2g\big(f(x)\big) \circ Df(x)\big)(v), D^2f(x)(v)\Big) \\
&= c\Big(Dg\big(f(x)\big), D^2f(x)(v)\Big) + c\Big(\big(D^2g\big(f(x)\big) \circ Df(x)\big)(v), Df(x)\Big) \\
&= \underbrace{\underbrace{Dg\big(f(x)\big)}_{\in \mathsf{L}(Y,Z)}\Big(\underbrace{D^2f(x)(v)}_{\in \mathsf{L}(X,Y)}\Big)}_{\in \mathsf{L}(X,Z)} + \underbrace{\Big(\underbrace{D^2g\big(f(x)\big)}_{\in \mathsf{L}\left(Y,\mathsf{L}(Y,Z)\right)}\Big)\Big(\underbrace{Df(x)(v)\big)}_{\in Y}\Big) \circ \underbrace{Df(x)}_{\in \mathsf{L}(X,Y)}\big)}_{\in \mathsf{L}(X,Z)},
\end{aligned}$$

where the second equality is just the differential of the composition operator. Simplifying gives

$$D^2(g \circ f)(x)(v,w) = Dg\big(f(x)\big)\big(D^2f(x)(v,w)\big) + D^2g\big(f(x)\big)\big(Df(x)(v), Df(x)(w)\big).$$

∎

The following examples of the chain rule are quite useful.

103 Lemma *Let U be an open convex subset of $\boldsymbol{R}^{\mathrm{n}}$ and let $f\colon U \to \boldsymbol{R}$ be continuous. Fix $x \in U$ and choose v so that $x \pm v \in U$. Define the function $g\colon [-1,1] \to \boldsymbol{R}$ by*

$$g(\lambda) = f(x + \lambda v).$$

If f is differentiable at x, then g is differentiable at 0 and

$$g'(0) = f'(x) \cdot v = \sum_{j=1}^{n} D_j f(x) v_j.$$

If f is twice differentiable at x, then g is twice differentiable at 0, and

$$g''(0) = D^2f(x)(v,v) = \sum_{i=1}^{n}\sum_{j=1}^{n} D_{i,j} f(x) v_i v_j = v \cdot f''(x) v.$$

A **path** in a vector space X is a function p from an interval of $\boldsymbol{R}$ into X, usually assumed to be continuous. The next result generalizes the previous.

104 Lemma (Second differential on a path) *Let U be an open convex subset of $\boldsymbol{R}^{\mathrm{n}}$ and let $f\colon U \to \boldsymbol{R}$ be continuous. Fix $x \in U$ and let $p\colon (-\varepsilon, \varepsilon) \to U$ be a path with $p(0) = x$. Define the function $g\colon [-1,1] \to \boldsymbol{R}$ by*

$$g(\lambda) = f\big(p(\lambda)\big).$$

If f is differentiable at x and p is differentiable at 0, then g is differentiable at 0 and

$$g'(0) = f'(x) \cdot p'(0).$$

If f is twice differentiable at x and p is twice differentiable at 0, then g is twice differentiable at 0, and

$$\begin{aligned} g''(0) &= p'(0) \cdot f''(x)p'(0) + f'(x) \cdot p''(0) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} D_{i,j}f(x)p_i'(0)p_j'(0) + \sum_{j=1}^{n} D_j f(x)p_j''(0). \end{aligned}$$

3.13 Taylor's Theorem

There are multivariable versions of Taylor's Theorem that correspond to Theorems 73 and 74. The first one is gotten from Theorems 73 by looking at the single variable function $h(t) = f(x + tv)$.

105 Theorem (Multivariate Taylor's Theorem) **********

I will state and prove Young's form of the multivariate Taylor Theorem for the case of second differentials, since this is the most important for what follows, and the general case is essentially the same, but notationally challenging. For more details see Loomis and Sternberg [100, p. 190ff].

106 Theorem (Young's Form of Multivariate Taylor's Theorem, $n = 2$) *Let f be a continuously differentiable real-valued function on an open subset U of $\boldsymbol{R}^{\mathrm{n}}$. Let x belong to U and assume that f is twice differentiable at x.*

Then for every v such that $x + v \in U$,

$$f(x + v) = f(x) + Df(x)(v) + \tfrac{1}{2}D^2f(x)(v, v) + \tfrac{r(v)}{2}\|v\|^2,$$

where $\lim_{v\to 0} r(v) = 0$.

Proof: To ease notation, let Q denote the quadratic form defined by $D^2f(x)$, that is,

$$Q(v) = D^2f(x)(v, v).$$

Since f is twice differentiable at x, applying the definition of differential to the first differential Df, we see that for every $\varepsilon > 0$, there exists some $\delta > 0$, such that if $\|v\| < \delta$, then

$$\|Df(x + v)(\cdot) - Df(x)(\cdot) - D^2f(x)(v, \cdot)\| \leqslant \varepsilon\|v\|.$$

Replacing v by λv for $0 \leqslant \lambda \leqslant 1$ lets us rewrite this as

$$\|Df(x + \lambda v)(\cdot) - Df(x)(\cdot) - \lambda D^2f(x)(v, \cdot)\| \leqslant \varepsilon\lambda\|v\|.$$

Now evaluating at y, and recalling that for any $p, y \in \boldsymbol{R}^{\mathrm{n}}$, we have $|p \cdot y| \leqslant \|p\| \cdot \|y\|$, so

$$\big|Df(x + \lambda v)(y) - Df(x)(y) - \lambda D^2f(x)(v, y)\big| \leqslant \varepsilon\lambda\|v\| \cdot \|y\|.$$

Setting $y = v$, we get

$$\big|Df(x + \lambda v)(v) - Df(x)(v) - \lambda Q(v)\big| \leqslant \varepsilon\lambda\|v\|^2,$$

or in other terms,

$$\lambda\big(Q(v) - \varepsilon\|v\|^2\big) \leqslant Df(x + \lambda v)(v) - Df(x)(v) \leqslant \lambda\big(Q(v) + \varepsilon\|v\|^2\big), \qquad (\star)$$

for $\|v\| < \delta$ and $0 \leqslant \lambda \leqslant 1$. Now define $h(\lambda) = f(x + \lambda v)$, so that $h'(\lambda) = Df(x + \lambda v)(v)$. Observe that h' is continuous, so by the Second Fundamental Theorem of Calculus 82,

$$f(x+v) - f(x) = h(1) - h(0) = \int_0^1 h'(\lambda)\, d\lambda = \int_0^1 Df(x+\lambda v)(v)\, d\lambda.$$

Thus integrating each term of $(\star)$ with respect to λ over the interval $[0,1]$ gives

$$\tfrac{1}{2}\big(Q(v) - \varepsilon\|v\|^2\big) \leqslant f(x+v) - f(x) - Df(x)(v) \leqslant \tfrac{1}{2}\big(Q(v) + \varepsilon\|v\|^2\big) \qquad (\star\star)$$

for all $\|v\| \leqslant \delta$. Now rearrange $(\star\star)$ above to produce the following version of Taylor's formula:

$$f(x+v) = f(x) + Df(x)(v) + \tfrac{1}{2}D^2 f(x)(v,v) + \tfrac{r(v)}{2}\|v\|^2,$$

where $|r(v)| \leqslant \varepsilon$, which implies $\lim_{v\to 0} r(v) = 0$. ∎

3.14 Extrema of functions of several variables

107 Theorem (Necessary First Order Conditions) *If U is an open subset of a normed space, and $x^* \in U$ is a local extremum of f, and f has directional derivatives at x^*, then for any nonzero v, the directional derivative satisfies $D_v f(x^*) = 0$. In particular, if f is differentiable at x^*, then $Df(x^*) = 0$.*

Proof: If f has a local maximum at x^*, then $f(x^*) - f(x) \geqslant 0$ for every x in some neighborhood of x^*. Fix some nonzero v. Since x^* is an interior point of U, there is an $\varepsilon > 0$ such that $x + \lambda v \in U$ for any $\lambda \in (-\varepsilon, \varepsilon)$.

$$\frac{f(x^*+\lambda v) - f(x^*)}{\lambda} \quad \text{is} \quad \begin{cases} \leqslant 0 & \text{for} \quad \lambda > 0 \\ \geqslant 0 & \text{for} \quad \lambda < 0. \end{cases}$$

Therefore $D_v f(x^*) = \lim_{\lambda\to 0} \frac{f(x^*+\lambda v)-f(x^*)}{\lambda} = 0$, since the limit exists. A similar argument applies if x^* is a local minimizer. ∎

It is also possible to derive first order conditions by reducing the multidimensional case to the one-dimensional case by means of the chain rule.

Proof using one-dimensional case: Since x^* is an interior point of U, there is an $\varepsilon > 0$ such that $x^* + \lambda v \in U$ for any $\lambda \in (-\varepsilon, \varepsilon)$ and any $v \in \boldsymbol{R}^{\mathrm{n}}$ with $\|v\| = 1$. Set $g_v(\lambda) = f(x^* + \lambda v)$. Then g_v has an extremum at $\lambda = 0$. Therefore $g_v'(0) = 0$. By the chain rule, $g_v'(\lambda) = Df(x^* + \lambda v)(v)$. Thus we see that $Df(x^*)(v) = 0$ for every v, so $Df(x^*) = 0$. ∎

It might seem possible to derive sufficient second order conditions using the same method. The problem is that looking only along segments of the form $x + \lambda v$ can be misleading, as the next example shows.

108 Example (Maximum on lines is not a maximum) This is an example of a function that achieves a maximum at zero along every line through the origin, yet nevertheless does not achieve a local maximum at zero. Hancock [71, pp. 31–32] attributes this example to Peano, as generalized by Goursat [63, vol. 1, p. 108]. It also appears in Apostol [6, p. 149] and Sydsaeter [146]. Define $f\colon \boldsymbol{R}^2 \to \boldsymbol{R}$ by

$$f(x,y) = -(y - x^2)(y - 2x^2).$$

This function is zero along the parabolas $y = x^2$ and $y = 2x^2$. For $y > 2x^2$ or $y < x^2$ we have $f(x,y) < 0$. For $x^2 < y < 2x^2$ we have $f(x,y) > 0$. See Figure 3.5. Note that $f(0,0) = 0$

and $f'(0,0) = (0,0)$. Now in every neighborhood of the origin f assumes both positive and negative values, so it is neither a local maximizer nor a local minimizer. On the other hand, for every straight line through the origin there is a neighborhood of zero for which $f(x,y) < 0$ if $(x,y) \neq (0,0)$ on that line. Thus zero is a strict local maximizer along any straight line through the origin. □

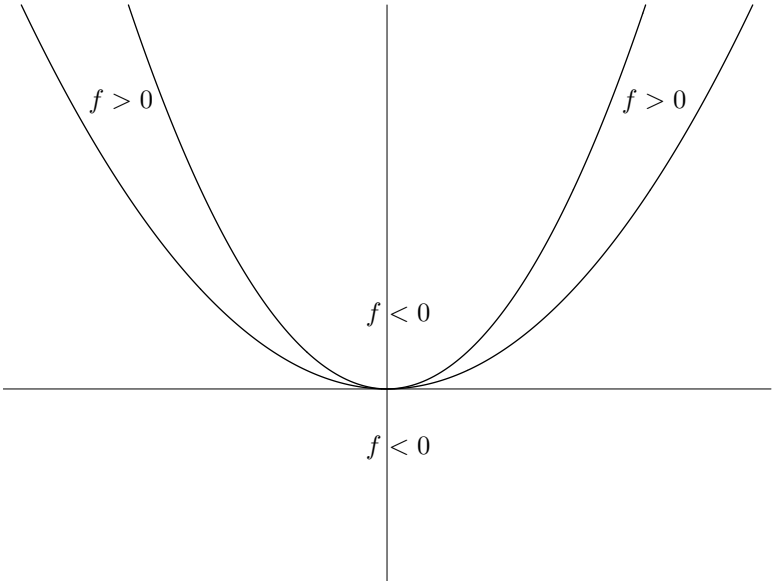

Figure 3.5. $f(x,y) = -(y - x^2)(y - 2x^2)$.

Even considering higher order polynomials than straight lines is not sufficient, as the next example, due to Hancock [71, p. 36] shows.

109 Example This is a modification of Example 108. This function achieves a maximum at zero along polynomial curve through the origin, yet nevertheless does not achieve a local maximum at zero. Define $f\colon \boldsymbol{R}^2 \to \boldsymbol{R}$ by

$$f(x,y) = -\big(y - \sin^2(x)\big)\big(y - \sin^2(x) - e^{\frac{-1}{x^2}}\big),$$

where we adopt the convention that $e^{\frac{-1}{x^2}} = 0$ when $x = 0$. (The function is continuous and differentiable this way.) This function is zero along the curves $y = \sin(x)^2$ and $y = \sin(x)^2 + e^{\frac{-1}{x^2}}$. Unfortunately a picture cannot help here, since near the origin the $e^{\frac{-1}{x^2}}$ term is smaller than the width of the lines used to draw one. As in Example 108, for $y > \sin(x)^2 + e^{\frac{-1}{x^2}}$ or $y < \sin(x)^2$ we have $f(x,y) < 0$. For $\sin(x)^2 < y < \sin(x)^2 + e^{\frac{-1}{x^2}}$ we have $f(x,y) > 0$. Therefore every neighborhood of $(0,0)$ has contains both positive and negative values of f, so again zero is neither a local maximizer nor a local minimizer of f. Note that $f(0,0) = 0$ and $f'(0,0) = (0,0)$.

Needs work.

********** □

Nevertheless, we can still derive sufficient second order conditions. The following theorem is translated from Loomis and Sternberg [100, Theorem 16.4, p. 190]. It roughly corresponds to Theorem 78.

110 Theorem (Sufficient second order conditions) *Let f be a continuously differentiable real-valued function on an open subset U of $\boldsymbol{R}^{\mathrm{n}}$. Let x^* belong to U and assume that $Df(x^*) = 0$ and that f is twice differentiable at x^*.*

Define critical point somewhere. What about the infinite-dimensional case?

Use $D^2 f(x)$.

If the Hessian matrix $f''(x^*)$ is positive definite, then x^* is a strict local minimizer of f.

If the Hessian matrix $f''(x^*)$ is negative definite, then x^* is a strict local maximizer of f.

If the Hessian is nonsingular but indefinite, then x^* is neither a local maximum, nor a local minimum.

Figure out your notation.

Proof: By Young's form of Taylor's Theorem for many variables 106, recalling that $Df(x^*) = 0$, we have

$$f(x^* + v) = f(x^*) + \tfrac{1}{2}D^2 f(x^*)(v, v) + \tfrac{r(v)}{2}\|v\|^2,$$

where $\lim_{v\to 0} r(v) = 0$. What this tells us is that the increment $f(x^* + v) - f(x^*)$ is bounded between two quadratic forms that can be made arbitrarily close to $Q(v) = D^2 f(x^*)(v, v)$. This is the source of conclusions.

The quadratic form Q achieves its maximum M and minimum m values on the unit sphere (and they are the maximal and minimal eigenvalues, see Proposition 304). If Q is positive definite, then $0 < m \leqslant M$, and homogeneity of degree 2 implies that $m\|v\|^2 \leqslant Q(v) \leqslant M\|v\|^2$ for all v. Choose $0 < \varepsilon < m$. Then there exist $\delta > 0$ such that $\|v\| < \delta$ implies $|r(v)| < \varepsilon$. The first inequality in $(\star\star)$ thus implies

$$0 < \frac{m - \varepsilon}{2}\|v\|^2 \leqslant f(x^* + v) - f(x^*),$$

for $\|v\| < \delta$, which shows that x^* is a strict local minimizer. Similarly if Q is negative definite, then x^* is a strict local maximizer. If Q is nonsingular, but neither negative or positive definite, then $\boldsymbol{R}^{\mathrm{n}}$ decomposes into two orthogonal nontrivial subspaces, and is positive definite on one and negative definite on the other. It follows then that x^* is neither a maximizer nor a minimizer. ∎

For one variable, the necessary second order conditions Corollary 79 followed from the sufficient conditions Theorem 78, since if not $f''(x) > 0$, then $f''(x) \leqslant 0$. The multivariable case is not quite so easy, since the negation of positive definiteness is not negative semidefiniteness.

111 Theorem (Necessary second order conditions) *Let f be a continuously differentiable real-valued function on an open subset U of $\boldsymbol{R}^{\mathrm{n}}$ and assume that f is twice differentiable at x^*, and define the quadratic form $Q(v) = D^2 f(x^*)(v, v)$. If x^* is a local maximizer, then Q is negative semidefinite. If x^* is a local minimizer, then Q is positive semidefinite.*

Proof: Assume that x^* is a local maximizer. By Theorem 107, we have $Df(x^*) = 0$, so as in the proof of Theorem 110, we have

$$f(x^* + v) = f(x^*) + \tfrac{1}{2}D^2 f(x^*)(v, v) + \tfrac{r(v)}{2}\|v\|^2,$$

where $\lim_{v\to 0} r(v) = 0$. Suppose v is an eigenvector of Q corresponding to an eigenvalue $\lambda > 0$. Then $Q(\alpha v, \alpha v) = \lambda\alpha^2\|v\|^2$, then again as in the proof of Theorem 110, for α small enough $|r(\alpha v)| < \lambda$, so

$$f(x^* + \alpha v) = f(x^*) + \frac{1}{2}\big(\lambda + r(\alpha v)\big)\alpha^2\|v\|^2 > f(x^*),$$

a contradiction. Thus all the eigenvalues of Q are nonpositive, so by Proposition 306 it is negative semidefinite.

A similar argument (with appropriate sign changes) works when x^* is a local minimizer. ∎

We can also use the chain rule to reduce the problem to the one-dimensional case.

This seems to work for the infinite dimensional case.

Proof using the chain rule: As in the proof of Theorem 107, define $g(\lambda) = f(x^* + \lambda v)$. By Corollary 79 $g''(0) \leqslant 0$. So by Lemma 103, using $Df(x^*) = 0$,

$$g''(0) = D^2 f(x^*)(v, v) \leqslant 0.$$

That is, Q is negative semidefinite. ∎

3.15 Implicit Function Theorems

The Implicit Function Theorem is a basic tool for analyzing extrema of differentiable functions.

112 Definition *An equation of the form*

$$f(x,p) = y \tag{3.4}$$

implicitly defines x as a function of p *on a domain P if there is a function ξ on P for which $f(\xi(p),p) = y$ for all $p \in P$. It is traditional to assume that $y = 0$, but not essential.*

The use of zero in the above equation serves to simplify notation. The condition $f(x,p) = y$ is equivalent to $g(x,p) = 0$ where $g(x,p) = f(x,p) - y$, and this transformation of the problem is common in practice.

The implicit function theorem gives conditions under which it is possible to solve for x as a function of p in the neighborhood of a known solution $(\bar{x},\bar{p})$. There are actually many implicit function theorems. If you make stronger assumptions, you can derive stronger conclusions. In each of the theorems that follows we are given a subset X of $\boldsymbol{R}^{\mathrm{n}}$, a metric space P (of parameters), a function f from $X \times P$ into $\boldsymbol{R}^{\mathrm{n}}$, and a point $(\bar{x},\bar{p})$ in the interior of $X \times P$ such that $D_x f(\bar{x},\bar{p})$ exists and is invertible. Each asserts the existence of neighborhoods U of $\bar{x}$ and W of $\bar{p}$ and a function $\xi\colon W \to U$ such that $f\big(\xi(p),p\big) = f(\bar{x},\bar{p})$ for all $p \in W$. They differ in whether ξ is uniquely defined (in U) and how smooth it is. The following table serves as a guide to the theorems. For ease of reference, each theorem is stated as a standalone result.

Theorem	Hypotheses	Conclusion
All	f is continuous on $X \times P$ $D_x f(\bar{x},\bar{p})$ is invertible	$f\big(\xi(p),p\big) = f\big(\bar{x},\bar{p}\big)$ for all p in W $\xi(\bar{p}) = \bar{x}$
113		ξ is continuous at $\bar{p}$
114	$D_x f$ is continuous on $X \times P$	ξ is unique in U ξ is continuous on W
115	$Df(\bar{x},\bar{p})$ (wrt x,p) exists	ξ is differentiable at $\bar{p}$
116	Df (wrt x,p) exists on $X \times P$ $D_x f$ is continuous on $X \times P$	ξ is unique in U ξ is differentiable on W
117	f is C^k on $X \times P$	ξ is unique in U ξ is C^k on W

The first result is due to Halkin [68, Theorem B].

113 Theorem (Implicit Function Theorem 0) *Let X be a subset of $\boldsymbol{R}^{\mathrm{n}}$, let P be a metric space, and let $f\colon X \times P \to \boldsymbol{R}^{\mathrm{n}}$ be continuous. Suppose the derivative $D_x f$ of f with respect to x exists at a point and that $D_x f(\bar{x},\bar{p})$ is invertible. Let*

$$\bar{y} = f(\bar{x},\bar{p}).$$

Then for any neighborhood U of $\bar{x}$, there is a neighborhood W of $\bar{p}$ and a function $\xi\colon W \to U$ such that:

a. $\xi(\bar{p}) = \bar{x}$.

b. $f\big(\xi(p),p\big) = \bar{y}$ *for all* $p \in W$.

c. ξ *is continuous at the point* $\bar{p}$.

However, it may be that ξ is neither continuous nor uniquely defined on any neighborhood of $\bar{p}$. There are two ways to strengthen the hypotheses and derive a stronger conclusion. One is to assume the derivative with respect to x exists and is continuous on $X \times P$. The other is to make P a subset of a Euclidean space and assume that f has a derivative with respect to (x, p) at the single point $(\bar{x}, \bar{p})$.

Taking the first approach allows us to conclude that the function ξ is uniquely defined and moreover continuous. The following result is Theorem 9.3 in Loomis and Sternberg [100, pp. 230–231].

114 Theorem (Implicit Function Theorem 1a) *Let X be an open subset $\boldsymbol{R}^{\mathrm{n}}$, let P be a metric space, and let $f\colon X \times P \to \boldsymbol{R}^{\mathrm{n}}$ be continuous. Suppose the derivative $D_x f$ of f with respect to x exists at each point (x, p) and is continuous on $X \times P$. Assume that $D_x f(\bar{x}, \bar{p})$ is invertible. Let*

$$\bar{y} = f(\bar{x}, \bar{p}).$$

Then there are neighborhoods $U \subset X$ and $W \subset P$ of $\bar{x}$ and $\bar{p}$, and a function $\xi\colon W \to U$ such that:

a. *$f(\xi(p); p) = \bar{y}$ for all $p \in W$.*

b. *For each $p \in W$, $\xi(p)$ is the unique solution to (3.4) lying in U. In particular, then*

$$\xi(\bar{p}) = \bar{x}.$$

c. *ξ is continuous on W.*

The next result, also due to Halkin [68, Theorem E] takes the second approach. It concludes that ξ is differentiable at a single point. Related results may be found in Hurwicz and Richter [83, Theorem 1], Leach [98, 99], Nijenhuis [120], and Nikaidô [121, Theorem 5.6, p. 81].

115 Theorem (Implicit Function Theorem 1b) *Let X be a subset of $\boldsymbol{R}^{\mathrm{n}}$, let P be an open subset of $\boldsymbol{R}^{\mathrm{m}}$, and let $f\colon X \times P \to \boldsymbol{R}^{\mathrm{n}}$ be continuous. Suppose the derivative Df of f with respect to (x, p) exists at $(\bar{x}, \bar{p})$. Write $Df(\bar{x}, \bar{p}) = (T, S)$, where $T\colon \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}^{\mathrm{n}}$ and $S\colon \boldsymbol{R}^{\mathrm{m}} \to \boldsymbol{R}^{\mathrm{m}}$, so that $Df(\bar{x}, \bar{p})(h, z) = Th + Sz$. Assume T is invertible. Let*

$$\bar{y} = f(\bar{x}, \bar{p}).$$

Then there is a neighborhood W of $\bar{p}$ and a function $\xi\colon W \to X$ satisfying

a. *$\xi(\bar{p}) = \bar{x}$.*

b. *$f\big(\xi(p), p\big) = \bar{y}$ for all $p \in W$.*

c. *ξ is differentiable (hence continuous) at $\bar{p}$, and*

$$D\xi(\bar{p}) = -T^{-1} \circ S.$$

The following result is Theorem 9.4 in Loomis and Sternberg [100, p. 231]. It strengthens the hypotheses of both Theorems 114 and 115. In return we get differentiability of ξ on W.

116 Theorem (Semiclassical Implicit Function Theorem) *Let $X \times P$ be an open subset of $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}^{\mathrm{m}}$, and let $f\colon X \times P \to \boldsymbol{R}^{\mathrm{n}}$ be differentiable. Suppose the derivative $D_x f$ of f with respect to x is continuous on $X \times P$. Assume that $D_x f(\bar{x}, \bar{p})$ is invertible. Let*

$$\bar{y} = f(\bar{x}, \bar{p}).$$

Then there are neighborhoods $U \subset X$ and $W \subset P$ of $\bar{x}$ and $\bar{p}$ on which equation (3.4) uniquely defines x as a function of p. That is, there is a function $\xi\colon W \to U$ such that:

a. $f(\xi(p);p) = \bar{y}$ *for all* $p \in W$.

b. *For each* $p \in W$, $\xi(p)$ *is the unique solution to (3.4) lying in* U. *In particular, then*

$$\xi(\bar{p}) = \bar{x}.$$

c. ξ *is differentiable on* W, *and*

$$\begin{bmatrix} \dfrac{\partial \xi^1}{\partial p_1} & \cdots & \dfrac{\partial \xi^1}{\partial p_m} \\ \vdots & & \vdots \\ \dfrac{\partial \xi^n}{\partial p_1} & \cdots & \dfrac{\partial \xi^n}{\partial p_m} \end{bmatrix} = - \begin{bmatrix} \dfrac{\partial f^1}{\partial x_1} & \cdots & \dfrac{\partial f^1}{\partial x_n} \\ \vdots & & \vdots \\ \dfrac{\partial f^n}{\partial x_1} & \cdots & \dfrac{\partial f^n}{\partial x_n} \end{bmatrix}^{-1} \begin{bmatrix} \dfrac{\partial f^1}{\partial p_1} & \cdots & \dfrac{\partial f^1}{\partial p_m} \\ \vdots & & \vdots \\ \dfrac{\partial f^n}{\partial p_1} & \cdots & \dfrac{\partial f^n}{\partial p_m} \end{bmatrix}.$$

The classical version maybe found, for instance, in Apostol [6, Theorem 7-6, p. 146], Rudin [134, Theorem 9.28, p. 224], or Spivak [145, Theorem 2-12, p. 41]. Some of these have the weaker statement that there is a unique function ξ within the class of continuous functions satisfying both $\xi(\bar{p}) = \bar{x}$ and $f(\xi(p);p) = 0$ for all p. Dieudonné [43, Theorem 10.2.3, p. 272] points out that the C^k case follows from the formula for $D\xi$ and the fact that the mapping from invertible linear transformations to their inverses, $A \mapsto A^{-1}$, is C^∞. (See Marsden [107, Lemma 2, p. 231].)

117 Classical Implicit Function Theorem *Let* $X \times P$ *be an open subset of* $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}^{\mathrm{m}}$, *and let* $f\colon X \times P \to \boldsymbol{R}^{\mathrm{n}}$ *be* C^k, *for* $k \geqslant 1$. *Assume that* $D_x f(\bar{x}, \bar{p})$ *is invertible. Let*

$$\bar{y} = f(\bar{x}, \bar{p}).$$

Then there are neighborhoods $U \subset X$ *and* $W \subset P$ *of* $\bar{x}$ *and* $\bar{p}$ *on which equation (3.4) uniquely defines* x *as a function of* p. *That is, there is a function* $\xi\colon W \to U$ *such that:*

a. $f(\xi(p);p) = \bar{y}$ *for all* $p \in W$.

b. *For each* $p \in W$, $\xi(p)$ *is the unique solution to (3.4) lying in* U. *In particular, then*

$$\xi(\bar{p}) = \bar{x}.$$

c. ξ *is* C^k *on* W, *and*

$$\begin{bmatrix} \dfrac{\partial \xi^1}{\partial p_1} & \cdots & \dfrac{\partial \xi^1}{\partial p_m} \\ \vdots & & \vdots \\ \dfrac{\partial \xi^n}{\partial p_1} & \cdots & \dfrac{\partial \xi^n}{\partial p_m} \end{bmatrix} = - \begin{bmatrix} \dfrac{\partial f^1}{\partial x_1} & \cdots & \dfrac{\partial f^1}{\partial x_n} \\ \vdots & & \vdots \\ \dfrac{\partial f^n}{\partial x_1} & \cdots & \dfrac{\partial f^n}{\partial x_n} \end{bmatrix}^{-1} \begin{bmatrix} \dfrac{\partial f^1}{\partial p_1} & \cdots & \dfrac{\partial f^1}{\partial p_m} \\ \vdots & & \vdots \\ \dfrac{\partial f^n}{\partial p_1} & \cdots & \dfrac{\partial f^n}{\partial p_m} \end{bmatrix}.$$

As a bonus, let me throw in the following result, which is inspired by Apostol [8, Theorem 7.21].

118 Theorem (Lipschitz Implicit Function Theorem) *Let* P *be a compact metric space and let* $f\colon \boldsymbol{R} \times P \to \boldsymbol{R}$ *be continuous and assume that there are real numbers* $0 < m < M$ *such that for each* p

$$m \leqslant \frac{f(x,p) - f(y,p)}{x - y} \leqslant M.$$

Then there is a unique function $\xi\colon P \to \boldsymbol{R}$ *satisfying* $f\big(\xi(p), p\big) = 0$. *Moreover,* ξ *is continuous.*

An interesting extension of this result to Banach spaces and functions with compact range may be found in Warga [159].

3.15.1 Proofs of Implicit Function Theorems

The proofs given here are based on fixed point arguments and are adapted from Halkin [67, 68], Rudin [134, pp. 220–227], Loomis and Sternberg [100, pp. 229–231], Marsden [107, pp. 230–237], and Dieudonné [43, pp. 265—273]. Another sort of proof, which is explicitly finite dimensional, of the classical case may be found in Apostol [6, p. 146] or Spivak [145, p. 41].

The first step is to show that for each p (at least in a neighborhood of $\bar{p}$) there is a zero of the function $f(x,p) - \bar{y}$, where $\bar{y} = f(\bar{x},\bar{p})$. As is often the case, the problem of finding a zero of $f - \bar{y}$ is best converted to the problem of finding a fixed point of some other function. The obvious choice is to find a fixed point of $\pi_X - (f - \bar{y})$ (where $\pi_X(x,p) = x$), but the obvious choice is not clever enough in this case. Let

$$T = D_x f(\bar{x},\bar{p}).$$

Define $\varphi\colon X \times P \to \boldsymbol{R}^{\mathrm{n}}$ by $\varphi = \pi_X - T^{-1}\big(f - \bar{y}\big)$. That is,

$$\varphi(x,p) = x - T^{-1}\big(f(x,p) - \bar{y}\big). \tag{3.5}$$

Note that $\varphi(x,p) = x$ if and only if $T^{-1}\big(f(x,p) - \bar{y}\big) = 0$. But the invertibility of T^{-1} guarantees that this happens if and only if $f(x,p) = \bar{y}$. Thus the problem of finding a zero of $f(\cdot,p) - \bar{y}$ is equivalent to that of finding a fixed point of $\varphi(\cdot,p)$. Note also that

$$\varphi(\bar{x},\bar{p}) = \bar{x}. \tag{3.6}$$

Observe that φ is continuous and also has a derivative $D_x\varphi$ with respect to x whenever f does. In fact,

$$D_x\varphi(x,p) = I - T^{-1}D_x f(x,p).$$

In particular, at $(\bar{x},\bar{p})$, we get

$$D_x\varphi(\bar{x},\bar{p}) = I - T^{-1}T = 0. \tag{3.7}$$

That is, $D_x\varphi(\bar{x},\bar{p})$ is the zero transformation.

Recall that for a linear transformation A, its operator norm $\|A\|$ is defined by $\|A\| = \sup_{|x|\leqslant 1}|Ax|$, and satisfies $|Ax| \leqslant \|A\| \cdot |x|$ for all x. If A is invertible, then $\|A^{-1}\| > 0$.

Proof of Theorem 113: Let X, P, and $f\colon X \times P \to \boldsymbol{R}^{\mathrm{n}}$ be as in the hypotheses of Theorem 113.

In order to apply a fixed point argument, we must first find a subset of X that is mapped into itself. By the definition of differentiability and (3.7) we can choose $r > 0$ so that

$$\frac{\big|\varphi(x,\bar{p}) - \varphi(\bar{x},\bar{p})\big|}{|x - \bar{x}|} \leqslant \frac{1}{2} \qquad \text{for all } x \in \bar{B}_r(\bar{x}).$$

Noting that $\varphi(\bar{x},\bar{p}) = \bar{x}$ and rearranging, it follows that

$$\big|\varphi(x,\bar{p}) - \bar{x}\big| \leqslant \frac{r}{2} \qquad \text{for all } x \in \bar{B}_r(\bar{x}).$$

For each p set $m(p) = \max_x\big|\varphi(x,p) - \varphi(x,\bar{p})\big|$ as x runs over the compact set $\bar{B}_r(\bar{x})$. Since φ is continuous (and $\bar{B}_r(\bar{x})$ is a fixed set), the Maximum Theorem 62 implies that m is continuous. Since $m(\bar{p}) = 0$, there is some $\varepsilon > 0$ such that $|m(p)| < \frac{r}{2}$ for all $p \in B_\varepsilon(\bar{p})$. That is,

$$\big|\varphi(x,p) - \varphi(x,\bar{p})\big| < \frac{r}{2} \qquad \text{for all } x \in \bar{B}_r(\bar{x}),\ p \in B_\varepsilon(\bar{p}).$$

For each $p \in B_\varepsilon(\bar{p})$, the function φ maps $\bar{B}_r(\bar{x})$ into itself, for

$$\begin{aligned}\big|\varphi(x,p) - \bar{x}\big| &\leqslant \big|\varphi(x,p) - \varphi(x,\bar{p})\big| + \big|\varphi(x,\bar{p}) - \bar{x}\big| \\ &< \frac{r}{2} + \frac{r}{2} \\ &= r.\end{aligned}$$

That is, $\varphi(x,p) \in \bar{B}_r(\bar{x})$. Since φ is continuous and $\bar{B}_r(\bar{x})$ is compact and convex, by the Brouwer Fixed Point Theorem (e.g., [30, Corollary 6.6, p. 29]), there is some $x \in \bar{B}_r(\bar{x})$ satisfying $\varphi(x,p) = x$, or in other words $f(x,p) = 0$.

We have just proven parts (a) and (b) of Theorem 113. That is, for every neighborhood X of $\bar{x}$, there is a neighborhood $W = \bar{B}_{\varepsilon(r)}(\bar{p})$ of $\bar{p}$ and a function ξ from $\bar{B}_{\varepsilon(r)}(\bar{p})$ into $\bar{B}_r(\bar{x}) \subset X$ satisfying $\xi(\bar{p}) = \bar{x}$ and $f\big(\xi(p),p\big) = 0$ for all $p \in W$. (Halkin actually breaks this part out as Theorem A.)

We can use the above result to construct a ξ that is continuous at $\bar{p}$. Start with a given neighborhood U of $\bar{x}$. Construct a sequence of $r_1 > r_2 > \cdots > 0$ satisfying $\lim r_n = 0$ and for each n consider the neighborhood $U_n = U \cap B_{r_n}(\bar{x})$. From the argument above there is a neighborhood W_n of $\bar{p}$ and a function ξ_n from W_n into $U_n \subset U$ satisfying $\xi_n(\bar{p}) = \bar{x}$ and $f\big(\xi_n(p),p\big) = 0$ for all $p \in W_n$. Without loss of generality we may assume $W_n \supset W_{n+1}$ (otherwise replace W_{n+1} with $W_n \cap W_{n+1}$), so set $W = W_1$. Define $\xi\colon W \to U$ by $\xi(p) = \xi_n(p)$ for $p \in W_n \setminus W_{n+1}$. Then ξ is continuous at $\bar{p}$, satisfies $\xi(\bar{p}) = \bar{x}$, and $f\big(\xi(p),p\big) = 0$ for all $p \in W$. ∎

Note that the above proof used in an essential way the compactness of $\bar{B}_r(\bar{x})$, which relies on the finite dimensionality of $\boldsymbol{R}^{\mathrm{n}}$. The compactness was used first to show that $m(p)$ is finite, and second to apply the Brouwer fixed point theorem.

Theorem 114 adds to the hypotheses of Theorem 113. It assumes that $D_x f$ exists everywhere on $X \times P$ and is continuous. The conclusion is that there are *some* neighborhoods U of $\bar{x}$ and W of $\bar{p}$ and a continuous function $\xi\colon W \to U$ such that $\xi(p)$ is the unique solution to $f\big(\xi(p),p\big) = 0$ lying in U. It is the uniqueness of $\xi(p)$ that puts a restriction on U. If U is too large, say $U = X$, then the solution need not be unique. (On the other hand, it is easy to show, as does Dieudonné [43, pp. 270–271], there is at most one continuous ξ, provided U is connected.) The argument we use here, which resembles that of Loomis and Sternberg, duplicates some of the proof of Theorem 113, but we do not actually need to assume that the domain of f lies in the finite dimensional space $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}^{\mathrm{m}}$, any Banach spaces will do, and the proof need not change. This means that we cannot use Brouwer's theorem, since closed balls are not compact in general Banach spaces. Instead, we will be able to use the Mean Value Theorem and the Contraction Mapping Theorem.

Proof of Theorem 114: Let X, P, and let $f\colon X \times P \to \boldsymbol{R}^{\mathrm{n}}$ obey the hypotheses of Theorem 114. Set $T = D_x f(\bar{x},\bar{p})$, and recall that T is invertible.

Again we must find a suitable subset of X so that each $\varphi(\cdot,p)$ maps this set into itself. Now we use the hypothesis that $D_x f$ (and hence $D_x\varphi$) exists and is continuous on $X \times P$ to deduce that there is a neighborhood $\bar{B}_r(\bar{x}) \times W_1$ of $(\bar{x},\bar{p})$ on which the operator norm $\|D_x\varphi\|$ is strictly less than $\frac{1}{2}$. Set $U = \bar{B}_r(\bar{x})$. Since $\varphi(\bar{x},\bar{p}) = \bar{x}$ and since φ is continuous (as f is), we can now choose W so that $\bar{p} \in W$, $W \subset W_1$, and $p \in W$ implies

$$\big|\varphi(\bar{x},p) - \bar{x}\big| < \frac{r}{2}.$$

We now show that for each $p \in W$, the mapping $x \mapsto \varphi(x,p)$ is a contraction that maps $\bar{B}_r(\bar{x})$ into itself. To see this, note that the Mean Value Theorem (or Taylor's Theorem) implies

$$\varphi(x,p) - \varphi(y,p) = D_x\varphi(z,p)(x-y),$$

for some z lying on the segment between x and y. If x and y lie in $\bar{B}_r(\bar{x})$, then z too must lie in $\bar{B}_r(\bar{x})$, so $\|D_x\varphi(z,p)\| < \frac{1}{2}$. It follows that

$$\big|\varphi(x,p) - \varphi(y,p)\big| < \frac{1}{2}\,|x-y| \qquad \text{for all } x,y \in \bar{B}_r(\bar{x}),\ p \in B_\varepsilon(\bar{p}), \tag{3.8}$$

so $\varphi(\cdot,p)$ is a contraction on $\bar{B}_r(\bar{x})$ with contraction constant $\frac{1}{2}$.

To see that $\bar{B}_r(\bar{x})$ is mapped into itself, let (x,p) belong to $\bar{B}_r(\bar{x}) \times W$ and observe that

$$\begin{aligned} \big|\varphi(x,p) - \bar{x}\big| &\leqslant \big|\varphi(x,p) - \varphi(\bar{x},p)\big| + \big|\varphi(\bar{x},p) - \bar{x}\big| \\ &< \frac{1}{2}\big|x - \bar{x}\big| + \frac{r}{2} \\ &< r. \end{aligned}$$

Thus $\varphi(x,p) \in \bar{B}_r(\bar{x})$.

Since $\bar{B}_r(\bar{x})$ is a closed subset of the complete metric space $\boldsymbol{R}^{\mathrm{n}}$, it is complete itself, so the Contraction Mapping Theorem guarantees that there is a unique fixed point of $\varphi(\cdot,p)$ in $\bar{B}_r(\bar{x})$. In other words, for each $p \in W$ there is a unique point $\xi(p)$ lying in $U = \bar{B}_r(\bar{x})$ satisfying $f(\xi(p),p) = 0$.

It remains to show that ξ must be continuous on W. This follows from a general result on parametric contraction mappings, presented as Lemma 119 below, which also appears in [100, Corollary 4, p. 230]. ∎

Note that the above proof nowhere uses the finite dimensionality of $\boldsymbol{R}^{\mathrm{m}}$, so the theorem actually applies to a general Banach space.

Proof of Theorem 115: For this theorem, in addition to the hypotheses of Theorem 113, we need P to be a subset of a Euclidean space (or more generally a Banach space), so that it makes sense to partially differentiate with respect to p. Now assume f is differentiable with respect to (x,p) at the point $(\bar{x},\bar{p})$.

There is a neighborhood W of $\bar{p}$ and a function $\xi\colon W \to X$ satisfying the conclusions of Theorem 113. It turns out that under the added hypotheses, such a function ξ is differentiable at $\bar{p}$.

We start by showing that ξ is locally Lipschitz continuous at $\bar{p}$. First set

$$\Delta(x,p) = f(x,p) - f(\bar{x},\bar{p}) - T(x - \bar{x}) - S(p - \bar{p}).$$

Since Df exists at $(\bar{x},\bar{p})$, there exists $r > 0$ such that $B_r(\bar{x}) \times B_r(\bar{p}) \subset X \times W$ and if $|x - \bar{x}| < r$ and $|p - \bar{p}| < r$, then

$$\frac{\big|\Delta(x,p)\big|}{|x - \bar{x}| + |p - \bar{p}|} < \frac{1}{2\,\|T^{-1}\|},$$

which in turn implies

$$\big|T^{-1}\Delta(x,p)\big| < \frac{1}{2}|x - \bar{x}| + \frac{1}{2}|p - \bar{p}|.$$

Since ξ is continuous at $\bar{p}$ and $\xi(\bar{p}) = \bar{x}$, there is some $r \geqslant \delta > 0$ such that $|p - \bar{p}| < \delta$ implies $|\xi(p) - \bar{x}| < r$. Thus

$$\big|T^{-1}\Delta\big(\xi(p),p\big)\big| < \frac{1}{2}\big|\xi(p) - \bar{x}\big| + \frac{1}{2}|p - \bar{p}| \qquad \text{for all } p \in B_\delta(\bar{p}). \tag{3.9}$$

But $f\big(\xi(p),p\big) - f(\bar{x},\bar{p}) = 0$ implies

$$\big|T^{-1}\Delta\big(\xi(p),p\big)\big| = \big|\big(\xi(p) - \bar{x}\big) + T^{-1}S(p - \bar{p})\big|. \tag{3.10}$$

Therefore, from the facts that $|a + b| < c$ implies $|a| < |b| + c$, and $\xi(\bar{p}) = \bar{x}$, equations (3.9) and (3.10) imply

$$\big|\xi(p) - \xi(\bar{p})\big| < \big|T^{-1}S(p - \bar{p})\big| + \frac{1}{2}\big|\xi(p) - \xi(\bar{p})\big| + \frac{1}{2}|p - \bar{p}| \quad \text{for all } p \in B_\delta(\bar{p})$$

or,

$$\big|\xi(p) - \xi(\bar{p})\big| < \big(2\|T^{-1}S\| + 1\big)|p - \bar{p}| \quad \text{for all } p \in B_\delta(\bar{p}).$$

That is, ξ satisfies a local Lipschitz condition at $\bar{p}$. For future use set $M = 2\|T^{-1}S\| + 1$.

Now we are in a position to prove that $-T^{-1}S$ is the differential of ξ at $\bar{p}$. Let $\varepsilon > 0$ be given. Choose $0 < r < \delta$ so that $|x - \bar{x}| < r$ and $|p - \bar{p}| < r$ implies

$$\frac{\big|\Delta(x,p)\big|}{|x-\bar{x}| + |p-\bar{p}|} < \frac{\varepsilon}{(M+1)\,\|T^{-1}\|},$$

so

$$\big|\big(\xi(p) - \xi(\bar{p})\big) + T^{-1}S(p-\bar{p})\big| = \big|T^{-1}\Delta\big(\xi(p),p\big)\big| < \frac{\varepsilon}{(M+1)}\big(|\xi(p) - \xi(\bar{p})| + |p-\bar{p}|\big) \leqslant \varepsilon|p-\bar{p}|,$$

for $|p - \bar{p}| < r$, which shows that indeed $-T^{-1}S$ is the differential of ξ at $\bar{p}$. ∎

Proof of Theorem 116: ************ ∎

Proof of Theorem 117: ************ ∎

Proof of Theorem 118: Let f satisfy the hypotheses of the theorem. Let $C(P)$ denote the set of continuous real functions on P. Then $C(P)$ is complete under the uniform norm metric, $\|f - g\| = \sup_p |f(p) - g(p)|$ [3, Lemma 3.97, p. 124]. For each p define the function $\psi_p \colon \boldsymbol{R} \to \boldsymbol{R}$ by

$$\psi_p(x) = (x) - \frac{1}{M} f(x,p).$$

Note that $\psi_p(x) = x$ if and only if $f(x,p) = 0$. If ψ_p has a unique fixed point, then we shall have shown that there is a unique function ξ satisfying $f\big(\xi(p),p\big) = 0$. It suffices to show that ψ_p is a contraction.

To see this, write

$$\begin{aligned}\psi_p(x) - \psi_p(y) &= x - y - \frac{f(x,p) - f(y,p)}{M} \\ &= \left(1 - \frac{1}{M}\frac{f(x,p) - f(y,p)}{x-y}\right)(x-y).\end{aligned}$$

By hypothesis

$$0 < m \leqslant \frac{f(x,p) - f(y,p)}{x-y} \leqslant M,$$

so

$$|\psi_p(x) - \psi_p(y)| \leqslant \big(1 - \frac{m}{M}\big)|x-y|.$$

This shows that ψ_p is a contraction with constant $1 - \frac{m}{M} < 1$.

To see that ξ is actually continuous, define the function $\psi \colon C(P) \to C(P)$ via

$$\psi g(p) = g(p) - \frac{1}{M} f\big(g(p),p\big).$$

(Since f is continuous, ψg is continuous whenever g is continuous.) The pointwise argument above is independent of p, so it also shows that $|\psi g(p) - \psi h(p)| \leqslant \big(1 - \frac{m}{M}\big)|g(p) - h(p)|$ for any functions g and h. Thus

$$\|\psi g - \psi h\| \leqslant \big(1 - \frac{m}{M}\big)\|g - h\|.$$

In other words ψ is a contraction on $C(P)$, so it has a unique fixed point $\bar{g}$ in $C(P)$, so $\bar{g}$ is continuous. But $\bar{g}$ also satisfies $f\big(\bar{g}(p),p\big)$, but since $\xi(p)$ is unique we have $\xi = \bar{g}$ is continuous. ∎

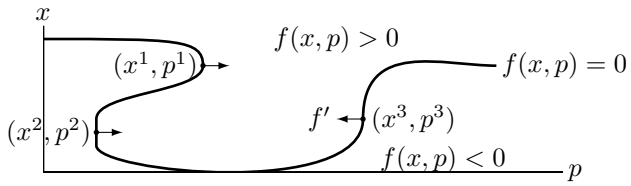

Figure 3.6. Looking for implicit functions.

119 Lemma (Continuity of fixed points) *Let $\varphi\colon X \times P \to X$ be continuous in p for each x, where X is a complete metric space under the metric d and P is a metrizable space. Suppose that φ is a uniform contraction in x. That is, there is some $0 \leqslant \alpha < 1$ such that*

$$d\big(\varphi(x,p) - \varphi(y,p)\big) \leqslant \alpha d(x,y)$$

for all x and y in X and all p in P. Then the mapping $\xi\colon P \to X$ from p to the unique fixed point of $\varphi(\cdot,p)$, defined by $\varphi\big(\xi(p),p\big) = \xi(p)$, is continuous.

Proof: Fix a point p in P and let $\varepsilon > 0$ be given. Let ρ be a compatible metric on P and using the continuity of $\varphi(x,\cdot)$ on P, choose $\delta > 0$ so that $\rho(p,q) < \delta$ implies that

$$d\big(\varphi\big(\xi(p),p\big), \varphi\big(\xi(p),q\big)\big) < (1-\alpha)\varepsilon.$$

So if $\rho(p,q) < \delta$, then

$$\begin{aligned} d\big(\xi(p),\xi(q)\big) &= d\big(\varphi\big(\xi(p),p\big), \varphi\big(\xi(q),q\big)\big) \\ &\leqslant d\big(\varphi\big(\xi(p),p\big), \varphi\big(\xi(p),q\big)\big) + d\big(\varphi\big(\xi(p),q\big), \varphi\big(\xi(q),q\big)\big) \\ &< (1-\alpha)\varepsilon + \alpha d\big(\xi(p),\xi(q)\big) \end{aligned}$$

so

$$(1-\alpha)d\big(\xi(p),\xi(q)\big) < (1-\alpha)\varepsilon$$

or

$$d\big(\xi(p),\xi(q)\big) < \varepsilon,$$

which proves that ξ is continuous at p. ∎

3.15.2 Examples

Figure 3.6 illustrates the Implicit Function Theorem for the special case $n = m = 1$, which is the only one I can draw. The figure is drawn sideways since we are looking for x as a function of p. In this case, the requirement that the differential with respect to x be invertible reduces to $\frac{\partial f}{\partial x} \neq 0$. That is, in the diagram the gradient of f may not be horizontal. In the figure, you can see that the points, (x^1,p^1), (x^2,p^2), and (x^3,p^3), the differentials $D_x f$ are zero. At (x^1,p^1) and (x^2,p^2) there is no way to define x as a continuous function of p locally. (Note however, that if we allowed a discontinuous function, we could define x as a function of p in a neighborhood of p^1 or p^2, but not uniquely.) At the point (x^3,p^3), we can uniquely define x as a function of p near p^3, but this function is not differentiable.

Another example of the failure of the conclusion of the Classical Implicit Function Theorem is provided by the function from Example 108.

120 Example (Differential not invertible) Define $f\colon \boldsymbol{R} \times \boldsymbol{R} \to \boldsymbol{R}$ by

$$f(x,p) = -(x - p^2)(x - 2p^2).$$

Consider the function implicitly defined by $f(x,p) = 0$. The function f is zero along the parabolas $x = p^2$ and $x = 2p^2$, and in particular $f(0,0) = 0$. See Figure 3.5 on page 51. The hypothesis of the Implicit Function Theorem is not satisfied since $\frac{\partial f(0,0)}{\partial x} = 0$. The conclusion also fails. The problem here is not that a smooth implicit function through $(x,p) = (0,0)$ fails to exist. The problem is that it is not unique. There are four distinct continuously differentiable implicitly defined functions. □

121 Example (Lack of continuous differentiability) Consider again the function $h(x) = x + 2x^2 \sin \frac{1}{x^2}$ from Example 70. Recall that h is differentiable everywhere, but not continuously differentiable at zero. Furthermore, $h(0) = 0$, $h'(0) = 1$, but h is not monotone on any neighborhood of zero. Now consider the function $f(x,p) = h(x) - p$. It satisfies $f(0,0) = 0$ and $\frac{\partial f(0,0)}{\partial x} \neq 0$, but it there is no unique implicitly defined function on any neighborhood, nor is there any continuous implicitly defined function.

To see this, note that $f(x,p) = 0$ if and only if $h(x) = p$. So a unique implicitly defined function exists only if h is invertible on some neighborhood of zero. But this is not so, for given any $\varepsilon > 0$, there is some $0 < p < \frac{\varepsilon}{2}$ for which there are $0 < x < x' < \varepsilon$ satisfying $h(x) = h(x') = p$. It is also easy to see that no continuous function satisfies $h\big(\xi(p)\big) = p$ either. □

If X is more than one-dimensional there are subtler ways in which $D_x f$ may fail to be continuous. The next example is taken from Dieudonné [43, Problem 10.2.1, p. 273].

122 Example Define $f\colon \boldsymbol{R}^2 \to \boldsymbol{R}^2$ by

$$f^1(x,y) = x$$

and

$$f^2(x,y) = \begin{cases} y - x^2 & 0 \leqslant x^2 \leqslant y \\ \dfrac{y^2 - x^2 y}{x^2} & 0 \leqslant y \leqslant x^2 \\ -f^2(x,-y) & y \leqslant 0. \end{cases}$$

Then f is everywhere differentiable on $\boldsymbol{R}^2$, and $Df(0,0)$ is the identity mapping, but Df is not continuous at the origin. Furthermore in every neighborhood of the origin there are distinct points (x,y) and (x',y') with $f(x,y) = f(x',y')$. Thus f has no local inverse, so the equation $f(x,y) - p = f(0,0)$ does not uniquely define a function. □ Work out the details.

3.15.3 Implicit vs. inverse function theorems

In this section we discuss the relationship between the existence of a unique implicitly defined function and the existence of an inverse function. These results are quite standard and may be found, for instance, in Marsden [107, p. 234].

First we show how the implicit function theorem can be used to prove an inverse function theorem. Suppose $X \subset \boldsymbol{R}^{\mathrm{n}}$ and $g\colon X \to \boldsymbol{R}^{\mathrm{n}}$. Let P be a neighborhood of $\bar{p} = g(\bar{x})$. Consider $f\colon X \times P \to \boldsymbol{R}^{\mathrm{n}}$ defined by

$$f(x,p) = g(x) - p.$$

Then $f(x,p) = 0$ if and only if $p = g(x)$. Thus if there is a unique implicitly defined function $\xi\colon P \to X$ implicitly defined by $f\big(\xi(p),p\big) = 0$, it follows that g is invertible and $\xi = g^{-1}$. Now compare the Jacobian matrix of f with respect to x and observe that it is just the Jacobian matrix of g. Thus each of the implicit function theorems has a corresponding inverse function theorem.

We could also proceed in the other direction, as is usually the case in textbooks. Let $X \times P$ be a subset of $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}^{\mathrm{m}}$, and let $f\colon X \times P \to \boldsymbol{R}^{\mathrm{n}}$, and suppose $f(\bar{x}, \bar{p}) = 0$. Define a function $g\colon X \times P \to \boldsymbol{R}^{\mathrm{n}} \times P$ by

$$g(x, p) = \big(f(x, p), p\big).$$

Suppose g is invertible, that is, it is one-to-one and onto. Define $\xi\colon P \to X$ by

$$\xi(p) = \pi_x\big(g^{-1}(0, p)\big).$$

Then ξ is the unique function implicitly defined by

$$\big(f\big(\xi(p), p\big), p\big) = 0$$

for all $p \in P$. Now let's compare hypotheses. The standard Inverse Function Theorem, e.g. [107, Theorem 7.1.1, p. 206], says that if g is continuously differentiable and has a nonsingular Jacobian matrix at some point, then there is a neighborhood of the point where g is invertible. The Jacobian matrix for $g(x, p) = \big(f(x, p), p\big)$ above is

$$\begin{bmatrix} \dfrac{\partial f^1}{\partial x_1} & \cdots & \dfrac{\partial f^1}{\partial x_n} & \dfrac{\partial f^1}{\partial p_1} & \cdots & \dfrac{\partial f^1}{\partial p_m} \\ \vdots & & \vdots & \vdots & & \vdots \\ \dfrac{\partial f^n}{\partial x_1} & \cdots & \dfrac{\partial f^n}{\partial x_n} & \dfrac{\partial f^n}{\partial p_1} & \cdots & \dfrac{\partial f^n}{\partial p_m} \\ 0 & \cdots & 0 & 1 & & 0 \\ \vdots & & \vdots & & \ddots & \\ 0 & \cdots & 0 & 0 & & 1 \end{bmatrix}.$$

Since this is block diagonal, it is easy to see that this Jacobian matrix is nonsingular at $(\bar{x}, \bar{p})$ if and only if the derivative $D_x f(\bar{x}, \bar{p})$, is invertible.

3.15.4 Global inversion

The inverse function theorems proven above are local results. Even if the Jacobian matrix of a function never vanishes, it may be that the function does not have an inverse everywhere. The following example is well known, see e.g., [107, Example 7.1.2, p. 208].

Add a section on the Gale–Nikaidô Theorem.

123 Example (A function without a global inverse) Define $f\colon \boldsymbol{R}^2 \to \boldsymbol{R}^2$ via

$$f(x, y) = (e^x \cos y, e^x \sin y).$$

Then the Jacobian matrix is

$$\begin{pmatrix} e^x \cos y & -e^x \sin y \\ e^x \sin y & e^x \cos y \end{pmatrix}$$

which has determinant $e^{2x}(\cos^2 y + \sin^2 y) = e^{2x} > 0$ everywhere. Nonetheless, f is not invertible since $f(x, y) = f(x, y + 2\pi)$ for every x and y. □

3.16 Applications of the Implicit Function Theorem

3.16.1 A fundamental lemma

A **curve** in $\boldsymbol{R}^{\mathrm{n}}$ is simply a function from an interval of $\boldsymbol{R}$ into $\boldsymbol{R}^{\mathrm{n}}$, usually assumed to be continuous.

124 Fundamental Lemma on Curves *Let U be an open set in $\boldsymbol{R}^{\mathrm{n}}$ and let $g\colon U \to \boldsymbol{R}^{\mathrm{m}}$. Let $x^* \in U$ satisfy $g(x^*) = 0$, and suppose g is differentiable at x^*. Assume that ${g_1}'(x^*), \dots, {g_m}'(x^*)$ are linearly independent. Let $v \in \boldsymbol{R}^{\mathrm{n}}$ satisfy*

$$ {g_i}'(x^*) \cdot v = 0, \quad i = 1, \dots, m. $$

Then there exists $\delta > 0$ and a curve $\hat{x}\colon (-\delta, \delta) \to U$ satisfying:

1. *$\hat{x}(0) = x^*$.*
2. *$g\big(\hat{x}(\alpha)\big) = 0$ for all $\alpha \in (-\delta, \delta)$.*
3. *$\hat{x}$ is differentiable at 0. Moreover, if g is C^k on U, then $\hat{x}$ is C^k on $(-\delta, \delta)$.*
4. *$\hat{x}'(0) = v$.*

Proof: Since the ${g_i}'(x^*)$s are linearly independent, $n \geqslant m$, and without loss of generality, we may assume the coordinates are numbered so that the $m \times m$ matrix

$$ \begin{bmatrix} \dfrac{\partial g^1}{\partial x_1} & \cdots & \dfrac{\partial g^1}{\partial x_m} \\ \vdots & & \vdots \\ \dfrac{\partial g^m}{\partial x_1} & \cdots & \dfrac{\partial g^m}{\partial x_m} \end{bmatrix} $$

is invertible at x^*.

Fix v satisfying ${g_i}'(x^*) \cdot v = 0$ for all $i = 1, \dots, m$. Rearranging terms we have

$$ \sum_{j=1}^{m} \frac{\partial g^i(x^*)}{\partial x_j} \cdot v_j = - \sum_{j=m+1}^{n} \frac{\partial g^i(x^*)}{\partial x_j} \cdot v_j \quad i = 1, \dots, m, $$

or in matrix terms

$$ \begin{bmatrix} \dfrac{\partial g^1}{\partial x_1} & \cdots & \dfrac{\partial g^1}{\partial x_m} \\ \vdots & & \vdots \\ \dfrac{\partial g^m}{\partial x_1} & \cdots & \dfrac{\partial g^m}{\partial x_m} \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} = - \begin{bmatrix} \dfrac{\partial g^1}{\partial x_{m+1}} & \cdots & \dfrac{\partial g^1}{\partial x_n} \\ \vdots & & \vdots \\ \dfrac{\partial g^m}{\partial x_{m+1}} & \cdots & \dfrac{\partial g^m}{\partial x_n} \end{bmatrix} \begin{bmatrix} v_{m+1} \\ \vdots \\ v_n \end{bmatrix}, $$

so

$$ \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} = - \begin{bmatrix} \dfrac{\partial g^1}{\partial x_1} & \cdots & \dfrac{\partial g^1}{\partial x_m} \\ \vdots & & \vdots \\ \dfrac{\partial g^m}{\partial x_1} & \cdots & \dfrac{\partial g^m}{\partial x_m} \end{bmatrix}^{-1} \begin{bmatrix} \dfrac{\partial g^1}{\partial x_{m+1}} & \cdots & \dfrac{\partial g^1}{\partial x_n} \\ \vdots & & \vdots \\ \dfrac{\partial g^m}{\partial x_{m+1}} & \cdots & \dfrac{\partial g^m}{\partial x_n} \end{bmatrix} \begin{bmatrix} v_{m+1} \\ \vdots \\ v_n \end{bmatrix}. $$

Observe that these conditions completely characterize v. That is, for any $y \in \boldsymbol{R}^{\mathrm{n}}$,

$$ \big({g_i}'(x^*) \cdot y = 0,\ i = 1, \dots, m, \text{ and } y_j = v_j,\ j = m{+}1, \dots, n\big) \quad \Longrightarrow \quad y = v. \tag{3.11} $$

Define the C^∞ function $f\colon \boldsymbol{R}^{\mathrm{m}} \times \boldsymbol{R} \to \boldsymbol{R}^{\mathrm{n}}$ by

$$ f(z, \alpha) = (z_1, \dots, z_m, x^*_{m+1} + \alpha v_{m+1}, \dots, x^*_n + \alpha v_n). $$

Set $z^* = (x^*_1, \dots, x^*_m)$ and note that $f(z^*, 0) = x^*$. Since x^* is an interior point of U, there is a neighborhood W of z^* and an interval $(-\eta, \eta)$ in $\boldsymbol{R}$ so that for every $z \in W$ and $\alpha \in (-\eta, \eta)$, the point $f(z, \alpha)$ belongs to $U \subset \boldsymbol{R}^{\mathrm{n}}$. Finally, define $h\colon W \times (-\eta, \eta) \to \boldsymbol{R}^{\mathrm{m}}$ by

$$ h(z, \alpha) = g\big(f(z, \alpha)\big) = g(z_1, \dots, z_m, x^*_{m+1} + \alpha v_{m+1}, \dots, x^*_n + \alpha v_n). $$

Observe that h possesses the same degree of differentiability as g, since h is the composition of g with the C^∞ function f.

Then for $j = 1, \dots, m$, we have $= \frac{\partial h^i}{\partial z_j}(z, 0) = \frac{\partial g_i}{\partial x_j}(x)$, where $z = (x_1, \dots, x_m)$. Therefore the m-vectors

$$\begin{bmatrix} \dfrac{\partial h^i(z^*;0)}{\partial z_1} \\ \vdots \\ \dfrac{\partial h^i(z^*;0)}{\partial z_m} \end{bmatrix}, \quad i = 1, \dots, m$$

are linearly independent.

But $h(z^*, 0) = 0$, so by the Implicit Function Theorem 115 there is an interval $(-\delta, \delta) \subset (-\eta, \eta)$ about 0, a neighborhood $V \subset W$ of z^* and a function $\zeta\colon (-\delta, \delta) \to V$ such that

$$\zeta(0) = z^*,$$

$$h\big(\zeta(\alpha), \alpha\big) = 0, \quad \text{for all } \alpha \in (-\delta, \delta),$$

and ζ is differentiable at 0. Moreover, by Implicit Function Theorem 117, if g is C^k on U, then h is C^k, so ζ is C^k on $(-\delta, \delta)$.

Define the curve $\hat{x}\colon (-\delta, \delta) \to U$ by

$$\hat{x}(\alpha) = f\big(\zeta(\alpha), \alpha\big) = \big(\zeta_1(\alpha), \dots, \zeta_m(\alpha), x^*_{m+1} + \alpha v_{m+1}, \dots, x^*_n + \alpha v_n\big). \tag{3.12}$$

Then $\hat{x}(0) = x^*$,

$$g\big(\hat{x}(\alpha)\big) = 0 \text{ for all } \alpha \in (-\delta, \delta),$$

and $\hat{x}$ is differentiable at 0, and if g is C^k, then $\hat{x}$ is C^k. So by the Chain Rule,

$$g_i{}'(x^*) \cdot \hat{x}'(0) = 0, \quad i = 1, \dots, m.$$

Now by construction (3.12), $\hat{x}_j'(0) = v_j$, for $j = m+1, \dots, n$. Thus (3.11) implies $\hat{x}'(0) = v$. ∎

3.16.2 A note on comparative statics

"Comparative statics" analysis tells us how equilibrium values of endogenous variables $x_1, \dots, x_n$ (the things we want to solve for) change as a function of the exogenous parameters $p_1, \dots, p_m$. (As such it is hardly unique to economics.) Typically we can write the equilibrium conditions of our model as the zero of a system of equations in the endogenous variables and the exogenous parameters:

$$\begin{aligned} F^1(x_1, \dots, x_n; p_1, \dots, p_m) &= 0 \\ &\vdots \\ F^n(x_1, \dots, x_n; p_1, \dots, p_m) &= 0 \end{aligned} \tag{3.13}$$

This implicitly defines x as a function of p, which we will explicitly denote $x = \xi(p)$, or

$$(x_1, \dots, x_n) = \big(\xi_1(p_1, \dots, p_m), \dots, \xi_n(p_1, \dots, p_m)\big).$$

This explicit function, if it exists, satisfies the implicit definition

$$F\big(\xi(p); p\big) = 0 \tag{3.14}$$

for at least a rectangle of values of p. The Implicit Function Theorem tells that such an explicit function exists whenever it is possible to solve for all its partial derivatives.

Setting $G(p) = F\big(\xi(p);p\big)$, and differentiating G^i with respect to p_j, yields, by equation (3.14),

$$\sum_k \frac{\partial F^i}{\partial x_k}\frac{\partial \xi^k}{\partial p_j} + \frac{\partial F^i}{\partial p_j} = 0 \tag{3.15}$$

for each $i = 1,\dots,n$, $j = 1,\dots,m$. In matrix terms we have

$$\begin{bmatrix} \frac{\partial F^1}{\partial x_1} & \cdots & \frac{\partial F^1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F^n}{\partial x_1} & \cdots & \frac{\partial F^n}{\partial x_n} \end{bmatrix} \begin{bmatrix} \frac{\partial \xi^1}{\partial p_1} & \cdots & \frac{\partial \xi^1}{\partial p_m} \\ \vdots & & \vdots \\ \frac{\partial \xi^n}{\partial p_1} & \cdots & \frac{\partial \xi^n}{\partial p_m} \end{bmatrix} + \begin{bmatrix} \frac{\partial F^1}{\partial p_1} & \cdots & \frac{\partial F^1}{\partial p_m} \\ \vdots & & \vdots \\ \frac{\partial F^n}{\partial p_1} & \cdots & \frac{\partial F^n}{\partial p_m} \end{bmatrix} = 0. \tag{3.16}$$

Provided $\begin{bmatrix} \frac{\partial F^1}{\partial x_1} & \cdots & \frac{\partial F^1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F^n}{\partial x_1} & \cdots & \frac{\partial F^n}{\partial x_n} \end{bmatrix}$ has an inverse (the hypothesis of the Implicit Function Theorem) we can solve this:

$$\begin{bmatrix} \frac{\partial \xi^1}{\partial p_1} & \cdots & \frac{\partial \xi^1}{\partial p_m} \\ \vdots & & \vdots \\ \frac{\partial \xi^n}{\partial p_1} & \cdots & \frac{\partial \xi^n}{\partial p_m} \end{bmatrix} = -\begin{bmatrix} \frac{\partial F^1}{\partial x_1} & \cdots & \frac{\partial F^1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F^n}{\partial x_1} & \cdots & \frac{\partial F^n}{\partial x_n} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial F^1}{\partial p_1} & \cdots & \frac{\partial F^1}{\partial p_m} \\ \vdots & & \vdots \\ \frac{\partial F^n}{\partial p_1} & \cdots & \frac{\partial F^n}{\partial p_m} \end{bmatrix} \tag{3.17}$$

The old-fashioned derivation (see, e.g., Samuelson [136, pp. 10–14]) of this same result runs like this: "Totally differentiate" the ith row of equation (3.13) to get

$$\sum_k \frac{\partial F^i}{\partial x_k}\,dx_k + \sum_\ell \frac{\partial F^i}{\partial p_\ell}\,dp_\ell = 0 \tag{3.18}$$

for all i. Now set all dp_ℓs equal to zero except p_j, and divide by dp_j to get

$$\sum_k \frac{\partial F^i}{\partial x_k}\frac{dx_k}{dp_j} + \frac{\partial F^i}{\partial p_j} = 0 \tag{3.19}$$

for all i and j, which is equivalent to equation (3.15). For further information on total differentials and how to manipulate them, see [6, Chapter 6].

Using Cramer's Rule (e.g. [8, pp. 93–94]), we see then that

$$\frac{dx_i}{dp_j} = \frac{\partial \xi^i}{\partial p_j} = -\frac{\begin{vmatrix} \frac{\partial F^1}{\partial x_1} & \cdots & \frac{\partial F^1}{\partial x_{i-1}} & \frac{\partial F^1}{\partial p_j} & \frac{\partial F^1}{\partial x_{i+1}} & \cdots & \frac{\partial F^1}{\partial x_n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \frac{\partial F^n}{\partial x_1} & \cdots & \frac{\partial F^n}{\partial x_{i-1}} & \frac{\partial F^n}{\partial p_j} & \frac{\partial F^n}{\partial x_{i+1}} & \cdots & \frac{\partial F^n}{\partial x_n} \end{vmatrix}}{\begin{vmatrix} \frac{\partial F^1}{\partial x_1} & \cdots & \frac{\partial F^1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F^n}{\partial x_1} & \cdots & \frac{\partial F^n}{\partial x_n} \end{vmatrix}}. \tag{3.20}$$

Or, letting Δ denote the determinant of $\begin{bmatrix} \dfrac{\partial F^1}{\partial x_1} & \cdots & \dfrac{\partial F^1}{\partial x_n} \\ \vdots & & \vdots \\ \dfrac{\partial F^n}{\partial x_1} & \cdots & \dfrac{\partial F^n}{\partial x_n} \end{bmatrix}$, and letting $\Delta_{i,j}$ denote the determinant of the matrix formed by deleting its i-th row and j-th column, we have

$$\frac{\partial \xi^i}{\partial p_j} = -\sum_{k=1}^{n} (-1)^{i+k} \frac{\partial F^k}{\partial p_j} \frac{\Delta_{k,i}}{\Delta}. \tag{3.21}$$

Section 4

Convex analysis

4.1 Convex sets

125 Definition *A subset of a vector space is* ***convex*** *if it includes the line segment joining any two of its points. That is, C is convex if for each pair x, y of points in C, the* ***line segment***

$$\{\lambda x + (1-\lambda)y : \lambda \in [0,1]\}$$

is included in C.

Picture.

By induction, if a set C is convex, then for every finite subset $\{x_1, \ldots, x_n\}$ and nonnegative scalars $\{\lambda_1, \ldots, \lambda_n\}$ with $\sum_{i=1}^n \lambda_i = 1$, the linear combination $\sum_{i=1}^n \lambda_i x_i$ lies in C. Such a linear combination is called a **convex combination**, and the coefficients λ_i are frequently called **weights**.

126 Definition *The* ***convex hull*** $\operatorname{co} A$ *of a set A is the intersection of all convex sets that include A.*

Picture.

The convex hull of a set A is the smallest convex set including A, in the sense that $\operatorname{co} A$ is convex, includes A, and if C is a convex set that includes A, then C includes $\operatorname{co} A$. The convex hull of A consists precisely of all convex combinations from A. (These claims are actually lemmas, and you are asked to prove them in the next exercise.)

127 Exercise (Properties of convex sets) Prove the following.

1. If a set C is convex, then for every finite subset $\{x_1, \ldots, x_n\}$ and nonnegative scalars $\{\lambda_1, \ldots, \lambda_n\}$ with $\sum_{i=1}^n \lambda_i = 1$, the linear combination $\sum_{i=1}^n \lambda_i x_i$ lies in C.
2. The sum of two convex sets is convex.
3. Scalar multiples of convex sets are convex.
4. A set C is convex if and only if
$$\alpha C + \beta C = (\alpha + \beta)C$$
for all nonnegative scalars α and β.
5. The intersection of an arbitrary family of convex sets is convex.
6. The convex hull of a set A is the set of all convex of elements of A. That is,
$$\operatorname{co} A = \{\sum_{i=1}^n \lambda_i x_i : \sum_{i=1}^n \lambda_i = 1,\ \lambda_i > 0,\ x_i \in A,\ i = 1, \ldots, n.$$

star The interior and closure of a convex set are also convex.

□

While the sum of two convex sets is necessarily convex, the sum of two non-convex sets may also be convex. For example, let A be the set of rationals in $\boldsymbol{R}$ and B be the union of 0 and the irrationals. Neither set is convex, but their sum is the set of all real numbers, which is of course convex.

128 Definition *A set C in a vector space is a* ***cone*** *if whenever x belongs to C so does λx for every $\lambda \geqslant 0$. The cone C is* ***pointed*** *if for any nonzero x belonging to C, the point $-x$ does not belong to C. A cone is* ***nondegenerate*** *if it contains a nonzero vector. A cone less 0 is called a* ***deleted cone****.*

Picture.

Every linear subspace is a cone by this definition, but not a pointed cone. The origin by itself is a degenerate cone. A cone is pointed if and only if it includes no linear subspace other than $\{0\}$. Under this definition, a cone C always contains 0, but we may have occasion to call a set of the form $x + C$ a cone with vertex at x.

Given a set A, the **cone generated by** $\boldsymbol{A}$ is $\{\lambda x : \lambda \geqslant 0,\ x \in A\}$. It is truly a cone, and is the intersection of all the cones that include A. A cone generated by a single nonzero point is called a **ray**. A **finite cone** is a cone generated by the convex hull of a finite set. A nondegenerate finite cone is the sum of finitely many rays.

Picture.

129 Exercise (Properties of cones) Prove the following.

1. Scalar multiples of cones are cones.
2. The intersection of an arbitrary family of cones is a cone.
3. The sum of two cones is a cone.
4. The cone generated by A is the smallest (with respect to inclusion) cone that includes A.
5. A cone is convex if and only if it is closed under addition.
6. The cone generated by a convex set is convex.
7. A nondegenerate finite cone is the sum of finitely many rays.
8. The finite cone generated by the finite set $\{x_1, \dots, x_n\}$ is the set of nonnegative linear combinations of the x_is. That is,

$$\Big\{\sum_{i=1}^{n} \lambda_i x_i : \lambda_i \geqslant 0,\ i = 1, \dots, n\Big\}.$$

□

4.2 Affine sets and functions

Recall that a linear subspace M of a vector space X is a nonempty set closed under linear combinations. That is, $x, y \in M$ and $\alpha, \beta \in \boldsymbol{R}$ imply that $\alpha x + \beta y \in M$. Well an **affine subspace** (sometimes called a **flat**) is a nonempty set that is closed under **affine combinations**, that is, linear combinations of the form $\alpha x + \beta y$ satisfying $\alpha + \beta = 1$. Another way to say this is that if two point x and y belong to M, then the line

$$\{\lambda x + (1 - \lambda)y : \lambda \in \boldsymbol{R}\}$$

also belongs to M. An affine combination differs from a convex combination in that one of the scalars is allowed to be negative in affine combination. We can take more than two points in an affine combination, as the next exercise asks you to prove. (It's not hard, but not trivial either.)

Picture.

130 Exercise (Affine combinations) Let A be an affine subspace. Prove that if $x_1, \dots, x_n$ belong to A and $\lambda_1, \dots, \lambda_n$ are scalars that sum to one, then $\lambda_1 x_1 + \cdots + \lambda_n x_n$ also belongs to A. □

Every affine set is a **translate** of a unique linear subspace. The next exercise asks you to prove this. If two affine subspaces are translates of the same linear subspace, then they are said to be **parallel**. The dimension of the linear subspace is also called the **dimension** of the affine subspace.

Picture.

131 Exercise (Affine subspaces) Let X be a vector space. Prove the following.

1. Let M be a linear subspace of X and let a be a vector in X. Then $M+a = \{x+a : x \in M\}$ is an affine subspace of X.
2. Let A be an affine subspace of X, and let a and b belong to A.
   (a) The set $A - a = \{x - a : x \in A\}$ is a linear subspace of X.
   (b) $A - a = A - b$.
3. Consequently, for every affine subspace A, there is a linear subspace M such that $A = M + a$.
4. If M and N are linear subspaces such $A = M + x = N + y$ for some $x, y \in A$, then $M = N$. This subspace is called the **linear subspace parallel to A**.
5. If M is a linear subspace
6. An affine subspace is a linear subspace if and only it contains 0.
7. Let M denote the unique linear subspace parallel to A. For $x \in M$ and $y \in A$ together imply that $x + y \in A$.

□

Given a set A, the **affine hull** aff A of A is the intersection of all affine subspaces that include A. It is the smallest affine subspace that includes A and consists of all affine combinations of points in A. (Prove this to yourself.)[1]

Recall that a **real linear function** on a vector space X is a function f that satisfies

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y) \qquad \text{for all } x, y \in X,\ \alpha, \beta \in \boldsymbol{R}.$$

When $X = \boldsymbol{R}^{\mathrm{n}}$, if f is linear, there is a unique $p \in \boldsymbol{R}^{\mathrm{n}}$ satisfying

$$f(x) = p \cdot x$$

for all x, namely $p_i = f(e^i)$, $i = 1, \dots, n$.

132 Definition *Let A be an affine subspace of the vector space X. A real function $f\colon X \to \boldsymbol{R}$ is **affine** if for every $x, y \in A$ and scalar λ,*

$$f\big(\lambda x + (1-\lambda)y\big) = \lambda f(x) + (1-\lambda) f(y).$$

[1] We could call the intersection of all linear subspaces that include A the **linear hull** of A. It is the smallest linear subspace that includes A and consists of all linear combinations of points in A. But instead, we traditionally call this set the **span** of A.

133 Exercise (Affine functions) Let A be an affine subspace of the vector space X. A real function f on A is affine if and only it is of the form $f(x) = g(x-a) - \gamma$, where a belongs to A and g is linear on the linear subspace $A - a$. Moreover, g is independent of the choice of a in A, and $\gamma = -f(a)$.

In particular, when $A = X$, then an affine function f on X can be written as $f(x) = g(x) - \gamma$, where g is linear on X and $\gamma = -f(0)$. □

4.3 Convex and concave functions

Interpret geometrically.

134 Definition *A function $f\colon C \to \boldsymbol{R}$ on a convex subset C of a vector space is:*

$$\begin{aligned} \textbf{\textit{concave}} \text{ if} \quad & f\big(\lambda x + (1-\lambda)y\big) \geqslant \lambda f(x) + (1-\lambda) f(y) \\ \textbf{\textit{strictly concave}} \text{ if} \quad & f\big(\lambda x + (1-\lambda)y\big) > \lambda f(x) + (1-\lambda) f(y) \\ \textbf{\textit{convex}} \text{ if} \quad & f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y) \\ \textbf{\textit{strictly convex}} \text{ if} \quad & f\big(\lambda x + (1-\lambda)y\big) < \lambda f(x) + (1-\lambda) f(y) \end{aligned}$$

for all x, y in C with $x \neq y$ and all $0 < \lambda < 1$.

It is easy to show that a function f is concave if and only if

$$f\Big(\sum_{i=1}^{n} \lambda_i x_i\Big) \geqslant \sum_{i=1}^{n} \lambda_i f(x_i)$$

for every convex combination $\sum_{i=1}^{n} \lambda_i x_i$.

135 Exercise Prove the following.

1. The sum of concave functions is concave.
2. A nonnegative multiple of a concave function is concave.
3. The pointwise limit of a sequence of concave functions is concave.
4. The pointwise infimum of a family of concave functions is concave.
5. A function is both concave and convex if and only if it is affine.

□

4.4 Talking convex analysis

Mathematicians who specialize in *convex analysis* use a different terminology from what you are likely to encounter in a calculus or real analysis text. In particular, the excellent books by Rockafellar [130] and Castaing and Valadier [36] can be incomprehensible if you pick one up and start reading in the middle. In this section I explain some of their terminology and how it relates to mine, which I shall call the conventional terminology.

First off, in convex analysis, concave functions are almost always taken to be defined everywhere on $\boldsymbol{R}^{\mathrm{n}}$ (or some general vector space), and are allowed to assume the extended values $+\infty$ and $-\infty$. (Actually convex analysts talk mostly about convex functions, not concave functions.) Let us use $\boldsymbol{R}^{\sharp}$ to denote the **extended real numbers**, $\boldsymbol{R}^{\sharp} = \boldsymbol{R} \cup \{+\infty, -\infty\}$. Given an extended real-valued function $f\colon \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}^{\sharp}$, the **hypograph** of f is the subset of $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$ defined by

$$\{(x, \alpha) \in \boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R} : \alpha \leqslant f(x)\}.$$

The **epigraph** is defined by reversing the inequality. Note well that the hypograph or epigraph of f is a subset of $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$, not of $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}^\sharp$. That is, each α in the definition of hypograph or epigraph is a real number, not an extended real. For example, the epigraph of the constant function $+\infty$ is the empty set.

To a convex analyst, an extended real-valued function is **concave** if its hypograph is a convex subset of $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$. Given a concave function $f\colon \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}^\sharp$, its **effective domain**, $\operatorname{dom} f$, is the projection of its hypograph on $\boldsymbol{R}^{\mathrm{n}}$, that is,

$$\operatorname{dom} f = \{x \in \boldsymbol{R}^{\mathrm{n}} : f(x) > -\infty\}.$$

The effective domain of a concave function is a (possibly empty) convex set.

Similarly, an extended real-valued function is **convex** if its epigraph is a convex subset of $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$. Given a convex function $f\colon \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}^\sharp$, its **effective domain** is

$$\operatorname{dom} f = \{x \in \boldsymbol{R}^{\mathrm{n}} : f(x) < \infty\}.$$

This terminology leaves the question of what is the effective domain of a function that is neither concave nor convex unanswered. However in convex analysis, the question never arises.

We can extend a conventional real-valued concave function f defined on a subset C of $\boldsymbol{R}^{\mathrm{n}}$ to an extended real-valued function $\hat{f}$ defined on all of $\boldsymbol{R}^{\mathrm{n}}$ by setting

$$\hat{f}(x) = \begin{cases} f(x) & x \in C \\ -\infty & x \notin C. \end{cases}$$

Note that $\hat{f}$ is concave in the convex analyst's sense if and only if f is concave in the conventional sense, in which case we also have that $\operatorname{dom} \hat{f} = C$. (We can similarly extend conventionally defined convex functions to $\boldsymbol{R}^{\mathrm{n}}$ by setting them equal to $+\infty$ where conventionally undefined.)

Proper functions

In the language of convex analysis, a concave function is **proper** if its effective domain is nonempty and its hypograph contains no vertical lines. (A **vertical line** in $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$ is a set of the form $\{x\} \times \boldsymbol{R}$ for some $x \in \boldsymbol{R}^{\mathrm{n}}$.) That is, f is proper if $f(x) > -\infty$ for at least one x and $f(x) < \infty$ for every x. Every proper concave function is gotten by taking a finite-valued concave function defined on some nonempty convex set and extending it to all of $\boldsymbol{R}^{\mathrm{n}}$ as above.

> *Thus every concave function in the conventional terminology corresponds to a proper concave function in the terminology of convex analysis.*

A convex function is proper if its effective domain is nonempty and its epigraph contains no vertical lines. A convex function f is proper if $-f$ is a proper concave function.

As an example of a nontrivial improper concave function, consider this one taken from Rockafellar [130, p. 24].

136 Example (A nontrivial improper concave function) The function $f\colon \boldsymbol{R} \to \boldsymbol{R}^\sharp$ defined by

$$f(x) = \begin{cases} +\infty & |x| < 1 \\ 0 & |x| = 1 \\ -\infty & |x| > 1 \end{cases}$$

is an improper concave function that is not constant. □

Some authors, in paticular, Hiriart-Urruty and Lemaréchal [76] do not permit convex functions to assume the value $-\infty$, so for them, properness is equivalent to nonemptyness of the effective domain.

Indicator functions

Another deviation from conventional terminology is the convex analysts' definition of the indicator function. The **indicator function** of the set C, denoted $\delta(\cdot \mid C)$, is defined by

$$\delta(x \mid C) = \begin{cases} 0 & x \in C \\ +\infty & x \notin C. \end{cases}$$

The indicator of C is a convex function if and only if C is a convex set. This is not to be confused with the probabilists' indicator function $\mathbf{1}_C$ defined by

$$\mathbf{1}_C(x) = \begin{cases} 1 & x \in C \\ 0 & x \notin C. \end{cases}$$

4.5 Affine sets and the relative interior of a convex set

An important difference between convex analysts and the rest of us is the use of the term "relative interior". When a convex analyst refers to the relative interior of a convex set, she does not mean the interior relative to the entire ambient vector space, but to explain it we need a few more definitions.

Recall that a set E in a vector space is **affine** if it includes all the lines (not just line segments) generated by its points. Every set E is included in a smallest affine set $\operatorname{aff} E$, called its **affine hull**. A set A is affine if and only if for each $x \in A$, the set $A - x$ is a linear subspace. In $\boldsymbol{R}^{\mathrm{n}}$, every linear subspace and so every affine subspace is closed, so as a result, in $\boldsymbol{R}^{\mathrm{n}}$, a subset E and its closure $\overline{E}$ have the same affine hull. This need not be true in an infinite dimensional topological vector space.

137 Definition *The **relative interior** of a convex set C in a topological vector space, denoted $\operatorname{ri} C$, is defined to be its topological interior relative to its affine hull $\operatorname{aff} C$.*

In other words, $x \in \operatorname{ri} C$ if and only if there is some open neighborhood U of x such that $y \in U \cap \operatorname{aff} C$ implies $y \in \operatorname{ri} C$.

Even a one point set has a nonempty relative interior in this sense, namely itself. The only convex set with an empty relative interior is the empty set. Similarly, the **relative boundary** of a convex set is the boundary relative to the affine hull. This turns out to be the closure minus the relative interior.

Note well that this is not the same relative interior that a topologist would mean. In particular, it is *not* true that $A \subset B$ implies $\operatorname{ri} A \subset \operatorname{ri} B$. For instance, consider a closed interval and one of its endpoints. The relative interior of the interval is the open interval and the relative interior of the singleton endpoint is itself, which is disjoint from the relative interior of the interval.

The most important property of the relative interior of a convex set for our purposes is the following simple proposition.

138 Proposition *Let C be a convex subset of a topological vector space. If $x \in \operatorname{ri} C$ and $y \in C$, then for $|\varepsilon|$ small enough, we have $x + \varepsilon(x - y) \in \operatorname{ri} C$.*

Proof: Let A be the affine hull of C. Then $\operatorname{ri} C$ is a relatively open subset of A in the topological sense. Define $h\colon \boldsymbol{R} \to A$ by $h(\lambda) = x + \lambda(x - y)$. (To verify that $h(\lambda)$ actually belongs to $A = \operatorname{aff} C$, note that both x and y belong to C and $h(\lambda)$ is the affine combination $(1 + \lambda)x + (-\lambda)y$.) Then h is continuous since scalar multiplication and vector addition are continuous functions. Moreover, $h(0) = x \in \operatorname{ri} C$. So by continuity, the inverse image $U = h^{-1}(\operatorname{ri} C)$ of the open subset $\operatorname{ri} C$ of A is an open set in $\boldsymbol{R}$ that includes 0. That is, there is some $\eta > 0$ such that $d(\varepsilon, 0) = |\varepsilon| < \eta$ implies that $\varepsilon \in U$, so $h(\varepsilon) \in \operatorname{ri} C$. ∎

We may use this fact without any special mention. Indeed it is the key to the proof of the next theorem.

139 Theorem *If f is an improper concave function, then $f(x) = \infty$ for every $x \in \operatorname{ri}\operatorname{dom} f$. If f is an improper convex function, then $f(x) = -\infty$ for every $x \in \operatorname{ri}\operatorname{dom} f$.*

Proof: There are two ways a concave function f can be improper. The first is that $\operatorname{dom} f$ is empty, in which case, the conclusion holds vacuously. The second case is that $f(x) = +\infty$ for some $x \in \operatorname{dom} f$. Let y belong to $\operatorname{ri}\operatorname{dom} f$. Then by the remark above, y is proper convex combination $\lambda x + (1-\lambda)z$ $(0 < \lambda < 1)$, where $z = x + \varepsilon(y - x)$ for some $\varepsilon < 0$ and $z \in \operatorname{dom} f$ (so that $f(z) > -\infty$). Then $f(y) \geqslant \lambda\infty + (1-\lambda)f(z) = +\infty$. ∎

4.6 Topological properties of convex sets

140 Lemma *In a topological vector space, both the interior and the closure of a convex set are convex.*

Proof: Let C be a convex subset of a topological vector space, let $0 \leqslant \lambda \leqslant 1$. Observe that for any $x, y \in \operatorname{int} C \subset C$, convexity of C implies that $\lambda x + (1-\lambda)y \in C$. In other words,

$$\lambda(\operatorname{int} C) + (1-\lambda)(\operatorname{int} C) \subset C.$$

Since $\operatorname{int} C$ is open, $\lambda(\operatorname{int} C) + (1-\lambda)(\operatorname{int} C)$ is open (Corollary 44). But the interior of a set includes every open subset, so $\lambda(\operatorname{int} C) + (1-\lambda)(\operatorname{int} C)$ is a subset of $\operatorname{int} C$. This shows that $\operatorname{int} C$ is convex.

To see that $\overline{C}$ is convex, we will deal with the case where M is a metric space.[2] Let x and y belong to $\overline{C}$ and let $0 \leqslant \lambda \leqslant 1$. Then there are sequences in C such that $x_n \to x$ and $y_n \to y$. Since C is convex, $\lambda x_n + (1-\lambda)y_n \in C$. Since $\lambda x_n + (1-\lambda)y_n \to \lambda x + (1-\lambda)y$, we have $\lambda x + (1-\lambda)y \in \overline{C}$. Thus $\overline{C}$ is convex. ∎

141 Lemma *If C is a convex subset of a topological vector space, and if x belongs to the interior of C and y belongs to the closure of C, then for $0 < \lambda \leqslant 1$ the convex combination $\lambda x + (1-\lambda)y$ belongs to the interior of C.*

Proof: This is obviously true for $\lambda = 1$, and is vacuously true if $\operatorname{int} C$ is empty, so assume that $\operatorname{int} C$ is nonempty, and let $x \in \operatorname{int} C$, $y \in \overline{C}$, and $0 < \lambda < 1$.

Since $x \in \operatorname{int} C$, there is an open neighborhood U of zero such that $x + U \subset C$. Then for any point $z \in C$, the set $V(z) = \bigcup_{0<\alpha\leqslant 1} \alpha(x+U) + (1-\alpha)z$ of convex combinations of z and points in $x + U$ is an open set and lies in C, so it is a subset of the interior of X. This region is shaded in Figure 4.1. (The figure is drawn for the case where $z \notin x + U$, and $\lambda = 1/3$. For ease of drawing U is shown as a disk, but it need not be circular or even convex.) It seems clear from the picture that if z is close enough to y, then $\lambda x + (1-\lambda)y$ belongs to the shaded region. We now express this intuition algebraically.

Since $y \in \overline{C}$, there is a point $z \in C$ that also belongs to the open neighborhood $y - \frac{\lambda}{1-\lambda}U$. Thus $(1-\lambda)(y-z)$ belongs to λU, so $\lambda x + (1-\lambda)y = \lambda x + (1-\lambda)z + (1-\lambda)(y-z)$ belongs to $\lambda x + (1-\lambda)z + \lambda U \subset V(z) \subset \operatorname{int} C$. ∎

142 Corollary *If C is a convex subset of a topological vector space with nonempty interior, then:*

1. *$\operatorname{int} C$ is dense in $\overline{C}$, so $\overline{C} = \overline{\operatorname{int} C}$.*

[2] The same proof applies to general topological vector spaces when sequences are replaced by *nets*, if you know what that means.

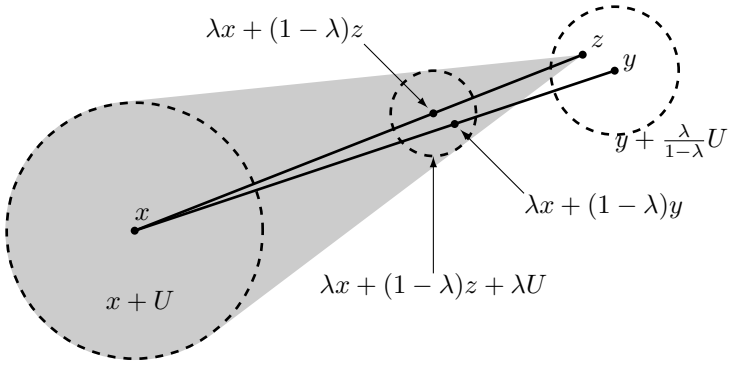

Figure 4.1. The open set $V(z)$ (shaded) contains $\lambda x + (1-\lambda)y$ if z is close to y.

2. $\operatorname{int} C = \operatorname{int} \overline{C}$.

Proof: (1) Let y belong to $\overline{C}$. Pick $x \in \operatorname{int} C$. Then for $0 < \lambda \leqslant 1$, the point $\lambda x + (1 - \lambda y$ belongs to $\operatorname{int} C$ and converges to y as $\lambda \to 0$. Thus $\operatorname{int} C$ is dense in $\overline{C}$, that is, $\overline{C} \subset \overline{\operatorname{int} C}$. But $\overline{C} \supset \overline{\operatorname{int} C}$, so we have equality.

(2) Fix $x \in \operatorname{int} C$ and $y \in \operatorname{int} \overline{C}$. Pick a neighborhood W of zero satisfying $y + W \subset \overline{C}$. Then for $0 < \varepsilon < 1$ small enough, $\varepsilon(y - x) \in W$, so $y + \varepsilon(y - x) \in \overline{C}$. By Lemma 141, we have $y - \varepsilon(y - x) = \varepsilon y + (1 - \varepsilon)y \in \operatorname{int} C$. But then, using Lemma 141 once more, we obtain $y = \frac{1}{2}[y - \varepsilon(y - x)] + \frac{1}{2}[y + \varepsilon(y - x)] \in \operatorname{int} C$. Therefore, $\operatorname{int} \overline{C} \subset \operatorname{int} C \subset \operatorname{int} \overline{C}$, so $\operatorname{int} C = \operatorname{int} \overline{C}$. ∎

Note that a convex set with an empty interior may have a closure with a nonempty interior. For instance, any dense (proper) vector subspace has this property.

143 Proposition *In $\boldsymbol{R}^{\mathrm{n}}$, the relative interior of a nonempty convex set is nonempty.*

Proof: If C is a singleton $\{x\}$, then $\{x\}$ is its relative interior, so assume C has at least two elements. Also, if $x \in C$, then $\operatorname{ri} C = x + \operatorname{ri}(C - x)$, so we may assume that C contains 0, so its affine hull is actually a (finite dimensional) linear subspace M of $\boldsymbol{R}^{\mathrm{n}}$. In this case, $\operatorname{ri} C$ is just the interior of C relative to M.

Tighten this up.

Since C has more than one element, it has a nonempty maximal linearly independent subset, $b_1, \dots, b_k$. This set is also a basis for M. Define a norm on M by $\|x\| = \max_i |\alpha_i|$ where $x = \sum_{i=1}^k \alpha_i b_i$. (This generates the topology on M. Why?)

Now any point of the form $\sum_{i=1}^k \alpha_i b_i$ with each $\alpha_i \geqslant 0$ and $\sum_{i=1}^k \alpha_i \leqslant 1$ belongs to C, as a convex combination 0 and the b_is. In particular, $e = \sum_{i=1}^k \frac{1}{2k} b_i$ belongs to C. In fact, it is an interior point of C. To see this, consider the open ball (in the norm defined above) centered at e with radius $\frac{1}{2k}$. Let $x = \sum_{i=1}^k \alpha_i b_i$ with $\max_i |\alpha_i - \frac{1}{2k}| < \frac{1}{2k}$. Then $0 < \sum_{i=1}^k \alpha_i < 1$, so x belongs to C. Thus e belongs to the interior of C relative to M, and so to $\operatorname{ri} C$. ∎

Add proofs

I now state without proof some additional properties of relative interiors in $\boldsymbol{R}^{\mathrm{n}}$.

144 Proposition (Rockafellar [130, Theorem 6.3, p. 46]) *For a convex subset C of $\boldsymbol{R}^{\mathrm{n}}$,*

$$\overline{\operatorname{ri} C} = \overline{C}, \qquad \text{and} \qquad \operatorname{ri}(\operatorname{ri} C) = \operatorname{ri} C.$$

A consequence of this is that the affine hull of $\operatorname{ri} C$ and of $\overline{C}$ coincide.

145 Proposition (Rockafellar [130, Theorem 6.5, p. 47]) *Let $\{C_i\}_{i\in I}$ be a family of convex subsets of $\boldsymbol{R}^{\mathrm{n}}$, and assume $\bigcap_i \operatorname{ri} C_i \neq \varnothing$. Then*

$$\overline{\bigcap_i C_i} = \bigcap_i \overline{C_i}, \qquad \text{and for finite } I, \quad \operatorname{ri} \bigcap_i C_i = \bigcap_i \operatorname{ri} C_i.$$

146 Proposition (Rockafellar [130, Corollaries 6.6.1, 6.6.2, pp. 48–49]) *For convex subsets C, C_1, C_2 of $\boldsymbol{R}^{\mathrm{n}}$, and $\lambda \in \boldsymbol{R}$,*

$$\operatorname{ri}(\lambda C) = \lambda \operatorname{ri} C, \qquad \operatorname{ri}(C_1 + C_2) = \operatorname{ri} C_1 + \operatorname{ri} C_2, \qquad \text{and} \qquad \overline{C_1 + C_2} \supset \overline{C_1} + \overline{C_2}.$$

147 Proposition (Rockafellar [130, Lemma 7.3, p. 54]) *For any concave function f,*

$$\operatorname{ri} \operatorname{hypo} f = \{(x,\alpha) \in \boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R} : x \in \operatorname{ri} \operatorname{dom} f,\ \alpha < f(x)\}.$$

For a convex function f the corresponding result is

$$\operatorname{ri} \operatorname{epi} f = \{(x,\alpha) \in \boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R} : x \in \operatorname{ri} \operatorname{dom} f,\ \alpha > f(x)\}.$$

4.7 Closed functions

If we adopt the convex analyst's approach that a concave function assumes the value $-\infty$ outside its effective domain, then we are sure to have discontinuities at the boundary of the domain. Thus continuity is too much to expect as a global property of concave or convex functions. However, we shall see below that proper convex and concave functions are Lipschitz continuous on the relative interiors of their effective domains. However, a concave or convex function can be very badly behaved on the boundary of its domain.

148 Example (Bad behavior, Rockafellar [130, p. 53]) Let D be the open unit disk in $\boldsymbol{R}^2$, and let S be the unit circle. Define f to be zero on the disk D, and let f assume any values in $[0,\infty]$ on S. Then f is convex. Note that f can have both types of discontinuities along S, so convexity or concavity places virtually no restrictions on the boundary behavior. In fact, by taking f to be the indicator of a non-Borel subset of S, convexity by itself does not even guarantee measurability. □

Given that global continuity on $\boldsymbol{R}^{\mathrm{n}}$ cannot be reasonably be imposed, it seems that global semicontinuity may be an appropriate regularity condition. And for proper functions it is. However, for possibly improper functions, a slightly different condition has proven to be appropriate, namely closedness.

149 Definition *A concave function on $\boldsymbol{R}^{\mathrm{n}}$ is **closed** if one of the following conditions holds.*

1. *The function is identically $+\infty$.*
2. *The function is identically $-\infty$.*
3. *The function is proper and its hypograph is closed in $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$.*

A convex function is closed if (1)*, or* (2)*, or*

3′. *The function is proper and its epigraph is closed in $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$.*

In light of Theorem 29, it is obvious that:

150 Lemma *A proper (extended real-valued) concave function on $\boldsymbol{R}^{\mathrm{n}}$ is closed if and only if it is upper semicontinuous. A proper convex function on $\boldsymbol{R}^{\mathrm{n}}$ is closed if and only if it is lower semicontinuous.*

That is, closedness and semicontinuity are equivalent for proper functions. The next example clarifies the difference between closed and semicontinuous improper functions.

151 Example (Closedness vs. semicontinuity for improper functions) Let C be a nonempty closed convex set, and define the improper concave function f by

$$f(x) = \begin{cases} +\infty & x \in C \\ -\infty & x \notin C. \end{cases}$$

Then the hypograph of f is a nonempty closed convex set and f is upper semicontinuous, but f does meet the definition of a closed function (unless $C = \boldsymbol{R}^{\mathrm{n}}$). □

In fact this is the only kind of upper semicontinuous improper function.

152 Proposition *An upper semicontinuous improper concave function has no finite values. Ditto for a lower semicontinuous improper convex function.*

Proof: By Theorem 139, if a concave f has $f(x) = +\infty$, then $f(y) = +\infty$ for all $\in \operatorname{ri}\operatorname{dom} f$. By upper semicontinuity, $f(y) = +\infty$ for all $y \in \overline{\operatorname{ri}\operatorname{dom} f} \supset \operatorname{dom} f$. By definition of the effective domain, $f(y) = -\infty$ for $y \notin \operatorname{dom} f$. (This shows that $\operatorname{dom} f$ is closed.) ∎

There are some subtleties in dealing with closed functions, particularly if you are used to the conventional approach. For instance, if C is nonempty convex subset of $\boldsymbol{R}^{\mathrm{n}}$ that is not closed, the conventional approach allows us to define a concave or convex function with domain C, and undefined elsewhere. Consider the constant function zero on C. It is not a closed function in the sense I have defined, because its hypograph is not closed. It is closed in $C \times \boldsymbol{R}$, but not closed in $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$. Regarded as a conventional function on C, it is both upper and lower semicontinuous on C, but regarded as a concave extended real-valued function on $\boldsymbol{R}^{\mathrm{n}}$, it is not upper semicontinuous. This is because at a boundary point x, we have $\limsup_{y \to x} = 0 > f(x) = -\infty$.

This is not really subtle, but I should point out that a function can be closed without having a closed effective domain. For example, the logarithm function (extended to be an extended real-valued concave function) is closed, but has $(0, \infty)$ as its effective domain.

You might ask why we don't want improper semicontinuous functions with nonempty domains to be called closed functions. That is a good question. The answer has to do in part with another way of defining closed functions. Rockafellar [130, p 52, pp. 307–308] makes the following definition. Recall from Definition 32 that the upper envelope of f is defined by $\overline{f}(x) = \inf_{\varepsilon > 0} \sup_{d(y,x) < \varepsilon} f(y)$, and that the upper envelope is real-valued if f is locally bounded, and is upper semicontinuous.

153 Definition *The* ***closure*** $\operatorname{cl} f$ *of a concave function f on $\boldsymbol{R}^{\mathrm{n}}$ is defined by*

1. $\operatorname{cl} f(x) = +\infty$ *for all $x \in \boldsymbol{R}^{\mathrm{n}}$ if $f(y) = +\infty$ for some y.*
2. $\operatorname{cl} f(x) = -\infty$ *for all $x \in \boldsymbol{R}^{\mathrm{n}}$ if $f(x) = -\infty$ for all x.*
3. $\operatorname{cl} f$ *is the upper envelope of f, if f is a proper concave function.*

The ***closure*** $\operatorname{cl} f$ *of a convex function f on $\boldsymbol{R}^{\mathrm{n}}$ is defined by*

1′. $\operatorname{cl} f(x) = -\infty$ *for all $x \in \boldsymbol{R}^{\mathrm{n}}$ if $f(y) = -\infty$ for some y.*

2′. $\operatorname{cl} f(x) = +\infty$ *for all $x \in \boldsymbol{R}^{\mathrm{n}}$ if $f(x) = +\infty$ for all x.*

3′. $\operatorname{cl} f$ *is the lower envelope of f, if f is a proper convex function.*

The next result is a definition in Rockafellar, but a (trivial) theorem using my definition.

154 Proposition *A concave (or convex) function is* ***closed*** *if and only if* $f = \operatorname{cl} f$.

These contorted definitions are required to make the Proposition 223 (that $\operatorname{cl} f = f^{**}$) hold for improper as well as proper concave functions.

It is obvious that $\operatorname{cl} f = -\operatorname{cl}(-f)$. A little less trivial is the following.

155 Proposition *If* $f\colon \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}^{\sharp}$ *is concave, then* $\operatorname{cl} f$ *is concave. If* f *is convex, then* $\operatorname{cl} f$ *is convex.*

Proof: I shall just prove the concave case. If f is concave and does not assume the value $+\infty$, Theorem 33 asserts that the hypograph of the closure of f is the closure of the hypograph of f, which is convex. If f does assume the value $+\infty$, then $\operatorname{cl} f$ is identically $+\infty$, so its hypograph is $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$, which is convex. Either way, the hypograph of $\operatorname{cl} f$ is convex. ∎

According to Rockafellar [130, p. 53], "the closure operation is a reasonable normalization which [sic] makes convex functions more regular by redefining their values at certain points where there are unnatural discontinuities. This is the secret of the great usefulness of the operation in theory and in applications." A more useful characterization of the closure is given in Theorem 170 below. It asserts that for any concave function, its closure is the infimum of all the continuous affine functions that dominate it. (In finite dimensional spaces, every real linear functional, and hence every affine function, is automatically continuous.) That is,

$$\operatorname{cl} f(x) = \inf\{h(x) : h \geqslant f \text{ and } h \text{ is affine and continuous}\}.$$

For if $f(x) = \infty$ for some x, then no affine function dominates f, so $\operatorname{cl} f(x) = \inf \varnothing = \infty$ everywhere. If $f(x) = -\infty$ everywhere, then $\operatorname{cl} f(x) = -\infty$ everywhere as well. The closure of a convex function is the supremum of the continuous affine functions it dominates,

$$\operatorname{cl} f(x) = \inf\{h(x) : f \geqslant h \text{ and } h \text{ is affine and continuous}\}.$$

If it had been up to me, I would have made this the definition of the closure and closedness of concave functions.

Implicit in Rockafellar's remark is that the closure operation only deals with bad behavior on the boundary of the domain. Indeed we have the following result.

156 Theorem *Let* f *be a proper concave function on* $\boldsymbol{R}^{\mathrm{n}}$. *Then* $\operatorname{cl} f$ *is a proper closed concave function, and* f *and* $\operatorname{cl} f$ *agree on* $\operatorname{ri} \operatorname{dom} f$.

See [130, Theorem 7.4, p. 56].

157 Corollary *If* f *is a proper concave or convex function and* $\operatorname{dom} f$ *is affine, then* f *is closed.*

See [130, Corollary 7.4.2, p. 56].

4.8 Separation Theorems

Separating hyperplane theorems are the heart and soul of convex analysis.

Given $p \in \boldsymbol{R}^{\mathrm{n}}$ and $\alpha \in \boldsymbol{R}$, let $[p \geqslant \alpha]$ denote the set $\{x \in \boldsymbol{R}^{\mathrm{n}} : p \cdot x \geqslant \alpha\}$. The sets $[p = \alpha]$, etc., are defined similarly. A **hyperplane** in $\boldsymbol{R}^{\mathrm{n}}$ is a set of the form $[p = \alpha]$, where $p \neq 0$.[3]

[3] In more general linear spaces, a hyperplane is a level set $[f = \alpha]$ of a nonzero real-valued linear function (or functional, as they are more commonly called). If the linear functional is not continuous, the hyperplane is dense. If the function is continuous, then the hyperplane is closed. Open and closed half spaces are topologically open and closed if and only if the functional is continuous.

The vector p can be thought of as a real-valued linear function on $\boldsymbol{R}^{\mathrm{n}}$, or as a vector normal (orthogonal) to the hyperplane at each point. My multiplying p and α by the same nonzero scalar does not change the hyperplane. Note that a hyperplane is an affine subspace.

A **closed half space** is a set of the form $[p \geqslant \alpha]$ or $[p \leqslant \alpha]$, while an open **open half space** is of the from $[p > \alpha]$ or $[p < \alpha]$. We say that nonzero p, or the hyperplane $[p = \alpha]$, **separates** A and B if either $A \subset [p \geqslant \alpha]$ and $B \subset [p \leqslant \alpha]$, or $B \subset [p \geqslant \alpha]$ and $A \subset [p \leqslant \alpha]$. Let us agree to write $p \cdot A \geqslant p \cdot B$ to mean $p \cdot x \geqslant p \cdot y$ for all x in A and y in B.

Note that if a set A belongs to the hyperplane $[p = \alpha]$, then the half spaces $[p \leqslant \alpha]$ and $[p \geqslant \alpha]$ separate A from A, so separation by itself is not very interesting. A better notion is porper separation. Say that nonzero p, or the hyperplane $[p = \alpha]$ **properly separates** A and B if it separates them and it is not the case that $A \cup B \subset [p = \alpha]$, that is, if there exists some x in A and y in B such that $p \cdot x \neq p \cdot y$.

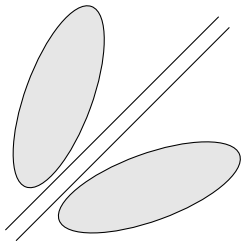

Figure 4.2. Strong separation.

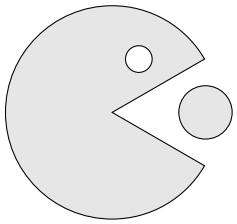

Figure 4.3. These sets cannot be separated by a hyperplane.

There are stronger notions of separation. The hyperplane $[p = \alpha]$ **strictly separates** A and B if A and B are in disjoint open half spaces, that is, $A \subset [p > \alpha]$ and $B \subset [p < \alpha]$ (or vice versa). It **strongly separates** A and B if A and B are in disjoint closed half spaces. That is, there is some $\varepsilon > 0$ such that $A \subset [p \geqslant \alpha + \varepsilon]$ and $B \subset [p \leqslant \alpha]$ (or vice versa). Another way to state strong separation is that $\inf_{x \in A} p \cdot x > \sup_{y \in B} p \cdot y$ (or swap A and B).

158 Lemma *Let A and B be nonempty convex subsets of a topological vector space, and let p be a continuous linear functional. If p properly separates A and B, then it properly separates* $\operatorname{ri} A$ *and* $\operatorname{ri} B$.

Proof: By Proposition 144, we know that $A \subset \overline{\operatorname{ri} A}$ and $B \subset \overline{\operatorname{ri} B}$ and $x \mapsto p \cdot x$ is continuous, if p does not properly separate the relative interiors, that is, if $p \cdot \operatorname{ri} A = p \cdot \operatorname{ri} B$, then $p \cdot A = p \cdot B$, and the separation of A and B is not proper. ∎

Here are some simple results that are used so commonly that they are worth noting.

159 Exercise Let A and B be disjoint nonempty convex subsets of $\boldsymbol{R}^{\mathrm{n}}$ and suppose nonzero p in $\boldsymbol{R}^{\mathrm{n}}$ properly separates A and B with $p \cdot A \geqslant p \cdot B$.

1. If A is a linear subspace, then p annihilates A. That is, $p \cdot x = 0$ for every x in A.
2. If A is a cone, then $p \cdot x \geqslant 0$ for every x in A.
3. If B is a cone, then $p \cdot x \leqslant 0$ for every x in B.
4. If A includes a set of the form $x + \boldsymbol{R}^{\mathrm{n}}_{++}$, then $p > 0$.
5. If B includes a set of the form $x - \boldsymbol{R}^{\mathrm{n}}_{++}$, then $p > 0$.

□

We now come to what I think of as the main result on separation of convex sets. I prove the result for Hilbert spaces of arbitrary dimension, since the proof is not much different from the proof for $\boldsymbol{R}^{\mathrm{n}}$, although it is true in general locally convex spaces.

Recall that a Hilbert space is a vector space with an inner product that induces a complete metric via $d(x,y) = \sqrt{(x-y)\cdot(x-y)}$ and a norm defined by $\|x\| = (x\cdot x)^{1/2}$, called the Euclidean metric and norm. Balls in the Euclidean metric are convex. Any vector p defines a continuous real-valued linear function defined by $x \mapsto p\cdot x$, and conversely: For any continuous real-valued linear function g on a Hilbert space, there is vector p satisfying $g(x) = p\cdot x$. The space $\boldsymbol{R}^{\mathrm{n}}$ is a Hilbert space under its usual inner product. The space ℓ_2 of square summable sequences is an infinite dimensional Hilbert space.

160 Strong Separating Hyperplane Theorem *Let K and C be disjoint nonempty convex subsets of a Hilbert space. Suppose K is compact and C is closed. Then there exists a nonzero p that strongly separates K and C.*

Proof: Define $f\colon K \to \boldsymbol{R}$ by

$$f(x) = \inf\{d(x,y) : y \in C\},$$

that is $f(x)$ is the distance from x to C. The function f is continuous. To see this, observe that for any y, the distance $d(x',y) \leqslant d(x',x) + d(x,y)$, and $d(x,y) \leqslant d(x,x') + d(x',y)$. Thus $|d(x,y) - d(x',y)| \leqslant d(x,x')$, so $|f(x) - f(x')| \leqslant d(x,x')$. Thus f is actually Lipschitz continuous.

Since K is compact, f achieves a minimum on K at some point $\bar{x}$.

I next claim that there is some point $\bar{y}$ in C such that $d(\bar{x},\bar{y}) = f(\bar{x}) = \inf\{d(\bar{x},y : y \in C\}$. That is, $\bar{y}$ achieves the infimum in the definition of f, so the infimum is actually a minimum. The proof of this is subtler than you might imagine (particularly in an arbitrary Hilbert space). To see that such a $\bar{y}$ exists, for each n, let $C_n = \{y \in C : d(\bar{x},y) \leqslant f(\bar{x}) + 1/n\}$. Then each C_n is a nonempty, closed, and convex subset of C, and $C_{n+1} \subset C_n$ for each n. Moreover $\inf\{d(\bar{x},y) : y \in C_n\} = \inf\{d(\bar{x},y) : y \in C\} = f(\bar{x})$, that is, if such a $\bar{y}$ exists, it must belong to C_n for every n. I now claim that $\operatorname{diam} C_n = \sup\{d(y_1,y_2) : y_1, y_2 \in C_n\} \to 0$ as $n \to \infty$. To see this, start with the **parallelogram identity**[4]

$$\|x_1 + x_2\|^2 = 2\|x_1\|^2 - \|x_1 - x_2\|^2 + 2\|x_2\|^2.$$

Now let y_1, y_2 belong to C_n. The distance from $\bar{x}$ to the midpoint of the segment joining y_1, y_2 is given by $d(\bar{x}, \frac{1}{2}y_1 + \frac{1}{2}y_2) = \|\frac{1}{2}(y_1 - \bar{x}) + \frac{1}{2}(y_2 - \bar{x})\|$. Evaluate the parallelogram identity for $x_i = \frac{1}{2}(y_i - \bar{x})$ to get

$$d(\bar{x}, \tfrac{1}{2}y_1 + \tfrac{1}{2}y_2)^2 = \tfrac{1}{2}d(y_1,\bar{x})^2 + \tfrac{1}{2}d(y_2,\bar{x})^2 - \tfrac{1}{4}d(y_1,y_2)^2.$$

so rearranging gives

$$d(y_1,y_2)^2 = 2\big[d(y_1,\bar{x})^2 - d(\bar{x}, \tfrac{1}{2}y_1 + \tfrac{1}{2}y_2)^2\big] + 2\big[d(y_2,\bar{x})^2 - d(\bar{x}, \tfrac{1}{2}y_1 + \tfrac{1}{2}y_2)^2\big]. \tag{4.1}$$

[4]This says that the sum of the squares of the lengths of the diagonals of a parallelogram is equal to the sum of the squares of the lengths of the sides. (Consider the parallelogram with vertices $0, x_1, x_2, x_1 + x_2$. Its diagonals are the segments $[0, x_1 + x_2]$ and $[x_1, x_2]$, and their lengths are $\|x_1 + x_2\|$ and $\|x_1 - x_2\|$. It has two sides of length $\|x_1\|$ and two of length $\|x_2\|$.) To prove this, note that

$$(x_1 + x_2)\cdot(x_1 + x_2) = x_1\cdot x_1 + 2x_1\cdot x_2 + x_2\cdot x_2$$

$$(x_1 - x_2)\cdot(x_1 - x_2) = x_1\cdot x_1 - 2x_1\cdot x_2 + x_2\cdot x_2.$$

Add these two equations and restate in terms of norms.

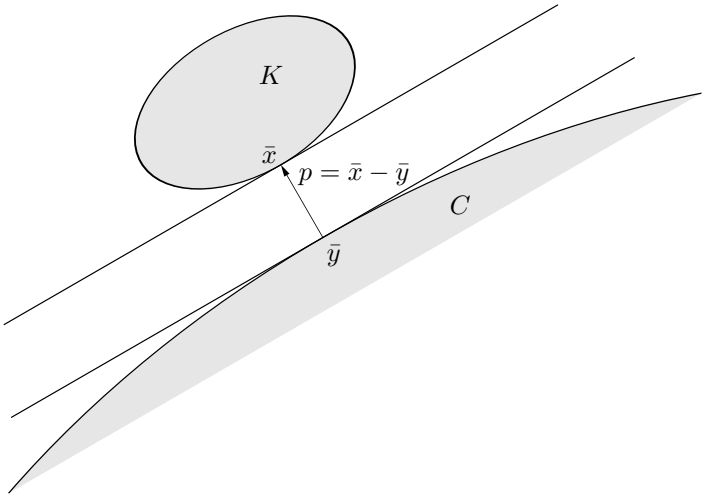

Figure 4.4. Minimum distance and separating hyperplanes.

Now for any point $y \in C_n$, we have $f(\bar{x}) \leqslant d(\bar{x}, y) \leqslant f(\bar{x}) + 1/n$, so $d(\bar{x}, y)^2 - f(\bar{x})^2 \leqslant \big(f(\bar{x}) + 1/n\big)^2 - f(\bar{x})^2 = 2f(\bar{x})/n + 1/n^2$. Now both y_1 and $\frac{1}{2}y_1 + \frac{1}{2}y_2$ belong to C_n, so

$$\begin{aligned} \big|d(y_1, \bar{x})^2 - d(\bar{x}, \tfrac{1}{2}y_1 + \tfrac{1}{2}y_2)^2\big| &= \big|d(y_1, \bar{x})^2 - f(\bar{x})^2 - \big(d(\bar{x}, \tfrac{1}{2}y_1 + \tfrac{1}{2}y_2)^2 - f(\bar{x})^2\big)\big| \\ &\leqslant \big|d(y_1, \bar{x})^2 - f(\bar{x})^2\big| + \big|d(\bar{x}, \tfrac{1}{2}y_1 + \tfrac{1}{2}y_2)^2 - f(\bar{x})^2\big| \\ &\leqslant 2\big(\tfrac{2}{n}f(\bar{x}) + \tfrac{1}{n^2}\big), \end{aligned}$$

and similarly for y_2. Substituting this in (4.1) gives

$$d(y_1, y_2)^2 \leqslant 8\big(\tfrac{2}{n}f(\bar{x}) + \tfrac{1}{n^2}\big) \to 0 \text{ as } n \to \infty,$$

so $\operatorname{diam} C_n \to 0$. In any Hilbert space the Euclidean metric is complete, so by the Cantor Intersection Theorem 22, the intersection $\bigcap_{n=1}^{\infty} C_n$ is a singleton $\{\bar{y}\}$. This $\bar{y}$ has the desired property. Whew! (It also follows that $\bar{y}$ is the unique point satisfying $f(\bar{x}) = d(\bar{x}, \bar{y})$, but we don't need to know that.) Maybe I should make this a separate lemma.

Put $p = \bar{x} - \bar{y}$. See Figure 4.4. Since K and C are disjoint, we must have $p \neq 0$. Then $0 < \|p\|^2 = p \cdot p = p \cdot (\bar{x} - \bar{y})$, so $p \cdot \bar{x} > p \cdot \bar{y}$. What remains to be shown is that $p \cdot \bar{y} \geqslant p \cdot y$ for all $y \in C$ and $p \cdot \bar{x} \leqslant p \cdot x$ for all $x \in K$:

So let y belong to C. Since $\bar{y}$ minimizes the distance (and hence the square of the distance) to $\bar{x}$ over C, for any point $z = \bar{y} + \lambda(y - \bar{y})$ (with $0 < \lambda \leqslant 1$) on the line segment between y and $\bar{y}$ we have

$$(\bar{x} - z) \cdot (\bar{x} - z) \geqslant (\bar{x} - \bar{y}) \cdot (\bar{x} - \bar{y}).$$

Rewrite this as

$$\begin{aligned} 0 &\geqslant (\bar{x} - \bar{y}) \cdot (\bar{x} - \bar{y}) - (\bar{x} - z) \cdot (\bar{x} - z) \\ &= (\bar{x} - \bar{y}) \cdot (\bar{x} - \bar{y}) - \big(\bar{x} - \bar{y} - \lambda(y - \bar{y})\big) \cdot \big(\bar{x} - \bar{y} - \lambda(y - \bar{y})\big) \\ &= (\bar{x} - \bar{y}) \cdot (\bar{x} - \bar{y}) - (\bar{x} - \bar{y}) \cdot (\bar{x} - \bar{y}) + 2\lambda(\bar{x} - \bar{y}) \cdot (y - \bar{y}) - \lambda^2(y - \bar{y}) \cdot (y - \bar{y}) \\ &= 2\lambda(\bar{x} - \bar{y}) \cdot (y - \bar{y}) - \lambda^2(y - \bar{y}) \cdot (y - \bar{y}) \\ &= 2\lambda p \cdot (y - \bar{y}) - \lambda^2(y - \bar{y}) \cdot (y - \bar{y}). \end{aligned}$$

Divide by $\lambda > 0$ to get

$$2p \cdot (y - \bar{y}) - \lambda(y - \bar{y}) \cdot (y - \bar{y}) \leqslant 0.$$

Letting $\lambda \downarrow 0$, we conclude $p \cdot \bar{y} \geqslant p \cdot y$.

A similar argument for $x \in K$ completes the proof. ∎

This proof is a hybrid of several others. The manipulation in the last series of inequalities appears in von Neumann and Morgenstern [158, Theorem 16.3, pp. 134–38], and is probably older. The role of the parallelogram identity in related problems is well known, see for instance, Hiriart-Urruty and Lemaréchal [76, pp. 41, 46] or Rudin [133, Theorem 12.3, p. 293]. A different proof for $\boldsymbol{R}^{\mathrm{n}}$ appears in Rockafellar [130, Corollary 11.4.2, p. 99].

Theorem 160 is true in general locally convex spaces, where p is interpreted as a continuous linear functional and $p \cdot x$ is replaced by $p(x)$. (But remember, compact sets can be rare in such spaces.) Roko and I give a proof of the general case in [3, Theorem 5.79, p. 207], or see Dunford and Schwartz [46, Theorem V.2.10, p. 417].

161 Corollary *Let C be a nonempty closed convex subset of a Hilbert space. Assume that the point x does not belong to C. Then there exists a nonzero p that strongly separates x and C.*

162 Definition *Let C be a set in a topological vector space and x a point belonging to C. The nonzero real-valued linear function p **supports C at x from below** if $p \cdot y \geqslant p \cdot x$ for all $y \in C$, and we may write $p \cdot C \geqslant p \cdot x$. We say that p **supports C at x from above** if $p \cdot y \leqslant p \cdot x$ for all $y \in C$, and we may write $p \cdot x \geqslant p \cdot C$. The hyperplane $\{y : p \cdot y = p \cdot x\}$ is a **supporting hyperplane for C at x**. The support is **proper** if $p \cdot y \neq p \cdot x$ for some y in C. We may also say that the half-space $\{z : p \cdot z \geqslant p \cdot x\}$ supports C at x if p supports C at from below, etc.*

163 Lemma *If p properly supports the convex set C at x, then the relative interior of C does not meet the supporting hyperplane. That is, if $p \cdot C \geqslant p \cdot x$, then $p \cdot y > p \cdot x$ for all $y \in \operatorname{ri} C$.*

Proof: Geometrically, this says that if z is in the hyperplane, and y is on one side, the line through y and z must go through to the other side. Algebraically, let p properly support C at x, say $p \cdot C \geqslant p \cdot x$. Then there exists $y \in C$ with $p \cdot y > p \cdot x$. Let z belong to $\operatorname{ri} C$. By separation $p \cdot z \geqslant p \cdot x$, so suppose by way of contradiction that $p \cdot z = p \cdot x$. Since z is in the relative interior of C, there is some $\varepsilon > 0$ such that $z + \varepsilon(z - y)$ belongs to C. Then $p \cdot \big(z + \varepsilon(z - y)\big) = p \cdot x - \varepsilon p \cdot (x - y) < p \cdot x$, a contradiction. ∎

164 Finite Dimensional Supporting Hyperplane Theorem *Let C be a convex subset of $\boldsymbol{R}^{\mathrm{n}}$ and let $\bar{x}$ belong to C. Then there is a hyperplane properly supporting C at $\bar{x}$ if and ony if $\bar{x} \notin \operatorname{ri} C$.*

Proof of Theorem 164: ($\Longrightarrow$) This is just Lemma 163.

($\Longleftarrow$) Without loss of generality, we can translate C by $-\bar{x}$, and thus assume $\bar{x} = 0$.

Assume $0 \notin \operatorname{ri} C$. (This implies that C is not a singleton, and also that $C \neq \boldsymbol{R}^{\mathrm{n}}$.) Define

$$A = \bigcup_{\lambda > 0} \lambda \operatorname{ri} C.$$

Clearly $\varnothing \neq \operatorname{ri} C \subset A$, $0 \notin A$ but $0 \in \overline{A}$, and A is a deleted cone. More importantly, A is convex (cf. Exercise 129), and A lies in the span of $\operatorname{ri} C$ (cf. Proposition 144).

Since $\boldsymbol{R}^{\mathrm{n}}$ is finite dimensional there exists a finite maximal collection of linearly independent vectors $v_1, \dots, v_k$ that lie in $\operatorname{ri} C$. Since $\operatorname{ri} C$ contains at least one nonzero point, we have $k \geqslant 1$. Let $v = \sum_{i=1}^{k} v_i$, and note that $(1/k)v$ belongs to $\operatorname{ri} C$. I claim that $-v \notin \overline{A}$.

To see this, assume by way of contradiction that $-v$ belongs to $\overline{A}$. Thus, there exists a sequence $\{x_n\}$ in A satisfying $x_n \to -v$. Since $v_1, \dots, v_k$ is a maximal independent set, we must

be able to write $x_n = \sum_{i=1}^k \lambda_i^n v_i$. By Corollary 49, $\lambda_i^n \xrightarrow[n\to\infty]{} -1$ for each i. In particular, for some n we have $\lambda_i^n < 0$ for each i. Now if for this n we let $\lambda = \sum_{i=1}^p \lambda_i^n < 0$, then

$$0 = \tfrac{1}{1-\lambda} x_n + \sum_{i=1}^m \left(-\tfrac{\lambda_i^n}{1-\lambda}\right) v_i \in A, \quad \text{as } A \text{ is convex,}$$

which is a contradiction. Hence $-v \notin \overline{A}$.

Now by Corollary 161 there exists some nonzero p strongly separating $-v$ from $\overline{A}$. That is, $p \cdot (-v) < p \cdot y$ for all $y \in \overline{A}$. Moreover, since $\overline{A}$ is a cone, $p \cdot y \geqslant 0 = p \cdot 0$ for all $y \in \overline{A}$, and $p \cdot (-v) < 0$ (Exercise 159). Thus p supports $\overline{A} \supset C$ at 0. Moreover, $p \cdot (1/k)v > 0$, so p properly supports C at 0. ∎

The next theorem yields only proper separation but requires only that the sets in question have disjoint relative interiors. In particular it applies whenever the sets themselves are disjoint. It is a strictly finite-dimensional result.

165 Finite Dimensional Separating Hyperplane Theorem *Two nonempty convex subsets of $\boldsymbol{R}^{\mathrm{n}}$ can be properly separated by a hyperplane if and only their relative interiors are disjoint.*

Proof: ($\Longleftarrow$) Let A and B be nonempty convex subsets of $\boldsymbol{R}^{\mathrm{n}}$ with $\operatorname{ri} A \cap \operatorname{ri} B = \varnothing$. Put $C = A - B$. By Proposition 146 $\operatorname{ri} C = \operatorname{ri} A - \operatorname{ri} B$, so $0 \notin \operatorname{ri} C$. It suffices to show that there exists some nonzero $p \in \boldsymbol{R}^{\mathrm{n}}$ satisfying $p \cdot x \geqslant 0$ for all $x \in C$, and $p \cdot y > 0$ for some $y \in C$. If $0 \notin \overline{C}$, this follows from Corollary 161. If $0 \in \overline{C}$, it follows from Theorem 164.

($\Longrightarrow$) If p properly separates A and B, then the same argument used in the proof of Theorem 164 shows that $\operatorname{ri} A \cap \operatorname{ri} B = \varnothing$. ∎

Now I'll state without proof some general theorems that apply in infinite dimensional spaces.

166 Infinite Dimensional Supporting Hyperplane Theorem *If C is a convex set with nonempty interior in a topological vector space, and x is a boundary point of C, then there is a nonzero continuous linear functional properly supporting C at x.*

For a proof see [3, Lemma 7.7, p. 259]. Properness of the support is not shown there, but it isi easy to verify. If the set has an empty interior, then it may fail to have supporting closed hyperplanes at boundary points. For example, in ℓ_1, the positive cone cannot be supported by a continuous linear functional at any strictly positive sequence, see [3, Example 7.8, p. 259]. In Banach spaces however we have the following result, the proof of which is in [3, Theorem 7.43, p. 284].

167 Bishop–Phelps Theorem *Let C be a nonempty closed convex subset of a Banach space. Then the set of points at which C is supported by a nonzero continuous linear functional is dense in the boundary of C.*

Finally, Theorem 166 can be used to prove the following.

168 Infinite Dimensional Separating Hyperplane Theorem *Two disjoint nonempty convex subsets of a topological vector space can be properly separated by a closed hyperplane (or continuous linear functional) if one of them has a nonempty interior.*

As an application we have the following result due to Fan, Glicksberg, and Hoffman [50].

169 Concave Alternative Theorem *Let C be a nonempty convex subset of a vector space, and let $f^1, \dots, f^m \colon C \to \boldsymbol{R}$ be concave. Letting $f = (f_1, \dots, f_m) \colon C \to \boldsymbol{R}^{\mathrm{m}}$, exactly one of the following is true.*

$$\exists \bar{x} \in C \quad f(\bar{x}) \gg 0. \tag{4.2}$$

Or (exclusive),

$$\exists p > 0 \ \forall x \in C \quad p \cdot f(x) \leqslant 0. \tag{4.3}$$

Proof: Clearly both cannot be true. Suppose (4.2) fails. Set

$$H = \{f(x) : x \in C\} \quad \text{and set} \quad \hat{H} = \{y \in \boldsymbol{R}^{\mathrm{m}} : \exists x \in C \quad y \leqq f(x)\}.$$

Since (4.2) fails, we see that H and $\boldsymbol{R}^{\mathrm{m}}_{++}$ are disjoint. Consequently $\hat{H}$ and $\boldsymbol{R}^{\mathrm{m}}_{++}$ are disjoint. Now observe that $\hat{H}$ is convex. To see this, suppose $y^1, y^2 \in \hat{H}$. Then $y^i \leqq f(x^i)$, $i = 1, 2$. Therefore, for any $\lambda \in (0, 1)$,

$$\lambda y^1 + (1 - \lambda) y^2 \leqq \lambda f(x^1) + (1 - \lambda) f(x^2) \leqq f\big(\lambda x^1 + (1 - \lambda) x^2\big),$$

since each f^j is concave. Therefore $\lambda y^1 + (1 - \lambda) y^2 \in \hat{H}$.

Thus, by the Separating Hyperplane Theorem 165, there is a nonzero vector $p \in \boldsymbol{R}^{\mathrm{m}}$ properly separating $\hat{H}$ and $\boldsymbol{R}^{\mathrm{m}}_{++}$. We may assume

$$p \cdot \hat{H} \leqslant p \cdot \boldsymbol{R}^{\mathrm{m}}_{++}. \tag{4.4}$$

By Exercise 159, $p > 0$. Evaluating (4.4) at $z = \varepsilon \mathbf{1}$ for $\varepsilon > 0$, we get $p \cdot y \leqslant \varepsilon p \cdot \mathbf{1}$. Since ε may be taken arbitrarily small, we conclude that $p \cdot y \leqslant 0$ for all y in $\hat{H}$. In particular, $p \cdot f(x) \leqslant 0$ for all x in C. ∎

4.9 Hyperplanes in $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$ and affine functions

Sometimes it is useful to think of the Euclidean space $\boldsymbol{R}^{n+1}$ as the product $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$. For instance, if $f \colon \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}$, it is natural to view the graph as a subset of the domain × range, $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$. I will refer to a typical element in $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$ as a point (x, α) where $x \in \boldsymbol{R}^{\mathrm{n}}$ and $\alpha \in \boldsymbol{R}$. I may call x the "vector component" and α the "real component," even when $n = 1$. A hyperplane in $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$ is defined in terms of its normal vector (p, λ). If the real component $\lambda = 0$, we say the hyperplane is **vertical**. If the hyperplane is not vertical, by homogeneity we can arrange for λ to be -1 (you will see why in just a moment).

Non-vertical hyperplanes in $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$ are precisely the graphs of affine functions on $\boldsymbol{R}^{\mathrm{n}}$. That is,

$$\operatorname{gr}\big(x \mapsto p \cdot x - \beta\big) \quad = \quad \text{the non-vertical hyperplane } \big\{(x, \alpha) \in \boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R} : (p, -1) \cdot (x, \alpha) = \beta\big\}.$$

And the non-vertical hyperplane

$$\big\{(x, \alpha) \in \boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R} : (p, \lambda) \cdot (x, \alpha) = \beta\big\} \quad \text{where } \lambda \neq 0 \quad = \quad \operatorname{gr}\big(x \mapsto -(1/\lambda) p \cdot x + \beta/\lambda\big).$$

4.10 Closed functions revisited

In this section, we give a more useful characterization of the closure of a concave (or convex) function. Recall that the real-valued function g **dominates** the real-valued function f on X, written $g \geqslant f$, if for every $x \in X$ we have $g(x) \geqslant f(x)$.

170 Theorem *Let $f\colon \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}^\sharp$ be concave (not necessarily proper). Then for every $x \in \boldsymbol{R}^{\mathrm{n}}$,*

$$\operatorname{cl} f(x) = \inf\{h(x) : h \geqslant f \text{ and } h \text{ is affine and continuous}\}.$$

If $f\colon \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}^\sharp$ is convex, then for every $x \in \boldsymbol{R}^{\mathrm{n}}$,

$$\operatorname{cl} f(x) = \sup\{h(x) : f \geqslant h \text{ and } h \text{ is affine and continuous}\}.$$

Proof: I shall prove the concave case. There are three subcases. If f is improper and assumes the value $+\infty$, then by definition $\operatorname{cl} f$ is the constant function $+\infty$. In this case, no affine function, which is (finite) real-valued, dominates f so the infimum is over the empty set, and thus $+\infty$. The second subcase is that f is the improper constant function $-\infty$. In this case every affine function dominates f, so the infimum is $-\infty$.

So assume we are in the third subcase, namely that f is proper. That is, $f(x) < \infty$ for all $x \in \boldsymbol{R}^{\mathrm{n}}$, and $\operatorname{dom} f$ is nonempty. Then by definition $\operatorname{cl} f$ is the upper envelope of f. That is,

$$\operatorname{cl} f(x) = \inf_{\varepsilon>0} \sup_{d(y,x)<\varepsilon} f(y).$$

Define $g(x) = \inf\{h(x) : h \geqslant f \text{ and } h \text{ is affine and continuous}\}$. If h is affine, continuous, and dominates f, then by Theorem 33, h dominates $\operatorname{cl} f$, so g dominates $\operatorname{cl} f$.

We now show that $\operatorname{cl} f \geqslant g$. It suffices to show that for any (x, α) with $\alpha > \operatorname{cl} f(x)$, there is an affine function h dominating f with $h(x) \leqslant \alpha$. Now $\alpha > \operatorname{cl} f(x) = \limsup_{y\to x} f(y)$ implies that (x, α) does not belong to the closure of the hypograph of f.

There are two cases to consider. The simpler case is that x belongs to $\operatorname{dom} f$. So assume now that $x \in \operatorname{dom} f$. Since f is concave, its hypograph and the closure thereof are convex, and since f is proper, its hypograph is nonempty. So by Corollary 161 there is a nonzero $(p, \lambda) \in \boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$ strongly separating (x, α) from the closure of the hypograph of f. In particular, for each $y \in \operatorname{dom} f$ the point $\big(y, f(y)\big)$ belongs to the hypograph of f. Thus strong separation implies that for some $\varepsilon > 0$, for any $y \in \operatorname{dom} f$,

$$p \cdot x + \lambda\alpha > p \cdot y + \lambda f(y) + \varepsilon. \tag{4.5}$$

Theorem 186 comes later!

The same argument as that in the proof of Theorem 186 shows that $\lambda \geqslant 0$. Moreover, taking $y = x$ (since $x \in \operatorname{dom} f$) shows that $\lambda \neq 0$. So dividing by λ gives

$$(1/\lambda)p \cdot (x - y) + \alpha > f(y) + (\varepsilon/\lambda)$$

for all $y \in \operatorname{dom} f$. Define

$$h(y) = (1/\lambda)p \cdot (x - y) + \alpha.$$

Then h is a continuous affine function satisfying

$$h(y) > f(y) + \eta \quad \text{for all } y \in \operatorname{dom} f,$$

where $\eta = (\varepsilon/\lambda) > 0$ and $h(x) = \alpha$, as desired.

The case where (x, α) satisfies $\alpha > \operatorname{cl} f(x)$, but $x \notin \operatorname{dom} f$ is more subtle. The reason the above argument does not work is that the hyperplane may be vertical ($\lambda = 0$), and hence not the graph of any affine function. So assume that $\lambda = 0$. Then (4.5) becomes

$$p \cdot x > p \cdot y + \varepsilon$$

for all $y \in \operatorname{dom} f$. Define the continuous affine function g by

$$g(y) = p \cdot (x - y) - \varepsilon/2,$$

and note that $g(x) < 0$, and $g(y) > 0$ for all $y \in \operatorname{dom} f$.

But we still have (from the above argument) a continuous affine function h satisfying

$$h(y) > f(y) \quad \text{for all } y \in \operatorname{dom} f.$$

Now for any $\gamma > 0$, we have

$$\gamma g(y) + h(y) > f(y) \quad \text{for all } y \in \operatorname{dom} f,$$

and for $y \notin \operatorname{dom} f$, $f(y) = -\infty$, so the inequality holds for all y in $\boldsymbol{R}^{\mathrm{n}}$. But since $g(x) < 0$, for γ large enough, $\gamma g(x) + h(x) < \alpha$, so this is the affine function we wanted.

I think that covers all the bases (and cases).

The case of a convex function is dealt with by replacing the epigraph with the hypograph and reversing inequalities. ∎

In light of Lemma 150, we have the following.

171 Corollary *An upper semicontinuous proper concave function is the pointwise infimum of the continuous affine functions that dominate it.*

A lower semicontinuous proper convex function is the pointwise supremum of the continuous affine functions that it dominates.

4.11 Sublinear functions

172 Definition *A function f from a (possibly deleted) convex cone C in a real vector space into $\boldsymbol{R}^\sharp$ is*

positively homogeneous *(of degree 1) if for every vector $x \in C$ and every real $\lambda > 0$,*

$$f(\lambda x) = \lambda f(x).$$

subadditive *if for all vectors x and y in C,*

$$f(x + y) \leqslant f(x) + f(y).$$

superadditive *if for all vectors x and y in C,*

$$f(x + y) \geqslant f(x) + f(y).$$

sublinear *if it is both positively homogeneous and subadditive.*

By these definitions we ought to say that f is **superlinear** if it is both positively homogeneous and superadditive, but no one does. Note that for a positively homogeneous fuction, since $\lambda 0 = 0$ for all real λ we must have $f(0) = 0$ or $f(0)$ is infinite.

173 Exercise A positively homogeneous function is subadditive if and only it is convex. It is superadditive if and only if it concave.

The hypograph of a positively homogeneous concave function is (possibly deleted) convex cone. The epigraph of a sublinear (positively homogeneous convex) function is (possibly deleted) convex cone. □

4.12 Support functions

The Separating Hyperplane Theorem 160 is the basis for a number of results concerning closed convex sets. Given any set A in $\boldsymbol{R}^n$ its **closed convex hull**, denoted $\overline{\text{co}}\, A$, is the intersection of all closed convex sets that include A. That is,

$$\overline{\text{co}}\, A = \bigcap \{C : A \subset C \text{ and } C \text{ is closed and convex}\}.$$

It is of course the smallest closed convex set that includes A. If A is empty, then it is closed and convex, so $\overline{\text{co}}\, A$ is empty. If A is nonempty, then $\overline{\text{co}}\, A$ is nonempty since $\boldsymbol{R}^n$ itself is closed and convex. Less obvious is the following.

174 Theorem *Let A be a subset of $\boldsymbol{R}^n$. Then*

$$\overline{\text{co}}\, A = \bigcap \{H : A \subset H \text{ and } H \text{ is a closed half space}\}.$$

In particular, a closed convex set is the intersection of all the closed half spaces that include it.

Proof: Clearly $\overline{\text{co}}\, A$ is included in the intersection since every closed half space is also a closed convex set. It is also clear that the result is true for $A = \varnothing$. So assume A, and hence $\overline{\text{co}}\, A$, is nonempty.

It suffices to show that if $x \notin \overline{\text{co}}\, A$, then there is a closed half space that includes $\overline{\text{co}}\, A$ but does not contain x. By the Separating Hyperplane Theorem 160 there is a nonzero p that strongly separates the closed convex set $\overline{\text{co}}\, A$ from the compact convex set $\{x\}$. But this clearly implies that there is a closed half space of the form $[p \geqslant \alpha]$ that includes $\overline{\text{co}}\, A$, but doesn't contain x. ∎

The **support function** μ_A of a set A is a handy way to summarize all the closed half spaces that included A. It is defined by[5]

$$\mu_A(p) = \inf\{p \cdot x : x \in A\}.$$

We allow for the case that $\mu_A(p) = -\infty$. Note that $\mu_\varnothing$ is the improper concave function $+\infty$. Also note that the infimum may not actually be attained even if it is finite. For instance, consider the closed convex set $A = \{(x, y) \in \boldsymbol{R}^2_{++} : xy \geqslant 1\}$, and let $p = (0, 1)$. Then $\mu_A(p) = 0$ even though $p \cdot (x, y) = y > 0$ for all $(x, y) \in A$. If A is compact, then of course μ_A is always finite, and there is some point in A where the infimum is actually a minimum.

Theorem 174 immediately implies yields the following description of $\overline{\text{co}}\, A$ in terms of μ_A.

175 Theorem *For any set A in $\boldsymbol{R}^n$,*

$$\overline{\text{co}}\, A = \{x \in \boldsymbol{R}^n : \forall p \in \boldsymbol{R}^n \; p \cdot x \geqslant \mu_A(p)\}.$$

Moreover, $\mu_A = \mu_{\overline{\text{co}}\, A}$.

Proof: Observe that

$$C := \{x \in \boldsymbol{R}^n : \forall p \in \boldsymbol{R}^n \; p \cdot x \geqslant \mu_A(p)\} = \bigcap \{[p \geqslant \mu_A(p)] : p \in \boldsymbol{R}^n\}$$

is an intersection of closed half spaces. By definition, if $x \in A$, then $p \cdot x \geqslant \mu_A(p)$, that is, $A \subset [p \geqslant \mu_A(p)]$. Thus by Theorem 174, $\overline{\text{co}}\, A \subset C$.

For the reverse inclusion, suppose $x \notin \overline{\text{co}}\, A$. By the Separating Hyperplane Theorem 160 there is a nonzero p such $\overline{\text{co}}\, A \subset [p \geqslant \alpha]$ and $p \cdot x < \alpha$. Since $A \subset \overline{\text{co}}\, A$ we have $\mu_A(p) = \inf\{p \cdot x : x \in A\} > p \cdot x$, so $x \notin C$.

To see that $\mu_A = \mu_{\overline{\text{co}}\, A}$ first note that $\mu_A \geqslant \mu_{\overline{\text{co}}\, A}$ since $A \subset \overline{\text{co}}\, A$. The first part of the theorem implies $\mu_{\overline{\text{co}}\, A} \geqslant \mu_A$. ∎

[5] Fenchel [51] and Roko and I [3, p. 288] define $h_A(p) = \sup\{p \cdot x : x \in A\}$, which makes it convex rather than concave, and $h_A(p) = -\mu_A(-p)$. The definition in these notes follows Mas-Colell, Whinston, and Green [108], and may be more useful to economists.

176 Lemma *The support function μ_A is concave and positively homogeneous of degree 1, that is, $\mu_A(\lambda p) = \lambda\mu_A(p)$ for all p and all $\lambda \geqslant 0$.*

Proof: Each x defines a linear (and therefore concave) function ℓ_x via $\ell_x\colon p \mapsto p \cdot x$. Thus by Exercise 135 (4), $\mu_A = \inf_{x\in A} \ell_x$ is concave. Homogeneity is obvious. ∎

The next result is the concave version of Rockafellar's [130, p. 114] Corollary 13.2.1.

177 Theorem *Let f be a positively homogeneous concave function on $\boldsymbol{R}^{\mathrm{n}}$. Then the closure of f is the support function μ_C of the closed convex set $C = \{p : \forall y\ p \cdot y \geqslant f(y)\}$.*

Proof: If $f(x) = \infty$ for some x, then $\operatorname{cl} f$ is identically ∞, and so is μ_C, as C is empty. If f is identically $-\infty$, so is $\operatorname{cl} f$, and so is μ_C as $C = \boldsymbol{R}^{\mathrm{n}}$. This leaves the case where f is proper, so $f(x)$ is finite for some x.

By Theorem 170, $\operatorname{cl} f(x) = \inf\{g(x) : g \geqslant f,\ g \text{ affine}\}$, and by definition, $\mu_C(x) = \inf\{p \cdot x : \forall y\ p \cdot y \geqslant f(y)\}$. Since $p\colon y \mapsto p \cdot y$ is affine, it suffices to prove that if the affine function $g\colon y \mapsto p \cdot y + \alpha$ dominates f, then the linear function $p\colon y \mapsto p \cdot y$ satisfies $g \geqslant p \geqslant f$.

Since g dominates f, and $f(x)$ is finite, we have $g(\lambda x) \geqslant f(\lambda x) = \lambda f(x)$ for all $\lambda > 0$. Letting $\lambda \downarrow 0$ we have $\alpha = \lim_{\lambda\downarrow 0} g(\lambda x) \geqslant \lim_{\lambda\downarrow 0} \lambda f(x) = 0$, so $\alpha \geqslant 0$. That is $g \geqslant p$. Moreover, for all y we have $\lambda p \cdot y + \alpha = g(\lambda y) \geqslant f(\lambda y) = \lambda f(y)$, so dividing by $\lambda > 0$ gives $p \cdot x + (\alpha/\lambda) \geqslant f(x)$, for all y. Letting $\lambda \to \infty$, we must have $p \cdot y \geqslant f(y)$ for all y. That is, $p \geqslant f$.

This shows that $\operatorname{cl} f = \mu_C$. ∎

4.13 The superdifferential of a concave function

There is a useful way to characterize the concavity of differentiable functions.

178 Theorem (Concave functions lie below tangent lines) *Suppose f is concave on a convex neighborhood $C \subset \boldsymbol{R}^{\mathrm{n}}$ of x, and differentiable at x. Then for every y in C,*

$$f(x) + f'(x) \cdot (y - x) \geqslant f(y). \tag{4.6}$$

Proof: Let $y \in C$. Rewrite the definition of concavity as

$$f\big(x + \lambda(y - x)\big) \geqslant f(x) + \lambda\big(f(y) - f(x)\big).$$

Rearranging and dividing by $\lambda > 0$,

$$\frac{f\big(x + \lambda(y - x)\big) - f(x)}{\lambda} \geqslant f(y) - f(x).$$

Letting $\lambda \downarrow 0$, the left hand side converges to $f'(x) \cdot (y - x)$. ∎

The converse is true as the following clever argument shows.

179 Theorem *Let f be differentiable on a convex open set $U \subset \boldsymbol{R}^{\mathrm{n}}$. Suppose that for every x and y in C, we have $f(x) + f'(x) \cdot (y - x) \geqslant f(y)$. Then f is concave.*

Proof: For each $x \in C$, define the function h_x by $h_x(y) = f(x) + f'(x) \cdot (y - x)$. Each h_x is concave, $f \leqslant h_x$ for each x, and $f(x) = h_x(x)$. Thus

$$f = \inf_{x\in C} h_x,$$

so by Exercise 135 (4), f is concave. ∎

Theorem 237 below provides a powerful generalization of this result.

180 Definition *Let $f\colon \boldsymbol{R}^n \to \boldsymbol{R}$ be concave. A vector p is a* ***supergradient*** *of f at the point x if for every y it satisfies the* ***supergradient inequality****,*

$$f(x) + p \cdot (y - x) \geqslant f(y).$$

Similarly, if f is convex, then p is a ***subgradient*** *of f at x if*

$$f(x) + p \cdot (y - x) \leqslant f(y)$$

for every y.

For concave f, the set of all supergradients of f at x is called the ***superdifferential*** *of f at x, and is denoted $\partial f(x)$. If the superdifferential is nonempty at x, we say that f is* ***superdifferentiable*** *at x. For convex f the same symbol $\partial f(x)$ denotes the set of subgradients and is called the* ***subdifferential****. If it is nonempty we say that f is* ***subdifferentiable****.*[6]

There is an equivalent geometric interpretation of the super/subgradients.

181 Lemma *If f is concave and $f(x)$ is finite, the vector p is a supergradient of f at x if and only if $(p, -1)$ supports the hypograph of f at $\big(x, f(x)\big)$ from below, that is*

$$(p, -1) \cdot \big(x, f(x)\big) \leqslant (p, -1)(y, \alpha) \qquad \text{for all } (y, \alpha) \in \operatorname{hypo} f.$$

The vector p is a subgradient of the convex function f at x if and only if $(p, -1)$ supports the epigraph of f at $\big(x, f(x)\big)$ from above, that is

$$(p, -1) \cdot \big(x, f(x)\big) \geqslant (p, -1)(y, \alpha) \qquad \text{for all } (y, \alpha) \in \operatorname{epi} f.$$

N.B. If $(p, -1)$ supports the hypograph of f from below, the supporting hyperplane lies *above* the hypograph in the usual sense. The word "below" refers to the fact that the number $(p, -1) \cdot \big(x, f(x)\big)$ lies below the numbers $(p, -1)(y, \alpha)$. Similarly, if $(p, -1)$ supports the epigraph of f from above, the supporting hyperplane lies below the epigraph.

Proof: I demonstrate only the concave case.

($\Longrightarrow$) Assume that p is a supergradient and let $\alpha \leqslant f(y)$. Then $f(x)+p\cdot(y-x) \geqslant f(y) \geqslant \alpha$. Multiplying by -1 and rearranging this yields

$$p \cdot x - f(x) \leqslant p \cdot y - \alpha,$$

or

$$(p, -1) \cdot \big(x, f(x)\big) \leqslant (p, -1) \cdot (y, \alpha) \qquad \text{whenever } f(y) \geqslant \alpha.$$

That is, $(p, -1)$ supports the hypograph of f at $\big(x, f(x)\big)$ from below.

($\Longleftarrow$) This follows by reversing the steps above. ∎

Here are some simple results.

182 Lemma *If a concave function f is superdifferentiable at a point x with $f(x)$ finite, then f is proper.*

Proof: If $f(x)$ is finite and f is superdifferentiable at x, the supergraient inequality implies that f does not assume the value $+\infty$. The effective domain contains x, so f is indeed proper. ∎

[6] Rockafellar [130, p. 308] suggests this terminology as being more appropriate than the terminology he actually uses, so I shall use it. He uses the term subgradient to mean both subgradient and supergradient, and subdifferential to mean both subdifferential and superdifferential.

183 Lemma *For a proper concave function f, if x does not belong to $\operatorname{dom} f$, then f is not superdifferentiable at x.*

Proof: If $x \notin \operatorname{dom} f$, so that $f(x) = -\infty$ and $y \in \operatorname{dom} f$, then no p can satisfy the supergradient inequality. ∎

The definition is potentially inconsistent for affine functions, which are both concave and convex, but it isn't thanks to the following result.

184 Lemma *The affine function $f(x) = p \cdot x + \alpha$ satisfies $\partial f(x) = \{p\}$, whether f is viewed as concave or convex.*

Proof: Clearly p satisfies both the supergradient and subgradient inequalities. Now suppose q satisfies the supergradient inequality $p \cdot x + \alpha + q \cdot (y - x) \geqslant p \cdot y + \alpha$ for all y. Pick any v and set $y = x + v$ and conclude $q \cdot v \geqslant p \cdot v$, and do the same for $-v$. This shows that $(p - q) \cdot v = 0$ for all v, so $q = p$. Thus p is the unique solution of the supergradient inequality. Ditto for the subgradient inequality. ∎

It is clear that if f is either concave or convex, then

$$\partial(-f)(x) = -\partial f(x),$$

where ∂ indicates the superdifferential when preceding a concave function and the subdifferential when preceding a convex function.

Theorem 178 clearly implies that the gradient of a concave function at a point of differentiability is also a supergradient. The gradient of a convex function at a point of differentiability is also a subgradient. In fact, if $\partial f(x)$ is a singleton, then f is differentiable at x and $\partial f(x) = \{f'(x)\}$, see Theorem 219 below.

185 Lemma *The superdifferential of a concave function (or the subdifferential of a convex function) at a point is a (possibly empty) closed convex set.*

Proof: This is immediate since it is the set of solutions to a system of weak linear inequalities, one for each y. ∎

Concave functions are superdifferentiable at relative interior points.

186 Theorem (Superdifferentiability) *A proper concave (or convex) function on $\boldsymbol{R}^{\mathrm{n}}$ is superdifferentiable at each point of the relative interior of its effective domain.*

Proof: Let f be a proper concave function, and let x belong to $\operatorname{ri}\operatorname{dom} f$. Observe that $\big(x, f(x)\big)$ belongs to the hypograph of f, but not to its relative interior. Since the hypograph is convex, the Supporting Hyperplane Theorem 164 asserts that there is a nonzero $(p, \lambda) \in \boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$ properly supporting the hypograph at $\big(x, f(x)\big)$ from below. That is,

$$p \cdot x + \lambda f(x) \leqslant p \cdot y + \lambda\alpha \qquad \text{for all } y \in \operatorname{dom} f \text{ and all } \alpha \leqslant f(y). \tag{4.7}$$

I claim that $\lambda < 0$: By considering very negative values of α, we conclude that $\lambda \leqslant 0$. Suppose momentarily that $\lambda = 0$. Since x belongs to the relative interior of $\operatorname{dom} f$, for any z in $\operatorname{dom} f$ there is some $\varepsilon > 0$ such that $x \pm \varepsilon(x - z)$ belong to $\operatorname{dom} f$. Then (4.7) (with $y = x \pm \varepsilon(x - z)$) implies $p \cdot (x - z) = 0$. Thus $(p, 0) \cdot (z, \alpha) = (p, 0) \cdot \big(x, f(x)\big)$ for all $(z, \alpha) \in \operatorname{hypo} f$. But this contradicts the properness of the support at $\big(x, f(x)\big)$. Therefore $\lambda < 0$.

Dividing (p, λ) by $-\lambda > 0$ implies that $\big((-1/\lambda)p, -1\big)$ also supports the hypograph from below, so $(-1/\lambda)p$ is a supergradient by Lemma 181. ∎

Non-superdifferentiability can however occur on the boundary of the domain.

187 Example (A non-superdifferentiable point) Define $f\colon [0,1] \to [0,1]$ by $f(x) = x^{\frac{1}{2}}$. Then f is clearly concave, but $\partial f(0) = \varnothing$, since the supergradient inequality implies $p \cdot x \geqslant f(x) - f(0) = x^{\frac{1}{2}}$, so $p \geqslant (\frac{1}{x})^{\frac{1}{2}}$ for all $0 < x \leqslant 1$. Clearly no real number p fills the bill. □

Following Fenchel [51] and Rockafellar [130], define the one-sided directional derivative

$$f'(x;v) = \lim_{\lambda \downarrow 0} \frac{f(x+\lambda v) - f(x)}{\lambda},$$

allowing the values ∞ and $-\infty$. (Phelps [125] uses the notation $d^+(x)(v)$.)

In Example 187, $f'(0,1) = \infty$, that is, the graph of the function becomes arbitrarily steep as we approach the boundary. This is the only way superdifferentiability fails. I prove it in Corollary 216 below.

4.14 Maxima of concave functions

Talk about minima as well. An interior minimum implies constancy.

Concave functions have two important properties. One is that any local maximum is a global maximum. The other is that first order conditions are sufficient as well as necessary for a maximum.

188 Theorem (Concave local maxima are global) *Let $f\colon C \to \boldsymbol{R}$ be a concave function (C convex). If x^* is a local maximizer of f, then it is a global maximizer of f over C.*

Proof: Let x belong to C. Then for small $\lambda > 0$, $f(x^*) \geqslant f\big(\lambda x + (1-\lambda)x^*\big)$. (Why?) By the definition of concavity,

$$f\big(\lambda x + (1-\lambda)x^*\big) \geqslant \lambda f(x) + (1-\lambda) f(x^*).$$

Thus $f(x^*) \geqslant \lambda f(x) + (1-\lambda) f(x^*)$, which implies $f(x^*) \geqslant f(x)$. ∎

189 Corollary *If f is strictly concave, a local maximum is a strict global maximum.*

190 Theorem (First order conditions for concave functions) *Suppose f is concave on a convex set $C \subset \boldsymbol{R}^{\mathrm{n}}$. A point x^* in C is a global maximum point of f if and only 0 belongs to the superdifferential $\partial f(x^*)$.*

Proof: Note that x^* is a global maximum point of f if and only if

$$f(x^*) + 0 \cdot (y - x^*) \geqslant f(y)$$

for all y in C, but this is just the supergradient inequality for 0. ∎

In particular, this result shows that f is superdifferentiable at any maximum point, even if it is not an interior point. The next result is immediate.

191 Corollary *Suppose f is concave on a convex neighborhood $C \subset \boldsymbol{R}^{\mathrm{n}}$ of x^*, and differentiable at x^*. If $f'(x^*) = 0$, then f has a global maximum over C at x^*.*

Note that the conclusion of Theorem 188 does not hold for quasiconcave functions. For instance,

$$f(x) = \begin{cases} 0 & x \leqslant 0 \\ x & x \geqslant 0, \end{cases}$$

has a local maximum at -1, but it is not a global maximum over $\boldsymbol{R}$. However, if f is explicitly quasiconcave, then we have the following.

Oops! explicit quasiconcavity is defined later.

192 Theorem (Local maxima of explicitly quasiconcave functions) *Let $f\colon C \to \boldsymbol{R}$ be an explicitly quasiconcave function (C convex). If x^* is a local maximizer of f, then it is a global maximizer of f over C.*

Proof: Let x belong to C and suppose $f(x) > f(x^*)$. Then by the definition of explicit quasiconcavity, for any $1 > \lambda > 0$, $f\big(\lambda x + (1-\lambda)x^*\big) > f(x^*)$. Since $\lambda x + (1-\lambda)x^* \to x^*$ as $\lambda \to 0$ this contradicts the fact that f has a local maximum at x^*. ∎

4.15 Supergradient of a support function

****

If the infimum of p is actually achieved at a point in A, we can say more. By Theorem 175 we might as well assume that A is closed and convex.

193 Theorem *Let C be a closed convex set. Then x is a supergradient of the support function μ_C at p if and only if x belongs to C and minimizes p over C. In other words,*

$$\partial\mu_C(p) = \{x \in C : p \cdot x = \mu_C(p)\}.$$

Proof: Recall that the supergradient inequality for this case is

$$\mu_C(p) + x \cdot (q - p) \geqslant \mu_C(q) \qquad \text{for all } q.$$

($\Longrightarrow$) I first claim that if x does not belong to C, it is not a supergradient of μ_C at p. For if $x \notin C$, then by Theorem 175 there is some q for which $q \cdot x < \mu_C(q)$. Thus for $\lambda > 0$ large enough, $\lambda q \cdot x < \mu_C(\lambda q) + \big(p \cdot x - \mu_C(p)\big)$. Rearranging terms violates the supergradient inequality applied to λq. Therefore, by contraposition, if x is a supergradient of the support function μ_C at p, then x belongs to C.

So let x be a supergradient of μ_C at p. Setting $q = 0$ in the supergradient inequality, we conclude that $\mu_C(p) \geqslant p \cdot x$. But x belongs to C, so x minimizes p over C, and $\mu_C(p) = p \cdot x$.

In other words, $\partial\mu_C(p) \subset \{x \in C : p \cdot x = \mu_C(p)\}$

($\Longleftarrow$) Suppose now that x belongs to C and $p \cdot x = \mu_C(p)$, that is, x minimizes p over C. By the definition of μ_C, for any $q \in \boldsymbol{R}^{\mathrm{n}}$, $q \cdot x \geqslant \mu_C(q)$. Now add $\mu_C(p) - p \cdot x = 0$ to the left-hand side of the inequality to obtain the supergradient inequality.

Thus $\{x \in C : p \cdot x = \mu_C(p)\} \subset \partial\mu_C(p)$, completing the proof. ∎

194 Corollary *Let C be a closed convex set. Suppose x belongs to C and strictly minimizes p over C. Then μ_C is differentiable at p and*

$$\mu'_C(p) = x.$$

Proof: This follows from Theorem 193 and Theorem 219. ∎

195 Example Let's look at $C = \{(x_1, x_2) \in \boldsymbol{R}^2_{++} : x_1 x_2 \geqslant 1\}$. This is a closed convex set and its support function is easily calculated. If $p \notin \boldsymbol{R}^2_+$, then $\mu_C(p) = -\infty$. For $p \geqq 0$, it not hard to see that $\mu_C(p) = 2\sqrt{p_1 p_2}$, which has no supergradient when $p_1 = 0$ or $p_2 = 0$.

(To see this, consider first the case $p \geqq 0$. The Lagrangean for the minimization problem is $p_1 x_1 + p_2 x_2 + \lambda(1 - x_1 x_2)$. By the Lagrange Multiplier Theorem 270, the first order conditions are $p_1 - \lambda x_1^* = 0$ and $p_2 - \lambda x_2^* = 0$. Thus $x_1^* x_2^* = \frac{p_1 p_2}{\lambda^2}$, so $\lambda = \sqrt{p_1 p_2}$. Thus $x_1^* = \sqrt{\frac{p_1}{p_2}}$ and $x_2^* = \sqrt{\frac{p_2}{p_1}}$ and $\mu_C(p) = p_1 x_1^* + p_2 x_2^* = 2\sqrt{p_1 p_2}$.

Now suppose some $p_i < 0$. For instance, suppose $p_2 < 0$. Then $p \cdot (\varepsilon, \frac{1}{\varepsilon}) \to -\infty$ as $\varepsilon \to 0$, so $\mu_C(p) = -\infty$.) □

4.16 Concavity and continuity

We shall see in a moment that concave functions on $\boldsymbol{R}^{\mathrm{n}}$ are continuous at interior points. The only discontinuities can be jumps downward at the boundary of the domain. This is not true for infinite dimensional domains, as witnessed by Example 52. However the next result is true in general topological vector spaces.

196 Theorem (Local continuity of convex functions) *If a convex function is defined and bounded above on a neighborhood of some point in a topological vector space, then it is continuous at that point.*

Proof: Let C be a convex set in a tvs, and let $f\colon C \to \boldsymbol{R}$ be convex. We begin by noting the following consequences of convexity. Fix $x \in C$ and suppose z satisfies $x+z \in C$ and $x-z \in C$. Let $\delta \in [0,1]$. Then $x + \delta z = (1-\delta)x + \delta(x+z)$, so $f(x+\delta z) \leqslant (1-\delta)f(x) + \delta f(x+z)$. Rearranging terms yields

$$f(x+\delta z) - f(x) \leqslant \delta\left[f(x+z) - f(x)\right], \tag{4.8}$$

and replacing z by $-z$ gives

$$f(x-\delta z) - f(x) \leqslant \delta\left[f(x-z) - f(x)\right]. \tag{4.9}$$

Also, since $x = \frac{1}{2}(x+\delta z) + \frac{1}{2}(x-\delta z)$, we have $f(x) \leqslant \frac{1}{2}f(x+\delta z) + \frac{1}{2}f(x-\delta z)$. Multiplying by two and rearranging terms we obtain

$$f(x) - f(x+\delta z) \leqslant f(x-\delta z) - f(x). \tag{4.10}$$

Combining (4.9) and (4.10) yields

$$f(x) - f(x+\delta z) \leqslant f(x-\delta z) - f(x) \leqslant \delta\left[f(x-z) - f(x)\right].$$

This combined with (4.8) implies

$$\left|f(x+\delta z) - f(x)\right| \leqslant \delta \max\left\{f(x+z) - f(x), f(x-z) - f(x)\right\}. \tag{4.11}$$

Now let $\varepsilon > 0$ be given. Since f is bounded above on an open neighborhood of x, there is a neighborhood V of zero, and a constant $M \geqslant 0$ such that $x + V \subset C$ and if $y \in x + V$, then $f(y) < f(x) + M$. Choosing $0 < \delta \leqslant 1$ so that $\delta M < \varepsilon$, equation (4.11) implies that if $y \in x + \delta V$, then $|f(y) - f(x)| < \varepsilon$. Thus f is continuous at x. ∎

197 Theorem (Global continuity of convex functions) *Let f be a convex function on an open convex set C in a topological vector space. The following are equivalent.*

1. *f is continuous on C.*
2. *f is upper semicontinuous on C.*
3. *f is bounded above on a neighborhood of some point in C.*
4. *f is continuous at some point in C.*

Proof: (1) $\Longrightarrow$ (2) Obvious.

(2) $\Longrightarrow$ (3) If f is upper semicontinuous and convex, then $\{y \in C : f(y) < f(x) + 1\}$ is a convex open neighborhood of x on which f is bounded.

(3) $\Longrightarrow$ (4) This is Theorem 196.

(4) $\Longrightarrow$ (1) Suppose f is continuous at x, and let y be any other point in C. Since scalar multiplication is continuous, $\{\beta \in \boldsymbol{R} : x + \beta(y-x) \in C\}$ includes an open neighborhood of 1. This implies that there is some point z in C such that $y = \lambda x + (1-\lambda)z$ with $0 < \lambda < 1$.

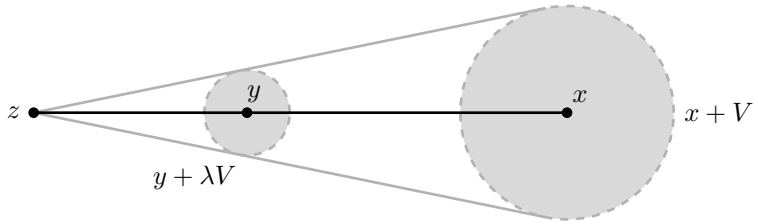

Figure 4.5. (4) $\Longrightarrow$ (1).

Also, since f is continuous at x, there is a circled neighborhood V of zero such that $x+V \subset C$ and f is bounded above on $x+V$, say by M. We claim that f is bounded above on $y+\lambda V$. To see this, let $v \in V$. Then $y+\lambda v = \lambda(x+v)+(1-\lambda)z \in C$. The convexity of f thus implies

$$f(y+\lambda v) \leqslant \lambda f(x+v) + (1-\lambda)f(z) \leqslant \lambda M + (1-\lambda)f(z).$$

That is, f is bounded above by $\lambda M + (1-\lambda)f(z)$ on $y+\lambda V$. By Theorem 196, f is continuous at y. ∎

198 Theorem *In a finite dimensional vector space, every convex function is continuous on the relative interior of its domain.*

Proof: Without loss of generality, we may translate the domain so its affine hull is a linear space, say $\boldsymbol{R}^{\mathrm{n}}$. Let $f\colon C \to \boldsymbol{R}$ be a convex function defined on a convex subset C of $\boldsymbol{R}^{\mathrm{n}}$, and let x be an relative interior point of C. Then there exist $a, b \in C$ with $a < b$ such that the box $[a,b] = \{y \in \boldsymbol{R}^{\mathrm{n}} : a \leq y \leq b\}$ is a neighborhood of x and satisfies $[a,b] \subset C$. Now $[a,b]$ is the convex hull of the 2^n vertexes v of the form $v_i \in \{a_i, b_i\}$, $i = 1, \dots, n$. Any point $y \in [a,b]$ can be written as a convex combination $y = \sum_{j=1}^{2^n} \lambda_j v^j$, where the v^js are vertexes. The convexity of f implies that $f(y) \leqslant \sum_{j=1}^{2^n} \lambda_j f(v^j)$, which is bounded above by $\max_j f(v^j)$. So by Theorem 197, f is continuous at x. ∎

4.17 Concavity and differentiability in one variable

We now examine the differentiability of concave functions. We start with the following simple, but fundamental, result for concave functions of one variable, cf. Fenchel [51, 2.16, p. 69], Phelps [125, Theorem 1.16, pp. 9–11], or Royden [132, Proposition 5.17, p. 113].

199 Lemma *Let f be a real-valued function defined on some interval I of $\boldsymbol{R}$. If f is concave, then for every $x < y < z$ in I,*

$$\frac{f(y)-f(x)}{y-x} \geqslant \frac{f(z)-f(x)}{z-x} \geqslant \frac{f(z)-f(y)}{z-y}.$$

Conversely, if one of the (three) inequalities is satisfied for every $x < y < z$ in I, then f is concave.

Equivalently,

$\dfrac{f(z)-f(x)}{z-x}$ *is decreasing in both x and z over $\{(x,z) : x < z\}$ if and only f is concave.*

200 Exercise Prove Lemma 199. □

You may find this written in the following form:

201 Corollary *Let f be a real-valued function defined on some interval I of $\boldsymbol{R}$. Then f is concave if and only if for every $x_1 < y_1$ and $x_2 < y_2$ in I, with $x_1 \leqslant x_1$ and $y_1 \leqslant y_2$ (that is, the interval $[x_2, y_2]$ lies to the right of $[x_1, y_1]$),*

$$\frac{f(y_1) - f(x_1)}{y_1 - x_1} \geqslant \frac{f(y_2) - f(x_1)}{y_2 - x_1} \geqslant \frac{f(y_2) - f(x_2)}{y_2 - x_2}.$$

Proof: Apply Lemma 199 to the case $x_1 < y_1 < y_2$ and $x_1 < x_2 < y_2$. (The cases $x_1 = x_2$ or $y_1 = y_2$ are trivial.) ∎

202 Corollary *Let f be a real-valued function defined on some interval I of $\boldsymbol{R}$. If f is concave, then the second difference function satisfies*

$$\Delta^2_{v,w} f(x) vw \leqslant 0$$

whenever it is defined.

203 Exercise Prove Corollary 202. □

Moreover the converse is also true.

Finally a consequence is that f is twice differentiable a.e., as a decreasing function is differentiable almost everywhere.

Lemma 199 has a number of consequences.

204 Corollary *Let f be a concave function defined on some interval I of $\boldsymbol{R}$. Then at every interior point x, f is continuous, and has a (finite) left-hand derivative $f'(x^-)$ and (finite) right-hand derivative $f'(x^+)$. Moreover, $f'(x^-) \geqslant f'(x^+)$, and both $f'(x^-)$ and $f'(x^+)$ are nonincreasing functions. Consequently, there are at most countably many points where $f'(x^-) > f'(x^+)$, that is, where f is nondifferentiable. Furthermore $f'(x^+)$ is lower semicontinuous and $f'(x^-)$ is upper semicontinuous, so on the set where $f'(x)$ exists it is continuous.*

Proof: ********** ∎

4.18 A digression on mid-concavity

A function f is called **mid-concave** if $f(\frac{1}{2}x + \frac{1}{2}y) \geqslant \frac{1}{2}f(x) + \frac{1}{2}f(y)$ for all points x and y in its domain. Mid-concavity does not imply concavity.

205 Example (Concavity vs. mid-concavity) Let D be the set of dyadic rational numbers, that is, rationals of the form $k/2^m$ for some integer k and some natural number m. These form a scalar field (closed under addition and multiplication, etc.), and we can view the real numbers $\boldsymbol{R}$ as a vector space over the field D. (Honest, we can—if x is a vector [that is, a real number] and d is a scalar, [that is, a dyadic rational], then dx is a real number, and hence a vector, etc.) Now every vector space has a (Hamel) basis (see, e.g., [3]), so let B be a basis for $\boldsymbol{R}$.

****

□

But if f is continuous, then mid-concavity does imply concavity.

206 Fact *Let $C \subset \boldsymbol{R}^n$ be an open convex set. If $f\colon C \to \boldsymbol{R}$ is continuous and mid-concave, then f is concave.*

This allows the following partial converse to Corollary 202.

207 Proposition *Let $C \subset \boldsymbol{R}^n$ be an open convex set, and let $f\colon C \to \boldsymbol{R}$. If the second difference function satisfies*

$$\Delta^2_{v,v} f(x) vv \leqslant 0$$

whenever it is defined, then f is mid-concave. Consequently, if f is also continuous, then it is concave.

Proof: Assume $\Delta^2_{v,v} f(x)vv \leqslant 0$ whenever defined. Let $x < y$ and set $v = w = (y-x)/2 > 0$, so

$$\Delta^2_{v,v} f(x) = f(x+y) - 2f\big((x+y)/2\big) + f(x) \leqslant 0,$$

so rearranging yields $f\big((x+y)/2\big) \geqslant \big(f(x)+f(y)\big)/2$. So f is mid-concave. ∎

4.19 Concavity and differentiability in more than one variable

The one-dimensional case has implications for the many dimensional case. The next results may be found in Fenchel [51, Theorems 33–34, pp. 86–87].

208 Theorem *Let f be a concave function on the open convex set C. For each direction v, $f'(x;v)$ is a lower semicontinuous function of x, and $\{x : f'(x;v) + f'(x;-v) < 0\}$ has Lebesgue measure zero. Thus $f'(x;v) + f'(x;-v) = 0$ almost everywhere, so f has a directional derivative in the direction v almost everywhere. Moreover, the directional derivative $Df(\cdot;v)$ is continuous on the set on the set on which it exists.*

Proof: Since f is concave, it is continuous (Theorem 198). Fix v and choose $\lambda_n \downarrow 0$. Then $g_n(x) := \frac{f(x+\lambda_n v)-f(x)}{\lambda_n}$ is continuous and by Lemma 199, $g_n(x) \uparrow f'(x;v)$ for each x. Thus Proposition 31 implies that $f'(x;v)$ is lower semicontinuous in x for any v.

Now $f'(x;v) + f'(x;-v) \leqslant 0$ by concavity, so let

$$A = \{x : f'(x;v) + f'(x;-v) < 0\}.$$

Note that since $f'(\cdot;v)$ and $f'(\cdot;-v)$ are lower semicontinuous, then A is a Borel subset of $\boldsymbol{R}^{\mathrm{n}}$. If $x \in A^c$, that is, if $f'(x;v) + f'(x;-v) = 0$, then $f'(x;v) = -f'(x;-v)$, so f has a directional derivative $D_v(x)$ in the direction v. And since $f'(\cdot;-v)$ is lower semicontinuous, the function $-f(\cdot;-v)$ is upper semicontinuous, $f'(\cdot;v)$ is actually continuous on A^c.

Thus we want to show that $A = \{x : f'(x;v) + f'(x;-v) < 0\}$ has Lebesgue measure zero.

If $v = 0$, then $f'(x;0) = -f'(x;-0) = 0$, so assume $v \neq 0$. Consider a line $L_y = \{y + \lambda v : \lambda \in \boldsymbol{R}\}$ parallel to v. By Corollary 204, $L_y \cap A = \{x \in L_y : f'(x;v) + f'(x;-v) < 0\}$ is countable, and hence of one-dimensional Lebesgue measure zero. Let M be the subspace orthogonal to v, so $M \times L = \boldsymbol{R}^{\mathrm{n}}$, where $L = L_0$ is the one-dimensional subspace spanned by v. Every $x \in \boldsymbol{R}^{\mathrm{n}}$ can be uniquely written as $x = (x_M, x_v)$, where $x_M \in M$ and $x_v \in L$. Then by Fubini's theorem,

$$\int \mathbf{1}_A(x)\, d\lambda^n(x) = \int_M \int_L \mathbf{1}_A(x_M, x_v)\, d\lambda(x_v)\, d\lambda^{n-1}(x_M) = \int_M 0\, d\lambda^{n-1}(x_M) = 0.$$

∎

209 Lemma *Let f be a concave function on the open convex set $C \subset \boldsymbol{R}^{\mathrm{n}}$. If all n partial derivatives of f exist at x, then f has a Gâteaux derivative at x. That is, all the directional derivatives exist and the mapping $v \mapsto D_v f(x)$ is linear.*

Proof: The mapping $v \mapsto f'(x;v)$ is itself concave, and since f has an i^{th} partial derivative, there is $\delta_i > 0$ so that $v \mapsto f'(x;v)$ is linear on the segment $L_i = (-\delta_i e^i, \delta_i e^i)$. Indeed $\lambda e^i \mapsto \frac{\partial f(x)}{\partial x_i}\lambda$. So by Lemma 212 below, the mapping $v \mapsto f'(x;v)$ is linear on $\operatorname{co}\bigcup_{i=1}^m L_i$. This makes it the Gâteaux derivative of f at x. ∎

210 Lemma *Let f be a concave function on the open convex set $C \subset \boldsymbol{R}^{\mathrm{n}}$. If f has a Gâteaux derivative at x, then it is a Fréchet derivative. (Cf. Fenchel [51, Property 32, p. 86], or Hiriart-Urruty–Lemaréchal [76, Proposition 4.2.1, p. 114].)*

Proof: Let $v \mapsto f'(x;v)$ be the Gâteaux derivative of f. We need to show that

$$\forall \varepsilon > 0 \ \exists \delta > 0 \ \forall 0 < \lambda < \delta \ \forall v_{\|v\|=1} \qquad \|f(x+\lambda v) - f(x) - \lambda f'(x;v)\| \leqslant \varepsilon\lambda.$$

Fix $\varepsilon > 0$. By definition, $f'(x;v) = \lim_{\lambda \downarrow 0} \big(f(x+\lambda v) - f(x)\big)/\lambda$, so for each v, there is a $\delta_v > 0$ such that for $0 < \lambda \leqslant \delta_v$,

$$\left| \frac{f(x+\lambda v) - f(x)}{\lambda} - f'(x;v) \right| < \varepsilon,$$

or multiplying by λ,

$$|f(x+\lambda v) - f(x) - \lambda f'(x;v)| < \varepsilon\lambda,$$

By Lemma 215 and the homogeneity of $f'(x;\cdot)$, for $\lambda > 0$ we have

$$f(x) + \lambda f'(x;v) - f(x+\lambda v) \geqslant 0.$$

Combining these two inequalities, for $0 < \lambda \leqslant \delta_v$, we have

$$0 \leqslant \lambda f'(x;v) - f(x+\lambda v) + f(x) < \varepsilon\lambda. \tag{$*$}$$

Once again consider the 2^n vectors $u^1, \dots, u^{2^n}$ with coordinates ± 1, and let $\delta = \min_j \delta_{u^j}$. Then $(*)$ holds with $v = u^j$ for any $0 < \lambda < \delta$.

Let $U = \operatorname{co}\{u^1, \dots, u^{2^n}\}$, which is a convex neighborhood of zero that includes all the vectors v with $\|v\| = 1$. Fixing λ, the function $h_\lambda(v) = \lambda f'(x;v) - f(x+\lambda v) + f(x)$ is convex in v, and any v in U can be written as a convex combination $v = \sum_{j=1}^{2^n} \alpha_j u^j$, so for any $0 < \lambda \leqslant \delta$,

$$0 \leqslant \lambda f'(x;v) - f(x+\lambda v) + f(x) = h_\lambda(v) \leqslant \sum_{j=1}^{2^n} \alpha_j h_\lambda(u^j) \leqslant \max_j h_\lambda(u^j) < \varepsilon\lambda.$$

Since this is true for every vector v of norm one, we are finished. ∎

211 Theorem *Let f be a concave function on the open convex set $C \subset \boldsymbol{R}^{\mathrm{n}}$. Then f is differentiable almost everywhere on C.*

Proof: By Theorem 208 for each i, the i^{th} partial derivative exists for almost every x. Therefore all n partial derivatives exist for almost every x. The result now follows from Lemma 209 and 210. ∎

This lemma is used in the proof of Theorem 211.

212 Lemma *Let g be concave on C and let $x \in \operatorname{ri} C$. Let $v^1, \dots, v^m$ be linearly independent and assume that g is affine on each of the segments $L_i = \{x + \lambda v^i : |\lambda| \leqslant \delta_i\} \subset C$, $i = 1, \dots, m$. Then g is affine on $A = \operatorname{co} \bigcup_{i=1}^m L_i$.*

Proof: By hypothesis, there is an α_i satisfying

$$g(x + \lambda v^i) = g(x) + \alpha_i \lambda \text{ on } L_i, \quad i = 1, \dots, m.$$

Define ℓ on the span of $v^1, \dots, v^m$ by $\ell(\lambda_1 v^1 + \cdots + \lambda_m v^m) = \alpha_1\lambda_1 + \cdots + \alpha_m\lambda_m$. Then ℓ is linear, so the function h on A defined by $h(y) = g(x) + \ell(y - x)$ is affine. Moreover h agrees with g on each segment L_i. In particular $g(x) - h(x) = 0$.

Now any point y in A can be written as a convex combination of points $x + \lambda_i v^i$ belonging to L_i. Since g is concave, for a convex combination $\sum_i \alpha_i (x + \lambda_i v^i)$ we have

$$g\big(\sum_i \alpha_i (x + \lambda_i v^i)\big) \geqslant \sum_i \alpha_i g(x + \lambda_i v^i) = \sum_i \alpha_i h(x + \lambda_i v^i) = h\big(\sum_i \alpha_i (x + \lambda_i v^i)\big),$$

where the final equality follows from the affinity of h. Therefore $g - h \geqslant 0$ on A. But $g - h$ is concave, x belongs to $\operatorname{ri} A$, and $(g - h)(x) = 0$. Therefore $g - h = 0$ on A. (To see this, let y belong to A. Since x in $\operatorname{ri} A$, for some $z \in A$ and some $0 < \lambda < 1$, we may write $x = \lambda y + (1 - \lambda)z$, so $0 = (g - h)(x) \geqslant \lambda(g - h)(y) + (1 - \lambda)(g - h)(z) \geqslant 0$, which can only happen if $(g - h)(y) = (g - h)(z) = 0$.)

Thus g is the affine function h on A. ∎

This result depends on the fact that x belongs to $\operatorname{ri} C$, and can fail otherwise. For instance, let $C = \boldsymbol{R}^2_+$ and $f(x, y) = xy$. Then f is linear (indeed zero) on the nonnegative x and y axes, which intersect at the origin, but f is not linear on the convex hull of the axes. Of course, the origin is not in the relative interior.

The next fact may be found in Fenchel [51, Theorem 35, p. 87ff], or Katzner [91, Theorems B.5-1 and B.5-2].

213 Fact *If $f\colon C \subset \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}$ is twice differentiable, then the Hessian H_f is everywhere negative semidefinite if and only if f is concave. If H_f is everywhere negative definite, then f is strictly concave.*

********** There are many ways to see this. One way is to look at the second difference $\Delta^2_{v,w} f = f(x + w + v) - f(x + w) - \big(f(x + v) - f(x)\big)$. By
****

4.20 Directional derivatives and supergradients

Given a point x in a convex set C, the set of directions v into C at x is

$$P_C(x) = \{v \in \boldsymbol{R}^{\mathrm{n}} : \exists \varepsilon > 0 \; x + \varepsilon v \in C\}$$

is a convex cone, but not necessarily a closed cone. (Think of this set for a point on the boundary of a disk—it is an open half space together with zero.) The set $x + P_C(x)$, a cone with vertex x, is what Fenchel [51, p. 41] calls the **projecting cone** of C from x.

The following is a simple consequence of Corollary 204.

214 Theorem *Let f be a concave function on $\boldsymbol{R}^{\mathrm{n}}$, and let f be finite at the point x. Then the difference quotient*

$$\frac{f(x + \lambda v) - f(x)}{\lambda} \quad 0 < \lambda \leqslant 1,$$

is a nonincreasing function of λ. Thus the possibly infinite directional derivative $f'(x; v)$ exists and

$$f'(x; v) = \sup_{0 < \lambda \leqslant 1} \frac{f(x + \lambda v) - f(x)}{\lambda}.$$

Moreover, $f'(x; v)$ is a positively homogeneous and concave function of v satisfying $f'(x; 0) = 0$, and

$$\operatorname{dom} f'(x; v) = P_{\operatorname{dom} f}(x).$$

Proof: From

$$\frac{f(x + \alpha\lambda v) - f(x)}{\lambda} = \alpha \frac{f(x + \alpha\lambda v) - f(x)}{\alpha\lambda},$$

we see that $f'(x; \alpha v) = \alpha f'(x; v)$ for all $\alpha \geqslant 0$. That is, $f'(x; v)$ is positively homogeneous of degree one in v. Note that $f'(x, 0) = 0$, as the difference quotients are all zero. Furthermore, if f is concave, then

$$f\big(x + \alpha\lambda u + (1 - \alpha)\lambda v\big) - f(x) \geqslant \alpha\big(f(x + \lambda u) - f(x)\big) + (1 - \alpha)\big(f(x + \lambda v) - f(x)\big),$$

so dividing by λ and taking limits shows that $f'(x;\cdot)$ is concave.

If $v \in P_{\operatorname{dom} f}(x)$, that is, $x + \varepsilon v \in \operatorname{dom} f$ for $\varepsilon > 0$ small enough, then $f(x + \varepsilon v) > -\infty$, so $f'(x; v) > -\infty$. ∎

There is an intimate relation between one-sided directional derivatives and the superdifferential, cf. Fenchel [51, Property 29, p. 81] or Rockafellar [130, Theorem 23.2, p. 216]. We start with the following extension of Theorem 178.

215 Lemma *Let f be a concave function on $\boldsymbol{R}^{\mathrm{n}}$ and let f be finite at x. Then for every $y \in \boldsymbol{R}^{\mathrm{n}}$,*

$$f(x) + f'(x; y - x) \geqslant f(y).$$

(If f is convex the inequality is reversed.)

Proof: If $y \notin \operatorname{dom} f$, then $f(y) = -\infty$, so the conclusion follows. If y belongs to the effective domain, then by concavity

$$\frac{f\big(x + \lambda(y - x)\big) - f(x)}{\lambda} \geqslant f(y) - f(x).$$

Letting $\lambda \downarrow 0$, the left hand side converges to $f'(x; y - x)$, which may be $+\infty$. ∎

Geometrically, this says that the hypograph of $y \mapsto f(x) + f'(x; y - x)$ includes the hypograph of f. We can use this to complete the description of subdifferentiability of f. The following result may be partially found in Fenchel [51, Property 31, p. 84] and more explicitly in Rockafellar [130, Theorem 23.3, p. 216] (which are stated for convex functions).

216 Corollary *Let f be a proper concave function on $\boldsymbol{R}^{\mathrm{n}}$, and let $x \in \operatorname{dom} f$. If $f'(x; v) < \infty$ for some v such that $x + v \in \operatorname{ri} \operatorname{dom} f$, then f is superdifferentiable at x.*

Proof: Let v satisfy $x + v \in \operatorname{ri} \operatorname{dom} f$ and $f'(x; v) < \infty$. Then, as in the proof of Theorem 186, there is $(p, -1)$ supporting the hypograph of $f'(x;\cdot)$ at the point $\big(v, f'(x; v)\big)$. That is,

$$p \cdot v - f'(x; v) \leqslant p \cdot u - f'(x, u) \quad \text{for all } u \in \operatorname{dom} f'(x;\cdot). \tag{4.12}$$

Taking $u = 0$ implies $p \cdot v - f'(x; v) \leqslant 0$. Taking $u = \lambda v$ for $\lambda > 0$ large implies $p \cdot v - f'(x; v) \geqslant 0$. Thus $f'(x; v) = p \cdot v$. Then (4.12) becomes

$$p \cdot u \geqslant f'(x, u) \quad \text{for all } u \in \operatorname{dom} f'(x;\cdot).$$

Adding $f(x)$ to both sides and applying Lemma 215, we get the supergradient inequality

$$f(x) + p \cdot u \geqslant f(x) + f'(x, u) \geqslant f(x + u)$$

for $u \in \operatorname{dom} f'(x;\cdot) = P_{\operatorname{dom} f}(x)$. For any u not in this set, $f(x + \lambda u) = -\infty$ for $\lambda > 0$ and the supergradient inequality holds trivially. Thus p is a supergradient of f at x. ∎

217 Lemma (The directional derivative is the support function of the superdifferential) *Let f be a concave function on $\boldsymbol{R}^{\mathrm{n}}$, and let $f(x)$ be finite. Then*

$$p \in \partial f(x) \iff \forall v \in \boldsymbol{R}^{\mathrm{n}} \quad p \cdot v \geqslant f'(x; v).$$

Proof: ($\Longrightarrow$) Let $p \in \partial f(x)$. By the supergradient inequality, for any $v \in \boldsymbol{R}^{\mathrm{n}}$,

$$f(x) + p \cdot (\lambda v) \geqslant f(x + \lambda v))$$

We may subtract the finite value $f(x)$ from the right hand side, even if $x + \lambda v \notin \operatorname{dom} f$. Thus

$$p \cdot (\lambda v) \geqslant f(x + \lambda v) - f(x).$$

Dividing by $\lambda > 0$ and letting $\lambda \downarrow 0$ gives

$$p \cdot v \geqslant f'(x; v)$$

.

($\Longleftarrow$) If $p \notin \partial f(x)$, then there is some v such that the supergradient inequality is violated, that is,

$$f(x) + p \cdot v < f(x + v). \tag{4.13}$$

Since $f(x+v) = -\infty$ if $x+v \notin \operatorname{dom} f$, we conclude $x+v \in \operatorname{dom} f$. By concavity, for $0 < \lambda \leqslant 1$,

$$f(x + \lambda v) \geqslant f(x) + \lambda\big[f(x + v) - f(x)\big]$$

or

$$\frac{f(x + \lambda v) - f(x)}{\lambda} \geqslant f(x + v) - f(x),$$

so by (4.13)

$$\frac{f(x + \lambda v) - f(x)}{\lambda} \geqslant f(x + v) - f(x) > p \cdot v,$$

so taking limits gives $f'(x; v) > p \cdot v$. The conclusion now follows by contraposition. ∎

The next result may be found in Rockafellar [130, Theorem 23.2, p. 216].

218 Corollary *Let f be a concave function on $\boldsymbol{R}^{\mathrm{n}}$, and let $f(x)$ be finite. Then the closure of the directional derivative at x (as a concave function of the direction) is the support function of the superdifferential at x. That is,*

$$\operatorname{cl} f'(x; \cdot) = \mu_{\partial f(x)}(\cdot).$$

Proof: Since $h : v \mapsto f'(x; v)$ is concave and homogeneous, by Theorem 177, $\operatorname{cl} h = \mu_C$, where $C = \{p : \forall v \; p \cdot v \geqslant h(v)\}$. By Lemma 217, $C = \partial f(x)$. ∎

The next result may be found in Rockafellar [130, Theorem 25.1, p. 242].

219 Theorem *Let f be a concave function defined on the convex set $C \subset \boldsymbol{R}^{\mathrm{n}}$. Then f is differentiable at the interior point $x \in C$ if and only if the superdifferential $\partial f(x)$ is a singleton, in which case $\partial f(x) = \{f'(x)\}$.*

Proof: ($\Longrightarrow$) Suppose f is differentiable at the interior point x. The for any v, $f'(x; v) = f'(x) \cdot v$. Moreover there is an $\varepsilon > 0$ such that for any v, $x + \varepsilon v \in C$. Now the superdifferential $f'(x)$ is nonempty, since $f'(x) \in \partial f(x)$, so by Lemma 217, if $p \in \partial f(x)$, then

$$p \cdot \varepsilon v \geqslant f'(x; \varepsilon v) = f'(x) \cdot \varepsilon v.$$

But this also holds for $-v$, so

$$p \cdot v = f'(x) \cdot v.$$

Since this holds for all v, we have $p = f'(x)$.

($\Longleftarrow$) Suppose $\partial f(x) = \{p\}$. Since x is interior there is an $\alpha > 0$ such that if $v \in \alpha B$, then $x + v \in C$, where B is the unit ball in $\boldsymbol{R}^{\mathrm{n}}$. Define $g \colon \alpha B \to \boldsymbol{R}$ by

$$g(v) = f(x + v) - f(x) - p \cdot v.$$

Note that f is differentiable at x if and only if g is differentiable at 0, in which case $g'(0) = f'(x) - p$.

Now the supergradient inequality asserts that $f(x)+p\cdot v \geqslant f(x+v)$, so $g \leqslant 0$. But $g(0) = 0$, that is, 0 maximizes g over αB, so by Theorem 190, $0 \in \partial g(0)$.

In fact, $\partial g(0) = \{0\}$. For if $q \in \partial g(0)$, we have

$$\begin{aligned} g(0) + q \cdot v &\geqslant g(v) \\ 0 + q \cdot v &\geqslant f(x+v) - f(x) - p \cdot v \\ f(x) + (p+q) \cdot v &\geqslant f(x+v), \end{aligned}$$

which implies $p + q \in \partial f(x)$, so $q = 0$.

Why is it closed? See Rockafellar 23.4.

By Lemma 218, the closure of $g'(x;\cdot)$ is the support function $\partial g(0) = \{0\}$, so $\operatorname{cl} g'(x;\cdot) = 0$. But this implies that $g'(x;\cdot)$ is itself closed, and so identically zero. But zero is a linear function, so by Lemma 210, g is differentiable at zero. ∎

4.21 Fenchel's conjugate duality

We saw in Section 4.12 that every closed convex set is the intersection of all the closed half spaces that include it, so the support function embodies all the information about the set. Now the hypograph of an upper semicontinuous concave function f is a closed convex set, so its support function embodies all the information about f. But the supporting hyperplanes to the hypograph of f are the graphs of affine functions that dominate f. Well, not quite. A hyperplane in $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$ of the form $\{(x,\alpha) : (p,-1)\cdot(x,\alpha) = \beta\}$, is the graph of the affine function $h(x) = p \cdot x - \beta$, but if the supporting hyperplane is vertical, then it does not define the graph of a function.

Let f be a convex function defined on $\boldsymbol{R}^{\mathrm{n}}$ as a convex analyst would define it. That is, the values $\pm\infty$ are allowed, and the effective domain $\operatorname{dom} f$ is defined to be the set where f is (finite) real-valued. It is proper if the effective domain is nonempty and doesn't take on the value $-\infty$. (That means it assumes the value ∞ outside the effective domain.) Regardless of whether f is proper, we may still enquire whether the affine function $h(x) = p \cdot x - \alpha$ is dominated by f.

220 Definition *For a convex function f, the* ***Fenchel (convex) conjugate****, or simply* ***conjugate****, f^* of f is defined by*

$$f^*(p) = \inf\{\alpha : f(x) \geqslant p \cdot x - \alpha \text{ for all } x \in \boldsymbol{R}^{\mathrm{n}}\}. \tag{4.14}$$

Note that if there is no α for which f dominates the affine function, then we have an infimum over the empty set, so $f^*(p) = \infty$. I always get confused working with suprema and infima, so it helps me to write out a number of equivalent statements.

$$\alpha \geqslant f^*(p) \iff p \cdot x - \alpha \leqslant f(x) \text{ for every } x. \tag{4.15}$$

$$\alpha \geqslant f^*(p) \iff p \cdot x - f(x) \leqslant \alpha \text{ for every } x.$$

$$f^*(p) = \sup_{x \in \boldsymbol{R}^{\mathrm{n}}} p \cdot x - f(x). \tag{4.16}$$

$$f^*(p) = \sup\{p \cdot x - \beta : \beta \geqslant f(x)\} = \sup\{p \cdot x - \beta : (x,\beta) \in \operatorname{epi} f\}. \tag{4.17}$$

In fact, (4.16) is usually taken to be the definition.

Following Rockafellar [130, p. 308], for a concave function g, we define its ***(concave) conjugate*** *by*

$$g^*(p) = \sup\{\alpha : p \cdot x - \alpha \geqslant g(x) \text{ for all } x \in \boldsymbol{R}^{\mathrm{n}}\}.$$

Equivalently,

$$\alpha \leqslant g^*(p) \iff p \cdot x - \alpha \geqslant g(x) \text{ for every } x. \tag{4.18}$$

$$\alpha \leqslant g^*(p) \iff p \cdot x - g(x) \geqslant \alpha \text{ for every } x.$$

$$g^*(p) = \inf_{x \in \boldsymbol{R}^n} p \cdot x - g(x). \tag{4.19}$$

$$g^*(p) = \inf\{p \cdot x - \beta : \beta \leqslant g(x)\} = \sup\{p \cdot x - \beta : (x, \beta) \in \mathit{hypograph}\, g\}. \tag{4.20}$$

221 Example Since an affine function f is both concave and convex, we have a problem, albeit a minor problem. The concave conjugate and the convex conjugate do not agree, but only differ outside their common effective domain. To see this, consider the affine function

$$f(x) = p \cdot x - \alpha.$$

For any affine function $g(x) = q \cdot x - \beta$, if $q \neq p$, then neither f nor g dominates the other.[7] Therefore, for $q \neq p$,

$$f^*(q) = \inf\{\alpha : f(x) \geqslant q \cdot x - \alpha \text{ for all } x \in \boldsymbol{R}^n\} = \inf \varnothing = \infty \text{ treating } f \text{ as convex,}$$

$$f^*(q) = \sup\{\alpha : f(x) \geqslant q \cdot x - \alpha \text{ for all } x \in \boldsymbol{R}^n\} = \sup \varnothing = -\infty \text{ treating } f \text{ as concave,}$$

but

$$f^*(p) = \alpha \text{ either way.}$$

So f^* on its effective domain does not depend on whether f is treated as concave or convex. □

From the definitions, when f is convex, then for any p, we have $f(x) \geqslant p \cdot x - f^*(p)$ for every x. That is,

$$f(x) + f^*(p) \geqslant p \cdot x \qquad \text{for all } x, p. \tag{Fenchel's Inequality}$$

This result is known as **Fenchel's inequality**. Moreover, it is also clear that $f^* = (\operatorname{cl} f)^*$. In the concave case, Fenchel's inequality becomes

$$p \cdot x \geqslant g(x) + g^*(p) \quad \text{for all } x, p.$$

If f is concave or convex, then $-f$ is convex or concave respectively, and it is easy to see that their conjugates satisfy

$$(-f)^*(p) = -(f^*)(-p). \tag{4.21}$$

222 Lemma *The convex conjugate of a convex function is a closed convex function. The concave conjugate of a concave function is a closed concave function.*

Proof: Note that the function $x \mapsto p \cdot x - f(x)$ is an affine function of p. The convex conjugate f^* of a convex function f is the pointwise supremum of this family and so closed and convex, and the concave conjugate f^* of a concave function f is the pointwise infimum and so closed and concave. ∎

Note that f^* then has a conjugate $(f^*)^*$, usually written as just f^{**}.

223 Proposition *If f is concave or convex, then* $\operatorname{cl} f = f^{**}$.

Proof: I will discuss only the convex case. Using (4.15) and Theorem 170,

$$\begin{aligned}
\operatorname{cl} f(y) &= \sup\{h(y) : f \geqslant h \text{ and } h \text{ is affine and continuous}\} \\
&= \sup\{p \cdot y - \alpha : \text{for all } x,\ p \cdot x - \alpha \leqslant f(x)\} \\
&= \sup\{p \cdot y - \alpha : \alpha \geqslant f^*(p)\} \\
&= \sup_p p \cdot y - f^*(p) \\
&= f^{**}(y)
\end{aligned}$$

where the last equality follows from (4.16) applied to f^*. ∎

[7] To see this, note that if $q \neq p$, then for some x, we have $(p - q) \cdot x = \gamma \neq 0$. Then $y_\delta = \frac{\delta}{\gamma} x$ satisfies $p \cdot y_\delta - q \cdot y_\delta = \delta$. By choosing δ appropriately we can obtain either $f(y_\delta) > g(y_\delta)$ or vice-versa.

224 Lemma *The convex conjugate of a convex function f is proper if and only if f is proper. The concave conjugate of a concave function g is proper if and only if g is proper.*

Proof: The improper constant functions $+\infty$ and $-\infty$ are clearly conjugate to each other. It follows from Proposition 223 that otherwise the conjugate is proper. ∎

An economic interpretation of the conjugate

Consider a multiproduct firm that can produce n different outputs. Let the convex function $f\colon \boldsymbol{R}^{\mathrm{n}}_{+} \to \boldsymbol{R}$ be its cost function. That is, $f(x)$ is the cost of producing the output vector $x \in \boldsymbol{R}^{\mathrm{n}}_{+}$. Convexity of the cost function captures the property of decreasing returns to scale in production. Let p the vector of output prices. Then $p \cdot x - f(x)$ is the firm's profit from choosing the output vector x. The convex conjugate f^* is just the firm's profit function, that is, $f^*(p)$ is the maximum profit the firm can make at prices p.

Or consider a firm that produces one good from n inputs, where the price of the outputs has been normalized to unity. Let the concave function g be its production function, so that $g(x)$ is the quantity (and value) of output from the input vector $x \in \boldsymbol{R}^{\mathrm{n}}_{+}$. Concavity of the production function again captures the property of decreasing returns to scale in production. Let p the vector of input prices. Then $g(x) - p \cdot x$ is the firm's profit from choosing the input vector x. The concave conjugate $g^*(p) = \inf_x p \cdot x - g(x) = -\sup_x g(x) - p \cdot x$ is just the negative of the firm's profit function, that is, $-g^*(p)$ is the maximum profit the firm can make at input prices p.

4.22 Subgradients and conjugates

The convex conjugate is defined as a supremum, $f^*(p) = \sup_x p \cdot x - f(x)$. Suppose this supremum is actually a maximum, that is, there is some x for which $f^*(p) = p \cdot x - f(x)$. Rearranging we see that Fenchel's inequality holds as an equality, $f^*(p) + f(x) = p \cdot x$. In fact, Fenchel's inequality holds as an equality if and only if the supremum is attained. This equality also characterizes subgradients.

Recall that p is a subgradient of the convex function f at x, written $p \in \partial f(x)$, if it satisfies the subgradient inequality:

$$f(y) \geqslant f(x) + p \cdot (y - x)$$

for every y. The right hand side is an affine function h of y, namely $h(y) = p \cdot y - \big(p \cdot x - f(x)\big)$ dominated by f and agreeing with f at the point x, $(h(x) = f(x))$. Thus the conjugate f^* evaluated at p satisfies

$$f^*(p) = p \cdot x - f(x),$$

so Fenchel's inequality is satisfied as an equality.

225 Lemma *If f is a proper convex function, then*

$$f(x) + f^*(p) = p \cdot x \quad \Longleftrightarrow \quad \big(f \text{ is subdifferentiable at } x \text{ and } p \in \partial f(x)\big),$$

where f^* is the convex conjugate of f.

If g is a proper concave function, then

$$g(x) + g^*(p) = p \cdot x \quad \Longleftrightarrow \quad \big(f \text{ is superdifferentiable at } x \text{ and } p \in \partial g(x)\big),$$

where g^* is the concave conjugate of g.

Proof: I'll discuss only the convex case. We have already observed the ($\Longleftarrow$) implication, so assume Fenchel's inequality holds as an equality, $f(x) + f^*(p) = p \cdot x$. Then $f^*(p) = p \cdot x - f(x)$. By definition of f^* this means the affine function $h(y) = p \cdot y - \big(p \cdot x - f(x)\big)$ is dominated by f, which is just the subgradient inequality for p. ∎

When f satisfies the additional property that it is closed, then we have the next result.

226 Corollary *If f is a proper closed convex (or concave) function, then $p \in \partial f(x)$ if and only if $x \in \partial f^*(p)$.*

Proof: By Lemma 225, $p \in \partial f(x)$ if and only if the Fenchel inequality holds with equality, that is, $f(x) + f^*(p) = p \cdot x$. But since f is a proper closed convex function, $f = \operatorname{cl} f = f^{**}$ by Proposition 223, so this is equivalent to $f^{**}(x) + f^*(p) = p \cdot x$, which by Lemma 225 applied to the proper convex function f^* holds if and only if $x \in \partial f^*(p)$. ∎

Using the fact that $f^{**} = f$ for closed functions (Proposition 223), we have the following.

227 Corollary *If f is a proper closed convex function, then the following are equivalent.*

1. $f(x) + f^*(p) = p \cdot x$.
2. $p \in \partial f(x)$.
3. $x \in \partial f^*(p)$.
4. $p \cdot x - f(x) = \max_y p \cdot y - f(y)$.
5. $p \cdot x - f^*(p) = \max_q q \cdot x - f^*(q)$.

If f is a proper closed concave function, then the following are equivalent.

1. $f(x) + f^*(p) = p \cdot x$.
2. $p \in \partial f(x)$.
3. $x \in \partial f^*(p)$.
4. $p \cdot x - f(x) = \min_y p \cdot y - f(y)$.
5. $p \cdot x - f^*(p) = \min_q q \cdot x - f^*(q)$.

(The conjugate is the concave conjugate in case f is concave.)

4.23 Support functions and conjugates

Recall that the convex analyst's indicator function is defined by

$$\delta(x \mid C) = \begin{cases} 0 & x \in C \\ +\infty & x \notin C. \end{cases}$$

The indicator of C is a convex function if and only if C is a convex set, and is a proper closed convex function if and only C is a nonempty closed convex set.

Now

$$p \cdot x - \delta(x \mid C) = \begin{cases} -\infty & x \notin C \\ p \cdot x & x \in C \end{cases}$$

so the conjugate $\delta^*(p \mid C)$ satisfies

$$\delta^*(p \mid C) = \sup_x p \cdot x - \delta(x \mid C) = \sup_{x \in C} p \cdot x,$$

so

$$\delta^*(p \mid C) = \pi_C(p),$$

where π_C is the economist's profit function. On the other hand if we look at the concave indicator function $-\delta(x \mid C)$, we get

$$(-\delta)^*(p \mid C) = -(\delta^*)(-p \mid C) = \mu_C(p).$$

There is another relation between conjugates and support functions. For a proper convex function f, (4.17) implies

$$\delta^*\big((p,-1) \mid \operatorname{epi} f\big) = f^*(p).$$

This justifies the remarks at the beginning of Section 4.21.

4.24 Conjugate functions and maximization: Fenchel's Duality Theorem

Let $f\colon C \to \boldsymbol{R}$ be a proper closed concave function on a convex subset of $\boldsymbol{R}$. Assume x^* maximizes f over C. If we extend f to be concave on all of $\boldsymbol{R}^{\mathrm{n}}$ by setting $f(x) = -\infty$ for x not in C, then x^* still maximizes f over $\boldsymbol{R}^{\mathrm{n}}$. Now by Theorem 190 we have $0 \in \partial f(x^*)$, so by Corollary 227, it follows that $x^* \in \partial f^*(0)$ and $f(x^*) = -f^*(0)$, where f^* is the (concave) conjugate of f. This (fortunately) agrees with the definition $f^*(p) = \sup\{\alpha : f(x) \geqslant p \cdot x - \alpha \text{ for all } x \in \boldsymbol{R}^{\mathrm{n}}\}$, which reduces to $f^*(0) = -\sup_x f(x)$.

But there is a more interesting relationship between conjugates and maximization. The next result is due to Fenchel [51, § 47–48, pp. 105–109]. It also appears in Rockafellar [130, Theorem 31.1, p. 327–329], who also provides a number of variations. It states that every concave maximization problem has a dual convex minimization problem, and the solutions to the two coincide.

228 Fenchel's Duality Theorem (Concave version) *Let f be a proper convex function and g be a proper concave function on $\boldsymbol{R}^{\mathrm{n}}$. If $\operatorname{ri}\operatorname{dom} f \cap \operatorname{ri}\operatorname{dom} g \neq \varnothing$, then*

$$\sup_x g(x) - f(x) = \inf_p f^*(p) - g^*(p),$$

where f^* is the convex conjugate of f and g^* is the concave conjugate of g. Moreover, the infimum is attained for some $\bar{p} \in \boldsymbol{R}^{\mathrm{n}}$.

If in addition f and g are closed and if $\operatorname{ri}\operatorname{dom} f^* \cap \operatorname{ri}\operatorname{dom} g^* \neq \varnothing$, then the supremum is attained at some $\bar{x} \in \operatorname{dom} f \cap \operatorname{dom} g$, and is finite.

Interchanging f and g and sup and inf gives the following.

229 Fenchel's Duality Theorem (Convex version) *Let f be a proper convex function and g be a proper concave function on $\boldsymbol{R}^{\mathrm{n}}$. If $\operatorname{ri}\operatorname{dom} f \cap \operatorname{ri}\operatorname{dom} g \neq \varnothing$, then*

$$\inf_x f(x) - g(x) = \sup_p g^*(p) - f^*(p),$$

where f^* is the convex conjugate of f and g^* is the concave conjugate of g. Moreover, the supremum is attained for some $\bar{p} \in \boldsymbol{R}^{\mathrm{n}}$. (Note that since the functions are extended real-valued, the supremum may be attained yet be infinite.)

If in addition, f and g are closed and if $\operatorname{ri}\operatorname{dom} f^* \cap \operatorname{ri}\operatorname{dom} g^* \neq \varnothing$, then the infimum is attained at some $\bar{x} \in \operatorname{dom} f \cap \operatorname{dom} g$, and is finite.

Proof of concave version: From Fenchel's Inequality, for every x and p,

$$f(x) + f^*(x) \geqslant p \cdot x \geqslant g(x) + g^*(p), \tag{4.22}$$

so

$$f^*(p) - g^*(p) \geqslant g(x) - f(x) \qquad \text{for all } x, p.$$

(Since we are subtracting extended real-valued functions, we should make sure that the meaningless expression $\infty - \infty$ does not occur. Now f and g are proper convex and concave functions respectively, so $f(x) > -\infty$ and $g(x) < \infty$ for all x, so $g(x) - f(x)$ is a well defined extended real number. By Fenchel's inequality, $g^*(p) \leqslant p \cdot x - g(x) < \infty$ for any $x \in \operatorname{dom} f \cap \operatorname{dom} g$. Similarly, we have $f^*(p) > -\infty$ for every p. Thus $f^*(p) - g^*(p)$ is a well defined extended real number.)

Therefore taking the infimum on the right and the supremum on the left,

$$\inf_p f^*(p) - g^*(p) \geqslant \sup_x g(x) - f(x). \tag{4.23}$$

We need now to show that there is no "duality gap," that is, the reverse inequality also holds. Consider first the case where the right hand side of this inequality is ∞. Then the left hand side is also ∞ and the infimum is attained for every p.

So assume $\alpha = \sup_{x \in C} g(x) - f(x) < \infty$. Then in fact, α is not $-\infty$ (look at any $x \in \operatorname{ri}\operatorname{dom} f \cap \operatorname{ri}\operatorname{dom} g$), and so finite. Moreover α satisfies

$$\begin{aligned} \alpha &= \sup_x g(x) - f(x) \\ &= \inf\{\beta : \forall\, x \ \beta \geqslant g(x) - f(x)\} \\ &= \inf\{\beta : \forall\, x \ f(x) + \beta \geqslant g(x)\}. \end{aligned}$$

Now consider the epigraph A of $f + \alpha$

$$A = \{(x, \beta) \in \boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R} : \beta \geqslant f(x) + \alpha\}$$

and the strict hypograph B of g,

$$B = \{(x, \beta) \in \boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R} : \beta < g(x)\}.$$

Then A and B are disjoint nonempty convex subsets of $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$, so by Theorem 165 there exists a nonzero $(\bar{p}, \lambda) \in \boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}$ that properly separates A and B, say $(\bar{p}, \lambda) \cdot A \leqslant (\bar{p}, \lambda) \cdot B$.

It follows then that $\lambda < 0$. To see this, suppose $\lambda = 0$. Then proper separation of A and B by $(\bar{p}, 0)$ implies $\inf_{x \in \operatorname{dom} f} \bar{p} \cdot x < \sup_{y \in \operatorname{dom} g} \bar{p} \cdot y$, which implies that $\bar{p}$ properly separates $\operatorname{dom} f$ and $\operatorname{dom} g$, which contradicts $\operatorname{ri}\operatorname{dom} f \cap \operatorname{ri}\operatorname{dom} g \neq \varnothing$ (Theorem 165). If $\lambda > 0$, then for large enough $\beta > 0$, we have $(\bar{p}, \lambda) \cdot (x, \beta) > (\bar{p}, \lambda) \cdot \big(x, g(x)\big)$, a contradiction of the separation inequality.

Thus without loss of generality we may take $\lambda = -1$. Then separation implies

$$\bar{p} \cdot x - f(x) - \alpha \leqslant \bar{p} \cdot x - g(x) \qquad \text{for all } x.$$

Taking the supremum on the left and the infimum on the right gives

$$f^*(\bar{p}) - \alpha = \sup_x \bar{p} \cdot x - f(x) - \alpha \leqslant \inf_x \bar{p} \cdot x - g(x) = g^*(\bar{p}).$$

Recalling the definition of α gives

$$\sup_x g(x) - f(x) = \alpha \geqslant f^*(\bar{p}) - g^*(\bar{p}) \geqslant \inf_p f^*(p) - g^*(p).$$

This proves the reverse inequality, so these are actually equalities, there is no gap, and $\bar{p}$ attains the infimum.

Now assume that f and g are closed, and that $\operatorname{ri}\operatorname{dom} f^* \cap \operatorname{ri}\operatorname{dom} g^* \neq \varnothing$. Apply the argument just used to the functions f^* and g^*, to get $\sup_p g^*(p) - f^*(p) = \inf_x f^{**}(x) - g^{**}(x)$ and is finite. Now use the fact that $f = f^{**}$ and $g = g^{**}$, to get that the infimum of $f - g$, and hence the supremum of $g - f$, is attained for some $\bar{x}$. ∎

4.25 The calculus of sub/superdifferentials

230 Theorem *(cf. Aubin [14, Theorem 3.4, p. 37].) Let f and g be proper closed concave (or convex) functions on $\boldsymbol{R}^{\mathrm{n}}$. If $\operatorname{ri}\operatorname{dom} f \cap \operatorname{ri}\operatorname{dom} g \neq \varnothing$, then for each $p \in \operatorname{dom} f^* \cap \operatorname{dom} g^*$, there is some q satisfying*

$$(f+g)^*(p) = f^*(p-q) + g^*(q).$$

Proof: I'll prove the concave case. By definition,

$$\begin{aligned}(f+g)^*(p) &= \inf_x \; p\cdot x - \big(f(x)+g(x)\big)\\ &= \inf_x \; (-g(x)) - (f(x) - p\cdot x)\\ &= \sup_q \; (f-p)^*(q) - (-g)^*(q),\end{aligned}$$

where the last equality is the convex version of Fenchel's Duality Theorem 229 applied to the convex function $-g$ and the concave function $x \mapsto f(x) - p\cdot x$. Moreover this supremum is attained for some $\tilde{q}$. Now recall that $(-g)^*(q) = -g^*(-q)$ ((4.21)), so define $\bar{q} = -\tilde{q}$. Furthermore,

$$(f-p)^*(q) = \sup_x \; q\cdot x - (f(x) - p\cdot x) = \sup_x \; (p+q)\cdot x - f(x) = f^*(p+q).$$

Substituting above yields

$$\begin{aligned}(f+g)^*(p) &= f^*(p+\tilde{q}) + g^*(-\tilde{q})\\ &= f^*(p-\bar{q}) + g^*(\bar{q}).\end{aligned}$$

∎

231 Theorem *Let f and g be proper closed concave (or convex) functions on $\boldsymbol{R}^{\mathrm{n}}$. If the point x belongs to $\operatorname{ri}\operatorname{dom} f \cap \operatorname{ri}\operatorname{dom} g$, then*

$$\partial(f+g)(x) = \partial f(x) + \partial g(x)$$

Proof: (cf. Aubin [14, Theorem 4.4, p. 52].) Note that x must belong to the relative interior of $\operatorname{dom}(f+g)$, so each of f, g, and $f+g$ is superdifferentiable at x. Moreover $f+g$ is a proper closed concave function, as it is easy to see that the sum of upper (or lower) semicontinuous functions is upper (or lower) semicontinuous.

It is easy to see that $\partial f(x) + \partial g(x) \subset \partial(f+g)(x)$—just add the supergradient inequalities. That is, if $p \in \partial f(x)$ and $q \in \partial g(x)$, for each y we have

$$f(x) + p\cdot(y-x) \geqslant f(y) \qquad \text{and} \qquad g(x) + q\cdot(y-x) \geqslant g(y),$$

so

$$(f+g)(x) + (p+q)\cdot(y-x) \geqslant (f_g)(y).$$

That is, $p+q \in \partial(f+g)(x)$. (By the way, the assumption that $x \in \operatorname{ri}\operatorname{dom} f \cap \operatorname{ri}\operatorname{dom} g$ is not needed for this part.)

For the reverse inclusion, let p belong to the superdifferential $\partial(f+g)(x)$. Then by Corollary 227

$$(f+g)(x) + (f+g)^*(p) = p\cdot x,$$

but by Theorem 230 (this is where the assumption that $x \in \operatorname{ri}\operatorname{dom} f \cap \operatorname{ri}\operatorname{dom} g$ is needed), there is a q satisfying $(f+g)^*(p) = f^*(p-q) + g^*(q)$, so we have

$$f(x) + f^*(p-q) + g(x) + g^*(q) = p\cdot x.$$

Subtracting $q \cdot x$ from both sides and rearranging gives

$$f(x) + f^*(p-q) + g(x) + g^*(q) - q \cdot x = (p-q) \cdot x$$
$$\big[f(x) + f^*(p-q) - (p-q) \cdot x\big] + \big[g(x) + g^*(q) - q \cdot x\big] = 0.$$

But by Fenchel's Inequality for concave functions, each of the two bracketed terms is nonpositive, so each must be zero. But then by Corollary 227, we have

$$p - q \in \partial f(x) \qquad \text{and} \qquad q \in \partial g(x).$$

Thus $p = (p-q) + q$ belongs to $\partial f(x) + \partial g(x)$. ∎

232 Theorem $\partial g(Ax) =$???

Proof: ********** ∎

Returning to Fenchel's Duality Theorem, the first order necessary condition for a maximum of $g - f$ at x^* is that $0 \in \partial(g-f)(x^*)$. When $x^* \in \operatorname{ri}\operatorname{dom} f \cap \operatorname{ri}\operatorname{dom} g$, so that $\partial(g-f)(x^*) = \partial g(x^*) - \partial f(x^*)$, we have $\partial f(x^*) \cap \partial g(x^*) \neq \varnothing$. A generalization of this condition that is also sufficient is given in the next theorem.

233 Theorem *Let f be a closed proper convex function and g be a closed proper concave function on $\boldsymbol{R}^{\mathrm{n}}$. Assume $\operatorname{ri}\operatorname{dom} f \cap \operatorname{ri}\operatorname{dom} g^* \neq \varnothing$ and $\operatorname{ri}\operatorname{dom} f^* \cap \operatorname{ri}\operatorname{dom} g \neq \varnothing$. Then the following conditions are equivalent.*

1. $\sup_{x \in C} g(x) - f(x) = g(\bar{x}) - f(\bar{x}) = f^*(\bar{p}) - g^*(\bar{p}) = \inf_p f^*(p) - g^*(p)$.
2. $\bar{p} \in \partial g(\bar{x})$ *and* $\bar{x} \in \partial f^*(\bar{p})$.
3. $\bar{p} \in \partial f(\bar{x})$ *and* $\bar{x} \in \partial g^*(\bar{p})$.
4. $\bar{p} \in \partial g(\bar{x}) \cap \partial f(\bar{x})$ *and* $\bar{x} \in \partial f^*(\bar{p}) \cap \partial g^*(\bar{p})$.

Proof: (1) $\implies$ (4): If

$$\sup_p g^*(p) - f^*(p) = g^*(\bar{p}) - f^*(\bar{p}) = f(\bar{x}) - g(\bar{x}) = \inf_x f(x) - g(x),$$

rearranging and using (4.22) gives

$$g^*(\bar{p}) + g(\bar{x}) = \bar{p} \cdot \bar{x} = f(\bar{x}) + f^*(\bar{p}).$$

Thus by Corollary 227,

$$\bar{x} \in \partial f^*(\bar{p}), \qquad \bar{x} \in \partial g^*(\bar{p}), \qquad \bar{p} \in \partial f(\bar{x}), \qquad \text{and} \quad \bar{p} \in \partial g(\bar{x}).$$

(2) $\implies$ (1): From Corollary 227 we have $\bar{p} \in \partial g(\bar{x})$ implies $g(\bar{x}) + g^*(\bar{p}) = \bar{p} \cdot \bar{x}$, and $\bar{x} \in \partial f^*(\bar{p})$ implies $f(\bar{x}) + f^*(\bar{p}) = \bar{p} \cdot \bar{x}$. Therefore

$$g(\bar{x}) + g^*(\bar{p}) = f(\bar{x}) + f^*(\bar{p})$$

so

$$g(\bar{x}) - f(\bar{x}) = f^*(\bar{p}) - g^*(\bar{p}).$$

Moreover by (4.23) we have

$$\inf_p f^*(p) - g^*(p) \geqslant \sup_x g(x) - f(x).$$

Thus

$$g(\bar{x}) - f(\bar{x}) = f^*(\bar{p}) - g^*(\bar{p}) \geqslant \inf_p f^*(p) - g^*(p) \geqslant \sup_x g(x) - f(x) \geqslant g(\bar{x}) - f(\bar{x}),$$

so $g(\bar{x}) - f(\bar{x}) = \sup_x g(x) - f(x)$. Similarly, $f^*(\bar{p}) - g^*(\bar{p}) = \inf_p f^*(p) - g^*(p)$.

The implication (3) $\implies$ (1) is similar, and (4) $\implies$ (3) and (4) $\implies$ (2) are trivial. ∎

234 Example (Fenchel's Duality Theorem and the Support Function) Recall that the support function μ_A of a nonempty closed convex set A is given by $\mu_A(p) = \inf_{x\in A} p \cdot x$. Recall that the indicator function $\delta(x \mid A)$ satisfies $\delta(x \mid A) = 0$ for $x \in A$, and $\delta(x \mid A) = \infty$ for $x \notin A$. Let

$$f(x) = q \cdot x \qquad \text{and} \qquad g(x) = -\delta(x \mid A).$$

Then f is a proper closed convex function and g is a proper closed concave function, and

$$\mu_A(q) = \inf_x f(x) - g(x).$$

The dual problem is to find $\sup_p g^*(p) - f^*(p)$.

Now $\operatorname{ri}\operatorname{dom} f = \boldsymbol{R}^{\mathrm{n}}$ and $\operatorname{ri}\operatorname{dom} g = \operatorname{ri} A$. By Theorem 229, the supremum is attained, and it is easy to see that it is attained at q. Thus $0 \in \partial(f^* - g^*)(q)$.

Recall from Example 221 that $f^*(q) = 0$ and $f^*(p) = \infty$ for $p \neq q$. Thus $\operatorname{ri}\operatorname{dom} f^* = \{q\}$. Also the concave conjugate of the concave function g satisfies $g^*(p) = \inf_x p\cdot x - g(p) = \inf_{x\in A} p\cdot x = \mu_A(p)$. So $\operatorname{dom} g^* = \{p : \mu_A(p) \text{ is finite}\}$.

In order to apply the remainder of Fenchel's Duality Theorem 229 or Theorem 233, we must have $q \in \operatorname{ri}\operatorname{dom}\mu_A$. Assume this for a moment. In that case, $\bar{x}$ achieves the infimum ($q \cdot \bar{x} = \mu_A(q)$ if and only if there exists $\bar{p}$ satisfying

$$\bar{x} \in \partial g^*(\bar{p}), \qquad \bar{p} \in \partial f(\bar{x}).$$

Now $\partial f(x) = q$ for any x, so $\bar{p} \in \partial f(\bar{x})$ if and only if $\bar{p} = q$. So $\bar{x}$ minimizes $q \cdot x$ over A if and only if $\bar{x} \in \partial\mu_A(q)$. This constitutes another proof of Theorem 193 for the case where $q \in \operatorname{ri}\operatorname{dom}\mu_A$.

Needs work.

Unfortunately the conditions under which $q \in \operatorname{ri}\operatorname{dom}\mu_A$ are not very simple to explain. See Rockafellar [130, Corollary 13.3.4, p. 117, and also p. 66]. □

4.26 Supergradients and cyclically monotone mappings

Recall that a real function $g\colon X \subset \boldsymbol{R} \to \boldsymbol{R}$ is increasing if $x \geqslant y$ implies $g(x) \geqslant g(y)$. Another way to say this is $\big[g(x) - g(y)\big](x - y) \geqslant 0$ for all x, y. Or equivalently, g is nondecreasing if

$$g(x)(y - x) + g(y)(x - y) \leqslant 0 \quad \text{for all } x, y.$$

More generally, a correspondence $\varphi : X \subset \boldsymbol{R}^{\mathrm{m}} \twoheadrightarrow \boldsymbol{R}^{\mathrm{m}}$ is **monotone (increasing)** if

$$(p_x - p_y) \cdot (x - y) \geqslant 0 \quad \text{for all } x, y \in X, \text{ and all } p_x \in \varphi(x), p_y \in \varphi(y).$$

We could also write this as $p_x \cdot (y-x) + p_y \cdot (x-y) \leqslant 0$. A mapping φ is **monotone decreasing** if the reverse inequality always holds.

There is a natural generalization of these conditions. A finite sequence $x_0, x_1, \ldots, x_n, x_{n+1}$ with $x_{n+1} = x_0$ is sometimes called a **cycle**. A mapping $g\colon U \subset \boldsymbol{R}^{\mathrm{m}} \to \boldsymbol{R}^{\mathrm{m}}$ is called **cyclically monotone (increasing)** if for every cycle $x_0, x_1, \ldots, x_n, x_{n+1} = x_0$ in U, we have

$$g(x_0) \cdot (x_1 - x_0) + g(x_1) \cdot (x_2 - x_1) + \cdots + g(x_n) \cdot (x_0 - x_n) \leqslant 0.$$

If the same sum is always $\geqslant 0$, we shall say that g is **cyclically monotone decreasing**.

More generally, a correspondence $\varphi\colon U \subset \boldsymbol{R}^{\mathrm{m}} \twoheadrightarrow \boldsymbol{R}^{\mathrm{m}}$ is called **cyclically monotone (increasing)**[8] if for every cycle $(x_0, p_0), (x_1, p_1), \ldots, (x_{n+1}, p_{n+1}) = (x_0, p_0)$ in the graph of φ, that

[8] Most authors define monotone and cyclically monotone correspondences to be increasing, and do not make a definition for decreasing monotonicity. This is because mathematicians find convex functions (such as norms) to be natural, and as we shall see below there is an important relationship between convex functions and (cyclically) monotone increasing mappings. Economists however find concave functions to be naturally occurring (as in production functions) so it seems natural to introduce a term for (cyclically) monotone decreasing mappings. Just keep in mind that for every statement about convex functions, there is a corresponding one for concave functions derived by replacing f by $-f$.

is, with $p_i \in \varphi(x_i)$ for all i, we have

$$p_0 \cdot (x_1 - x_0) + p_1 \cdot (x_2 - x_1) + \cdots + p_n \cdot (x_0 - x_n) \leqslant 0.$$

We mention that if $m = 1$ ($\boldsymbol{R}^{\mathrm{m}} = \boldsymbol{R}$) then a function g is cyclically monotone if and only if it is monotone. For $m \geqslant 2$, there are monotone functions that are not cyclically monotone.

235 Example (Monotonicity vs. cyclical monotonicity) This example is based on a remark of Rockafellar [130, p. 240]. Define the function $g\colon \boldsymbol{R}^2 \to \boldsymbol{R}^2$ by

$$g(x, y) = (2y - x, -y).$$

Then g is monotone (decreasing):

$$\begin{aligned} g(x_0, y_0) \cdot (x_1 - x_0, y_1 - y_0) + g(x_1, y_1) \cdot (x_0 - x_1, y_0 - y_1) & \\ = \ & (2y_0 - x_0, -y_0) \cdot (x_1 - x_0, y_1 - y_0) + (2y_1 - x_1, -y_1) \cdot (x_0 - x_1, y_0 - y_1) \\ = \ & (2y_0 - x_0, -y_0) \cdot (x_1 - x_0, y_1 - y_0) - (2y_1 - x_1, -y_1) \cdot (x_1 - x_0, y_1 - y_0) \\ = \ & (2y_0 - x_0 - 2y_1 + x_1, y_1 - y_0) \cdot (x_1 - x_0, y_1 - y_0) \\ = \ & (x_1 - x_0)^2 - 2(y_1 - y_0)(x_1 - x_0) + (y_1 - y_0)^2 \\ = \ & \big((x_1 - x_0) - (y_1 - y_0)\big)^2 \\ \geqslant \ & 0. \end{aligned}$$

But g is not cyclically monotone (decreasing): Consider the cycle $(0, -2)$, $(2, -2)$, $(3, 0)$, $(0, -2)$. Then

$$\begin{aligned} g(0,-2) \cdot \big((2,-2) - (0,-2)\big) + g(2,-2) \cdot \big((3,0) - (2,-2)\big) + g(3,0) \cdot \big((0,-2) - (3,0)\big) & \\ = \ & (-4, 2) \cdot (2, 0) + (-6, 2) \cdot (1, 2) + (-3, 0) \cdot (-3, -2) \\ = \ & -8 - 2 + 9 \\ = \ & -1. \end{aligned}$$

In fact, Rockafellar asserts the following. Let $g\colon \boldsymbol{R}^{\mathrm{n}} \to \boldsymbol{R}^{\mathrm{n}}$ be linear, that is, $g(x) = Ax$, where A is an $n \times n$ matrix. If A is negative quasi-semidefinite but not symmetric, then g is monotone decreasing, but not cyclically monotone decreasing. □

The next result is a simple corollary of Theorem 178.

236 Corollary (Cyclical monotonicity of the derivative) *Let f be concave and differentiable on a convex open set $U \subset \boldsymbol{R}^{\mathrm{m}}$. Then the gradient mapping $x \mapsto f'(x)$ is cyclically monotone (decreasing). That is, for any cycle $x_0, x_1, \ldots, x_n, x_{n+1}$ in U with $x_{n+1} = x_0$, we have*

$$\sum_{k=0}^{n} f'(x_k) \cdot (x_{k+1} - x_k) \geqslant 0.$$

Proof: By Theorem 178, $f'(x_k) \cdot (x_{k+1} - x_k) \geqslant f(x_{k+1}) - f(x_k)$. Summing both sides gives

$$\sum_{k=0}^{n} f'(x_k) \cdot (x_{k+1} - x_k) \geqslant \sum_{k=0}^{n} \big[f(x_{k+1}) - f(x_k)\big] = 0,$$

where the last equality follows from the fact that $x_{n+1} = x_0$. ∎

Note that the gradient of a convex function is cyclically monotone (increasing).

The remarkable fact is that the supergradient correspondence is characterized by cyclical monotonicity. The next result is due to Rockafellar, and may be found (in different terminology) in his book [130, Theorem 24.8, p. 238].

237 Theorem (Rockafellar) *Let $C \subset \boldsymbol{R}^{\mathrm{m}}$ be a nonempty convex set and let $\varphi : C \twoheadrightarrow \boldsymbol{R}^{\mathrm{m}}$ be a correspondence with nonempty values. Then φ is cyclically monotone decreasing if and only if there is an upper semicontinuous concave function $f\colon C \to \boldsymbol{R}$ satisfying*

$$\varphi(x) \subset \partial f(x) \quad \text{for every } x \in C.$$

Proof: If $\varphi(x) \subset \partial f(x)$ for a concave f, then the definition of $\partial f(x)$ and the same argument used to prove Corollary 236 show that φ is cyclically monotone decreasing.

For the converse, assume φ is cyclically monotone decreasing. Fix any point x_0 in C and fix $p_0 \in \varphi(x_0)$. Given any finite sequence $(x_1, p_1), \dots, (x_n, p_n)$ in $\boldsymbol{R}^{\mathrm{m}} \times \boldsymbol{R}^{\mathrm{m}}$, define the affine function $g^{x_1,\dots,x_n}_{p_1,\dots,p_n}$ by

$$g^{x_1,\dots,x_n}_{p_1,\dots,p_n}(y) = p_0 \cdot (x_1 - x_0) + \cdots + p_n \cdot (y - x_n).$$

The construction of such functions $g^{x_1,\dots,x_n}_{p_1,\dots,p_n}$ is illustrated in Figures 4.6 and 4.7.

Now define the function $f\colon C \to \boldsymbol{R}$ to be the pointwise infimum of the $g^{x_1,\dots,x_n}_{p_1,\dots,p_n}$ as $(x_1, p_1), \dots, (x_n, p_n)$ ranges over all finite sequences in the graph of φ. That is,

$$f(y) = \inf\{g^{x_1,\dots,x_n}_{p_1,\dots,p_n}(y) : \forall i,\ x_i \in C,\ p_i \in \varphi(x_i)\}.$$

Since f is the pointwise infimum of a collection of continuous affine functions, it is concave by part 4 of Exercise 135, and upper semicontinuous by Proposition 31.

Cyclical monotonicity implies that the infimum defining f is finite, that is, $f(y) > -\infty$ for every $y \in C$. To see this, fix some p in $\varphi(y)$. Then by cyclical monotonicity

$$g^{x_1,\dots,x_n}_{p_1,\dots,p_n}(y) + p \cdot (x_0 - y) = p_0 \cdot (x_1 - x_0) + \cdots + p_n \cdot (y - x_n) + p \cdot (x_0 - y) \geqslant 0.$$

Rearranging gives

$$g^{x_1,\dots,x_n}_{p_1,\dots,p_n}(y) \geqslant p \cdot (y - x_0).$$

Therefore $f(y) \geqslant p \cdot (y - x_0) > -\infty$ for any y.

We claim that f is the desired function. That is, any x, y in C and any $p \in \varphi(x)$ satisfy the supergradient inequality

$$f(x) + p \cdot (y - x) \geqslant f(y).$$

To see this, let $\varepsilon > 0$ be given. Then by the definition of f, since $f(x)$ is finite, there is a finite sequence $(x_0, p_0), \dots, (x_n, p_n)$ in the graph of φ with

$$f(x) + \varepsilon > g^{x_1,\dots,x_n}_{p_1,\dots,p_n}(x).$$

Extend this sequence by appending (x, p). Again by the definition of f, for all y,

$$g^{x_1,\dots,x_n,x}_{p_1,\dots,p_n,p}(y) \geqslant f(y).$$

But

$$g^{x_1,\dots,x_n,x}_{p_1,\dots,p_n,p}(y) = p_0 \cdot (x_1 - x_0) + \cdots + p_n \cdot (x - x_n) + p \cdot (y - x) = g^{x_1,\dots,x_n}_{p_1,\dots,p_n}(x) + p \cdot (y - x).$$

Combining these gives

$$f(x) + \varepsilon + p \cdot (y - x) > g^{x_1,\dots,x_n}_{p_1,\dots,p_n}(x) + p \cdot (y - x) = g^{x_1,\dots,x_n,x}_{p_1,\dots,p_n,p}(y) \geqslant f(y).$$

Since $\varepsilon > 0$ is arbitrary, we conclude that $f(x) + p \cdot (y - x) \geqslant f(y)$, so indeed $\varphi(x) \subset \partial f(x)$. ∎

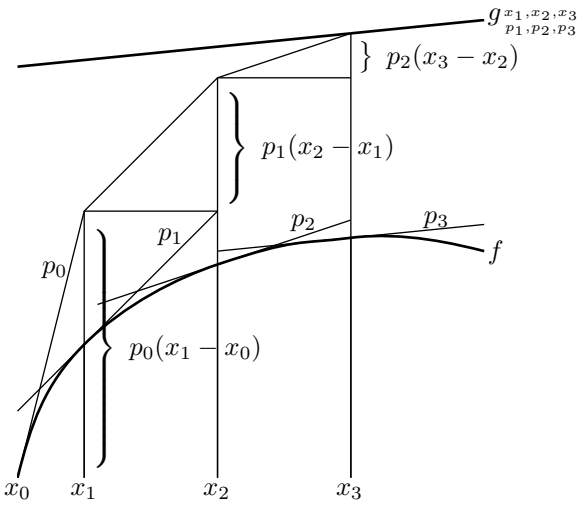

Figure 4.6. The function $g^{x_1,x_2,x_3}_{p_1,p_2,p_3}(y) = p_0 \cdot (x_1 - x_0) + p_1 \cdot (x_2 - x_1) + p_2 \cdot (x_3 - x_2) + p_3 \cdot (y - x_3)$, where each p_i is taken from $\partial f(x_i)$.

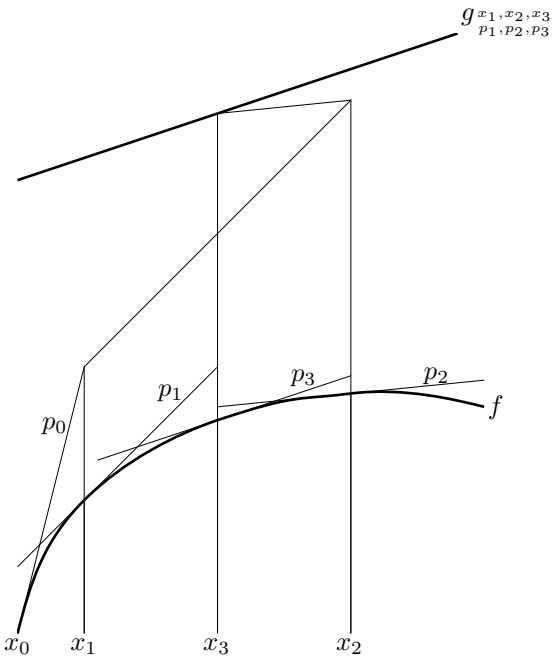

Figure 4.7. Another version of $g^{x_1,x_2,x_3}_{p_1,p_2,p_3}(y) = p_0 \cdot (x_1 - x_0) + p_1 \cdot (x_2 - x_1) + p_2 \cdot (x_3 - x_2) + p_3 \cdot (y - x_3)$, where the x_i have been reordered.

4.27 Monotonicity and second derivatives

From Corollary 236 we know that the gradient of a concave function $f\colon C \to \boldsymbol{R}$, where C is an open convex set in $\boldsymbol{R}^{\mathrm{n}}$, is cyclically monotone (decreasing). That is, it satisfies

$$\sum_{k=0}^{n} f'(x_k) \cdot (x_{k+1} - x_k) \geqslant 0.$$

Therefore it is monotone (decreasing). That is,

$$f'(x_0) \cdot (x_1 - x_0) + f'(x_1) \cdot (x_0 - x_1) \geqslant 0,$$

which can be rearranged as

$$\big(f'(x_1) - f'(x_0)\big) \cdot (x_1 - x_0) \leqslant 0.$$

This is enough to show that the second differential (if it exists) is negative semidefinite.

Consider a point x in C and choose v so that $x \pm v$ to C. Then by monotonicity with $x_0 = x$ and $x_1 = x + \lambda v$,

$$\big(f'(x + \lambda v) - f'(x)\big) \cdot (\lambda v) \leqslant 0.$$

Dividing by the positive quantity λ^2 implies

$$v \cdot \frac{\big(f'(x + \lambda v) - f'(x)\big)}{\lambda} \leqslant 0.$$

Define the function $g\colon (-1, 1) \to \boldsymbol{R}$ by

$$g(\lambda) = v \cdot f'(x + \lambda v).$$

In particular, if f is twice differentiable, then by the Chain Rule

$$D^2 f(x)(v, v) = g'(0) = \lim_{\lambda \to 0} v \cdot \frac{g(\lambda) - g(0)}{\lambda} = \lim_{\lambda \to 0} v \cdot \frac{\big(f'(x + \lambda v) - f'(x)\big)}{\lambda} \leqslant 0.$$

Thus the Hessian matrix $f''(x)$ is negative semidefinite, which gives another proof of half of Fact 213.

238 Remark At this point I was a bit confused. If you are not confused, you may not wish to read this.

We have just shown that if a twice differentiable function has a monotone gradient, then it has negative semidefinite Hessian, so it is concave, and therefore its gradient is actually cyclically monotone. Thus every monotone gradient is cyclically monotone. Now Theorem 237 says that every cyclically monotone vector field is a selection from the subdifferential of a concave function. I am embarrassed to admit it, but I thought for a while therefore that the argument above allowed me to conclude that every monotone vector field is a selection from the subdifferential of a concave function, which is a stronger claim (and not true).

What the argument above shows is this: Every monotone vector field that also happens to be a gradient of a twice differentiable function is indeed cyclically monotone. But, there are differentiable vector fields that are not gradients of twice differentiable functions. (A vector field is just a function from $\boldsymbol{R}^{\mathrm{n}}$ into $\boldsymbol{R}^{\mathrm{n}}$. If it is the gradient of a real function f, then f is called the potential of the field.) The reason for this is that second differentials are symmetric (Corollary 101). So if $x \mapsto g(x)$ is a gradient of a twice differentiable function f, then

$$D_j g_i(x) = D_i D_j f(x) = D_j D_i f(x) = D_i g_j(x).$$

Now consider the vector field of Example 235, namely $g\colon \boldsymbol{R}^2 \to \boldsymbol{R}^2$ defined by

$$g(x, y) = (2y - x, -y).$$

This vector field is continuously differentiable, but

$$D_1 g_2 = 0, \qquad D_2 g_1 = 2,$$

so g cannot be the gradient of any twice differentiable function. However, as we saw in Example 235, g is monotone (decreasing), but not cyclically monotone (decreasing).

By the way, this is analogous to the "integrability problem" in demand theory. The Weak Axiom of Revealed Preference can be used to show that the Slutsky matrix is negative quasidefinite (negative without necessarily being symmetric), see, e.g., Samuelson [136, pp. 109–111] or Kihlstrom, MasColell, and Sonnenschein [92], but it takes the Strong Axiom to show symmetry: Gale [56], Houthakker [78], Uzawa [152].

Now let's return to support functions.

239 Lemma *Suppose $x(p)$ minimizes $p \cdot x$ over the nonempty set A. Suppose further that it is the unique minimizer of $p \cdot x$ over $\overline{\text{co}}\, A$. If $\frac{\partial^2 \mu_C(p)}{\partial p_i^2}$ exists (or equivalently $\frac{\partial x(p)}{\partial p_i}$ exists), then*

$$\frac{\partial x_i(p)}{\partial p_i} \leqslant 0.$$

Do we need twice differentiability or just the existence of the second partial????

Proof: This follows from Corollary 194 and the discussion above. ∎

This, by the way, summarizes almost everything interesting we now about cost minimization.

4.28 Solutions of systems of equalities and inequalities

In this section we present some basic results on the existence of solutions to linear equalities and inequalities. These results are in the form of alternatives, that is, "an opportunity for choice between two things, courses, or propositions, either of which may be chosen, but not both" [117]. While it is possible to prove these results in inequalities in a purely algebraic fashion (cf. Gale [57, Chapter 2]), the geometric approach is illuminating.

> *In this section we shall adopt David Gale's [57] notation, which does not distinguish between row and column vectors. This means that if A is an $n \times m$ matrix, and x is a vector, and I write Ax, you infer that x is an m-dimensional column vector, and if I write yA, you infer that y is an n-dimensional row vector. The notation yAx means that x is an m-dimensional column vector, y is an n-dimensional row vector, and yAx is the scalar $yA \cdot x = y \cdot Ax$.*

Consider the system of linear equations

$$Ax = b.$$

If A has an inverse, then this system always has a unique solution, namely $x = A^{-1}b$. But even if A does not have an inverse, the system may have a solution, possibly several. This brings up the question of how to characterize the existence of a solution. The answer is given by the Fredhom Alternative, which may be found in Gale [57, Theorem 2.5] or Franklin [54, Example 4, p. 57].

240 Theorem (Fredholm Alternative) *Exactly one of the two following alternatives holds.*

$$\exists x \quad Ax = b. \tag{4.24}$$

$$\exists y \quad yA = 0 \text{ and } y \cdot b > 0. \tag{4.25}$$

Proof: It is easy to see that both (4.24) and (4.25) cannot be true, for then we would have $0 = 0 \cdot x = yAx = y \cdot b > 0$, a contradiction. Let M be the subspace spanned by the columns of A, and suppose (4.24) is false. That is, b does not belong to M. Then by the strong Separating Hyperplane Theorem 160 there is a nonzero vector y strongly separating the compact convex set $\{b\}$ from the closed convex set M, that is, $y \cdot b > y \cdot z$ for each $z \in M$. Since M is a subspace we have $y \cdot z = 0$ for every $z \in M$, and in particular for each column of A, so $yA = 0$ and $y \cdot b > 0$, which is just (4.25). ∎

The following corollary about linear functions is true in quite general linear spaces, see Aliprantis and Border [3, Theorem 5.91, p. 212], but we shall provide another proof using some of the special properties of $\boldsymbol{R}^{\mathrm{n}}$. Wim Luxemburg refers to this result as the Fundamental Theorem of Duality.

241 Corollary *Let $p^0, p^1, \dots, p^m \in \boldsymbol{R}^{\mathrm{n}}$ and suppose that $p^0 \cdot v = 0$ for all v such that $p^i \cdot v = 0$, $i = 1, \dots, m$. Then p^0 is a linear combination of $p^1, \dots, p^m$. That is, there exist scalars $\mu_1, \dots, \mu_m$ such that $p^0 = \sum_{i=1}^m \mu_i p^i$.*

Proof: Consider the matrix A whose columns are $p^1, \dots, p^m$, and set $b = p^0$. By hypothesis alternative (4.25) of Theorem 240 is false, so alternative (4.24) must hold. But that is precisely the conclusion of this theorem. ∎

Proof using orthogonal decomposition: Let $M = \operatorname{span}\{p^1, \dots, p^m\}$ and orthogonally project p^0 onto M to get $p^0 = p^0_M + p^0_\perp$, where $p^0_M \in M$ and $p^0_\perp \perp M$. That is, $p^0_\perp \cdot p = 0$ for all $p \in M$. In particular, $p^i \cdot p^0_\perp = 0$, $i = 1, \dots, m$. Consequently, by hypothesis, $p^0 \cdot p^0_\perp = 0$ too. But

$$0 = p^0 \cdot p^0_\perp = p^0_M \cdot p^0_\perp + p^0_\perp \cdot p^0_\perp = 0 + \|p^0_\perp\|.$$

Thus $p^0_\perp = 0$, so $p^0 = p^0_M \in M$. That is, p^0 is a linear combination of $p^1, \dots, p^m$. ∎

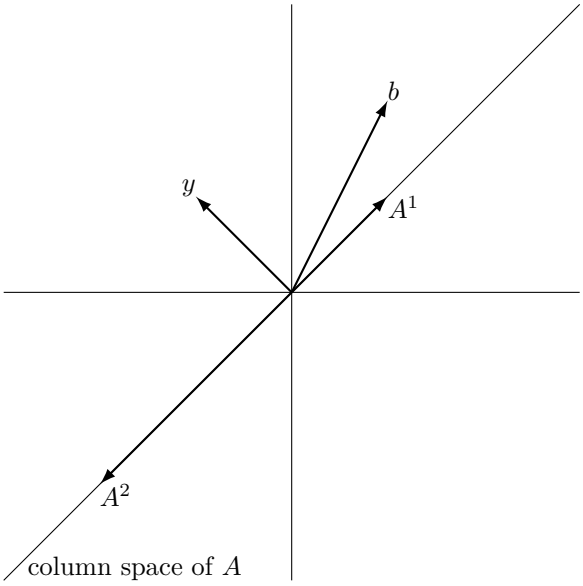

Figure 4.8. Geometry of the Fredholm Alternative

Proof using extension of linear functionals: cf. Aliprantis and Border [3, Theorem 5.91, p. 212].

Let us define the kernel by $\ker p = \{x : p \cdot x = 0\} = \{p\}_\perp$. If $p^0 = \sum_{i=1}^m \lambda_i p^i$, then clearly $\bigcap_{i=1}^m \ker p^i \subset \ker p^0$. To prove the converse, assume that $\bigcap_{i=1}^m \ker p^i \subset \ker p^0$. Define the linear operator $T\colon X \to \boldsymbol{R}^{\mathrm{m}}$ by $T(x) = \big(p^1 \cdot x, \dots, p^m \cdot x\big)$. On the range of T define the linear functional $\varphi\colon T(X) \to \boldsymbol{R}$ by $\varphi\big(p^1 \cdot x, \dots, p^m \cdot x\big) = p^0 \cdot x$. The hypothesis $\bigcap_{i=1}^m \ker p^i \subset \ker p^0$ guarantees that φ is well defined. Now note that φ extends to all of $\boldsymbol{R}^{\mathrm{m}}$, so there exist scalars $\lambda_1, \dots, \lambda_m$ such that $p^0(x) = \sum_{i=1}^m \lambda_i p^i(x)$ for each $x \in X$, as desired. ∎

To study inequalities, we start out with some preliminary results on finite cones.

Let $\{x_1, \dots, x_n\}$ be a set of vectors in $\boldsymbol{R}^{\mathrm{m}}$. Let us say that the linear combination $\sum_{i=1}^n \lambda_i x_i$ **depends** on the subset A if $A = \{x_i : \lambda_i \neq 0\}$.

The next result appears, for instance, in Gale [57, Theorem 2.11, p. 50]. It is true for general (not necessarily finite dimensional) vector spaces.

242 Lemma *A nonnegative linear combination of a set of vectors can be replaced by a nonnegative linear combination depending on an independent subset.*

That is, if $x_1, \dots, x_n$ belong to an arbitrary vector space and $y = \sum_{i=1}^n \lambda_i x_i$ where each λ_i is nonnegative, then there exist nonnegative $\beta_1, \dots, \beta_n$ such that $y = \sum_{i=1}^n \beta_i x_i$ and $\{x_i : \beta_i > 0\}$ is independent.

Proof: Recall that the empty set is independent. This covers the case where the x_is are independent and $y = 0$. We proceed by induction on the number of vectors x_i on which y depends. The case $n = 1$ is obvious.

So suppose the conclusion holds whenever y depends on no more than $n-1$ of the x_is, and suppose $\lambda_i > 0$ for each $i = 1, \dots, n$. If $x_1, \dots, x_n$ itself constitutes an independent set, there is nothing to prove, just set $\beta_i = \lambda_i$ for each i. On the other hand, if $x_1, \dots, x_n$ are dependent, then there exist numbers $\alpha_1, \dots, \alpha_n$, not all zero, such that

$$\sum_{i=1}^n \alpha_i x_i = 0.$$

We will cleverly rescale these α_is and use them to eliminate one of the x_is.

Without loss of generality we may assume that at least one $\alpha_i > 0$ (otherwise we could multiply them all by -1). This implies that $M = \max_i \frac{\alpha_i}{\lambda_i} > 0$. Renumbering if necessary, we can assume without loss of generality that $M = \frac{\alpha_n}{\lambda_n}$. Then $\lambda_i \geqslant \frac{1}{M}\alpha_i$ for all i and $\lambda_n = \frac{1}{M}\alpha_n$. Thus we can eliminate x_n from the linear combination and still keep all the remaining coefficients nonnegative:

$$\begin{aligned} y &= \sum_{i=1}^{n} \lambda_i x_i - \tfrac{1}{M} \sum_{i=1}^{n} \alpha_i x_i \\ &= \sum_{i=1}^{n} \big(\lambda_i - \tfrac{1}{M}\alpha_i\big) x_i \\ &= \sum_{i=1}^{n-1} \big(\lambda_i - \tfrac{1}{M}\alpha_i\big) x_i. \end{aligned}$$

So we have shown that y can be written as a nonnegative linear combination of no more than $n-1$ vectors, so by the induction hypothesis it depends on an independent subset. ∎

As an application of Lemma 242 we digress to prove Carathéodory's theorem on convex hulls in finite dimensional spaces.

243 Carathéodory Convexity Theorem *In $\boldsymbol{R}^{\mathrm{n}}$, every vector in the convex hull of a nonempty set can be written as a convex combination of at most $n+1$ vectors from the set.*

Proof: Let A be a nonempty subset of $\boldsymbol{R}^{\mathrm{n}}$, and let x belong to the convex hull of A. Then we can write x as a convex combination $x = \sum_{i=1}^{m} \lambda_i x_i$ of points x_i belonging to A. For any vector y in $\boldsymbol{R}^{\mathrm{n}}$ consider the "augmented" vector $\hat{y}$ in $\boldsymbol{R}^{\mathrm{n}+1}$ defined by $\hat{y}_j = y_j$ for $j = 1, \dots, n$ and $\hat{y}_{n+1} = 1$. Then it follows that $\hat{x} = \sum_{i=1}^{m} \lambda_i \hat{x}_i$ since $\sum_{i=1}^{m} \lambda_i = 1$. Renumbering if necessary, by Lemma 242, we can write $\hat{x} = \sum_{i=1}^{k} \alpha_i \hat{x}_i$, where $x_1, \dots, x_k$ are independent and $\alpha_i > 0$ for all i. Since an independent set in $\boldsymbol{R}^{\mathrm{n}+1}$ has at most $n+1$ members, $k \leqslant n+1$. But this reduces to the two equations $x = \sum_{i=1}^{k} \alpha_i x_i$ and $1 = \sum_{i=1}^{k} \alpha_i$. In other words, x is a convex combination of $k \leqslant n+1$ vectors of A. ∎

The next application of Lemma 242 is often asserted to be obvious, but is not so easy to prove. It is true in general Hausdorff topological vector spaces.

244 Lemma *Every finite cone is closed.*

Proof for the finite dimensional case: Consider $C = \{\sum_{i=1}^{k} \lambda_i x_i : \lambda_i \geqslant 0,\ i = 1, \dots, k\}$, the finite cone generated by the vectors $x_1, \dots, x_k$ in the finite dimensional space $\boldsymbol{R}^{\mathrm{m}}$. Let y be the limit of some sequence $\{y_n\}$ in C. In light of Lemma 242, we can write each y_n as a nonnegative linear combination of an independent subset of the x_is. Since there are only finitely many such subsets, by passing to a subsequence we may assume without loss of generality that each y_n depends on the same independent subset, say $x_1, \dots, x_p$. Writing $y_n = \sum_{i=1}^{p} \lambda_i^n x_i$, by Corollary 49, we have $y = \sum_{i=1}^{p} \lambda_i x_i$, where $\lambda_i = \lim_{n\to\infty} \lambda_i^n$. Since $\lambda_i^n \geqslant 0$, so is λ_i. Therefore y belongs to C. ∎

Sketch of proof for general case: Let X be a Hausdorff topological vector space. Then any finite dimensional subspace of X is closed, e.g., Aliprantis and Border [3, Corollary 5.22, p. 178], so it suffices to show that the cone generated by a finite set of vectors is a closed subset of their (finite dimensional) linear span. Furthermore, by the same arguments as in the finite dimensional case, we may assume without loss of generality that we have a cone generated by an independent set.

So let $C = \{\sum_{i=1}^p \lambda_i x_i : \lambda_i \geqslant 0 \ i = 1, \dots, p\}$ be the cone generated by the independent vectors $x_1, \dots, x_p$. Let L denote the p-dimensional subspace spanned by this basis. For each y in L there is a unique p-vector $\lambda(y)$ such that $y = \sum_{1=1}^p \lambda_i x_i$. (In the finite dimensional case, it is given by $\lambda(y) = (X'X)^{-1}X'y$.) It is straightforward to verify that the mapping $\lambda\colon L \to \boldsymbol{R}^{\mathrm{p}}$ is a linear homeomorphism and that C is the inverse image under λ of the closed set $\boldsymbol{R}^{\mathrm{p}}_+$, which shows that C is closed in L, and so in X too. ∎

The next theorem is one of many more or less equivalent results on the existence of solutions to linear inequalities.

245 Farkas' Lemma *Exactly one of the following alternatives holds. Either*

$$xA = b \tag{4.26}$$

for some $x \geqq 0$. *OR (exclusive)*

$$Ay \geqq 0 \quad b \cdot y < 0 \tag{4.27}$$

for some y.

Proof: Postmultiplying (4.26) by y, we get $xAy = b \cdot y < 0$, but premultiplying by x in (4.27), we have $xAy \geqslant 0$ whenever $x \geqq 0$, so the alternatives are inconsistent.

Let $C = \{xA : x \geqq 0\}$. If (4.26) fails, then b does not belong to C. By Lemma 244, the convex cone C is closed, so by the Strong Separating Hyperplane Theorem 160 there is some nonzero y such that $z \cdot y \geqslant 0$ for all $z \in X$ and $b \cdot y < 0$. But this is just (4.27). ∎

For a purely algebraic proof of Farkas' Lemma, see Gale [57, Theorem 2.6, p. 44]. The next result is a variant of Farkas' Lemma, see Gale [57, Theorem 2.9, p. 49].

246 Theorem *Exactly one of the following alternatives holds. Either*

$$xA \leqq b \tag{4.28}$$

for some $x \geqq 0$. *OR (exclusive)*

$$Ay \geqq 0 \quad b \cdot y < 0 \tag{4.29}$$

for some $y \geqq 0$.

Proof: It is easy to see that the alternatives are inconsistent, so suppose that (4.28) fails. This means that there is no nonnegative solution (x, z) to the equalities $xA + zI = [x, z]\begin{bmatrix} A \\ I \end{bmatrix} = b$, so by Farkas' Lemma 245 there is some y such that $\begin{bmatrix} A \\ I \end{bmatrix} y = \begin{bmatrix} Ay \\ Iy \end{bmatrix} \geqq 0$ and $b \cdot y < 0$. But this just says that (4.29) holds for some $y \geqq 0$. ∎

Each of these results on solutions of linear inequalities has a corresponding result for more general concave or convex functions and vice versa. For instance, the general Concave Alternative 169 has a linear formulation, which we shall present in just a moment. But first we mention the following obvious fact about nonnegative vectors.

$$\boxed{x \geqq 0 \iff \big(x \cdot y \geqslant 0 \text{ for every } y \geqq 0\big).}$$

We can now state the analog of Concave Alternative 169 for linear inequalities, perhaps due to von Neumann and Morgenstern [158, § 16.4, p. 140] who called it the Theorem of the Alternative for Matrices. It can also be found in Gale [57, Theorem 2.10, p. 49].

247 Theorem *Exactly one of the following alternatives holds. Either*

$$xA \leqq 0 \tag{4.30}$$

for some $x > 0$. OR (exclusive)

$$Ay \gg 0 \tag{4.31}$$

for some $y > 0$.

Proof: Define the function $f\colon C = \boldsymbol{R}^{\mathrm{m}}_{+} \to \boldsymbol{R}^{\mathrm{n}}$ by $f(y) = Ay$. By the Concave Alternative, either there exists some $\bar{y} \in C = \boldsymbol{R}^{\mathrm{m}}_{+}$ with $f(\bar{y}) = A\bar{y} \gg 0$, or (exclusive) there are nonnegative $x_1, \dots, x_n$, not all zero, that is $x > 0$, such that $\sum_{i=1}^{n} x_i f^i(y) = xAy \leqq 0$ for all $y \in C = \boldsymbol{R}^{\mathrm{m}}_{+}$. But by the obvious fact mentioned above, this just means $xA \leqq 0$. ∎

Another related result is the Stiemke Alternative. It has important applications in the theory of no-arbitrage pricing of financial assets.

248 Stiemke's Theorem *Let A be an $n \times m$ matrix. Either*

(1) the system of inequalities

$$Ax > 0$$

has a solution $x \in \boldsymbol{R}^{\mathrm{m}}$,

Or else

(2) the system of equations

$$yA = 0$$

has a strictly positive solution $y \gg 0$ in $\boldsymbol{R}^{\mathrm{n}}$

(but not both).

Proof: Clearly both (1) and (2) cannot be true, for then we must have both $yAx = 0$ (as $yA = 0$) and $yAx > 0$ (as $y \gg 0$ and $Ax > 0$). So it suffices to show that if (1) fails, then (2) must hold.

In geometric terms, the negation of (1) asserts that the span M of the columns $\{A^1, \dots, A^n\}$ intersects the nonnegative orthant of $\boldsymbol{R}^{\mathrm{n}}$ only at the origin. Thus the unit simplex Δ in $\boldsymbol{R}^{\mathrm{n}}$ is disjoint from M if and only if (1) fails, where $\Delta = \{x \in \boldsymbol{R}^{\mathrm{n}} : x \geqq 0 \text{ and } \sum_{j=1}^{n} x_j = 1\}$.[9]

So assume that condtion (1) fails. Then since Δ is compact and convex and M is closed and convex, by Theorem 160, there is a hyperplane strongly separating Δ and M. That is, there is some nonzero $y \in \boldsymbol{R}^{\mathrm{n}}$ and some $\varepsilon > 0$ satisfying

$$y \cdot x + \varepsilon < y \cdot z \quad \text{for all } x \in M,\ z \in \Delta.$$

Since M is a linear subspace, we must have $y \cdot x = 0$ for all $x \in M$. Consequently $y \cdot z > \varepsilon > 0$ for all $z \in \Delta$. Since the j^{th} unit coordinate vector e^j belongs to Δ, we see that $y_j = y \cdot e^j > 0$. That is, $y \gg 0$.

Since each $A^i \in M$, we have that $y \cdot A^i = 0$, i.e.,

$$yA = 0.$$

This completes the proof. ∎

The next alternative is a variation on Stiemke's.

[9] Note that since M is a linear subspace, if M intersects the nonnegative orthant at a nonzero point x, then $\frac{1}{\sum_i x_i} x$ belongs to $M \cap \Delta$.

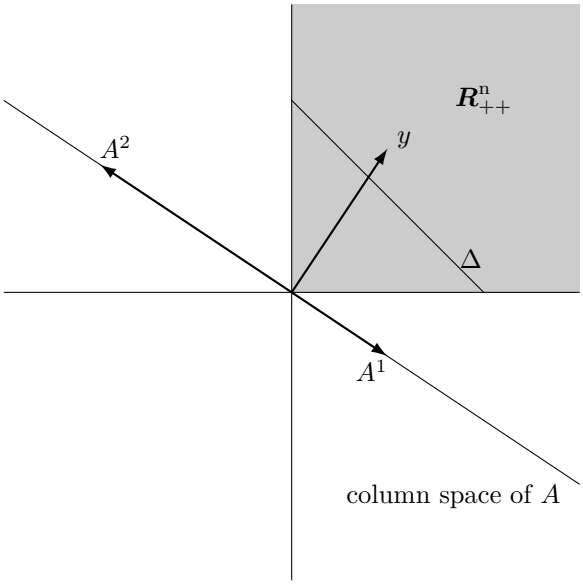

Figure 4.9. Geometry of the Stiemke Alternative

249 Theorem *Let A be an $n \times m$ matrix. Either*

(1) the system of inequalities

$$Ax \gg 0$$

has a solution $x \in \boldsymbol{R}^{\mathrm{m}}$,

Or else

(2) the system of equations

$$yA = 0, \quad y \cdot \mathbf{1} = 1$$

has a semi-positive solution $y > 0$ in $\boldsymbol{R}^{\mathrm{n}}$

(but not both).

Proof: Clearly both (1) and (2) cannot be true, for then we must have both $yAx = 0$ (as $yA = 0$) and $yAx > 0$ (as $y > 0$ and $Ax \gg 0$). So it suffices to show that if (1) fails, then (2) must hold.

Assume that condtion (1) fails. Then the span M of the columns $\{A^1, \dots, A^n\}$ is disjoint from the strictly positive orthant $\boldsymbol{R}^{\mathrm{n}}_{++}$. By Theorem 165, there is a nonzero y separating $\boldsymbol{R}^{\mathrm{n}}_{++}$, and M. As in the above arguments we must have $yA = 0$, and $y > 0$, so it can be normalized to satisfy $y \cdot \mathbf{1} = 1$. ∎

Finally we come to another alternative, Motzkin's Transposition Theorem [113], proven in his 1934 Ph.D. thesis. This statement is take from his 1951 paper [114].[10]

250 Motzkin's Transposition Theorem *Let A be an $m \times n$ matrix, let B be an $\ell \times n$ matrix, and let C be an $r \times n$ matrix, where B or C may be omitted (but not A). Exactly one of the following alternatives holds. Either there exists $x \in \boldsymbol{R}^{\mathrm{n}}$ satisfying*

$$\begin{aligned} Ax &\gg 0 \\ Bx &\geqq 0 \\ Cx &= 0 \end{aligned} \tag{4.32}$$

[10] Motzkin [114] contains an unfortunate typo. The condition $Ax \gg 0$ is erroneously given as $Ax \ll 0$.

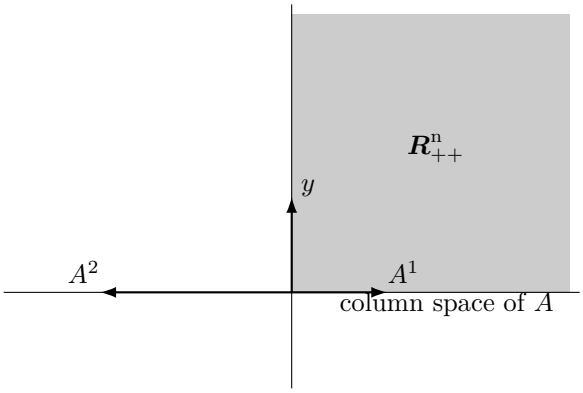

Figure 4.10. Geometry of Theorem 249

or else there exist $p^1 \in \boldsymbol{R}^{\mathrm{m}}$, $p^2 \in \boldsymbol{R}^{\ell}$, *and* $p^3 \in \boldsymbol{R}^{\mathrm{r}}$ *satisfying*

$$\begin{aligned} p^1A + p^2B + p^3C &= 0 \\ p^1 &> 0 \\ p^2 &\geqq 0. \end{aligned} \tag{4.33}$$

Motzkin expressed (4.33) in terms of the transpositions of A, B, and C.

251 Exercise Prove the Transposition Theorem. Hint: If x satisfies (4.32), it can be scaled to satisfy

$$\begin{bmatrix} -A \\ -B \\ C \\ -C \end{bmatrix} x \leqq \begin{bmatrix} -\mathbf{1} \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Apply the variant Farkas' Lemma 246. □

4.29 Constrained maxima and Lagrangean saddlepoints

In this section we discuss the relation between constrained maxima of concave functions and saddlepoints of the so-called Lagrangean.

252 Definition *Let $\varphi\colon X \times Y \to \boldsymbol{R}$. A point (x^*, y^*) in $X \times Y$ is a* ***saddlepoint of $\boldsymbol{\varphi}$ (over $\boldsymbol{X \times Y}$)*** *if it satisfies*

$$\varphi(x, y^*) \leqslant \varphi(x^*, y^*) \leqslant \varphi(x^*, y) \quad \text{for all} \quad x \in X,\ y \in Y.$$

That is, (x^*, y^*) is a saddlepoint of φ if x^* maximizes $\varphi(\cdot, y^*)$ over X and y^* minimizes $\varphi(x^*, \cdot)$ over Y. Saddlepoints of a function have the following nice interchangeability property.

253 Lemma (Interchangeability of saddlepoints) *Let $\varphi\colon X \times Y \to \boldsymbol{R}$, and let (x_1, y_1) and (x_2, y_2) be saddlepoints of φ. Then*

$$\varphi(x_1, y_1) = \varphi(x_2, y_1) = \varphi(x_1, y_2) = \varphi(x_2, y_2).$$

Consequently (x_1, y_2) and (x_2, y_1) are also saddlepoints.

Proof: We are given that

$$\varphi(x, y_1) \underset{(4.34a)}{\leqslant} \varphi(x_1, y_1) \underset{(4.34b)}{\leqslant} \varphi(x_1, y) \qquad x \in X,\ y \in Y, \tag{4.34}$$

and

$$\varphi(x, y_2) \underset{(4.35a)}{\leqslant} \varphi(x_2, y_2) \underset{(4.35b)}{\leqslant} \varphi(x_2, y) \qquad x \in X,\ y \in Y. \tag{4.35}$$

Evaluating (4.34a) at $x = x_2$ yields

$$\varphi(x_2, y_1) \leqslant \varphi(x_1, y_1) \tag{4.36}$$

evaluating (4.34b) at $y = y_2$ yields

$$\varphi(x_1, y_1) \leqslant \varphi(x_1, y_2) \tag{4.37}$$

evaluating (4.35a) at $x = x_1$ yields

$$\varphi(x_1, y_2) \leqslant \varphi(x_2, y_2) \tag{4.38}$$

and evaluating (4.35b) at $y = y_1$ yields

$$\varphi(x_2, y_2) \leqslant \varphi(x_2, y_1). \tag{4.39}$$

Combining these yields

$$\varphi(x_2, y_1) \underset{(4.36)}{\leqslant} \varphi(x_1, y_1) \underset{(4.37)}{\leqslant} \varphi(x_1, y_2) \underset{(4.38)}{\leqslant} \varphi(x_2, y_2) \underset{(4.39)}{\leqslant} \varphi(x_2, y_1)$$

which implies that

$$\varphi(x_2, y_1) = \varphi(x_1, y_1) = \varphi(x_1, y_2) = \varphi(x_2, y_2). \tag{4.40}$$

To see that (x_2, y_1) is a saddlepoint, observe

$$\varphi(x, y_1) \underset{(4.34a)}{\leqslant} \varphi(x_1, y_1) \underset{(4.40)}{=} \varphi(x_2, y_1) \underset{(4.40)}{=} \varphi(x_2, y_2) \underset{(4.35b)}{\leqslant} \varphi(x_2, y) \qquad x \in X,\ y \in Y.$$

Similarly, (x_1, y_2) is also a saddlepoint. ∎

Saddlepoints play an important rôle in the sis of constrained maximum problems via Lagrangean functions.

254 Definition *Given $f, g_1, \dots, g_m \colon X \to \boldsymbol{R}$, the associated* ***Lagrangean*** *$L \colon X \times \Lambda \to \boldsymbol{R}$ is defined by*

$$L(x, \lambda) = f(x) + \sum_{j=1}^{m} \lambda_j g_j(x) = f(x) + \lambda \cdot g(x),$$

where Λ is an appropriate subset of $\boldsymbol{R}^{\mathrm{m}}$. (Usually $\Lambda = \boldsymbol{R}^{\mathrm{m}}$ or $\boldsymbol{R}^{\mathrm{m}}_{+}$.) The components of λ are called ***Lagrange multipliers****.*

The first result is that saddlepoints of Lagrangeans are constrained maxima. This result makes no restrictive assumptions on the domain or the functions.

255 Theorem (Lagrangean saddlepoints are constrained maxima) *Let X be an arbitrary set, and let $f, g_1, \dots, g_m \colon X \to \boldsymbol{R}$. Suppose that (x^*, λ^*) is a saddlepoint of the Lagrangean $L(x, \lambda) = f + \lambda \cdot g$ (over $X \times \boldsymbol{R}^{\mathrm{m}}_{+}$). That is,*

$$L(x, \lambda^*) \underset{(4.41a)}{\leqslant} L(x^*, \lambda^*) \underset{(4.41b)}{\leqslant} L(x^*, \lambda) \qquad x \in X,\ \lambda \geqq 0. \tag{4.41}$$

Then x^* maximizes f over X subject to the constraints $g_j(x) \geqslant 0$, $j = 1, \dots, m$, and furthermore

$$\lambda^*_j g_j(x^*) = 0 \quad j = 1, \dots, m. \tag{4.42}$$

Proof: Inequality (4.41*b*) implies $\lambda^* \cdot g(x^*) \leqslant \lambda \cdot g(x^*)$ for all $\lambda \geqq 0$. Therefore $g(x^*) \geqq 0$ (why?), so x^* satisfies the constraints. Setting $\lambda = 0$, we see that $\lambda^* \cdot g(x^*) \leqslant 0$. This combined with $\lambda \geqq 0$ and $g(x^*) \geqq 0$ implies $\lambda^* \cdot g(x^*) = 0$. Indeed it implies $\lambda^*_j g_j(x^*) = 0$ for $j = 1, \dots, m$.

Now note that (4.41*a*) implies $f(x) + \lambda^* \cdot g(x) \leqslant f(x^*)$ for all x. Therefore, if x satisfies the constraints, $g(x) \geqq 0$, we have $f(x) \leqslant f(x^*)$, so x^* is a constrained maximizer. ∎

Condition (4.42) implies that if the multiplier λ^*_j is strictly positive, then the corresponding constraint is **binding**, $g_j(x^*) = 0$; and if a constraint is **slack**, $g_j(x^*) > 0$, then the corresponding multiplier satisfies $\lambda^*_j = 0$. These conditions are sometimes called the **complementary slackness** conditions.

The converse of Theorem 255 is not quite true, but almost. To state the correct result we now introduce the notion of a generalized Lagrangean.

256 Definition *A* ***generalized Lagrangean*** *$L_\mu \colon X \times \Lambda \to \boldsymbol{R}$, where $\mu \geqslant 0$, is defined by*

$$L_\mu(x, \lambda) = \mu f(x) + \sum_{j=1}^{m} \lambda_j g_j(x),$$

where Λ is an appropriate subset of $\boldsymbol{R}^{\mathrm{m}}$.

Note that each choice of μ generates a different generalized Lagrangean. However for $\Lambda = \boldsymbol{R}^{\mathrm{m}}_{+}$, as long as $\mu > 0$, a point (x, λ) is a saddlepoint of the Lagrangean if and only if it is a saddlepoint of the generalized Lagrangean. Thus the only case to worry about is $\mu = 0$.

The next results state that for concave functions satisfying a regularity condition, constrained maxima are saddlepoints of some generalized Lagrangean.

257 Theorem (Concave constrained maxima are nearly Lagrangean saddlepoints) *Let $C \subset \boldsymbol{R}^{\mathrm{n}}$ be convex, and let $f, g_1, \dots, g_m \colon C \to \boldsymbol{R}$ be concave. Suppose x^* maximizes f subject to the constraints $g_j(x) \geqslant 0$, $j = 1, \dots, m$. Then there exist real numbers $\mu^*, \lambda_1^*, \dots, \lambda_m^* \geqslant 0$, not all zero, such that (x^*, λ^*) is a saddlepoint of the generalized Lagrangean L_{μ^*}. That is,*

$$\mu^* f(x) + \sum_{j=1}^m \lambda_j^* g_j(x) \underset{(4.43a)}{\leqslant} \mu^* f(x^*) + \sum_{j=1}^m \lambda_j^* g_j(x^*) \underset{(4.43b)}{\leqslant} \mu^* f(x^*) + \sum_{j=1}^m \lambda_j g_j(x^*) \tag{4.43}$$

for all $x \in C$ and all $\lambda_1, \dots, \lambda_m \geqslant 0$. Furthermore

$$\sum_{j=1}^m \lambda_j^* g_j(x^*) = 0. \tag{4.44}$$

Proof: Since x^* is a constrained maximizer there is no $x \in C$ satisfying $f(x) - f(x^*) > 0$ and $g(x) \geqq 0$. Therefore the Concave Alternative 169 implies the existence of nonnegative $\mu^*, \lambda_1^*, \dots, \lambda_m^*$, not all zero, satisfying

$$\mu^* f(x) + \sum_{j=1}^m \lambda_j^* g_j(x) \leqslant \mu^* f(x^*) \quad \text{for every } x \in C.$$

Evaluating this at $x = x^*$ yields $\sum_{j=1}^m \lambda_j^* g_j(x^*) \leqslant 0$. But each term in this sum is the product of two nonnegative terms, so (4.44) holds. This in turn implies (4.43a). Given that $g_j(x^*) \geqslant 0$ for all j, (4.44) also implies (4.43b). ∎

So as not lose sight of the forest for the trees, here is an immediate but importance consequence of the Saddlepoint Theorem. It essentially asserts that the Lagrange multipliers on the constraints are conversion factors between the values of the constraint functions and the objective function so that a constrained maximizer of the objective is an unconstrained maximizer of the Lagrangean. In Section 5.10 we will see another connection between Lagrange multipliers and conversion factors.

258 Corollary (Constrained maximizers maximize the generalized Lagrangean) *Let $C \subset \boldsymbol{R}^{\mathrm{n}}$ be convex, and let $f, g_1, \dots, g_m \colon C \to \boldsymbol{R}$ be concave. Suppose x^* maximizes f subject to the constraints $g_j(x) \geqslant 0$, $j = 1, \dots, m$. Then there exist real ns umbers $\mu^*, \lambda_1^*, \dots, \lambda_m^* \geqslant 0$, not all zero, such that x^* maximizes the generalized Lagrangean L_{μ^*}. That is,*

$$\mu^* f(x) + \sum_{j=1}^m \lambda_j^* g_j(x) \leqslant \mu^* f(x^*) + \sum_{j=1}^m \lambda_j^* g_j(x^*).$$

for all $x \in C$.

259 Corollary (When constrained maxima are true Lagrangean saddlepoints) *Under the hypotheses of Theorem 257 suppose in addition that **Slater's Condition**,*

$$\exists \bar{x} \in C \quad g(\bar{x}) \gg 0, \tag{$\boldsymbol{S}$}$$

is satisfied. Then $\mu^* > 0$, and may be taken equal to 1. Consequently $x^*, \lambda_1^*, \dots, \lambda_m^*$ is a saddlepoint of the Lagrangean for $x \in C$, $\lambda \geqq 0$. That is,

$$L(x, \lambda^*) \leqslant L(x^*, \lambda^*) \leqslant L(x^*, \lambda) \quad x \in C, \quad \lambda \geqq 0, \tag{4.45}$$

where $L(x, \lambda) = f(x) + \lambda \cdot g(x)$.

Proof: Suppose $\mu^* = 0$. Then evaluating (4.43) at $x = \bar{x}$ yields $\lambda^* \cdot g(\bar{x}) \leqslant 0$, but $g(\bar{x}) > 0$ implies $\lambda_j^* = 0$, $j = 1, \dots, m$. Thus $\mu = 0$ and $\lambda_j = 0$, $j = 1, \dots, m$, a contradiction. Therefore $\mu > 0$, and by dividing the Lagrangean by μ, we may take $\mu = 1$. The remainder is then just Theorem 257. ∎

Again, we state the obvious.

260 Corollary (Constrained maximizers maximize the Lagrangean) *Under the hypotheses of Corollary 259, The point x^* maximizes the Lagrangean, that is,*

$$L^*(x) \leqslant L^*(x^*), \quad x \in C.$$

where $L^*(x) = f(x) + \lambda^* \cdot g(x)$.

The above result may fail when the functions are not concave, even though they may be otherwise well behaved. See Example 269 below.

Karlin [89, vol. 1, Theorem 7.1.1, p. 201] proposed the following alternative to Slater's Condition:

$$\forall \lambda > 0 \ \ \exists \bar{x}(\lambda) \in C \qquad \lambda \cdot g\big(\bar{x}(\lambda)\big) > 0,$$

which we may as well call **Karlin's condition**.

261 Theorem *Let $C \subset \boldsymbol{R}^{\mathrm{n}}$ be convex, and let $g_1, \dots, g_m \colon C \to \boldsymbol{R}$ be concave. Then g satisfies Slater's Condition if and only it satisfies Karlin's Condition.*

Proof: Clearly Slater's Condition implies Karlin's. Now suppose g violates Slater's Condition. Then by the Concave Alternative Theorem 169, it must also violate Karlin's. ∎

The next example shows what can go wrong when Slater's Condition fails.

262 Example In this example, due to Slater [144], $C = \boldsymbol{R}$, $f(x) = x$, and $g(x) = -(1-x)^2$. Note that Slater's Condition fails because $g \leqslant 0$. The constraint set $[g \geqslant 0]$ is the singleton $\{1\}$. Therefore f attains a constrained maximum at $x^* = 1$. There is however no saddlepoint over $\boldsymbol{R} \times \boldsymbol{R}_+$ at all of the Lagrangean

$$L(x, \lambda) = x - \lambda(1-x)^2 = -\lambda + (1 + 2\lambda)x - \lambda x^2.$$

To see that L has no saddlepoint, consider first the case $\lambda = 0$. Then $L(x, 0) = x$, so there is no maximizer with respect to x. On the other hand if $\lambda > 0$, the first order condition for a maximum in x is $\frac{\partial L}{\partial x} = 0$, or $1 + 2\lambda - 2\lambda x = 0$, which implies $x = 1 + (1/2\lambda) > 1$. But for $x > 1$, $\frac{\partial L}{\partial \lambda} = -(1-x)^2 < 0$, so no minimum with respect to λ exists. □

4.29.1 The rôle of Slater's Condition

In this section we present a geometric argument that illuminates the rôle of Slater's Condition in the saddlepoint theorem. The saddlepoint theorem was proved by invoking the Concave Alternative Theorem 169, so let us return to the underlying argument used in its proof. In the framework of Theorem 257, define the function $h \colon C \to \boldsymbol{R}^{m+1}$ by

$$h(x) = \big(g_1(x), \dots, g_m(x), f(x) - f(x^*)\big)$$

and set

$$H = \{h(x) : x \in C\} \quad \text{and} \quad \hat{H} = \{y \in \boldsymbol{R}^{m+1} : \exists x \in C \quad y \leqq h(x)\}.$$

Then $\hat{H}$ is a convex set bounded in part by H. Figure 4.11 depicts the sets H and $\hat{H}$ for Slater's example 262, where $f(x) - f(x^*)$ is plotted on the vertical axis and $g(x)$ is plotted on

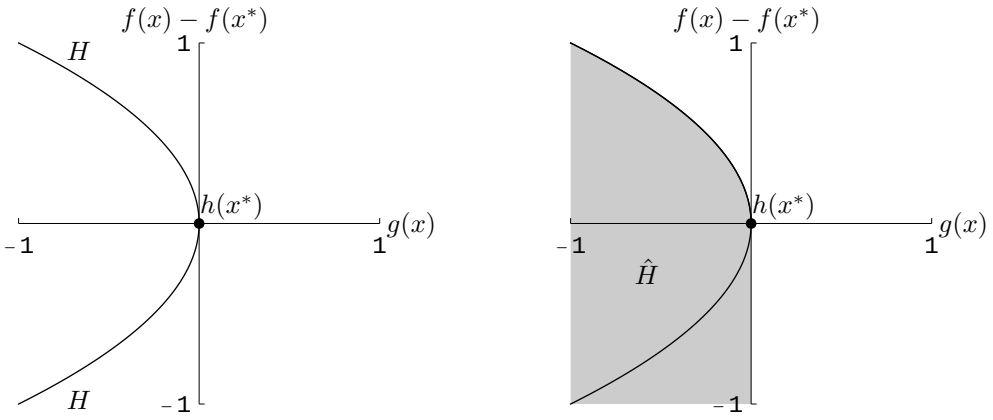

Figure 4.11. The sets H and $\hat{H}$ for Slater's example.

the horizontal axis. Now if x^* maximizes f over the convex set C subject to the constraints $g_j(x) \geqslant 0$, $j = 1, \dots, m$, then $h(x^*)$ has the largest vertical coordinate among all the points in H whose horizontal coordinates are nonnegative.

The semipositive $m+1$-vector $\hat{\lambda}^* = (\lambda_1^*, \dots, \lambda_m^*, \mu^*)$ from Theorem 257 is obtained by separating the convex set $\hat{H}$ and $\boldsymbol{R}^{m+1}_{++}$. It has the property that

$$\hat{\lambda}^* \cdot h(x) \leqslant \hat{\lambda}^* h(x^*)$$

for all $x \in C$. That is, the vector $\hat{\lambda}^*$ defines a hyperplane through $h(x^*)$ such that the entire set $\hat{H}$ lies in one half-space. It is clear in the case of Slater's example that the hyperplane is a vertical line, since it must be tangent to H at $h(x^*) = (0, 0)$. The fact that the hyperplane is vertical means that μ^* (the multiplier on f) must be zero.

If there is a non-vertical hyperplane through $h(x^*)$, then μ^* is nonzero, so we can divide by it and obtain a full saddlepoint of the true Lagrangean. This is where Slater's condition comes in.

In the one dimensional, one constraint case, Slater's Condition reduces to the existence of $\bar{x}$ satisfying $g(\bar{x}) > 0$. This rules out having a vertical supporting line through x^*. To see this, note that the vertical component of $h(x^*)$ is $f(x^*) - f(x^*) = 0$. If $g(x^*) = 0$, then the vertical line through $h(x^*)$ is simply the vertical axis, which cannot be, since $h(\bar{x})$ lies to the right of the axis. If $g(x^*) > 0$, then $\hat{H}$ includes every point below $h(x^*)$, so the only line separating $\hat{H}$ and $\boldsymbol{R}^2_{++}$ is horizontal, not vertical. See Figure 4.12.

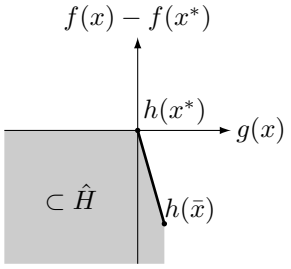

Figure 4.12. Slater's condition guarantees a non-vertical supporting line.

In Figure 4.12, the shaded area is included in $\hat{H}$. For instance, let $C = (-\infty, 0]$, $f(x) = x$, and $g(x) = x + 1$. Then the set $\hat{H}$ is just $\{y \in \boldsymbol{R}^2 : y \leqq (0, 1)\}$.

In the next section we shall see that if f and the g_js are linear, then Slater's Condition is not needed to guarantee a non-vertical supporting line. Intuitively, the reason for this is that for the linear programming problems considered, the set $\hat{H}$ is polyhedral, so even if $g(x^*) = 0$, there is still a non-vertical line separating $\hat{H}$ and $\boldsymbol{R}^{\mathrm{m}}_{++}$. The proof of this fact relies on our earlier results on linear inequalities. It is subtle because Slater's condition rules out a vertical supporting line. In the linear case, there may be a vertical supporting line, but if there is, there is also a non-vertical supporting line that yields a Lagrangean saddlepoint. As a case in point, consider $C = (-\infty, 0]$, $f(x) = x$, and $g(x) = x$. Then the set $\hat{H}$ is just $\{y \in \boldsymbol{R}^2 : y \leqq 0\}$, which is separated from $\boldsymbol{R}^2_{++}$ by every semipositive vector.

4.30 The saddlepoint theorem for linear programming

The material for this handout is based largely on the beautifully written book by David Gale [57].

A **maximum linear program in inequality form**[11] is a constrained maximization problem of the form

$$\max_x \ p \cdot x$$

subject to

$$\begin{aligned} xA &\leqq q && (4.46)\\ x &\geqq 0 && (4.47) \end{aligned}$$

where x and p belong to $\boldsymbol{R}^{\mathrm{n}}$, q belongs to $\boldsymbol{R}^{\mathrm{m}}$, and A is $n \times m$. The program is **feasible** if there is some x satisfying the constraints (4.46) and (4.47). Every maximum linear program in inequality form has a **dual program**, which is the minimization problem:

$$\min_y \ q \cdot y \qquad (4.48)$$

subject to

$$\begin{aligned} Ay &\geqq p && (4.49)\\ y &\geqq 0. && (4.50) \end{aligned}$$

The original maximum linear program may be called the **primal** program to distinguish it from the dual.

Let us start by examining the Lagrangean for the primal program. Write (4.46) as $q_j - (xA)_j \geqslant 0$, $j = 1, \dots, m$, and let y_j denote the Lagrange multiplier for this constraint. Incorporate (4.47) by setting the domain $X = \boldsymbol{R}^{\mathrm{n}}_+$. The Lagrangean is then

$$L(x, y) = p \cdot x + q \cdot y - xAy. \qquad (4.51)$$

Treating the dual as the problem of maximizing $-q \cdot y$ subject to $Ay - p \geqq 0$, and using x to denote the vector of Lagrange multipliers, the Lagrangean for the dual is:

$$-q \cdot y + xAy - x \cdot p,$$

which is just the negative of (4.51). Consequently, by the Saddlepoint Theorem, if $(\bar{x}, \bar{y})$ is a saddlepoint of $L(x, y) = p \cdot x + q \cdot y - xAy$ over $\boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}^{\mathrm{m}}$, then $\bar{x}$ is optimal for the primal program

[11] Gale [57] refers to this as the "standard form" of a linear program. However Dantzig [39] uses the term standard form in a different fashion. The expression "inequality form" has the virtue of being descriptive.

and $\bar{y}$ is optimal (minimal) for the dual program. In particular, if there is a saddlepoint, then both programs are feasible. If we knew that both programs satisfied Slater's Condition, then the Saddlepoint Theorem would assert that any pair of optimal solutions would be a saddlepoint of the Lagrangean. *Remarkably, for the linear programming case, we do not need Slater's Condition.*

263 Saddlepoint Theorem for Linear Programming *The primal and dual are simultaneously feasible if and only if the function*

$$L(x, y) = p \cdot x + q \cdot y - xAy$$

has a saddlepoint in $\boldsymbol{R}^{\mathrm{n}}_{+} \times \boldsymbol{R}^{\mathrm{m}}_{+}$. *A pair* $(\bar{x}, \bar{y})$ *is a saddlepoint if and only if* $\bar{x}$ *is optimal for the primal and* $\bar{y}$ *is optimal for the dual.*

A simple consequence worth noting is that the primal has an optimum if and only if the dual does. The proof is broken down into a series of lemmas.

264 Lemma *If* x *is feasible for the primal program and* y *is feasible for its dual, then* $p \cdot x \leqslant q \cdot y$.

Proof: Suppose x satisfies (4.46) and $y \geqq 0$. Then $xAy \leqslant q \cdot y$. Likewise if y satisfies (4.49) and $x \geqq 0$, then $xAy \geqslant x \cdot p$. Combining these pieces proves the lemma. ∎

This allows us to immediately conclude the following.

265 Corollary (Optimality Criterion for LP) *If* x *is feasible for the primal program and* y *is feasible for the dual, and if* $p \cdot x = q \cdot y = xAy$, *then* x *is optimal and* y *is optimal for the dual program.*

A consequence of this is the following result that Gale refers to as the **Equilibrium Theorem**. It is also known as the **Complementary Slackness Theorem**.

266 Complementary Slackness Theorem *Suppose* x *and* y *are feasible for the primal and dual respectively. They are optimal if and only if both*

$$(xA)_j < q_j \implies y_j = 0 \tag{4.52}$$

and

$$(Ay)_i > p_i \implies x_i = 0. \tag{4.53}$$

Proof: Suppose x and y are feasible for the primal and dual respectively. From $xA \leqq q$ and $y \geqq 0$, we have $xAy \leqslant q \cdot y$ with equality if and only if (4.52) holds. Similarly, (4.53) holds if and only if $xAy = p \cdot x$. The conclusion now follows from Corollary 265). ∎

The gap remaining is to show that if $\bar{x}$ is optimal, then the dual has an optimal solution $\bar{y}$ and that $p \cdot \bar{x} = q \cdot \bar{y}$ (instead of $p \cdot \bar{x} < q \cdot \bar{y}$). This brings us to the following.

267 Fundamental Duality Theorem of LP *If both a maximum linear program in inequality form and its dual are feasible, then both have optimal solutions, and the values of the two programs are the same. If one of the programs is infeasible, neither has a solution.*

Proof: (Gale [57]) Start by assuming both programs are feasible. We already know that if x and y are feasible for the primal and dual respectively, then $p \cdot x \leqslant q \cdot y$. Thus it suffices to find a solution $(x, y) \geqq 0$ to the inequalities

$$\begin{aligned} xA &\leqq q \\ -yA' &\leqq -p \\ y \cdot q - x \cdot p &\leqslant 0, \end{aligned}$$

or, in matrix form

$$[x, y] \begin{bmatrix} A & 0 & -p \\ 0 & -A' & q \end{bmatrix} \leqq [q, -p, 0]. \tag{4.54}$$

Either these inequalities have a solution, or by a theorem of the alternative, there is a nonnegative vector $\begin{bmatrix} u \\ v \\ \alpha \end{bmatrix} \geqq 0$, where $u \in \boldsymbol{R}^{\mathrm{m}}_+$, $v \in \boldsymbol{R}^{\mathrm{n}}_+$, and $\alpha \in \boldsymbol{R}_+$, satisfying

$$\begin{bmatrix} A & 0 & -p \\ 0 & -A & q \end{bmatrix} \begin{bmatrix} u \\ v \\ \alpha \end{bmatrix} \geqq \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{4.55}$$

and

$$[q, -p, 0] \begin{bmatrix} u \\ v \\ \alpha \end{bmatrix} < 0. \tag{4.56}$$

We shall show that this latter set of inequalities does not have a solution: Suppose by way of contradiction that (4.55) and (4.56) have a nonnegative solution. Rewriting (4.55), we have

$$Au \geqq \alpha p \tag{4.57}$$

and

$$vA \leqq \alpha q, \tag{4.58}$$

while (4.56) becomes

$$q \cdot u < p \cdot v. \tag{4.59}$$

Let $\bar{x} \geqq 0$ be feasible for the primal, that is, $\bar{x}A \leqq q$. Then

$$\bar{x}Au \leqslant q \cdot u \tag{4.60}$$

since $u \geqq 0$. Similarly let $\bar{y} \geqq 0$ be feasible for the dual, that is, $A\bar{y} \geqq p$. Then

$$vA\bar{y} \geqslant v \cdot p \tag{4.61}$$

since $v \geqq 0$.

We next show that $\alpha \neq 0$. For suppose $\alpha = 0$. Then (4.57) becomes $Au \geqq 0$, which implies

$$\bar{x}Au \geqslant 0,$$

since $\bar{x} \geqq 0$. Also (4.58) implies

$$vA\bar{y} \leqslant 0,$$

since $\bar{y} \geqq 0$. Combining this with (4.60) and (4.61) yields

$$q \cdot u \geqslant \bar{x}Au \geqslant 0 \geqslant vA\bar{y} \geqslant v \cdot p,$$

which contradicts (4.59).

This shows that $\alpha > 0$, so we may without loss of generality assume $\alpha = 1$. In this case, (4.57) becomes $Au \geqq p$ and (4.58) becomes $vA \leqq q$, which imply that v is feasible for the primal program and u is feasible for the dual. Therefore, by Lemma 264, $q \cdot u \geqslant p \cdot v$, which again contradicts (4.59). This contradiction shows that if both programs are feasible, then both have optimal solutions and both programs have the same value.

If either program is infeasible, then certainly it cannot have an optimal solution. So suppose that the primal program is infeasible, but the dual is feasible. That is, $xA \leqq q$ has no nonnegative solution, so by the theorem of the alternative again, there is a nonnegative y satisfying $Ay \geqq 0$ and $q \cdot y < 0$. Let z be any feasible nonnegative solution to the dual. Then $z + \alpha y$ is feasible for any $\alpha \geqslant 0$, and $q \cdot (z + \alpha y) = q \cdot z + \alpha q \cdot y \to -\infty$ as $\alpha \to \infty$. Therefore no optimal solution exists for the dual.

A similar argument works if the dual is infeasible, but the primal is feasible. ∎

You may suspect that it is possible to combine linear constraints with more general concave constraints that satisfy Slater's Condition. This is indeed the case as Uzawa [151] has shown. (See also Moore [111].)

4.30.1 Other formulations

Remember that the dual of a maximum linear program in inequality form is a linear program of the form

$$\begin{aligned} &\underset{y}{\text{minimize}} && q \cdot y \\ &\text{subject to} \\ &&Ay &\geqq p \\ &&y &\geqq 0 \end{aligned}$$

where x and p belong to $\boldsymbol{R}^{\mathrm{n}}$, q belongs to $\boldsymbol{R}^{\mathrm{m}}$, and A is $n \times m$. Let us call this a **minimum linear program in inequality form**. Now the dual program itself can be rewritten as the following maximum LP in inequality form:

$$\underset{y}{\text{maximize}} \quad -q \cdot y$$

subject to

$$\begin{aligned} y(-A') &\leqq -p \\ y &\geqq 0, \end{aligned}$$

where A' is the transpose of A. The dual of this program is:

$$\underset{x}{\text{minimize}} \quad -p \cdot x$$

subject to

$$\begin{aligned} -A'x &\geqq -p \\ x &\geqq 0, \end{aligned}$$

or

$$\underset{x}{\text{maximize}} \quad p \cdot x$$

subject to

$$\begin{aligned} xA &\leqq p \\ x &\geqq 0, \end{aligned}$$

which is our primal. Thus the dual of a minimum LP in inequality form is a maximum LP in inequality form, and vice-versa. Moreover the dual of the dual is the primal.

Not every linear program comes to us already in inequality form, nor is the inequality form always the easiest to work with. There are other forms, some of which have names, and all of which can be translated into one another. In fact, we just translated a minimum inequality form into a maximum inequality form above. Each of these forms also has a dual, and the program and its dual satisfy the Fundamental Duality Theorem of LP 267. That is, if both a linear program (in any form) and its dual are feasible, then both have optimal solutions, and

the values of the two programs are the same. If one of the programs is infeasible, neither has a solution.

Let us start with a linear program in **general maximum form**, which allows for linear inequalities and equations, and optional nonnegativity constraints.

$$\max_x \ p \cdot x = \sum_{i=1}^n p_i x_i$$

subject to

$$\begin{aligned} (xA)_j &\leqslant q_j \quad j \notin E \\ (xA)_j &= q_j \quad j \in E \\ x_i &\geqslant 0 \quad i \in N \end{aligned}$$

where N, the set of nonnegativity constraints on components of x is a subset of $\{1, \dots, n\}$, and E is a subset of $\{1, \dots, m\}$. Note that $(xA)_j = x \cdot A^j$ (where A^j is the j^{th} column of A) so by replacing A^j by $-A^j$ and q_j by $-q_j$, we can convert $\geqslant$ constraints to $\leqslant$ constraints, so this form is reasonably general.

We can translate this into inequality maximum form as follows. First we add a vector $z \in \boldsymbol{R}^{\text{n}}$ of **slack variables** and require $x \geqq 0$ and $z \geqq 0$. We replace x in the inequalities by $x - z$, which has components unrestricted in sign. To capture the requirement that $x_i - z_i \geqslant 0$ for $i \in N$, we add the inequality $z \cdot e^i \leqslant 0$, where e^i is the i^{th} unit coordinate vector in $\boldsymbol{R}^{\text{n}}$. (Do you see why this works?) Now by replacing each equality with a pair of inequalities,

$$(xA)_j = q_j \quad \iff \quad x \cdot A^j \leqslant q_j \text{ and } x \cdot (-A^j) \leqslant -q_j,$$

we have the following inequality maximum problem

$$\max_x \ (p, -p) \cdot (x, z) = p \cdot (x - z)$$

subject to

$$\begin{aligned} [x, z] \begin{bmatrix} A & -A_E & 0 \\ -A & A_E & D \end{bmatrix} &\leqq [q, d, 0] \\ [x, z] &\geqq 0 \end{aligned}$$

where

$$\begin{aligned} A_E &\text{ is the } n \times |E| \text{ matrix whose columns are } A^j,\ j \in E, \\ D &\text{ is the } n \times |N| \text{ matrix whose columns are } e^i,\ i \in N, \end{aligned}$$

and d is the $|E|$-vector whose components are $-q_j$, $j \in E$.

The dual of this is the inequality minimum problem

$$\min_{\hat{y}, u, v} \ [q, d, 0] \cdot [\hat{y}, u, v]$$

subject to

$$\begin{aligned} \begin{bmatrix} A & -A_E & 0 \\ -A & A_E & D \end{bmatrix} \begin{bmatrix} \hat{y} \\ u \\ v \end{bmatrix} &\geqq \begin{bmatrix} p \\ -p \end{bmatrix} \\ [\hat{y}, u, v] &\geqq 0, \end{aligned}$$

where $u \in \boldsymbol{R}^{|E|}$ and $v \in \boldsymbol{R}^{|N|}$. Now define y by

$$y_j = \begin{cases} \hat{y}_j - u_j & j \in E \\ \hat{y}_j & j \notin E \end{cases}$$

and observe that the objective function can be rewritten as $q \cdot y$ and the constraints as

$$\begin{aligned} Ay &\geqslant p \\ -Ay + Dv &\geqslant -p \\ y_j &\geqslant 0 \quad j \notin E \\ v &\geqq 0. \end{aligned}$$

Now $(Dv)_i = \begin{cases} v_i & i \in N \\ 0 & i \notin N \end{cases}$, so for $i \notin N$ we must have $(Ay)_i = p_i$ and $(Ay)_i \geqslant p_i$ otherwise. In other words the dual can be written:

$$\begin{array}{c} \underset{y}{\text{minimize}} \ q \cdot y \\ \text{subject to} \\ \begin{aligned} (Ay)_i &\geqslant p_i \quad i \in N \\ (Ay)_i &= p_i \quad i \notin N \\ y_j &\geqslant 0 \quad j \notin E. \end{aligned} \end{array}$$

Recall that the variables in the dual are the Lagrange multipliers for the primal. Thus we see that, the Lagrange multipliers associated with the equality constraints ($i \in E$) are not a priori restricted in sign, while the multipliers for the inequality constraints ($i \notin E$) are nonnegative. Since the primal variables are the Lagrange multipliers for the dual program, the nonnegativity constraints ($i \in N$) on the primal correspond to inequality constraints in the dual, and the unrestricted primal variable are associated with equality constraints in the dual.

There is one more useful form for linear programs, which is called the **equality form**.[12] In it, all the constraints are equations, and all the variables are nonnegative. An LP is in **equality maximum form** if it is written as:

$$\begin{array}{c} \underset{x}{\text{maximize}} \ p \cdot x \\ \text{subject to} \\ \begin{aligned} xA &= q \\ x &\geqq 0 \end{aligned} \end{array}$$

To transform an inequality form into the equality form, introduce slack variables $x \in \boldsymbol{R}^{\mathrm{m}}$ and observe that

$$xA \leqq q \quad \Longleftrightarrow \quad xA + z = q, \ z \geqq 0.$$

I leave it to you to verify that the dual program can be written as the decidedly non-equality minimum problem

[12] The equality form is what Dantzig [39] calls the standard form, and what Gale [57] calls the canonical form. Dantzig uses the term canonical in a different fashion.

Primal program		**Dual program**
Inequality maximum form		Inequality minimum form
$\underset{x}{\text{maximize}}\ p \cdot x$ subject to $xA \leqq q$ $x \geqq 0$		$\underset{y}{\text{minimize}}\ q \cdot y$ subject to $Ay \geqq p$ $y \geqq 0$
Equality maximum form		
$\underset{x}{\text{maximize}}\ p \cdot x$ subject to $xA = q$ $x \geqq 0$		$\underset{y}{\text{minimize}}\ q \cdot y$ subject to $Ay \geqq p$
Equality minimum form		
$\underset{y}{\text{minimize}}\ q \cdot y$ subject to $Ay = p$ $y \geqq 0$		$\underset{x}{\text{maximize}}\ p \cdot x$ subject to $xA \leqq q$
General maximum form		General minimum form
$\underset{x}{\text{maximize}}\ p \cdot x$ subject to $(xA)_j \leqslant q_j \quad j \notin E$ $(xA)_j = q_j \quad j \in E$ $x_i \geqq 0 \quad i \in N$		$\underset{y}{\text{minimize}}\ q \cdot y$ subject to $(Ay)_i \geqslant p_i \quad i \in N$ $(Ay)_i = p_i \quad i \notin N$ $y_j \geqslant 0 \quad j \notin E$

Table 4.1. Forms of linear programs and their duals.

$$\underset{y}{\text{minimize}}\ q \cdot y$$

subject to

$$Ay \geqq p$$

Note the lack of sign restrictions on y.

Table 4.1 summarizes these forms and their dual programs.

4.31 Linear equations as LPs

It is possible to recast the problem of solving linear equations and inequalities as LP problems. Consider the problem of finding a nonnegative solution to a system of equations. That is, find x such that

$$xA = q$$

$$x \geqq 0.$$

Consider the linear program in equality minimum form:

$$\underset{x,z}{\text{minimize}} \ \mathbf{1} \cdot z$$

subject to

$$\begin{aligned} xA + z &= q \\ [x, z] &\geqq 0 \end{aligned}$$

Here $\mathbf{1}$ is the vector whose components are all 1. Without loss of generality we may assume $q \geqq 0$, for if $q_j < 0$ we may multiply A^j and q^j by -1 without affecting the solution set. Then note that this program is feasible, since $x = 0$, $z = q$ is a nonnegative feasible solution. Since we require $z \geqq 0$, we have $\mathbf{1} \cdot z \geqslant 0$ and $\mathbf{1} \cdot z = 0$ if and only if $z = 0$, in which case $xA = q$. Thus, if this linear program has value 0 if and only $xA = q$, $x \geqq 0$ has a solution, and any optimal (x, z) provides a nonnegative solution to the equation.

At this point you might be inclined to say "so what?" In another handout, I will describe the simplex algorithm, which is a special version of Gauss–Jordan elimination, that is a reasonably efficient and easily programmable method for solving linear programs. In other words, it also finds nonnegative solutions to linear equations when they exist.

Section 5

Lagrange multiplier theory

5.1 Classical Lagrange Multiplier Theorem

Recall the following definition.

268 Definition *A point x^* is a **constrained local maximizer** of f subject to the constraints $g_1(x) = \alpha_1$, $g_2(x) = \alpha_2$,..., $g_m(x) = \alpha_m$ in some neighborhood W of x^* if x^* satisfies the constraints and also satisfies $f(x^*) \geqslant f(x)$ for all $x \in W$ that also satisfy the constraints.*

The classical Lagrange Multiplier Theorem on constrained optima for differentiable functions has a simple geometric interpretation, which is easiest to see with a single constraint. Consider a point that maximizes $f(x)$ subject to the equality constraint $g(x) = \alpha$. It should be clear from Figure 5.1 that at a point where a local maximum occurs, the level curves of f and g must

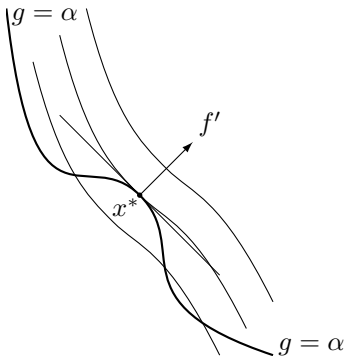

Figure 5.1. Constrained Maximum with an Equality Constraint.

be tangent. Since the gradient vectors are always perpendicular to the tangent line, they must be colinear.[1] Algebraically, this means that there are coefficients μ^* and λ^* (multipliers, if you will), not both zero, satisfying

$$\mu^* f'(x^*) + \lambda^* g'(x^*) = 0.$$

In general, this is all that can be said. But if the gradient g' is nonzero, then, as we shall see, the multiplier on f' can be taken to be unity, and we get the more familiar condition, $f' + \lambda^* g' = 0$.

[1] I know that this is usually spelled *collinear*, but my dictionary [117] lists *colinear* as a standard English word (bottom section, p. 524). The other spelling derives from assimilation of the prefix *com-*, derived from the Latin *cum*. The rules of assimilation change *com-* to *col-* before *l*, to *cor-* before *r*, to *con-* before any consonant except *b, h, l, m, p, r*, and *w*, to *co-* before a vowel, *h, w*, and sometimes before other consonants. All this seems unAmerican to me, so I prefer *colinear*. On the other hand, I still write *correspondence*.

Note that this does not imply that λ itself is nonzero, since f' may be zero itself. Also note that in general we cannot say anything about the sign of λ^*. That is, there is nothing to tell us if g' points in the same direction as f', or the opposite direction. This changes when we have an inequality constraint. If there is a local maximum of f subject to $g(x) \geqslant \alpha$, then the gradient of g points into $[g > \alpha]$, and the gradient of f points out. See Figure 5.2. This means

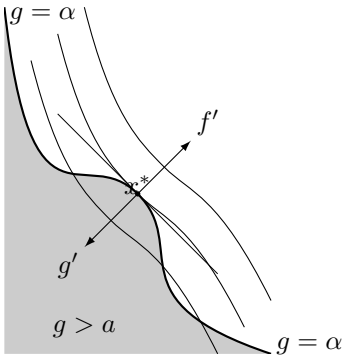

Figure 5.2. Constrained Maximum with an Inequality Constraint.

that we can take $\mu^*, \lambda^* \geqslant 0$. Even if $[g > \alpha]$ is empty, then $g' = 0$ (why?), so we can take $\mu^* = 0$ and $\lambda^* = 1$. That's really all there is to it, so keep these pictures in mind through all the complications needed to express these ideas formally.

In Section 4.29 we also saw that the Lagrange multipliers could serve as conversion factors so that (at least in the concave case) a constrained maximizer is also an unconstrained maximizer of the Lagrangean. Unfortunately, without concave functions this may not be the case. The following example may be found in Sydsaeter [146].

269 Example (Constained maximizers may not maximize the Lagrangean) Consider maximizing $f(x, y) = xy$ subject to $g(x, y) = 2 - x - y = 0$. The objective function f is not concave, although the constraint function g is. The constrained maximizer is $(x^*, y^*) = (1, 1)$ and for $\lambda^* = 1$ we have $f'(x^*, y^*) + \lambda^* g'(x^*, y^*) = 0$ but $L^*(x, y) = xy + 2 - x - y$ is not maximized at $(1, 1)$. (To see this, note that $L^*(1 + \varepsilon, 1 + \varepsilon) = (1 + \varepsilon)(1 + \varepsilon) + 2 - (1 + \varepsilon) - (1 + \varepsilon) = 1 + \varepsilon^2 > 1 = L^*(1, 1)$ for $\varepsilon \neq 0$.) Thus $(1, 1)$ is a maximizer along the constraint line, but is a minimizer of L^* along the ray from the origin, which is orthogonal to the constraint. □

The proofs of the Lagrange Multiplier Theorem make use of the Implicit Function Theorem and its corollaries, which I discuss in section 3.15 and 3.16. The main result is the Fundamental Lemma on Curves 124, which says that if x^* satisfies the m constraints $g_1(x), \dots, g_m(x) = 0$, and if v is orthogonal to the gradient of each of the independent constraints at x^*, then there is a differentiable curve $(\hat{x})$ through x^* satisfying the constraints with derivative equal to v at x^*.

270 Lagrange Multiplier Theorem I *Let $X \subset \boldsymbol{R}^n$, and let $f, g_1, \dots, g_m \colon X \to \boldsymbol{R}$ be continuous. Let x^* be an interior constrained local maximizer of f subject to $g(x) = 0$. Suppose f, $g_1, \dots, g_m$ are differentiable at x^*, and that $g_1'(x^*), \dots, g_m'(x^*)$ are linearly independent.*

Then there exist real numbers $\lambda_1^*, \dots, \lambda_m^*$, such that

$$f'(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i'(x^*) = 0.$$

Proof: Let $v \in \boldsymbol{R}^n$ satisfy

$$g_1'(x^*) \cdot v = \cdots = g_m'(x^*) \cdot v = 0.$$

By the Fundamental Lemma on Curves 124 there is a curve $\hat{x}\colon (-\delta,\delta) \to X$ that is differentiable at x^* and satisfying $g\big(\hat{x}(\alpha)\big) = 0$, $\hat{x}(0) = x^*$, and $\hat{x}'(0) = v$. Define $\tilde{f}(\alpha) = f\big(\hat{x}(\alpha)\big)$. Since $\tilde{f}$ achieves a local maximum at $\alpha = 0$ (why?), $\tilde{f}'(0) = 0$. Thus

$$0 = \tilde{f}'(0) = \sum_{j=1}^{n} D_j f(x^*)\hat{x}_j'(0) = f'(x^*) \cdot v.$$

Therefore

$$g_i{}'(x^*) \cdot v = 0 \quad i = 1,\dots,m \quad \Longrightarrow \quad f'(x^*) \cdot v = 0$$

so by Corollary 241 of the Fredholm Alternative, $f'(x^*)$ is a linear combination of the $g_i{}'(x^*)$s,

$$f'(x^*) = \sum_{i=1}^{m} \mu_i g_i{}'(x^*).$$

Thus setting $\lambda_i^* = -\mu_i$, we get $f'(x^*) + \sum_{i=1}^m \lambda_i^* g_i{}'(x^*) = 0$. ∎

The next result is provides a different version of the Lagrange Multiplier Theorem that includes the first as a special case. The argument is essentially that of Carathéodory [35, Theorem 11.1, pp. 175–177].

271 Lagrange Multiplier Theorem II *Let $X \subset \boldsymbol{R}^{\mathrm{n}}$, and let $f, g_1, \dots, g_m\colon X \to \boldsymbol{R}$ be continuous. Let x^* be an interior constrained local maximizer of f subject to $g(x) = 0$. Suppose f, $g_1, \dots, g_m$ are differentiable at x^*.*

Then there exist real numbers $\mu^*, \lambda_1^*, \dots, \lambda_m^*$, not all zero, such that

$$\mu^* f'(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i{}'(x^*) = 0.$$

Furthermore, if $g_1{}'(x^*), \dots, g_m{}'(x^*)$, are linearly independent, we may take μ^* to be unity.

Proof: Set $\alpha^* = f(x^*)$. Define $h\colon X \times \boldsymbol{R} \to \boldsymbol{R}^{\mathrm{m+1}}$ by

$$\begin{aligned} h^0(x;\alpha) &= f(x) - \alpha \\ h^i(x;\alpha) &= g_i(x), \quad i = 1,\dots,m. \end{aligned}$$

Start by observing that each h^i, $i = 0,\dots,m$, is differentiable in x at (x^*, α^*). I claim that

$$\begin{bmatrix} D_1 h^0(x^*;\alpha^*) \\ \vdots \\ D_n h^0(x^*;\alpha^*) \end{bmatrix}, \dots, \begin{bmatrix} D_1 h^m(x^*;\alpha^*) \\ \vdots \\ D_n h^m(x^*;\alpha^*) \end{bmatrix}$$

are linearly dependent.

To see this first note that if $m \geqslant n$, then the gradients must be dependent because they lie in $\boldsymbol{R}^{\mathrm{n}}$. So consider the case $m < n$, and suppose by way of contradiction that they are independent. Then by renumbering the coordinates if necessary, we may assume that

$$\begin{bmatrix} D_1 h^0(x^*;\alpha^*) & \cdots & D_{m+1} h^0(x^*;\alpha^*) \\ \vdots & & \vdots \\ D_1 h^m(x^*;\alpha^*) & \cdots & D_{m+1} h^m(x^*;\alpha^*) \end{bmatrix}$$

is invertible.

Let $U \times W$ be a neighborhood of x^* on which it is a local maximum point, where $U \subset \boldsymbol{R}^{m+1}$. Then, treating $x^*_{m+2}, \dots, x^*_n$ as fixed, the Implicit Function Theorem 113 implies that there is a neighborhood $V \subset R$ of α^* and a function $\xi\colon V \to U$ satisfying

$$\begin{aligned} \xi(\alpha^*) &= (x^*_1, \dots, x^*_{m+1}) \\ h^i\big(\xi(\alpha), x^*_{m+2}, \dots, x^*_n; \alpha\big) &= 0 \quad i = 0, \dots, m \end{aligned}$$

for all $\alpha \in V$.

Thus for $\alpha \in V$, $g_i\big(\xi(\alpha), x^*_{m+2}, \dots, x^*_n\big) = a_i$, $i = 1, \dots, m$, but for $\alpha \in V$ satisfying $\alpha > \alpha^*$, we have $f\big(\xi(\alpha), x^*_{m+2}, \dots, x^*_n\big) = \alpha > \alpha^* = f(x^*)$, contradicting the hypothesis that x^* maximizes f over $U \times W$. Therefore the gradients are dependent.

Thus there are real numbers $\mu^*, \lambda^*_1, \dots \lambda^*_m$, not all zero, such that

$$\mu^* \begin{bmatrix} D_1 h^0(x^*) \\ \vdots \\ D_n h^0(x^*) \end{bmatrix} + \sum_{i=1}^m \lambda^*_i \begin{bmatrix} D_1 h^i(x^*) \\ \vdots \\ D_n h^i(x^*) \end{bmatrix} = 0.$$

But from the definitions of h^i, we get

$$\mu^* f'(x^*) + \sum_{i=1}^m \lambda^*_i {g_i}'(x^*) = 0.$$

Suppose now that ${g_1}'(x^*), \dots, {g_m}'(x^*)$, are linearly independent. Suppose by way of contradiction that $\mu^* = 0$. Then $\sum_{i=1}^m \lambda^*_i {g_i}'(x^*) = 0$ and not all $\lambda^*_i = 0$, $i = 1, \dots, m$. This contradicts the linear independence of the ${g_i}'$=s.

Since $\mu^* \neq 0$, we can divide by it. Replacing λ^*_i by $\frac{\lambda^*_i}{\mu^*}$, $i = 1, \dots, m$, we have

$$f'(x^*) + \sum_{i=1}^m \lambda^*_i {g_i}'(x^*) = 0.$$

∎

Let us now consider some examples.

272 Example (Multipliers are zero) The Lagrange Multiplier Theorem does not guarantee that all the multipliers on the constraints will be nonzero. In fact the multipliers on the constraints may all be zero. For instance consider the constrained maximum of

$$f(x, y) = -(x^2 + y^2)$$

subject to the single constraint

$$g(x, y) = y = 0.$$

Observe that $g'(x, y) = (0, 1) \neq 0$, so the gradient is linearly independent. The point $(0, 0)$ is a constrained maximizer of f, but $f'(x, y) = (-2x, -2y)$ is equal to zero at $(0, 0)$. Thus the only way to solve $f'(0, 0) + \lambda^* g'(0, 0)$ is to set $\lambda^* = 0$. □

273 Example (Dependent constraint gradients) If you are like me, you may be tempted to think that if the gradients of the constraints are linearly dependent, then one of them may be redundant. This is not true. Consider the constrained maximum of

$$f(x, y) = x$$

subject to the two constraints

$$\begin{aligned} g_1(x,y) &= y - x^2 = 0 \\ g_2(x,y) &= y + x^2 = 0. \end{aligned}$$

It is easy to see that $(0,0)$ is the only point satisfying both constraints, and

$$g_1'(0,0) = (0,1) = g_2'(0,0).$$

Thus the gradients of the constraints are dependent at the maximizer. Since $f' = (1,0)$, there is no solution to $f'(0,0) + \lambda_1^* g_1(0,0) + \lambda_2^* g_2(0,0)$. There is however a nonzero solution to $\lambda_0^* f'(0,0) + \lambda_1^* g_1(0,0) + \lambda_2^* g_2(0,0)$, namely $\lambda_0^* = 0$, $\lambda_1^* = 1$, and $\lambda_2^* = -1$.

Notice that neither constraint is redundant, since if one of them is dropped, there are no constrained maxima. □

5.2 Second Order Conditions for a Constrained Extremum

The Fundamental Lemma 124 allows us to investigate the second order conditions for a constrained local maximum.

274 Theorem (Necessary Second Order Conditions) *Let $U \subset \boldsymbol{R}^{\mathrm{n}}$ and let $x^* \in \operatorname{int} U$. Let $f, g_1, \dots, g_m \colon U \to \boldsymbol{R}$ be C^2, and suppose x^* is a local constrained maximizer of f subject to $g(x) = 0$. Define the Lagrangean $L(x,\lambda) = f(x) + \sum_{i=1}^m \lambda_i g_i(x)$. Assume that ${g_1}'(x^*), \dots, {g_m}'(x^*)$ are linearly independent, so the conclusion of the Lagrange Multiplier Theorem holds, that is, there are $\lambda_1^*, \dots, \lambda_m^*$ satisfying the first order conditions*

$$L_x'(x^*, \lambda^*) = f'(x^*) + \sum_{i=1}^m \lambda_i^* {g_i}'(x^*) = 0.$$

Then

$$\sum_{i=1}^n \sum_{j=1}^n D_{i,j} L(x^*, \lambda^*) v_i v_j \leqslant 0,$$

for all $v \neq 0$ satisfying ${g_i}'(x^*) \cdot v = 0$, $i = 1, \dots, m$.

Proof: Let $v \neq 0$ satisfy ${g_i}'(x^*) \cdot v = 0$, $i = 1, \dots, m$. By Lemma 124 there is a C^2 function $\hat{x} \colon (-\delta, \delta) \to \boldsymbol{R}^{\mathrm{n}}$ satisfying $g_i\big(\hat{x}(\alpha)\big) = 0$, for $i = 1, \dots, m$, $\hat{x}(0) = x^*$, and $\hat{x}'(0) = v$. Define $\tilde{f}(\alpha) = f\big(\hat{x}(\alpha)\big) = L\big(\hat{x}(\alpha), \lambda^*\big)$. Then $\tilde{f}$ is C^2 and since $g\big(\hat{x}(\alpha)\big) = 0$, $\tilde{f}$ assumes its maximum at $\alpha = 0$.

Thus from our one dimensional theory we know $\tilde{f}'(0) = 0$ and $\tilde{f}''(0) \leqslant 0$. But

$$\tilde{f}'(\alpha) = \sum_{j=1}^n D_j L\big(\hat{x}(\alpha), \lambda^*\big) \frac{d\hat{x}_j(\alpha)}{d\alpha}$$

so

$$\tilde{f}''(\alpha) = \sum_{i=1}^n \left(\sum_{j=1}^n D_{i,j} L\big(\hat{x}(\alpha), \lambda^*\big) \frac{d\hat{x}_i(\alpha)}{d\alpha} \frac{d\hat{x}_j \cdot (\alpha)}{d\alpha} + D_j L\big(\hat{x}(\alpha), \lambda^*\big) \frac{d^2 \hat{x}_j(\alpha)}{d(\alpha^2)} \right).$$

But by the first order conditions

$$D_j L\big(\hat{x}(0), \lambda^*\big) = 0.$$

Using this and $\hat{x}_j'(0) = v_j$ gives

$$0 \geqslant \tilde{f}''(0) = \sum_{i=1}^n \sum_{j=1}^n D_{i,j} L(x^*, \lambda^*) v_i v_j.$$

∎

Sufficient conditions can be derived much as in the unconstrained case. In progress.

275 Theorem (Sufficient Second Order Conditions) *Let $U \subset \boldsymbol{R}^{\mathrm{n}}$ and let $x^* \in \operatorname{int} U$. Let $f, g_1, \dots, g_m \colon U \to \boldsymbol{R}$ be continuously differentiable on U and twice differentiable at x^*. Assume $g(x^*) = 0$, and that there exist $\lambda_1^*, \dots, \lambda_m^*$ satisfying the first order conditions*

$$L'_x(x^*, \lambda^*) = f'(x^*) + \sum_{i=1}^m \lambda_i^* g_i'(x^*) = 0.$$

Assume further that the strong second order condition holds, that is,

$$\sum_{i=1}^n \sum_{j=1}^n D_{i,j} L(x^*, \lambda^*) v_i v_j < 0,$$

for all $v \neq 0$ satisfying $g_i'(x^*) \cdot v = 0$, $i = 1, \dots, m$. Then x^* is a strict local maximizer of f subject to the constraints $g(x) = 0$.

Proof: By Young's form of Taylor's Theorem for many variables 106, recalling that $Df(x^*) = 0$, we have

$$f(x^* + v) = f(x^*) + Df(x^*)(v) + \tfrac{1}{2} D^2 f(x^*)(v, v) + \tfrac{r(v)}{2} \|v\|^2,$$

where $\lim_{v \to 0} r(v) = 0$. What this tells us is that the increment $f(x^* + v) - f(x^*)$ is bounded between two quadratic forms that can be made arbitrarily close to $Q(v) = D^2 f(x^*)(v, v)$. This is the source of conclusions.

********** The quadratic form Q achieves its maximum M and minimum m values on the unit sphere (and they are the maximal and minimal eigenvalues, see Proposition 304). If Q is positive definite, then $0 < m \leqslant M$, and homogeneity of degree 2 implies that $m\|v\|^2 \leqslant Q(v) \leqslant M\|v\|^2$ for all v. Choose $0 < \varepsilon < m$. Then there exist $\delta > 0$ such that $\|v\| < \delta$ implies $|r(v)| < \varepsilon$. The first inequality in $(\star\star)$ thus implies

$$0 < \frac{m - \varepsilon}{2} \|v\|^2 \leqslant f(x^* + v) - f(x^*),$$

for $\|v\| < \delta$, which shows that x^* is a strict local minimizer. Similarly if Q is negative definite, then x^* is a strict local maximizer. If Q is nonsingular, but neither negative or positive definite, then $\boldsymbol{R}^{\mathrm{n}}$ decomposes into two orthogonal nontrivial subspaces, and is positive definite on one and negative definite on the other. It follows then that x^* is neither a maximizer nor a minimizer. ∎

5.3 Constrained Minimization

Since minimizing f is the same as maximizing $-f$, we do not need any new results for minimization, but there a few things worth pointing out.

The Lagrangean for maximizing $-f$ subject to $g_i = 0$, $i = 1, \dots, m$ is

$$-f(x) + \sum_{i=1}^m \lambda_i g_i(x),$$

The second order condition for maximizing $-f$ is that

$$\sum_{i=1}^n \sum_{j=1}^n \left(-D_{ij} f(x^*) + \sum_{i=1}^m \lambda^* D_{ij} g(x^*) \right) v_i v_j \leqslant 0,$$

for all $v \neq 0$ satisfying $g_i'(x^*) \cdot v = 0$, $i = 1, \dots, m$. This can be rewritten as

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \left(D_{ij} f(x^*) - \sum_{i=1}^{m} \lambda^* D_{ij} g(x^*) \right) v_i v_j \geqslant 0,$$

which suggests that it is more convenient to define the Lagrangean for a minimization problem as

$$L(x, \lambda) = f(x) - \sum_{i=1}^{m} \lambda_i g_i(x).$$

The first order conditions will be exactly the same. For the second order conditions we have the following.

276 Theorem (Necessary Second Order Conditions for a Minimum) *Let $U \subset \boldsymbol{R}^{\mathrm{n}}$ and let $x^* \in \operatorname{int} U$. Let $f, g_1, \dots, g_m \colon U \to \boldsymbol{R}$ be C^2, and suppose x^* is a local constrained minimizer of f subject to $g(x) = 0$. Define the Lagrangean*

$$L(x, \lambda) = f(x) - \sum_{i=1}^{m} \lambda_i g_i(x).$$

Assume that $g_1'(x^*), \dots, g_m'(x^*)$ are linearly independent, so the conclusion of the Lagrange Multiplier Theorem holds, that is, there are $\lambda_1^*, \dots, \lambda_m^*$ satisfying the first order conditions

$$L'_x(x^*, \lambda^*) = f'(x^*) - \sum_{i=1}^{m} \lambda_i^* g_i'(x^*) = 0.$$

Then

$$\sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij} L(x^*, \lambda^*) v_i v_j \geqslant 0,$$

for all $v \neq 0$ satisfying $g_i'(x^*) \cdot v = 0$, $i = 1, \dots, m$.

5.4 Inequality constraints

The classical Lagrange Multiplier Theorem deals only with equality constraints. Now we take up inequality constraints. We start by transforming the problem into one involving only equality constraints. I learned this approach from Quirk [128].

277 Theorem *Let $U \subset \boldsymbol{R}^{\mathrm{n}}$ be open, and let $f, g_1, \dots, g_m \colon U \to \boldsymbol{R}$ be twice continuously differentiable on U. Let x^* be a constrained local maximizer of f subject to $g(x) \geqq 0$ and $x \geqq 0$.*

Let $B = \{i : g_i(x^*) = 0\}$, the set of binding constraints, and let $Z = \{j : x_j = 0\}$, the set of binding nonnegativity constraints. Assume that $\{g_i'(x^*) : i \in B\} \cup \{e^j : j \in Z\}$ is linearly independent. Then there exists $\lambda^* \in \boldsymbol{R}^{\mathrm{m}}$ such that

$$f'(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i'(x^*) \leqq 0. \tag{5.1}$$

$$x^* \cdot \left(f'(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i'(x^*) \right) = 0 \tag{5.2}$$

$$\lambda^* \geqq 0. \tag{5.3}$$

$$\lambda^* \cdot g(x^*) = 0. \tag{5.4}$$

Proof: Introduce $m+n$ **slack variables** $y_1,\dots,y_m$ and $z_1,\dots,z_n$, and consider the equality constrained maximization problem:

$$\text{maximize } \bar f(x) \text{ subject to } g^i(x) - y_i^2 = 0,\ i=1,\dots,m, \text{ and } x_j - z_j^2 = 0,\ j=1,\dots,n.$$

Define y^* and z^* by

$$y_i^* = \sqrt{g_i(x^*)} \tag{5.5}$$

$$z_j^* = \sqrt{x_j^*}, \tag{5.6}$$

Observe that (x^*, y^*, z^*) solves the equality constrained maximization problem.

So on $U \times \boldsymbol{R}^{\mathrm{m}} \times \boldsymbol{R}^{\mathrm{n}}$ define

$$\begin{aligned}
\bar f(x,y,z) &= f(x),\\
\bar g_i(x,y,z) &= g_i(x) - y_i^2, \quad i=1,\dots,m,\\
\bar h_j(x,y,z) &= x_j - z_j^2, \quad j=1,\dots,n.
\end{aligned}$$

Note that these functions are also twice continuously differentiable. Then (x^*, y^*, z^*) solves the revised equality constrained maximization problem:

$$\text{maximize } \bar f(x,y,z) \text{ subject to } \bar g_i(x,y,z) = 0,\ i=1,\dots,m, \text{ and } \bar h_j(x,y,z) = 0,\ j=1,\dots,n.$$

In order to apply the Lagrange Multiplier Theorem to this revised equality constrained problem, we need to verify that the gradients of $\bar g_i'(x^*,y^*,z^*)$, $i=1,\dots,m$ and $\bar h_j'(x^*,y^*,z^*)$, $j=1,\dots,n$ of the constraints with respect to the variables x, y, z are linearly independent. So suppose $\sum_{i=1}^m \alpha_i \bar g_i' + \sum_{j=1} n\beta_j \bar h_j' = 0$. Now

$$\bar g_i'(x^*,y^*,z^*) = \begin{bmatrix} g_i{}'(x^*) \\ -2y_i^* e^i \\ 0 \end{bmatrix} \qquad \text{and} \qquad \bar h_j'(x^*,y^*,z^*) = \begin{bmatrix} e^j \\ 0 \\ -2z_j^* e^j \end{bmatrix}. \tag{5.7}$$

So the y_i component of this sum is just $-2\alpha_i y_i^*$. Therefore

$$i \notin B \iff y_i^* > 0 \implies \alpha_i = 0.$$

Similarly the z_j component is $-2\beta_j z_j^*$, so

$$j \notin Z \iff z_j^* > 0 \implies \beta_j = 0.$$

Given this, the x component is just

$$\sum_{i\in B} \alpha_i \bar g_i'(x^*) + \sum_{j\in Z} \beta_j e^j = 0.$$

By hypothesis, these vectors are linearly independent, so we conclude that $\alpha_i = 0$, $i \in B$, and $\beta_j = 0$, $j \in Z$, which proves the linear independence of the constraint gradients of the revised equality problem.

So form the Lagrangean

$$\begin{aligned}
\bar L(x,y,z;\lambda,\mu) &= \bar f(x,y,z) + \sum_{i=1}^m \lambda_i \bar g_i(x,y,x) + \sum_{j=1}^n \mu_j \bar h_j(x,y,z).\\
&= f(x) + \sum_{i=1}^m \lambda_i \big(g_i(x) - y_i^2\big) + \sum_{j=1}^n \mu_j (x_j - z_j^2).
\end{aligned}$$

Then by the Lagrange Multiplier Theorem 270 there are multipliers λ_i^*, $i = 1, \dots, m$ and μ_j^*, $j = 1, \dots, n$ such that the following first order conditions are satisfied.

$$\frac{\partial f(x^*)}{\partial x_j} + \sum_{i=1}^{m} \lambda_i^* \frac{\partial g_i(x^*)}{\partial x_j} + \mu_j^* = 0 \qquad j = 1, \dots, n, \tag{5.8}$$

$$-2\lambda_1^* y_i^* = 0 \qquad i = 1, \dots, m, \tag{5.9}$$

$$-2\mu_j^* z_j^* = 0 \qquad j = 1, \dots, n. \tag{5.10}$$

Now the Hessian of the Lagrangean $\bar{L}$ (with respect to (x, y, z) and evaluated at $(x^*, y^*, z^*; \lambda^*, \mu^*)$) is block diagonal:

$$\left[\begin{array}{ccc|ccc|ccc}
\frac{\partial^2 L}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 L}{\partial x_1 \partial x_n} & & & & & & \\
\vdots & & \vdots & & & & & & \\
\frac{\partial^2 L}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 L}{\partial x_n \partial x_n} & & & & & & \\
\hline
 & & & -2\lambda_1^* & & & & & \\
 & & & & \ddots & & & & \\
 & & & & & -2\lambda_m^* & & & \\
\hline
 & & & & & & -2\mu_1^* & & \\
 & & & & & & & \ddots & \\
 & & & & & & & & -2\mu_n^*
\end{array}\right]$$

From the second order conditions (Theorem 274) for the revised equality constrained problem, we know that this Hessian is negative semidefinite under constraint. That is, if a vector is orthogonal to the gradients of the constraints, then the quadratic form in the Hessian is nonpositive. In particular, consider a vector of the form $v = (0, e^k, 0) \in \boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}^{\mathrm{m}} \times \boldsymbol{R}^{\mathrm{n}}$. It follows from (5.7) that for $i \neq k$ this vector v is orthogonal to $\bar{g}_i'(x^*, y^*, z^*)$, and also orthogonal to $\bar{h}_j'(x^*, y^*, z^*)$, $j = 1, \dots, n$. The vector v is orthogonal to $\bar{g}_k'(x^*, y^*, z^*)$ if and only if $y_k^* = 0$, that is, when $k \in B$. Thus for $k \in B$ the second order conditions imply $-2\lambda_k^* \leqslant 0$, so

$$g_i(x^*) = 0 \implies \lambda_i^* \geqslant 0.$$

Next consider a vector of the form $u = (0, 0, e^k) \in \boldsymbol{R}^{\mathrm{n}} \times \boldsymbol{R}^{\mathrm{m}} \times \boldsymbol{R}^{\mathrm{n}}$. It follows from (5.7) that this vector u is orthogonal to each $\bar{g}_i'(x^*, y^*, z^*)$ each $\bar{h}_j'(x^*, y^*, z^*)$ for $j \neq k$. The vector u is orthogonal to $\bar{h}_k'(x^*, y^*, z^*)$ if and only if $z_k^* = 0$, that is, $j \in Z$. Again the second order conditions imply that the quadratic form in u, which has value $-2\mu_k^*$ is nonnegative for $k \in Z$, so

$$x_j^* = 0 \implies \mu_j^* \geqslant 0.$$

Now if $i \notin B$, that is, $g_i(x^*) > 0$, so that $y_i^* > 0$, then from the first order condition (5.9) we have $\lambda_i^* = 0$. Also, from (5.10), if $x_j^* > 0$, so that $z_j^* > 0$, then $\mu_j^* = 0$. That is,

$$g_i(x^*) > 0 \implies \lambda_i^* = 0 \qquad \text{and} \qquad x_j^* > 0 \implies \mu_j^* = 0$$

Combining this with the paragraph above we see that $\lambda^* \geqq 0$ and $\mu^* \geqq 0$. Thus (5.8) implies conclusion (5.1). A little more thought will show you that we have just deduced conditions (5.2) through (5.4) as well. ∎

There is a simple variation on the slack variable approach that applies to mixed inequality and equality constraints. To prove the next result, simply omit the slack variables for the equality constraints and follow the same proof as in Theorem 277.

278 Corollary *Let $U \subset \boldsymbol{R}^{\mathrm{n}}$ be open, and let $f, g_1, \dots, g_m \colon U \to \boldsymbol{R}$ be twice continuously differentiable on U. Let x^* be a constrained local maximizer of f subject to*

$$\begin{aligned} g_i(x) &= 0 \quad i \in E, \\ g_i(x) &\geqslant 0 \quad i \in E^c, \\ x_j &\geqslant 0 \quad j \in N. \end{aligned}$$

Let $B = \{i \in E^c : g_i(x^*) = 0\}$ (binding inequality constraints), and let $Z = \{j \in N : x_j = 0\}$ (binding nonnegativity constraints). Assume that

$$\{g_i'(x^*) : i \in E \cup B\} \cup \{e^j : j \in Z\} \text{ is linearly independent,}$$

then there exists $\lambda^* \in \boldsymbol{R}^{\mathrm{m}}$ such that

$$\begin{aligned} \frac{\partial f(x^*)}{\partial x_j} + \textstyle\sum_{j=1}^m \lambda_i^* \dfrac{\partial g_j(x^*)}{\partial x_j} &\leqslant 0 \quad j \in N, \\ \frac{\partial f(x^*)}{\partial x_i} + \textstyle\sum_{j=1}^m \lambda_i^* \dfrac{\partial g_j(x^*)}{\partial x_i} &= 0 \quad j \in N^c, \\ \lambda_i^* &\geqslant 0 \quad i \in E^c. \\ x^* \cdot \Big(f'(x^*) + \textstyle\sum_{j=1}^m \lambda_i^* g_j'(x^*) \Big) &= 0 \\ \lambda^* \cdot g(x^*) &= 0. \end{aligned}$$

We now translate the result for minimization.

279 Theorem (Minimization) *Let $U \subset \boldsymbol{R}^{\mathrm{n}}$ be open, and let $f, g_1, \dots, g_m \colon U \to \boldsymbol{R}$ be twice continuously differentiable on U. Let x^* be a constrained local minimizer of f subject to $g(x) \geqq 0$ and $x \geqq 0$.*

Let $B = \{i : g_i(x^*) = 0\}$, the set of binding constraints, and let $Z = \{j : x_j = 0\}$, the set of binding nonnegativity constraints. Assume that $\{g_i'(x^*) : i \in B\} \cup \{e^j : j \in Z\}$ is linearly independent. Then there exists $\lambda^* \in \boldsymbol{R}^{\mathrm{m}}$ such that

$$f'(x^*) - \sum_{i=1}^m \lambda_i^* g_i'(x^*) \geqq 0. \tag{5.11}$$

$$x^* \cdot \left(f'(x^*) - \sum_{i=1}^m \lambda_i^* g_i'(x^*) \right) = 0 \tag{5.12}$$

$$\lambda^* \geqq 0. \tag{5.13}$$

$$\lambda^* \cdot g(x^*) = 0. \tag{5.14}$$

Proof: As in the proof of Theorem 277, introduce $m+n$ slack variables $y_1, \dots, y_m$ and $z_1, \dots, z_n$, and define $\bar{f}(x, y, z) = f(x)$, $\bar{g}_i(x, y, z) = g_i(x) - y_i^2$, $i = 1, \dots, m$ and $\bar{h}_j(x, y, z) = x_j - z_j^2$, $j = 1, \dots, n$. Again define y^* by $y_i^* = \sqrt{g_i(x^*)}$ and z^* by $z_j^* = \sqrt{x_j^*}$. Observe that (x^*, y^*, z^*) solves the revised equality constrained minimization problem:

minimize $\bar{f}(x, y, z)$ subject to $\bar{g}_i(x, y, z) = 0$, $i = 1, \dots, m$, and $\bar{h}_j(x, y, z) = 0$, $j = 1, \dots, n$.

The proof of the linear independence of the constraint gradients of the revised equality problem is the same as in Theorem 277.

So form the Lagrangean

$$\bar{L}(x, y, z; \lambda, \mu) = f(x) - \sum_{i=1}^{m} \lambda_i \big(g_i(x) - y_i^2\big) - \sum_{j=1}^{n} \mu_j (x_j - z_j^2).$$

Then by the Lagrange Multiplier Theorem 270 there are multipliers λ_i^*, $i = 1, \dots, m$ and μ_j^*, $j = 1, \dots, n$ such that the following first order conditions are satisfied.

$$\frac{\partial f(x^*)}{\partial x_j} - \sum_{i=1}^{m} \lambda_i^* \frac{\partial g_i(x^*)}{\partial x_j} - \mu_j^* = 0 \qquad j = 1, \dots, n, \tag{5.15}$$

$$2\lambda_1^* y_i^* = 0 \qquad i = 1, \dots, m, \tag{5.16}$$

$$2\mu_j^* z_j^* = 0 \qquad j = 1, \dots, n. \tag{5.17}$$

The Hessian of the Lagrangean $\bar{L}$ (with respect to (x, y, z) and evaluated at $(x^*, y^*, z^*; \lambda^*, \mu^*)$) is:

$$\begin{bmatrix} \frac{\partial^2 L}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 L}{\partial x_1 \partial x_n} & & & & \\ \vdots & & \vdots & & & & \\ \frac{\partial^2 L}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 L}{\partial x_n \partial x_n} & & & & \\ & & & 2\lambda_1^* & & & \\ & & & & \ddots & & \\ & & & & & 2\lambda_m^* & \\ & & & & & & 2\mu_1^* \\ & & & & & & & \ddots \\ & & & & & & & & 2\mu_n^* \end{bmatrix}$$

From the second order conditions for minimization (Theorem 276) for the revised equality constrained problem, we know that this Hessian is positive semidefinite under constraint. In particular, as in the proof of Theorem 277, we have that

$$g_i(x^*) = 0 \implies \lambda_i^* \geqslant 0.$$

$$x_j^* = 0 \implies \mu_j^* \geqslant 0.$$

From the first order conditions, if $i \notin B$, that is, $g_i(x^*) > 0$, so that $y_i^* = 0$, then $\lambda_i^* = 0$. Also if $x_j^* > 0$, so that $z_j^* = 0$, then $\mu_j^* = 0$. That is,

$$g_i(x^*) > 0 \implies \lambda_i^* = 0 \qquad \text{and} \qquad x_j^* > 0 \implies \mu_j^* = 0$$

Combining this with the paragraph above we see that $\lambda^* \geqq 0$ and $\mu^* \geqq 0$. Thus (5.15) implies conclusion (5.11). A little more thought will show you that we have just deduced conditions (5.12) through (5.14) as well.

∎

280 Corollary *Let $U \subset \boldsymbol{R}^{\mathrm{n}}$ be open, and let $f, g_1, \dots, g_m \colon U \to \boldsymbol{R}$ be twice continuously differentiable on U. Let x^* be a constrained local minimizer of f subject to*

$$g_i(x) = 0 \quad i \in E,$$

$$g_i(x) \geqslant 0 \quad i \in E^c,$$

$$x_j \geqslant 0 \quad j \in N.$$

Let $B = \{i \in E^c : g_i(x^*) = 0\}$ (binding inequality constraints), and let $Z = \{j \in N : x_j = 0\}$ (binding nonnegativity constraints). Assume that

$$\{g_i'(x^*) : i \in E \cup B\} \cup \{e^j : j \in Z\} \text{ is linearly independent,}$$

then there exists $\lambda^* \in \boldsymbol{R}^{\mathrm{m}}$ such that

$$\begin{aligned}
\frac{\partial f(x^*)}{\partial x_j} - \textstyle\sum_{j=1}^m \lambda_i^* \dfrac{\partial g_j(x^*)}{\partial x_j} &\leqslant 0 \quad j \in N,\\
\frac{\partial f(x^*)}{\partial x_i} + \textstyle\sum_{j=1}^m \lambda_i^* \dfrac{\partial g_j(x^*)}{\partial x_i} &= 0 \quad j \in N^c,\\
\lambda_i^* &\geqslant 0 \quad i \in E^c.\\
x^* \cdot \Big(f'(x^*) + \textstyle\sum_{j=1}^m \lambda_i^* g_j'(x^*)\Big) &= 0\\
\lambda^* \cdot g(x^*) &= 0.
\end{aligned}$$

5.5 Karush–Kuhn–Tucker Theory

A drawback of the slack variable approach is that it assumes twice continuous differentiability in order to apply the second order conditions and thus conclude $\lambda^* \geqq 0$ and $\mu^* \geqq 0$. Fortunately, Karush [90] and Kuhn and Tucker [95] provide another approach that remedies this shortcoming. They only assume differentiability, and replace the independence condition on gradients by a weaker but more obscure condition called the Karush–Kuhn–Tucker Constraint Qualification.

281 Definition *Let $f, g_1, \dots, g_m \colon \boldsymbol{R}^{\mathrm{n}}_+ \to \boldsymbol{R}$. Let*

$$C = \{x \in \boldsymbol{R}^{\mathrm{n}} : x \geqq 0,\ g_i(x) \leqslant 0,\ i = 1, \dots, m\}.$$

In other words, C is the constraint set. Consider a point $x^* \in C$ and define

$$B = \{i : g_i(x^*) = 0\} \text{ and } Z = \{j : x_j = 0\},$$

*the set of binding constraints and binding nonnegativity constraints, respectively. The point x^* satisfies the **Karush–Kuhn–Tucker Constraint Qualification** if $f, g_1, \dots, g_m$ are differentiable at x^*, and for every $v \in \boldsymbol{R}^{\mathrm{n}}$ satisfying*

$$\begin{aligned}
v_j = v \cdot e^j &\leqslant 0 \qquad j \in Z,\\
v \cdot g_i'(x^*) &\leqslant 0 \qquad i \in B,
\end{aligned}$$

there is a continuous curve $\xi \colon [0, \varepsilon) \to \boldsymbol{R}^{\mathrm{n}}$ satisfying

$$\begin{aligned}
\xi(0) &= x^*,\\
\xi(t) \in C &\quad \text{for all } t \in [0, \varepsilon),\\
D\xi(0) &= v,
\end{aligned}$$

Consistent notation?

where $D\xi(0)$ is the one-sided directional derivative at 0.

This condition is actually a little weaker than Kuhn and Tucker's condition. They assumed that the functions $f, g_1, \dots, g_m$ were differentiable everywhere and required ξ to be differentiable everywhere. You can see that it may be difficult to verify it in practice.

282 Theorem (Karush–Kuhn–Tucker) *Let $f, g_1, \dots, g_m \colon \boldsymbol{R}^{\mathrm{n}}_+ \to \boldsymbol{R}$ be differentiable at x^*, and let x^* be a constrained local maximizer of f subject to $g(x) \leqq 0$ and $x \geqq 0$.*

Let $B = \{i : g_i(x^*) = 0\}$, the set of binding constraints, and let $Z = \{j : x_j = 0\}$, the set of binding nonnegativity constraints. Assume that x^* satisfies the Karush–Kuhn–Tucker Constraint Qualification. Then there exists $\lambda^* \in \boldsymbol{R}^{\mathrm{m}}$ such that

$$f'(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i'(x^*) \leqq 0,$$

$$x^* \cdot \left(f'(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i'(x^*) \right) = 0,$$

$$\lambda^* \geqq 0,$$

$$\lambda^* \cdot g(x^*) = 0.$$

To better understand the hypotheses of the theorem, let's look at a classic example of its failure (cf. Kuhn and Tucker [95]).

283 Example (Failure of the Karush–Kuhn–Tucker Constraint Qualification) Consider the functions $f\colon \boldsymbol{R}^2 \to \boldsymbol{R}$ via $f(x, y) = x$ and $g\colon \boldsymbol{R}^2 \to \boldsymbol{R}$ via $g(x, y) = (1 - x)^3 - y$. The curve $g = 0$ is shown in Figure 5.3, and the constraint set in Figure 5.4.

Clearly $(x^*, y^*) = (1, 0)$ maximizes f subject to $(x, y) \geqq 0$ and $g \geqslant 0$. At this point we have $g'(1, 0) = (0, -1)$ and $f' = (1, 0)$ everywhere. Note that no λ (nonnegative or not) satisfies

$$(1, 0) + \lambda(0, -1) \leqq (0, 0).$$

Fortunately for the theorem, the Constraint Qualification fails at $(1, 0)$. To see this, note that the constraint $g \geqslant 0$ binds, that is $g(1, 0) = 0$ and the second coordinate of (x^*, y^*) is zero. Suppose $v = (v_x, v_y)$ satisfies

$$v \cdot g'(1, 0) = v \cdot (0, -1) = -v_y \leqslant 0 \qquad \text{and} \qquad v \cdot e^2 = v_y \geqslant 0,$$

that is, $v_y = 0$. For instance, take $v = (1, 0)$. The constraint qualification requires that there is a path starting at $(1, 0)$ in the direction $(1, 0)$ that stays in the constraint set. Clearly no such path exists, so the constraint qualification fails. □

Proof of Theorem 282: Let x^* be a constrained maximizer, and define the sets B and Z of indices as in the statement of the theorem, and let $v \in \boldsymbol{R}^{\mathrm{n}}$ satisfy

$$\begin{aligned} v_j = v \cdot e^j &\geqslant 0 \qquad j \in Z, \\ v \cdot g_i'(x^*) &\geqslant 0 \qquad i \in B. \end{aligned}$$

By the constraint qualification there is a continuous curve $\xi\colon [0, \varepsilon) \to \boldsymbol{R}^{\mathrm{n}}$ satisfying

$$\begin{aligned} \xi(0) &= x^*, \\ \xi(t) \in C &\quad \text{for all } t \in [0, \varepsilon), \\ D\xi(0) &= v. \end{aligned}$$

By hypothesis, x^* is a local maximizer, so $g(t) = f \circ \xi(t)$ attains its maximum at $t = 0$. It follows from Lemma 71 that $f'(x^*) \cdot v = f'(x^*) \cdot \xi'(0) \leqslant 0$.

We now use a theorem of the alternative in a manner similar to that in the proof of Theorem 270. We have just shown that the system

$$\begin{aligned} f'(x^*) \cdot v &> 0 \\ g_i'(x^*) \cdot v &\geqslant 0 && i \in B \\ e^j \cdot v &\geqslant 0 && j \in Z \end{aligned}$$

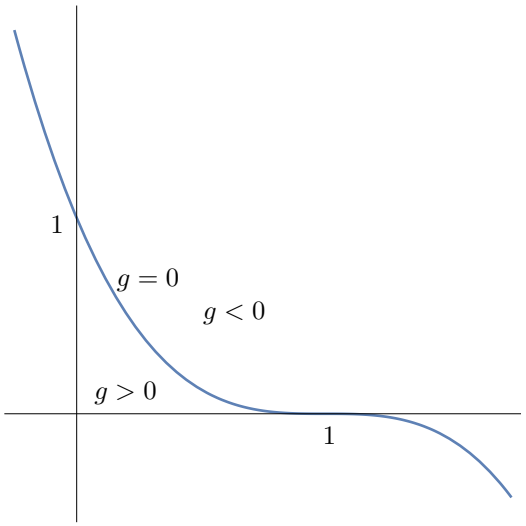

Figure 5.3. The function $g(x, y) = (1 - x)^3 - y$.

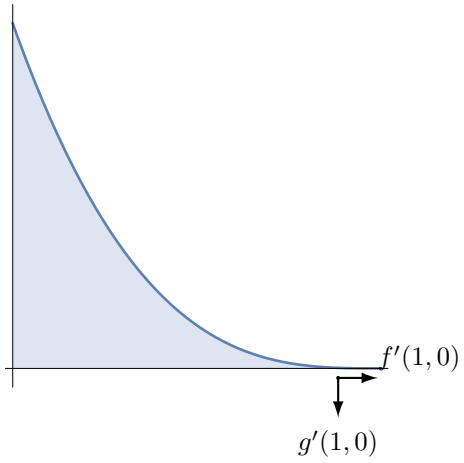

Figure 5.4. This constraint set violates the Constraint Qualification. (Note: f' and g' are not to scale.)

has no solution v. Therefore by Motzkin's Transposition Theorem 250 there exist $p > 0$, and $\lambda_i \geqslant 0$, $i \in B$ and $\mu_j \geqslant 0$, $j \in Z$ such that

$$pf'(x^*) + \sum_{i \in B} \lambda_i g_i'(x*) + \sum_{j \in Z} \mu_j e^j = 0.$$

Since $p > 0$, we may rescale these numbers so that without loss of generality $p = 1$: Create the vector $\lambda^* \in \boldsymbol{R}^{\mathrm{m}}$ by setting $\lambda_i^* = \lambda_i / p$ for $i \in B$ and $\lambda_i^* = 0$ for $i \notin B$, and define $\mu^* \in \boldsymbol{R}^{\mathrm{n}}$ by setting $\mu_j^* = \mu_j / p$ for $j \in Z$, and $\mu_j^* = 0$ for $j \notin Z$. Then dividing by $p > 0$ we have

$$f'(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i'(x*) + \sum_{j=1}^{n} \mu_j^* e^j = 0.$$

Since each $\mu_j^* \geqslant 0$, this says that

$$f'(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i'(x*) \leqq 0,$$

with component j being < 0 only if $\mu_j^* > 0$, which could happen only if $j \in Z$. Since $x^* \geqq 0$, these two facts imply

$$x_j^* \left(f'(x^*) + \sum_{i=1}^{m} \lambda_i^* {g_i}'(x^*) \right)_j = 0, \qquad j = 1, \dots, n.$$

Also, by construction $\lambda^* \geqq 0$, and $\lambda_i^* > 0$ can occur only if $i \in B$ (and not necessarily even then), so

$$\lambda_i^* g_i(x^*) = 0, \qquad i = 1, \dots, m.$$

This completes the proof. ∎

The next result, which may be found in Arrow, Hurwicz, and Uzawa [12, Corollaries 1, 4, 6, pp. 183–184], provides a tractable sufficient condition for the KTCQ.

284 Theorem *In Theorem 282, the KTCQ may be replaced by any of the conditions below.*

1. *Each g_i is convex. (This includes the case where each is linear.)*
2. *Each g_i is concave and there exists some $\hat{x} \gg 0$ for which each $g_i(\hat{x}) > 0$.*
3. *The set $\{e^j : j \in Z\} \cup \{{g_i}'(x^*) : i \in B\}$ is linearly independent.*

5.6 Karush–Kuhn–Tucker Theorem for Minimization

285 Theorem (Karush–Kuhn–Tucker) *Let $f, g_1, \dots, g_m \colon \boldsymbol{R}^{\mathrm{n}}_+ \to \boldsymbol{R}$ be differentiable at x^*, and let x^* be a constrained local minimizer of f subject to $g(x) \leqq 0$ and $x \geqq 0$.*

Let $B = \{i : g_i(x^*) = 0\}$, the set of binding constraints, and let $Z = \{j : x_j = 0\}$, the set of binding nonnegativity constraints. Assume that x^* satisfies the Karush–Kuhn–Tucker Constraint Qualification. Then there exists $\lambda^* \in \boldsymbol{R}^{\mathrm{m}}$ such that

$$f'(x^*) - \sum_{i=1}^{m} \lambda_i^* {g_i}'(x^*) \leqq 0,$$

$$x^* \cdot \left(f'(x^*) - \sum_{i=1}^{m} \lambda_i^* {g_i}'(x^*) \right) = 0,$$

$$\lambda^* \geqq 0,$$

$$\lambda^* \cdot g(x^*) = 0.$$

Proof: Minimizing f is the same as maximizing $-f$. The Karush–Kuhn–Tucker conditions for this imply that there exists $\lambda^* \in \boldsymbol{R}^{\mathrm{m}}_+$ such that

$$-f'(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i'(x^*) \leqq 0,$$

and the conclusion follows by multiplying this by -1. ∎

5.7 Quasiconcave functions

There are weaker notions of convexity that are commonly applied in economic theory.

286 Definition *A function $f\colon C \to \boldsymbol{R}$ on a convex subset C of a vector space is:*

- ***quasiconcave*** *if for all x, y in C with $x \neq y$ and all $0 < \lambda < 1$*

$$f\big(\lambda x + (1-\lambda)y\big) \geqslant \min\{f(x), f(y)\}.$$

- ***strictly quasiconcave*** *if for all x, y in C with $x \neq y$ and all $0 < \lambda < 1$*

$$f\big(\lambda x + (1-\lambda)y\big) > \min\{f(x), f(y)\}.$$

- ***explicitly quasiconcave*** *or* ***semistrictly quasiconcave*** *if it is quasiconcave and in addition, for all x, y in C with $x \neq y$ and all $0 < \lambda < 1$*

$$f(x) > f(y) \implies f\big(\lambda x + (1-\lambda)y\big) > \min\{f(x), f(y)\} = f(y).$$

- ***quasiconvex*** *if for all x, y in C with $x \neq y$ and all $0 < \lambda < 1$*

$$f\big(\lambda x + (1-\lambda)y\big) \leqslant \max\{f(x), f(y)\}.$$

- ***strictly quasiconvex*** *if for all x, y in C with $x \neq y$ and all $0 < \lambda < 1$*

$$f\big(\lambda x + (1-\lambda)y\big) < \max\{f(x), f(y)\}.$$

- ***explicitly quasiconvex*** *or* ***semistrictly quasiconvex*** *if it is quasiconvex and in addition, for all x, y in C with $x \neq y$ and all $0 < \lambda < 1$*

$$f(x) < f(y) \implies f\big(\lambda x + (1-\lambda)y\big) < \max\{f(x), f(y)\} = f(y).$$

There are other choices we could have made for the definition based on the next lemma.

287 Lemma *For a function $f\colon C \to \boldsymbol{R}$ on a convex set, the following are equivalent:*

1. *The function f is quasiconcave.*
2. *For each $\alpha \in \boldsymbol{R}$, the strict upper contour set $[f(x) > \alpha]$ is convex, but possibly empty.*
3. *For each $\alpha \in \boldsymbol{R}$, the upper contour set $[f(x) \geq \alpha]$ is convex, but possibly empty.*

Proof: (1) $\implies$ (2) If f is quasiconcave and x, y in C satisfy $f(x) > \alpha$ and $f(y) > \alpha$, then for each $0 \leqslant \lambda \leqslant 1$ we have

$$f\big(\lambda x + (1-\lambda)y\big) \geq \min\{f(x), f(y)\} > \alpha.$$

(2) $\implies$ (3) Note that

$$[f \geqslant \alpha] = \bigcap_{n=1}^{\infty} [f > \alpha - \tfrac{1}{n}],$$

and recall that the intersection of convex sets is convex.

(3) $\implies$ (1) If $[f \geqslant \alpha]$ is convex for each $\alpha \in \boldsymbol{R}$, then for $y, z \in C$ put $\alpha = \min\{f(y), f(z)\}$ and note that $f\big(\lambda y + (1-\lambda)z\big)$ belongs to $[f \geqslant \alpha]$ for each $0 \leqslant \lambda \leqslant 1$. ∎

288 Corollary *A concave function is quasiconcave. A convex function is quasiconvex.*

289 Lemma *A strictly quasiconcave function is also explicitly quasiconcave. Likewise a strictly quasiconvex function is also explicitly quasiconvex.*

Of course, not every quasiconcave function is concave.

290 Example (Explicit quasiconcavity) This example sheds some light on the definition of explicit quasiconcavity. Define $f\colon \boldsymbol{R} \to [0,1]$ by

$$f(x) = \begin{cases} 0 & x = 0 \\ 1 & x \neq 0. \end{cases}$$

If $f(x) > f(y)$, then $f\big(\lambda x + (1-\lambda)y\big) > f(y)$ for every $\lambda \in (0,1)$ (since $f(x) > f(y)$ implies $y = 0$). But f is not quasiconcave, as $\{x : f(x) \geqslant 1\}$ is not convex. □

291 Example (Sum of quasiconcave functions is not quasiconcave) Define f and g on the real line by $f(x) = x^+ = x \vee 0$ and $g(x) = x^- = -x \vee 0$. Then both f and g are quasiconcave, but the sum $(f+g)(x) = |x|$ is not quasiconcave. (Observe that both f and g are convex functions as well!) □

The next result has applications to production functions. (Cf. Jehle [85, Theorem 5.2.1, pp. 224–225] and Shephard [139, pp. 5–7].)

292 Theorem *Let $f\colon \boldsymbol{R}^{\mathrm{n}}_{+} \to \boldsymbol{R}_{+}$ be nonnegative, nondecreasing, quasiconcave, and positively homogeneous of degree k where $0 < k \leqslant 1$. Then f is concave.*

Proof: Let $x, y \in \boldsymbol{R}^{\mathrm{n}}$ and suppose first that $f(x) = \alpha > 0$ and $f(y) = \beta > 0$. (The case $\alpha = 0$ and/or $\beta = 0$ will be considered in a moment.) Then by homogeneity,

$$f\Big(\frac{x}{\alpha^{\frac{1}{k}}}\Big) = f\Big(\frac{y}{\beta^{\frac{1}{k}}}\Big) = 1$$

By quasiconcavity,

$$f\Big(\lambda \frac{x}{\alpha^{\frac{1}{k}}} + (1-\lambda)\frac{y}{\beta^{\frac{1}{k}}}\Big) \geqslant 1$$

for $0 \leqslant \lambda \leqslant 1$. So setting $\lambda = \frac{\alpha^{\frac{1}{k}}}{\alpha^{\frac{1}{k}} + \beta^{\frac{1}{k}}}$, we have

$$f\Big(\frac{x}{\alpha^{\frac{1}{k}} + \beta^{\frac{1}{k}}} + \frac{y}{\alpha^{\frac{1}{k}} + \beta^{\frac{1}{k}}}\Big) \geqslant 1.$$

By homogeneity,

$$f(x+y) \geqslant (\alpha^{\frac{1}{k}} + \beta^{\frac{1}{k}})^k = \left[f(x)^{\frac{1}{k}} + f(y)^{\frac{1}{k}}\right]^k. \tag{5.18}$$

Observe that since f is nonnegative and nondecreasing, (5.18) holds even if $f(x) = 0$ or $f(y) = 0$. Now replace x by μx and y by $(1-\mu)y$ in (5.18), where $0 \leqslant \mu \leqslant 1$, to get

$$\begin{aligned} f\big(\mu x + (1-\mu)y\big) &\geqslant \left[f(\mu x)^{\frac{1}{k}} + f\big((1-\mu)y\big)^{\frac{1}{k}}\right]^k \\ &= \left[\mu f(x)^{\frac{1}{k}} + (1-\mu) f(y)^{\frac{1}{k}}\right]^k \\ &\geqslant \mu\big(f(x)^{\frac{1}{k}}\big)^k + (1-\mu)\big(f(y)^{\frac{1}{k}}\big)^k \\ &= \mu f(x) + (1-\mu) f(y), \end{aligned}$$

where the last inequality follows from the concavity of $\gamma \mapsto \gamma^k$. Since x and y are arbitrary, f is concave. ∎

5.8 Quasiconcavity and Differentiability

Quasiconcavity has implications for derivatives.

293 Proposition *Let $C \subset \boldsymbol{R}^n$ be convex and let $f\colon C \to \boldsymbol{R}$ be quasi-concave. Let y belong to C and assume that f has a one-sided directional derivative $f'(x; y-x) = \lim_{\lambda \downarrow 0} \frac{f\left(x+\lambda(y-x)\right)-f(x)}{\lambda}$. Then*

Notation? Definition?

$$f(y) \geqslant f(x) \quad \Longrightarrow \quad f'(x; y-x) \geqslant 0.$$

In particular, if f is differentiable at x, then $f'(x) \cdot (y-x) \geqslant 0$ whenever $f(y) \geqslant f(x)$.

Proof: If $f(y) \geqslant f(x)$, then $f\big(x + \lambda(y-x)\big) = f\big((1-\lambda)x + \lambda y\big) \geqslant f(x)$ for $0 < \lambda \leqslant 1$ by quasiconcavity. Rearranging implies $\frac{f\left(x+\lambda(y-x)\right)-f(x)}{\lambda} \geqslant 0$ and taking limits gives the desired result. ∎

Converse???

294 Theorem *Let $C \subset \boldsymbol{R}^n$ be open and let $f\colon C \to \boldsymbol{R}$ be quasiconcave and twice-differentiable at $x \in C$. Then*

$$\sum_{i=1}^n \sum_{j=1}^n D_{i,j} f(x) v_i v_j \leqslant 0 \qquad \text{for any } v \text{ satisfying } f'(x) \cdot v = 0.$$

Proof: Pick $v \in \boldsymbol{R}^n$ and define $g(\lambda) = f(x + \lambda v)$. Then $g(0) = f(x)$, $g'(0) = f'(x) \cdot v$, and $g''(0) = \sum_{i=1}^n \sum_{j=1}^n D_{i,j} f(x) v_i v_j$. What we have to show is that if $g'(0) = 0$, then $g''(0) \leqslant 0$. Assume for the sake of contradiction that $g'(0) = 0$ and $g''(0) > 0$. Then by Theorem 78, g has a strict local minimum at zero. That is, for $\varepsilon > 0$ small enough, $f(x + \varepsilon v) > f(x)$ and $f(x - \varepsilon v) > f(x)$. But by quasiconcavity,

$$f(x) = f\big(\tfrac{1}{2}(x+\varepsilon v) + \tfrac{1}{2}(x-\varepsilon v)\big) \geqslant \min\{f(x+\varepsilon v), f(x-\varepsilon v)\} > f(x),$$

a contradiction. ∎

Converse???

5.9 Quasiconcavity and First Order Conditions

The following theorem and its proof may be found in Arrow and Enthoven [9].

295 Theorem (Arrow–Enthoven) *Let $f, g_1, \dots, g_m \colon \boldsymbol{R}^{\mathrm{n}}_+ \to \boldsymbol{R}$ be differentiable and quasiconcave. Suppose $x^* \in \boldsymbol{R}^{\mathrm{n}}_+$ and $\lambda^* \in \boldsymbol{R}^{\mathrm{m}}$ satisfy the constraints $g(x^*) \geqq 0$ and $x^* \geqq 0$ and the Karush–Kuhn–Tucker–Lagrange first order conditions:*

$$f'(x^*) + \sum_{j=1}^m \lambda_i^* g_j{}'(x^*) \leqq 0$$
$$x^* \cdot \left(f'(x^*) + \sum_{j=1}^m \lambda_i^* g_j{}'(x^*) \right) = 0$$
$$\lambda^* \geqq 0$$
$$\lambda^* \cdot g(x^*) = 0.$$

*Say that a variable x_j is **relevant** if it may take on a strictly positive value in the constraint set. That is, if there exists some $\hat{x} \geqq 0$ satisfying $\hat{x}_j > 0$ and $g(\hat{x}) \geqq 0$.*

Suppose one of the following conditions is satisfied:

1. *$D_{j_0} f(x^*) < 0$ for some relevant variable x_{j_0}.*
2. *$D_{j_1} f(x^*) > 0$ for some relevant variable x_{j_1}.*
3. *$f'(x^*) \neq 0$ and f is twice differentiable in a neighborhood of x^*.*
4. *f is concave.*

Then x^* maximizes $f(x)$ subject to the constraints $g(x) \geqq 0$ and $x \geqq 0$.

5.10 Value functions and envelope theorems

In this section we analyze parametrized maxima. Given a constrained maximizer x^* of $f(x,p)$ subject to the constraint $g(x,p)=0$, we denote the maximum value of f by $V(p)$. The function V is known as the **value function**. Theorems on the derivatives of the value function are called **envelope theorems**. Here's why.

Given a one-dimensional parametrized family of curves, $f_\alpha\colon [0,1] \to \boldsymbol{R}$, where α runs over some interval, a curve $h\colon [0,1] \to \boldsymbol{R}$ is the **envelope** of the family if each point on the curve h is tangent to one of the curves f_α and each curve f_α is tangent to h (see, e.g., Apostol [7, p. 342] for this definition). That is, for each α, there is some t and also for each t, there is some α, satisfying $f_\alpha(t) = h(t)$ and $f'_\alpha(t) = h'(t)$. For example, long-run cost curve is the envelope of the short-run cost curves, a result sometimes referred to as the "Wong–Viner Theorem."[2]

Consider now an unconstrained parametrized maximization problem. Let $x^*(p)$ be the value of the control variable x that maximizes $f(x,p)$, where p is our parameter of interest. For fixed x, the function

$$\varphi_x(p) = f(x,p)$$

defines a curve (or more generally a surface). The value function $V(p)$ satisfies

$$V(p) = f\big(x^*(p),p\big) = \max_x \varphi_x(p).$$

296 Informal Statement of the Envelope Theorem *Under appropriate conditions, the graph of the value function V is the envelope of the family of graphs of φ_x.*

To get a picture of this result, imagine a plot of the graph of f. It is the surface $z = f(x,p)$ in (x,p,z)-space. Orient the graph so that the x-axis is perpendicular to the page and the p-axis runs horizontally across the page, and the z-axis is vertical. The high points of the surface (minus perspective effects) determine the graph of the value function V. Here is an example:

297 Example Let

$$f(x,p) = p - (x-p)^2 + 1, \quad 0 \leqslant x,p \leqslant 2.$$

See Figure 5.5. Then given p, the maximizing x is given by $x^*(p) = p$, and $V(p) = p+1$. The side-view of this graph in Figure 5.6 shows that the high points do indeed lie on the line $z = 1+p$. For each x, the function φ_x is given by

$$\varphi_x(p) = p - (x-p)^2 + 1.$$

The graphs of these functions and of V are shown for selected values of x in Figure 5.7. Note that the graph of V is the envelope of the family of graphs φ_x. Consequently the slope of V is the slope of the φ_x to which it is tangent, that is,

$$V'(p) = \frac{\partial f}{\partial p}\Big|_{x=x^*(p)=p} = 1 + 2(x-p)\Big|_{x=p} = 1.$$

This last observation is an example of the Envelope Theorem. □

Most of the envelope theorems in these notes do not require differentiability with respect to the control variables, only with respect to the parameters (state variables). I am not aware of any statement of Theorem 298 in the literature, although the technique of proof is standard (cf. Myerson [116]). It provides a unifying approach to many related problems.

[2] According to Samuelson [136, p. 34], Jacob Viner asked his draftsman, one Mr. Wong, to draw the long run cost curve passing through the minimum of each short run cost curve, and tangent to it. Mr. Wong argued that this was impossible, and that the correct interpretation was that the long run curve was the envelope of the short run curves. See also Viner [157].

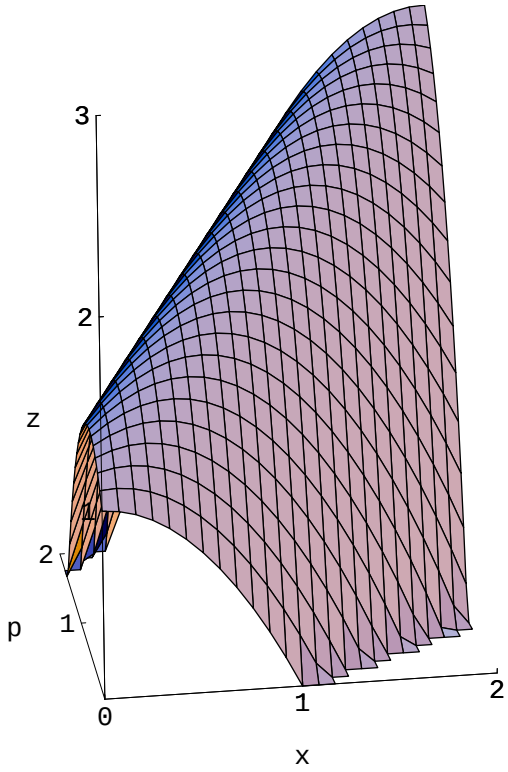

Figure 5.5. Graph of $f(x,p) = p - (x-p)^2 + 1$.

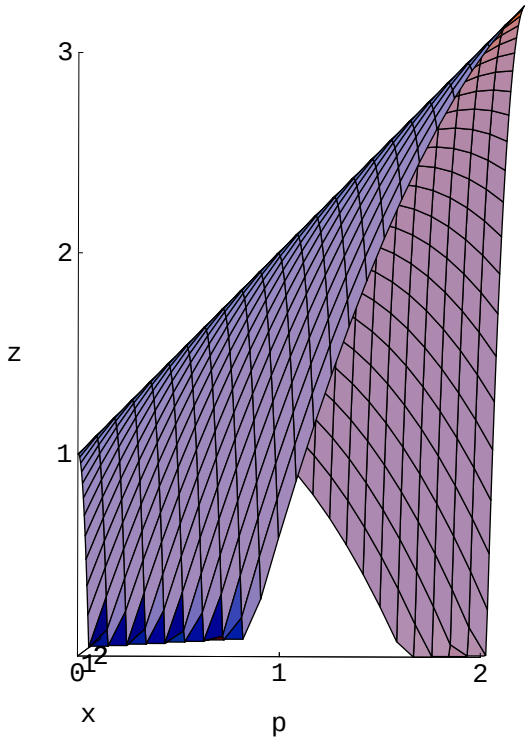

Figure 5.6. Graph of $f(x,p) = p - (x-p)^2 + 1$ viewed from the side.

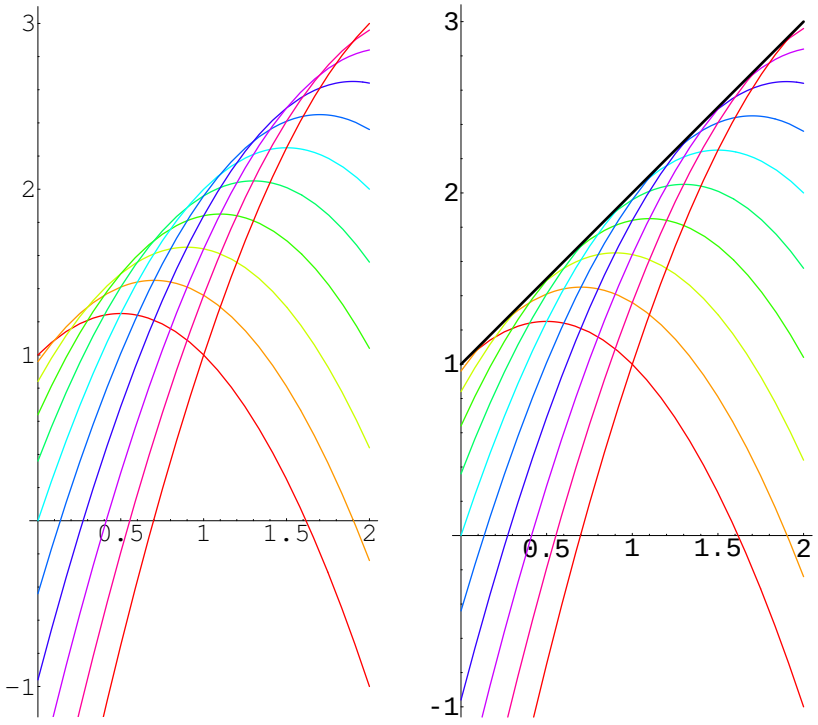

Figure 5.7. Graphs of φ_x for $x = 0, .2, \ldots, 2$ (left), and with the graph of $V(p) = p + 1$ as the envelope of the family $\{\varphi_x(p) \colon x \in [0, 2]\}$ (right), where $\varphi_x(p) = p - (x - p)^2 + 1 = f(x, p)$.

5.10.1 An envelope theorem for saddlepoints

298 Saddlepoint Envelope Theorem *Let X and Y be metric spaces, and let P be an open subset of R^n. Let*

$$L\colon X \times Y \times P \to R$$

and assume that the partial derivative $\frac{\partial L}{\partial p}$ exists and is jointly continuous on $X \times Y \times P$. For each p, let $\big(x^*(p), y^*(p)\big)$ be a saddlepoint of L in $X \times Y$. That is, for every $p \in P$

Does L need to be continuous?

$$L\big(x, y^*(p), p\big) \leqslant L\big(x^*(p), y^*(p), p\big) \leqslant L\big(x^*(p), y, p\big)$$

for all $x \in X$ and $y \in Y$. Set

$$V(p) = L\big(x^*(p), y^*(p), p\big).$$

Assume that x^* and y^* are continuous functions. Then V is continuously differentiable and

$$DV(p) = \frac{\partial}{\partial p} L\big(x^*(p), y^*(p), p\big).$$

Proof: Since P is finite dimensional, it suffices to show that V has continuous directional derivatives. Let h be a nonzero vector in R^n small enough so that $[p, p+h] \subset P$, where $[p, p+h] = \{p + th : 0 \leqslant t \leqslant 1\}$.

By definition,

$$V(p+h) - V(p) = L\big(x^*(p+h), y^*(p+h), p+h\big) - L\big(x^*(p), y^*(p), p\big).$$

Adding and subtracting a few terms that net out to zero, we have:

$$\begin{aligned} V(p+h) - V(p) = & \\ & L\big(x^*(p+h), y^*(p+h), p+h\big) - L\big(x^*(p+h), y^*(p), p+h\big) && (5.19)\\ + \quad & L\big(x^*(p+h), y^*(p), p+h\big) - L\big(x^*(p+h), y^*(p), p\big) && (5.20)\\ + \quad & L\big(x^*(p+h), y^*(p), p\big) - L\big(x^*(p), y^*(p), p\big) && (5.21) \end{aligned}$$

The saddle point property of (x^*, y^*) implies that terms (5.19) and (5.21) are nonpositive. Applying the Mean Value Theorem to term (5.20), we have:

$$\frac{\partial}{\partial p} L\big(x^*(p+h), y^*(p), p_1(h)\big) \cdot h \geqslant V(p+h) - V(p) \tag{5.22}$$

for some $p_1(h) \in [p, p+h]$.

Similarly:

$$\begin{aligned} V(p+h) - V(p) = & \\ & L\big(x^*(p+h), y^*(p+h), p+h\big) - L\big(x^*(p), y^*(p+h), p+h\big) \\ + \quad & L\big(x^*(p), y^*(p+h), p+h\big) - L\big(x^*(p), y^*(p+h), p\big) \\ + \quad & L(x^*(p), y^*(p+h), p) - L(x^*(p), y^*(p), p). \\ \geqslant \quad & \frac{\partial}{\partial p} L\big(x^*(p), y^*(p+h), p_2(h)\big) \cdot h && (5.23) \end{aligned}$$

for some $p_2(h) \in [p, p+h]$.

Combining (5.22) and (5.23) yields

$$\begin{aligned} \frac{\partial}{\partial p} L\big(x^*(p+h), y^*(p), p_1(h)\big) \cdot h \quad & \geqslant \quad V(p+h) - V(p) \\ & \geqslant \quad \frac{\partial}{\partial p} L\big(x^*(p), y^*(p+h), p_2(h)\big) \cdot h. \end{aligned}$$

Since $\frac{\partial}{\partial p}L$ is jointly continuous, replacing h by th, dividing by $\|th\|$ and letting $t \downarrow 0$ shows that the directional derivative of V in the direction h is:

$$\frac{\partial}{\partial p}L\big(x^*(p), y^*(p), p\big) \cdot h.$$

Thus

$$DV(p) = \frac{\partial}{\partial p}L\big(x^*(p), y^*(p), p\big).$$

∎

5.10.2 An envelope theorem for unconstrained maximization

The following corollary is extremely useful in the design of optimal revelation mechanisms, and indeed has been proven without statement many times over. It follows immediately from Theorem 298.

299 Corollary *Let X be a metric space and P an open subset of R^n. Let $w\colon X \times P \to \boldsymbol{R}$ and assume $\frac{\partial w}{\partial p}$ exists and is continuous in $X \times P$. For each $p \in P$, let $x^*(p)$ maximize $w(x,p)$ over X. Set*

$$V(p) = w\big(x^*(p), p\big).$$

Assume that $x^*\colon P \to X$ is a continuous function. Then V is continuously differentiable and

$$DV(p) = \frac{\partial w}{\partial p}\big(x^*(p), p\big).$$

Proof: Set $L(x,y,p) = w(x,p)$ and apply Theorem 298. ∎

5.10.3 Classical Envelope Theorem

300 Theorem *Let $X \subset \boldsymbol{R}^{\mathrm{n}}$ and $P \subset \boldsymbol{R}^{\ell}$ be open, and let $f, g_1, \dots, g_m\colon X \times P \to \boldsymbol{R}$ be C^1. For each $p \in P$, let $x^*(p)$ be an interior constrained local maximizer of $f(x,p)$ subject to $g(x,p) = 0$. Define the Lagrangean*

$$L(x, \lambda; p) = f(x,p) + \sum_{i=1}^{m} \lambda_i g_i(x,p),$$

and assume that the conclusion of the Lagrange Multiplier Theorem holds for each p, that is, Notation!!!! *there exist real numbers $\lambda_1^*(p), \dots, \lambda_m^*(p)$, such that the first order conditions*

$$\frac{\partial L\big(x^*(p), \lambda^*(p), p\big)}{\partial x} = f'_x\big(x^*(p), p\big) + \sum_{i=1}^{m} \lambda_i^*(p) g_{ix}'\big(x^*(p), p\big) = 0$$

are satisfied. Assume that $x^*\colon P \to X$ and $\lambda^*\colon P \to \boldsymbol{R}^{\mathrm{m}}$ are C^1. Set

$$V(p) = f\left(x^*(p), p\right).$$

Then V is C^1 and

$$\frac{\partial V(p)}{\partial p_j} = \frac{\partial L\big(x^*(p), \lambda^*(p), p\big)}{\partial p_j} = \frac{\partial f(x^*, p)}{\partial p_j} + \sum_{i=1}^{m} \lambda_i^*(p) \frac{\partial g_i(x^*, p)}{\partial p_j}.$$

Proof: Clearly V is C^1 as the composition of C^1 functions. Since x^* satisfies the constraints, we have

$$V(p) = f\left(x^*(p), p\right) = f\left(x^*(p), p\right) + \sum_{i=1}^{m} \lambda_i^*(p) g_i(x^*, p).$$

Therefore by the chain rule,

$$\begin{aligned}
\frac{\partial V(p)}{\partial p_j} &= \left(\sum_{k=1}^{n} \frac{\partial f(x^*,p)}{\partial x_k}\frac{\partial x^{*k}}{\partial p_j}\right) + \frac{\partial f(x^*,p)}{\partial p_j} \\
&\quad + \sum_{i=1}^{m}\left\{\frac{\partial \lambda_i^*(p)}{\partial p_j} g_i(x^*,p) + \lambda^*(p)\left[\left(\sum_{k=1}^{n}\frac{\partial g_i(x^*,p)}{\partial x_k}\frac{\partial x^{*k}}{\partial p_j}\right) + \frac{\partial g_i(x^*,p)}{\partial p_j}\right]\right\} \\
&= \frac{\partial f(x^*,p)}{\partial p_j} + \sum_{i=1}^{m}\lambda_i^*(p)\frac{\partial g_i(x^*,p)}{\partial p_j} \\
&\quad + \sum_{i=1}^{m}\frac{\partial \lambda_i^*(p)}{\partial p_j} g_i(x^*,p) \qquad (5.24) \\
&\quad + \sum_{k=1}^{n}\left(\frac{\partial f(x^*,p)}{\partial x_k} + \sum_{i=1}^{m}\lambda^*(p)\frac{\partial g_i(x^*,p)}{\partial x_k}\right)\frac{\partial x^{*k}}{\partial p_j}. \qquad (5.25)
\end{aligned}$$

The theorem now follows from the fact that both terms (5.24) and (5.25) are zero. Term (5.24) is zero since x^* satisfies the constraints, and term (5.25) is zero, since the first order conditions imply that each $\frac{\partial f(x^*,p)}{\partial x_k} + \sum_{i=1}^{m}\lambda^*(p)\frac{\partial g_i(x^*,p)}{\partial x_k} = 0$. ∎

5.10.4 Another Envelope Theorem

The previous theorem assume only that the conclusion of Lagrange Multiplier Theorem held. This version requires the assumptions of the Lagrange Multiplier Theorem to hold, but dispenses with the assumption that the multipliers are a C^1 function of the parameters. At the moment, there is an uncomfortable gap in the proof, so label it a conjecture.

301 Conjecture *Let $X \subset \boldsymbol{R}^{\mathrm{n}}$ and $P \subset \boldsymbol{R}^{\ell}$ be open, and let $f, g_1, \dots, g_m \colon X \times P \to \boldsymbol{R}$ be C^1. For each $p \in P$, let $x^*(p)$ be an interior constrained local maximizer of $f(x,p)$ subject to $g(x,p) = 0$. Assume that for each p, the gradients (with respect to x) g_{ix}' are linearly independent at $\big(x^*(p),p\big)$. Assume that $x^* \colon P \to X$ is C^1.*

Define the Lagrangean

$$L(x,\lambda;p) = f(x,p) + \sum_{i=1}^{m}\lambda_i g_i(x,p).$$

Then for each p there exist real numbers $\lambda_1^*(p), \dots, \lambda_m^*(p)$, such that the first order conditions Notation!!!!

$$\frac{\partial L\big(x^*(p),\lambda^*(p),p\big)}{\partial x} = f_x'\big(x^*(p),p\big) + \sum_{i=1}^{m}\lambda_i^*(p) g_i{}_x'\big(x^*(p),p\big) = 0$$

are satisfied. Set

$$V(p) = f\left(x^*(p),p\right).$$

Then V is C^1 and

$$\frac{\partial V(p)}{\partial p_j} = \frac{\partial L\big(x^*(p),\lambda^*(p),p\big)}{\partial p_j} = \frac{\partial f(x^*,p)}{\partial p_j} + \sum_{i=1}^{m}\lambda_i^*(p)\frac{\partial g_i(x^*,p)}{\partial p_j}.$$

The main idea of this proof appears in many places, e.g., Silberberg [141, 142], Clarke et. al. [38], and Diamond and McFadden [42] who attribute it to Gorman.

Proof: As in the proof of Theorem 300, the function V is clearly C^1. Now observe that we can embed our maximization in the family of problems

$$\text{maximize } f(x,p) \text{ subject to } g(x,p) - \alpha = 0 \qquad (\mathrm{P}(\alpha))$$

where α ranges over a neighborhood 0 in $\boldsymbol{R}^{\mathrm{m}}$. The first thing we have to show is that for each α, there is some (x,p) satisfying $g(x,p)+\alpha = 0$. We have already assumed that for each p there is some x_p satisfying $g(x_p,p) = 0$. Indeed $x_p = x^*(p)$ works. Now consider the function

$$h_p(x,\alpha) = g(x,p) - \alpha.$$

By hypothesis $h_p(x_p,0) = 0$. The Jacobian of h with respect to x is just the Jacobian of g, which is of full rank by our linear independence hypothesis. Therefore by the Implicit Function Theorem 117, there is a neighborhood U of 0 in $\boldsymbol{R}^{\mathrm{m}}$ such that $\alpha \in U$ implies the existence of some $\hat{x}_p(\alpha)$ such that $h_p\big(x\hat{x}_p(\alpha),\alpha\big) = 0$. Thus each problem $\mathrm{P}(\alpha)$ is feasible.

One gap in the proof is to show that in fact each $\mathrm{P}(\alpha)$m has an optimal solution. Assume for now that this is so, and let $x^*(p,\alpha) = \hat{x}_p(\alpha)$ be the optimum. Another gap is to show that x^* is a differentiable function of both p and α. Modify the definition of V so that

$$V(p,\alpha) = f\big(x^*(p,\alpha),p\big).$$

Now for any x and p, if we set $\alpha = g(x,p)$, then x satisfies $g(x,p)+\alpha = 0$. In particular, the value $f(x,p)$ is less than or equal to the optimal value $V\big(p,g(x,p)\big)$. In other words,

$$h(x,p) = V\big(p,g(x,p)\big) - f(x,p) \geqslant 0,$$

and is equal to zero for $x = x^*\big(p,g(x,p)\big)$. Thus minima of h occur whenever $x = x^*(p,0)$. The first order conditions for this minimum are that

$$\begin{aligned} \frac{\partial h}{\partial x_j} &= 0 \qquad j = 1,\dots,n,\\ \frac{\partial h}{\partial p_i} &= 0 \qquad i = 1,\dots,m. \end{aligned}$$

The first group of first order conditions imply

$$\frac{\partial h}{\partial x_j} = \sum_{k=1}^{m} \frac{\partial V}{\partial \alpha_i}\frac{\partial g_k}{\partial x_j} - \frac{\partial f}{\partial x_j} = 0,$$

which tells us that

$$\lambda_i^* = -\frac{\partial V}{\partial \alpha_i}$$

are the desired Lagrange multipliers. The second group of first order conditions imply

$$\frac{\partial h}{\partial p_i} = \frac{\partial V}{\partial p_i} + \sum_{k=1}^{m} \frac{\partial V}{\partial \alpha_i}\frac{\partial g_k}{\partial p_i} - \frac{\partial f}{\partial p_i} = 0,$$

or using the Lagrange multipliers defined above

$$\frac{\partial V}{\partial p_i} = \frac{\partial f}{\partial p_i} + \sum_{k=1}^{m} \lambda^* \frac{\partial g_k}{\partial p_i},$$

where of course the partials are evaluated at the optimizers. ∎

Section 6

Quadratic forms

In this section, subscripts typically denote the coordinates of a vector and superscripts typically are used to enumerate vectors or indicate exponents of a scalar.

6.1 Eigenvectors, eigenvalues, and characteristic roots

Let A be an $n \times n$ matrix. A scalar λ is an **eigenvalue** of A if there is a nonzero vector x in $\boldsymbol{R}^{\mathrm{n}}$ such that $Ax = \lambda x$. The vector x is called an **eigenvector** of A associated with λ. Since we are dealing only with real matrices, every eigenvalue is a real number. Note that the vector 0 is by definition *not* an eigenvector of A. Consequently at exactly one eigenvalue can be associated to an eigenvector (as $\alpha x = Ax = \lambda x$ and $x \neq 0$ imply $\lambda = \alpha$). While the vector 0 is never an eigenvector, the scalar 0 may be an eigenvalue. Indeed 0 is the eigenvalue associated with any nonzero vector in the null space of A.

It is not hard to show that eigenvectors corresponding to distinct eigenvalues must be linearly independent, so that every $n \times n$ matrix has at most n eigenvalues. But there are matrices with no eigenvalues. For instance, the 2×2 matrix $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ has no eigenvalues. (In order to satisfy $Ax = \lambda x$ we must have $\lambda x_1 = -x_2$ and $\lambda x_2 = x_1$. This cannot happen for any nonzero x and real λ.) On the other hand, the identity matrix has an eigenvalue 1, associated with every nonzero vector. We shall see below that symmetric matrices always have real eigenvalues. Now any eigenvalue must be associated with many eigenvectors, for if x is an eigenvector associated with λ, so is any nonzero scalar multiple of x. More generally, a linear combination of eigenvectors corresponding to an eigenvalue is also an eigenvector corresponding to the same eigenvalue (provided the linear combination does not equal the zero vector). The span of the set of eigenvectors associated with the eigenvalue λ is called the **eigenspace** of A corresponding to λ. Every nonzero vector in the eigenspace is an eigenvector associated with λ. The dimension of the eigenspace is sometimes called the **multiplicity** of λ.

Recall that the **characteristic polynomial** f of a square matrix A is defined by $f(\lambda) = \det(\lambda I - A)$. Roots of this polynomial, even complex roots, are called **characteristic roots** of A.[1]

302 Lemma *Every eigenvalue of a matrix is a characteristic root, and every real characteristic root is an eigenvalue.*

[1]Some authors, notably Carathéodory [35, p. 178] and Gantmacher [59, pp. 69–70], write the characteristic polynomial as $\det(A - \lambda I)$. For an $n \times n$ matrix this differs from the more common definition by a factor of -1^n, and so has the same roots. Interestingly, Gantmacher changes to the more common definition twelve pages later on page 82.

Proof: To see this note that if λ is an eigenvalue with eigenvector $x \neq 0$, then $(\lambda I - A)x = \lambda x - Ax = 0$, so $(\lambda I - A)$ is singular, so $\det(\lambda I - A) = 0$. That is, λ is a characteristic root of A.

Conversely, if $\det(\lambda I - A) = 0$, then there is some nonzero x with $(\lambda I - A)x = 0$, or $Ax = \lambda x$. ∎

303 Lemma *The determinant of a square matrix is the product of its characteristic roots.*

Proof: (cf. Apostol [8, p. 106]) Let A be an $n \times n$ square matrix and let f be its characteristic polynomial. Then $f(0) = \det(-A) = (-1)^n \det A$. On the other hand, we can factor f as

$$f(\lambda) = (\lambda - \lambda_1) \cdots (\lambda - \lambda_n)$$

where $\lambda_1, \ldots, \lambda_n$ are its characteristic roots. Thus $f(0) = (-1)^n \lambda_1 \cdots \lambda_n$. ∎

6.2 Quadratic forms

Let A be an $n \times n$ symmetric matrix, and let x be an n-vector. Then $x \cdot Ax$ is a scalar,

$$x \cdot Ax = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j.$$

The mapping $Q\colon x \mapsto x \cdot Ax$ is the **quadratic form** defined by A.[2]

The quadratic form $Q(x) = x \cdot Ax$ is positively homogeneous of degree 2 in x, so it is completely determined by its values on the unit sphere $S = \{x \in \boldsymbol{R}^{\mathrm{n}} : x \cdot x = 1\}$. Moreover Q is a continuous function, so it achieves a maximum and minimum on the unit sphere, which is compact. Every maximizer turns out to be an eigenvector of A, and the value of the maximum is its corresponding eigenvalue. This eigenvalue also turns out to be the Lagrange Multiplier for the constraint that the maximizer lies on the sphere. We can say even more, for if we restrict attention to the subspace orthogonal to the eigenvector and look for a maximizer, we get another eigenvector and eigenvalue. We can repeat this procedure until we have found them all:

304 Proposition (Extrema of quadratic forms on the sphere) *Let A be an $n \times n$ symmetric matrix. Define the vectors $x^1, \ldots, x^n$ recursively so that x^{k+1} maximizes the quadratic form $Q(x) = x \cdot Ax$ over $S_k = S \cap M_{k\perp}$, where S is the unit sphere in $\boldsymbol{R}^{\mathrm{n}}$, and M_k denotes the span of $x^1, \ldots, x^k$, with $M_0 = \{0\}$. Then each x^k, $k = 1, \ldots, n$ is an eigenvector of A, and $\lambda_k = Q(x^k)$ is its corresponding eigenvalue.*

Note that by construction $S_{k+1} \subset S_k$, so $\lambda_1 \geqslant \cdots \geqslant \lambda_n$. Indeed λ_n is the minimum of Q on the unit sphere. The sequence of eigenvalues and eigenvectors can be obtained in reverse order by minimizing rather than maximizing.

Proof: Let $S = \{x \in \boldsymbol{R}^{\mathrm{n}} : x \cdot x = 1\}$ denote the unit sphere. Set $M_0 = \{0\}$ and define $S_0 = S \cap M_{0\perp}$, where $M_{0\perp}$ is the orthogonal complement of M_0. (This is a little silly, since $M_{0\perp} = \boldsymbol{R}^{\mathrm{n}}$, so $S_0 = S$, but the reason will become apparent soon.) Since Q is continuous, it has a maximizer on S_0, which is compact. (This maximizer cannot be unique, since $Q(-x) = Q(x)$, and indeed if $A = I$, then Q is constant on S.) Fix a maximizer x^1 of Q over S_0.

[2] The term *form* refers to a polynomial function in several variables where each term in the polynomial has the same degree. (The *degree* of the term is the sum of the exponents. For example, in the expression $xyz+x^2y+xz+z$, the first two terms have degree three, the third term has degree two and the last one has degree one. It is thus not a form.) This is most often encountered in the phrases *linear form* (each term has degree one) or *quadratic form* (each term has degree two). Tom Apostol once remarked at a cocktail party that mathematicians evidently don't know the difference between form and function.

Proceed recursively for $k = 1, \dots, n-1$. Let M_k denote the span of $x^1, \dots, x^k$, and set $S_k = S \cap M_{k\perp}$. Let x^{k+1} maximize Q over S_k. By construction, $x^{k+1} \in M_{k\perp}$, so the x^k's are orthogonal, indeed orthonormal.

The quadratic form $Q(x) = x \cdot Ax$ is continuously differentiable and $Q'(x) = 2Ax$. Since x^1 maximizes Q on $S = S_0$, it maximizes Q subject to the constraint $1 - x \cdot x = 0$. Now the gradient of this constraint function is $-2x$, which is clearly nonzero (hence linearly independent) on S. It is a nuisance to have these 2s popping up, so let us agree to maximize $\frac{1}{2}x \cdot Ax$ subject $\frac{1}{2}(1 - x \cdot x) = 0$ instead. Therefore by the Lagrange Multiplier Theorem 270, there exists λ_1 satisfying

$$Ax^1 - \lambda_1 x^1 = 0.$$

This obviously implies that the Lagrange multiplier λ_1 is an eigenvalue of A and x^1 is a corresponding eigenvector. Further, it is the value of the maximum:

$$Q(x^1) = x^1 \cdot Ax^1 = \lambda_1 x^1 \cdot x^1 = \lambda_1,$$

since $x^1 \cdot x^1 = 1$.

We now proceed by induction on k. Let $x^1, \dots, x^n$ be recursively defined as above and assume that for $i = 1, \dots, k$, each x^i is an eigenvector of A and that $\lambda_i = Q(x^i)$ is its corresponding eigenvalue. We wish to show that x^{k+1} is an eigenvector of A and $\lambda_{k+1} = Q(x^{k+1})$ is its corresponding eigenvalue.

By hypothesis, x^{k+1} maximizes $\frac{1}{2}Q(x)$ subject to the constraints $\frac{1}{2}(1 - x \cdot x) = 0$, $x \cdot x^1 = 0, \dots, x \cdot x^k = 0$. The gradients of these constraint functions are $-x$ and $x^1, \dots, x^k$ respectively. By construction, $x^1, \dots, x^{k+1}$ are orthonormal, so at x^{k+1} the constraint gradients are linearly independent. Therefore by the Lagrange Multiplier Theorem there exist multipliers λ_{k+1} and $\mu_1, \dots, \mu_k$ satisfying

$$Ax^{k+1} - \lambda_{k+1}x^{k+1} + \mu_1 x^1 + \cdots + \mu_k x^k = 0. \tag{6.1}$$

Therefore

$$Q(x^{k+1}) = x^{k+1} \cdot Ax^{k+1} = \lambda_{k+1}x^{k+1} \cdot x^{k+1} - \mu_1 x^{k+1} \cdot x^1 - \cdots - \mu_k x^{k+1} \cdot x^k = \lambda_{k+1},$$

since $x^1, \dots, x^{k+1}$ are orthonormal. That is, the multiplier λ_{k+1} is the maximum value of Q over S_k.

By the induction hypothesis, $Ax^i = \lambda_i x^i$ for $i = 1, \dots, k$. Then since A is symmetric,

$$x^i \cdot Ax^{k+1} = x^{k+1} \cdot Ax^i = x^{k+1} \cdot (\lambda_i x^i) = 0, \quad i = 1, \dots, k.$$

That is, $x^{k+1} \in M_{k\perp}$, so $Ax^{k+1} - \lambda_{k+1}x^{k+1} \in M_{k\perp}$, so equation (6.1) implies

$$Ax^{k+1} - \lambda_{k+1}x^{k+1} = 0 \quad \text{and} \quad \mu_1 x^1 + \cdots + \mu_k x^k = 0.$$

(Recall that if $x \perp y$ and $x+y = 0$, then $x = 0$ and $y = 0$. *Hint*: This follows from $(x+y)\cdot(x+y) = x \cdot x + 2x \cdot y + y \cdot y = x \cdot x + y \cdot y$ when $x \cdot y = 0$.) We conclude therefore that $Ax^{k+1} = \lambda_{k+1}x^{k+1}$, so that x^{k+1} is an eigenvector of A and λ_{k+1} is the corresponding eigenvalue.

Note that since the first order conditions for a minimum are the same as for a maximum that by minimizing rather than maximizing, we can construct the sequence of eigenvectors in the reverse order by minimizing. The values of the minima are once again eigenvalues of A. Since an eigenvector can have only one associated eigenvalue, the sequence of eigenvalues is reproduced in reverse order as well. ∎

305 Corollary *Let A be an $n \times n$ symmetric matrix. Then $\boldsymbol{R}^{\mathrm{n}}$ has an orthonormal basis of eigenvectors of A. There are n eigenvalues, counting each eigenvalue at its multiplicity.*

6.3 Definite and semidefinite quadratic forms

A symmetric matrix A (or its associated quadratic form) is called

- **positive definite** if $x \cdot Ax > 0$ for all nonzero x.
- **negative definite** if $x \cdot Ax < 0$ for all nonzero x.
- **positive semidefinite** if $x \cdot Ax \geqslant 0$ for all x.
- **negative semidefinite** if $x \cdot Ax \leqslant 0$ for all x.

If A is not symmetric, then $\frac{A+A'}{2}$ is symmetric (where A' denotes the transpose of A) and $x \cdot Ax = x \cdot (\frac{A+A'}{2})x$ for any x, so we do not lose much applicability by this assumption. Some authors use the term **quasi-(semi)definite** when they do not wish to impose symmetry.

306 Proposition (Eigenvalues and definiteness) *The symmetric matrix A is*

1. *positive definite if and only if all its eigenvalues are strictly positive.*
2. *negative definite if and only if all its eigenvalues are strictly negative.*
3. *positive semidefinite if and only if all its eigenvalues are nonnegative.*
4. *negative semidefinite if and only if all its eigenvalues are nonpositive.*

First proof: I'll prove only the first statement. All the eigenvalues are strictly positive if and only if the least eigenvalue is strictly positive if and only the quadratic form is strictly positive on the unit sphere (Proposition 304) if and only the quadratic form is positive definite (homogeneity). ∎

Second proof: Let $\{x^1, \dots, x^n\}$ be an orthonormal basis for $\boldsymbol{R}^{\mathrm{n}}$ consisting of eigenvectors of A. (See Corollary 305.) Let λ_i be the eigenvalue corresponding to x^i. That is,

$$Ax^i = \lambda_i x^i.$$

Writing $y = \sum_{i=1}^n \alpha_i x^i$, we see that

$$y \cdot Ay = \sum_{i=1}^n \sum_{j=1}^n (\alpha_i x^i) \cdot A(\alpha_j x^j) = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \lambda_j x^i \cdot x^j = \sum_{k=1}^n (\alpha_k)^2 \lambda_k,$$

where the last equality follows from the orthonormality of $\{x^1, \dots, x^n\}$. All the statements above follow from this equation and the fact that $(\alpha_k)^2 \geqslant 0$ for all k. ∎

307 Proposition (Definiteness of the inverse) *If A is positive definite (negative definite), then A^{-1} exists and is also positive definite (negative definite).*

Proof: First off, how do we know the inverse of A exists? Suppose $Ax = 0$. Then $x \cdot Ax = x \cdot 0 = 0$. Since A is positive definite, we see that $x = 0$. Therefore A is invertible. Here are two proofs of the proposition.

First proof. Since $(Ax = \lambda x) \implies (x = \lambda A^{-1}x) \implies (A^{-1}x = \frac{1}{\lambda}x)$, the eigenvalues of A and A^{-1} are reciprocals, so they must have the same sign. Apply Proposition 306.

Second proof.

$$x \cdot A^{-1}x = y \cdot Ay \quad \text{where} \quad y = A^{-1}x.$$

∎

The proof of the next theorem may be found in Debreu [40] or Gantmacher [59, pp. 306–308].

308 Theorem *For a symmetric matrix A:*

1. *A is positive definite if and only if all its NW principal minors are strictly positive.*
2. *A is negative definite if and only if all its k^{th}-order NW principal minors have sign $(-1)^k$.*
3. *A is positive semidefinite if and only if all its principal minors are nonnegative.*
4. *A is negative semidefinite if and only if all its k^{th}-order principal minors have sign $(-1)^k$ or 0.*

Proof: We start with the necessity of the conditions on the minors.

First note that every principal submatrix of a matrix A inherits its definiteness. To see this let $I \subset \{1, \dots, n\}$ be the (nonempty) set of indices of rows and columns for the submatrix. Let x be any nonzero vector with $x_k = 0$ for $k \notin I$. Then

$$x \cdot Ax = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j = \sum_{i \in I} \sum_{j \in I} a_{ij} x_i x_j,$$

so the quadratic form defined by the submatrix cannot have a different sign from the quadratic form defined by A.

By Proposition 306, if a matrix is positive definite, all its eigenvalues are positive, so by Lemma 303 its determinant must be positive, as the product of the eigenvalues. Thus every principal submatrix of a positive definite matrix has a strictly positive determinant. Similarly, every principal submatrix of a positive semidefinite matrix has a nonnegative determinant.

The results for negative (semi)definiteness stem from the observation that a matrix A is negative (semi)definite if and only if $-A$ is positive (semi)definite, and that the determinant of a k^{th} order submatrix of $-A$ is $(-1)^k$ times the corresponding subdeterminant of A.

The sufficiency part is harder. To see why such a result might be true, consider first the case $n = 2$. Then, completing the square, we get

$$\begin{aligned} x \cdot Ax &= a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 \\ &= a_{11}\big(x_1 + \frac{a_{12}}{a_{11}}\, x_2\big)^2 + \frac{a_{11}a_{22} - a_{12}^2}{a_{11}}\, x_2^2 \\ &= D_1 y_1^2 + \frac{D_2}{D_1}\, y_2^2, \end{aligned}$$

where

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & \frac{a_{12}}{a_{11}} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

$D_1 = a_{11}$, the determinant of the 1×1 NW principal minor of A, and $D_2 = \det A$, the determinant of the 2×2 NW principal minor. In this case it is easy to see that $D_1 > 0$ and $D_2 > 0$ imply that A is positive definite.

Lagrange noticed that this technique could be generalized. That is, if $D_1 \neq 0, \dots, D_n \neq 0$ there is always a nonsingular upper triangular matrix U (with 1s on the main diagonal), so that

$$x \cdot Ax = \sum_{i=1}^{n} \frac{D_i}{D_{i-1}} y_i^2,$$

where $y = Ux$, $D_0 = 1$, and D_i is the determinant of the $i \times i$ NW principal minor of A. Given this decomposition, known as Jacobi's formula, it is easy to see why the conditions $D_1 > 0, \dots, D_n > 0$ guarantee that A is positive definite. The matrix U is computed by using Gaussian elimination on A. For details, see, e.g., Gantmacher [59, pp. 33–41, 300–302]. This proves parts (1) and (2).

To prove parts (3) and (4), we use the fact that if A has rank k, then there is a permutation matrix P so that $\hat{A} = P'AP$ satisfies $\hat{D}_1 > 0, \dots, \hat{D}_k > 0$ and $\hat{D}_{k+1} = \cdots = \hat{D}_n = 0$. Furthermore, each $\hat{D}_i$ is some $i \times i$ minor subdeterminant of the original A. Thus there is an upper triangular matrix $\hat{U}$ such that

$$x \cdot Ax = x \cdot P\hat{A}P'x = P'x \cdot \hat{A}P'x = \sum_{i=1}^{k} \frac{\hat{D}_i}{\hat{D}_{i-1}} y_i^2,$$

where $y = \hat{U}P'x$. Again see Gantmacher [59, pp. 33—41] for details. ∎

6.4 Quadratic forms under constraint

Proposition 304 considered the extrema of a quadratic form restricted to a subspace orthogonal to a set of eigenvectors. In this section we will generalize this problem to morte general subspaces.

A matrix A is **positive definite under the orthogonality constraints $b^1, \dots, b^m$** if it is symmetric and

$$x \cdot Ax > 0 \quad \text{for all} \quad x \neq 0 \quad \text{satisfying} \quad b^i \cdot x = 0, \quad i = 1, \dots, m.$$

The notions of negative definiteness and semidefiniteness under constraint are defined in the obvious analogous way. Notice that we can replace $b^1, \dots, b^m$ by any basis for the span of $b^1, \dots, b^m$, so without loss of generality we may assume that $b^1, \dots, b^m$ are linearly independent, or even orthonormal.

309 Theorem *Suppose A is an $n \times n$ symmetric matrix that is negative definite under constraint for the linearly independent constraint vectors $b^1, \dots, b^m$. That is, $x \cdot Ax < 0$ for all nonzero x satisfying $B'x = 0$, where B is the $n \times m$ matrix whose j^{th} column is b^j. Then:*

1. *The matrix*

$$\left[\begin{array}{c|c} A & B \\ \hline B' & 0 \end{array}\right]$$

is invertible.

2. *Write*

$$\left[\begin{array}{c|c} A & B \\ \hline B' & 0 \end{array}\right]^{-1} = \left[\begin{array}{c|c} C & D \\ \hline D' & E \end{array}\right].$$

Then C is negative semidefinite of rank $n - m$, with $Cx = 0$ if and only if x is a linear combination of $b^1, \dots, b^m$.

Proof: (cf. Samuelson [136, pp. 378–379], Quirk [128, pp. 22–25], and Diewert and Woodland [44, Appendix, Lemma 3])

(1) Observe that

$$\left[\begin{array}{c|c} x' & z' \end{array}\right] \left[\begin{array}{c|c} A & B \\ \hline B' & 0 \end{array}\right] \left[\begin{array}{c} x \\ \hline z \end{array}\right] = \left[\begin{array}{c|c} x' & z' \end{array}\right] \left[\begin{array}{c} Ax + Bz \\ \hline B'x \end{array}\right] = x'Ax + x'Bz + z'B'x.$$

Now suppose $\left[\begin{array}{c|c} A & B \\ \hline B' & 0 \end{array}\right] \left[\begin{array}{c} x \\ \hline z \end{array}\right] = 0$. Then

$$Ax + Bz = 0 \tag{6.2}$$

and

$$B'x = 0, \tag{6.3}$$

so

$$0 = \left[\, x' \,\middle|\, z' \,\right] \left[\begin{array}{c|c} A & B \\ \hline B' & 0 \end{array}\right] \left[\begin{array}{c} x \\ \hline z \end{array}\right] = x \cdot Ax. \tag{6.4}$$

Since A is definite under constraint, (6.3) and (6.4) imply that $x = 0$. Thus (6.2) implies $Bz = 0$. Since B has linearly independent columns, this implies $z = 0$.

Thus $\left[\begin{array}{c|c} A & B \\ \hline B' & 0 \end{array}\right] \left[\begin{array}{c} x \\ \hline z \end{array}\right] = 0$ implies $\left[\begin{array}{c} x \\ \hline z \end{array}\right] = 0$. Therefore $\left[\begin{array}{c|c} A & B \\ \hline B' & 0 \end{array}\right]$ is invertible.

(2) So write

$$\left[\begin{array}{c|c} A & B \\ \hline B' & 0 \end{array}\right]^{-1} = \left[\begin{array}{c|c} C & D \\ \hline D' & E \end{array}\right]$$

and observe that

$$\left[\begin{array}{c|c} C & D \\ \hline D' & E \end{array}\right] \left[\begin{array}{c|c} A & B \\ \hline B' & 0 \end{array}\right] = \left[\begin{array}{c|c} I_n & 0 \\ \hline 0 & I_m \end{array}\right].$$

Expanding this yields

$$\begin{aligned} CA + DB' &= I && (6.5)\\ CB &= 0 && (6.6)\\ D'A + EB' &= 0 && (6.7)\\ D'B &= I && (6.8) \end{aligned}$$

Now premultiply (6.5) by x' and postmultiply by Cx to get

$$x'CACx + x'D \underbrace{B'C}_{\substack{=0 \\ \text{by } 6.6}} x = x'Cx.$$

Now again by (6.6), we have $B'Cx = 0$, so Cx is orthogonal to each column of B. That is, Cx satisfies the constraints, so $x \cdot CACx \leqslant 0$ with < 0 if $Cx \neq 0$. Thus $x \cdot Cx \leqslant 0$ with < 0 if $Cx \neq 0$. That is, C is negative semidefinite.

To see that C has rank $n - m$, we show that $Cx = 0$ if and only if x is a linear combination of the columns of the m independent columns of B. Equation (6.6) already implies that $x = Bz$ implies $Cx = 0$. Now suppose $Cx = 0$. Premultiply (6.5) by x' to get

$$x'CA + x'DB' = x'.$$

Thus $x'C = 0$ implies $(x'D)B' = x'$, or $x = Bz$, where $z = Dx$.

Thus $Cx = 0$ if and only if x is a linear combination of the columns of B. Therefore the null space of C has dimension equal to the rank of B, which is m, so the rank of C equals $n - m$. ∎

The next result is a partial converse to Theorem 309.

310 Theorem *Suppose A is an $n \times n$ symmetric matrix that is negative semidefinite under constraint for the linearly independent constraint vectors $b^1, \dots, b^m$. That is, $x \cdot Ax \leqslant 0$ for all nonzero x satisfying $B'x = 0$, where B is the $n \times m$ matrix whose j^{th} column is b^j. Suppose also that the matrix*

$$\left[\begin{array}{c|c} A & B \\ \hline B' & 0 \end{array}\right]$$

is invertible. Then A is actually negative definite under constraint. That is, $x \cdot Ax < 0$ for all nonzero x satisfying $B'x = 0$.

Note that if B has full rank, then there are no nonzero x with $B'x = 0$. In that case the theorem is trivially true.

Proof: Suppose

$$\bar{x} \cdot A\bar{x} = 0 \quad \text{and} \quad B'\bar{x} = 0.$$

Then $\bar{x}$ maximizes the quadratic form $\frac{1}{2}x \cdot Ax$ subject to the constraints $B'x = 0$. Since the columns of B are independent, the constraint qualification is satisfied, so by the Lagrange Multiplier Theorem 270, there is a vector $\lambda \in \boldsymbol{R}^{\mathrm{m}}$ satisfying the first order conditions:

$$A\bar{x} + B\lambda = 0.$$

Thus

$$\left[\begin{array}{c|c} A & B \\ \hline B' & 0 \end{array}\right] \left[\begin{array}{c} \bar{x} \\ \hline \lambda \end{array}\right] = \left[\begin{array}{c} A\bar{x} + B\lambda \\ \hline B'\bar{x} \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline 0 \end{array}\right].$$

Since $\left[\begin{array}{c|c} A & B \\ \hline B' & 0 \end{array}\right]$ is invertible, we see that $\bar{x} = 0$ (and $\lambda = 0$). Thus $B'x = 0$ and $x \neq 0$ imply $x \cdot Ax < 0$. ∎

6.4.1 Determinantal conditions

Now consider the problem of maximizing the quadratic form $Q(x) = \frac{1}{2}x \cdot Ax$ over the unit sphere subject to the constraints that $b^1 \cdot x = 0, \dots, b^m \cdot x = 0$, where $m < n$ and $b^1, \dots, b^m$ are linearly independent. Note that the constraint set is a closed subset of the unit sphere (hence compact) so a maximizer exists. Let x^* be such a constrained maximizer. (It is not unique as at least $-x^*$ is also a maximizer.)

We want to apply the Lagrange Multiplier Theorem, so we need verify the linear independence of the gradients of the constraints. Write the unit sphere constraint as $\frac{1}{2}(1 - x \cdot x) = 0$ to avoid unsightly fractions. The gradient of this constraint is $-x$, and the gradient of $b^i \cdot x$ is b^i. Thus we need to show that x^*, b^1, ..., b^m are linearly independent. Since x^* is on the unit sphere, it is nonzero, and since it is orthogonal to each b^i, it cannot be a linear combination of them, so the set of gradients is independent.

Thus there exist Lagrange multipliers λ^*, $\mu_1^*, \dots, \mu_m^*$ satisfying the first-order conditions

$$Ax^* - \lambda^* x + \mu_1^* b^1 + \cdots + \mu_m^* b^m = 0. \tag{6.9}$$

Premultiplying equation (6.9) by x^*, and using the fact that x^* is orthogonal to each b^i, we get

$$Q(x^*) = x^* \cdot Ax^* = \lambda^* x^* \cdot x^* = \lambda^*.$$

That is, the Lagrange multiplier λ^* is the maximum value of Q.

We can combine equation (6.9) with the orthogonality conditions in one big matrix equation:

$$\left[\begin{array}{c|c} A - \lambda^* I & B \\ \hline B' & 0 \end{array}\right] \left[\begin{array}{c} x^* \\ \hline \mu^* \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline 0 \end{array}\right],$$

where B is the matrix whose columns are $b^1, \dots, b^m$ and μ^* is the vector with components $\mu_1^*, \dots, \mu_m^*$. Since x^* is nonzero (it lies on the unit sphere), the matrix $\left[\begin{array}{c|c} A - \lambda^* I & B \\ \hline B' & 0 \end{array}\right]$ must be singular, so

$$\det \left[\begin{array}{c|c} A - \lambda^* I & B \\ \hline B' & 0 \end{array}\right] = 0.$$

The next result is can be found in Hancock [71, p. 106], who attributes it to Zajaczkowski [163].

311 Proposition (Hancock) *Let A be an $n \times n$ symmetric matrix and let $\{b^1, \dots, b^m\}$ be linearly independent. Let*

$$f(\lambda) = \det \left[\begin{array}{c|c} A - \lambda I & B \\ \hline B' & 0 \end{array}\right].$$

If all the coefficients of f have the same sign, then A is negative semidefinite under constraint.

If the coefficients of f alternate in sign, then A is positive semidefinite under constraint. (Here we must consider the zero coefficients to be alternating in sign.)

If in addition, $f(0) = \det \left[\begin{array}{c|c} A & B \\ \hline B' & 0 \end{array}\right] \neq 0$, then A is actually definite under constraint.

Proof: Even without resort to Descartes' infamous Rule of Signs the following fact is easy to see: *If all the nonzero coefficients of a nonzero polynomial f have the same sign, then f has no strictly positive roots.* For if all the coefficients of a polynomial f are nonnegative, then $f(0) \geqslant 0$ and f is nondecreasing on $(0, \infty)$, so it has no positive roots. Likewise if all the coefficients are nonpositive, then $f(0) \leqslant 0$ and f is nonincreasing on $(0, \infty)$, so it has no positive roots. Trivially if $f(0) \neq 0$, then 0 is not a root.

From the discussion preceding the proposition, λ^*, the maximum value of $x'Ax$ on the unit sphere, is a root of f. If the coefficients of f do not change sign, then $\lambda^* \leqslant 0$. That is, A is negative semidefinite under the constraints, and is actually definite if $f(0) \neq 0$.

The results on positive (semi)definiteness follow from the fact that λ^* is a negative root of $f(\lambda)$ if and only if $-\lambda^*$ is a positive root of $f(-\lambda)$. ∎

The problem with applying this result is that he does not provide a simple formula for the coefficients.

6.4.2 Bordered matrices and quadratic forms

If A is some kind of definite under constraint, we define matrices of the form

$$\left[\begin{array}{ccccccc} a_{11} & \dots & a_{1r} & b_1^1 & \dots & b_1^m \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{r,1} & \dots & a_{rr} & b_r^1 & \dots & b_r^m \\ b_1^1 & \dots & b_r^1 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ b_1^m & \dots & b_r^m & 0 & \dots & 0 \end{array}\right]$$

to be ***r*-th order bordered minors of *A***. Note that the r refers to the number of rows and columns from A. The actual r-th order minor has $m + r$ rows and columns, where m is the number of constraint vectors. The proof of the following result may be found in Debreu [40, Theorems 4 and 5] or Mann [106]. Note that Mann errs in the statement of part 2. A proof may also be found sketched in Samuelson [136, pp. 376–378].

312 Theorem *Let A be an $n \times n$ symmetric matrix and let $\{b^1, \dots, b^m\}$ be linearly independent.*

1. *A is positive definite under constraint subject to $b^1 \dots, b^m$ if and only if*

$$(-1)^m \left|\begin{array}{ccccccc} a_{11} & \dots & a_{1r} & b_1^1 & \dots & b_1^m \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{r,1} & \dots & a_{rr} & b_r^1 & \dots & b_r^m \\ b_1^1 & \dots & b_r^1 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ b_1^m & \dots & b_r^m & 0 & \dots & 0 \end{array}\right| > 0$$

for $r = m{+}1, \dots, n$. That is, if and only if every r^{th}-order NW bordered principal minor has sign $(-1)^m$ for $r > m$.

2. A is negative definite under constraint subject to $b^1 \dots, b^m$ if and only if

$$(-1)^r \begin{vmatrix} a_{11} & \dots & a_{1r} & b_1^1 & \dots & b_1^m \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{r,1} & \dots & a_{rr} & b_r^1 & \dots & b_r^m \\ b_1^1 & \dots & b_r^1 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ b_1^m & \dots & b_r^m & 0 & \dots & 0 \end{vmatrix} > 0$$

for $r = m{+}1, \dots, n$. That is, if and only if every r^{th}-order NW bordered principal minor has sign $(-1)^r$ for $r > m$.

Note that for positive definiteness under constraint all the NW bordered principal minors of order greater than m have the same sign, the sign depending on the number of constraints. For negative definiteness the NW bordered principal minors alternate in sign. For the case of one constraint ($m = 1$) if A is positive definite under constraint these minors are negative. Again with one constraint if A is negative definite under constraint, then the minors of even order are positive and of odd order are negative.

To see how to derive statement (2) from statement (1), observe that A is negative definite under constraint if and only if $-A$ is positive definite under constraint, which by statement (1) is equivalent to

$$(-1)^m \begin{vmatrix} -a_{11} & \dots & -a_{1r} & b_1^1 & \dots & b_1^m \\ \vdots & & \vdots & \vdots & & \vdots \\ -a_{r,1} & \dots & -a_{rr} & b_r^1 & \dots & b_r^m \\ b_1^1 & \dots & b_r^1 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ b_1^m & \dots & b_r^m & 0 & \dots & 0 \end{vmatrix} > 0$$

for $r = m{+}1, \dots, n$. But

$$(-1)^m \begin{vmatrix} -a_{11} & \dots & -a_{1r} & b_1^1 & \dots & b_1^m \\ \vdots & & \vdots & \vdots & & \vdots \\ -a_{r,1} & \dots & -a_{rr} & b_r^1 & \dots & b_r^m \\ b_1^1 & \dots & b_r^1 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ b_1^m & \dots & b_r^m & 0 & \dots & 0 \end{vmatrix} = (-1)^{m+r} \begin{vmatrix} a_{11} & \dots & a_{1r} & -b_1^1 & \dots & -b_1^m \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{r,1} & \dots & a_{rr} & -b_r^1 & \dots & -b_r^m \\ b_1^1 & \dots & b_r^1 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ b_1^m & \dots & b_r^m & 0 & \dots & 0 \end{vmatrix}$$

$$= (-1)^{2m+r} \begin{vmatrix} a_{11} & \dots & a_{1r} & b_1^1 & \dots & b_1^m \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{r,1} & \dots & a_{rr} & b_r^1 & \dots & b_r^m \\ b_1^1 & \dots & b_r^1 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ b_1^m & \dots & b_r^m & 0 & \dots & 0 \end{vmatrix},$$

and $(-1)^{2m+r} = (-1)^r$, so (2) follows.

Answers to selected exercises

Chapter 2

Exercise 30 (p. 15)

Prove that if both the epigraph and hypograph of a function are closed, then the graph is closed.

Proof: This is easy, as

$$\operatorname{graph} f = \{(x,\alpha) : \alpha = f(x)\} = \{(x,\alpha) : \alpha \geqslant f(x)\} \bigcap \{(x,\alpha) : \alpha \leqslant f(x)\}$$

and the intersection closed sets is closed. ∎

Give an example to show that the converse is not true.

Example: Define $f\colon \boldsymbol{R} \to \boldsymbol{R}$ by

$$f(x) = \begin{cases} \frac{1}{x} & x \neq 0 \\ 0 & x = 0. \end{cases}$$

Clearly the graph is closed, but neither the hypograph nor epigraph is closed. ∎

Chapter 4

Exercise 130 (Affine combinations, p. 69)

Let A be an affine subspace. Prove that if $x_1, \ldots, x_n$ belong to A and $\lambda_1, \ldots, \lambda_n$ are scalars that sum to one, then $\lambda_1 x_1 + \cdots + \lambda_n x_n$ also belongs to A.

Proof: The proof is by induction on n. The result is clearly true for $n = 1$, since in that case we must have $\lambda_1 = 1$. So let $n > 1$, and assume that the result holds for $n-1$, and let $x_1, \ldots, x_n$ belong to A and $\lambda_1 + \cdots + \lambda_n = 1$. By renumbering if necessary we may assume that $\gamma = \lambda_1 + \cdots + \lambda_{n-1} = 1 - \lambda_n \neq 0$. (This is because if this sum is zero, then $\lambda_n = 1$, and if this is true for all subsets of size $n-1$, then $\lambda_i = 1$ for each i, which contradicts their summing to 1.)

Then $\sum_{i=1}^{n-1} \lambda_i/\gamma = 1$, so by the induction hypothesis, the point $y = \sum_{i=1}^{n-1} \frac{\lambda_i}{\gamma} x_i$ belongs to A. Therefore the affine combination

$$\gamma y + \lambda_n x_n = \sum_{i=1}^{n} \lambda_i x_i$$

belongs to A. ∎

Exercise 131 (Affine subspaces, p. 69)

Let X be a vector space. Prove the following.

1. Let M be a linear subspace of X and let a be a vector in X. Then $M+a = \{x+a : x \in M\}$ is an affine subspace of X.

   Proof: Let α and β be scalars with $\alpha + \beta = 1$, and let x and y belong to $M + a$. We wish to show that $\alpha x + \beta y$ belongs to $M + a$.

   Let $x = x' + a$ and $y = y' + a$, where x' and y' belong to M. Then

   $$\alpha x + \beta y = \alpha(x' + a) + \beta(y' + a) = \alpha x' + \beta y' + (\alpha + \beta)a = \alpha x' + \beta y' + a.$$

   Since M is a linear subspace $\alpha x' + \beta y'$ belong to M, so $\alpha x + \beta y$ belongs to $M + a$. ∎

2. Let A be an affine subspace of X, and let a and b belong to A.

   (a) The set $A - a = \{x - a : x \in A\}$ is a linear subspace of X.

   Proof: Let x and y belong to $A-a$, and let α and β be scalars, and let $\gamma = 1-(\alpha+\beta)$. We wish to show that $\alpha x + \beta y$ belongs to $A - a$. Now $x + a$, $y + a$, and a belong to A, which is closed under affine combinations, so

   $$\alpha(x + a) + \beta(y + a) + \gamma a = \alpha x + \beta y + a \quad \text{belongs to } A.$$

   That is, $\alpha x + \beta y$ belongs to $A - a$. ∎

   (b) $A - a = A - b$.

   Proof: Let x belong to A. We may write

   $$x - a = (x - b) - (a - b).$$

   But both $x - b$ and $a - b$ belong to $A - b$, which is a linear subspace, so $x - a$ belongs to $A - b$. This proves that $A - a \subset A - b$. Interchanging the roles of a and b proves the reverse inclusion, so $A - a = A - b$. ∎

3. Consequently, for every affine subspace A, there is a linear subspace M such that $A = M + a$.

   Proof: Set $M = A - a$. Then M is a linear subspace and $A = M = a$. ∎

4. If M and N are linear subspaces such $A = M + a = N + b$ for some $a, b \in A$, then $M = N$. This subspace is called the **linear subspace parallel to A**.

   Proof: We know that $A-a = A-b$, but $A-a = (M+a)-a = M$ and $A-b = (N+b)-b = N$, so $M = N$. ∎

5. An affine subspace is a linear subspace if and only it contains 0.

   Proof: If an affine subspace is also a linear subspace, it must contain 0. Conversely, if A is an affine subspace that contains 0, then by the above $A - 0 = A$ is a linear subspace. ∎

6. Let M denote the unique linear subspace parallel to A. For $x \in M$ and $y \in A$ together imply that $x + y \in A$.

   Proof: If $y \in A$, then by the above, $A - y = M$. Since $x \in M$, we have $x \in A - y$, or $x + y \in A$. ∎

Exercise 133 (Affine functions, p. 70)

A function f on an affine subspace A is affine if and only it is of the form $f(x) = g(x-a) - \gamma$, where a belongs to A and g is linear on the linear subspace $A - a$. Moreover, g is independent of the choice of a in A.

In particular, when $A = X$, then an affine function f on X can be written as $f(x) = g(x) - \gamma$, where g is linear on X and $\gamma = -f(0)$.

Proof: ($\Longleftarrow$) Fix some point a in A, a scalar γ, and a linear function g on $A - a$. Define f on A by $f(x) = g(x-a) - \gamma$. We need to show that f is affine. So let α and β be scalars satisfying $\alpha + \beta = 1$. Then

$$\begin{aligned} f(\alpha x + \beta y) &= f\big(\alpha(x-a) + \beta(y-a) + a\big) \\ &= g\big(\alpha(x-a) + \beta(y-a)\big) - \gamma \\ &= \alpha g(x-a) + \beta g(y-a) - \gamma \\ &= \alpha\big[g(x-a) - \gamma\big] + \beta\big[g(y-a) - \gamma\big] \\ &= \alpha f(x) + \beta f(y), \end{aligned}$$

which proves that f is affine.

($\Longrightarrow$) Let f be an affine function on A. Pick some point a in A, and define g on $A - a$ by $g(x) = f(x+a) - f(a)$. (Thus $f(y) = g(y-a) - \gamma$, where $\gamma = -f(a)$ for $y \in A$). We need to show that g is linear on $A - a$.

Let x belong to $A - a$. We first show that

$$g(\alpha x) = \alpha g(x) \qquad \text{for all } \alpha. \tag{1}$$

To see this, write

$$\begin{aligned} g(\alpha x) &= g\big(\alpha x + (1-\alpha)0\big) \\ &= f\big(\alpha x + (1-\alpha)0 + a\big) - f(a) \\ &= f\big(\alpha(x+a) + (1-\alpha)a\big) - f(a) \\ &= \alpha f(x+a) + (1-\alpha)f(a) - f(a) \\ &= \alpha\big[f(x+a) - f(a)\big] + (1-\alpha)\big[f(a) - f(a)\big] \\ &= \alpha g(x). \end{aligned}$$

Next we show that g is additive, that is,

$$g(x+y) = g(x) + g(y) \qquad \text{for all } x, y \in A - a. \tag{2}$$

Now

$$\begin{aligned} g(x+y) &= g\big(2[\tfrac{1}{2}x + \tfrac{1}{2}y]\big) \\ &= 2g\big(\tfrac{1}{2}x + \tfrac{1}{2}y\big) \qquad \text{by (1)} \\ &= 2\left\{f\big([\tfrac{1}{2}x + \tfrac{1}{2}y] + a\big) - f(a)\right\} \qquad \text{defn. of } g \\ &= 2\left\{f\big(\tfrac{1}{2}[x+a] + \tfrac{1}{2}[y+a]\big) - f(a)\right\} \\ &= 2\left\{\tfrac{1}{2}f(x+a) + \tfrac{1}{2}f(y+a) - f(a)\right\} \qquad \text{since } f \text{ is affine} \\ &= 2\left\{\tfrac{1}{2}\big[f(x+a) - f(a)\big] + \tfrac{1}{2}\big[f(y+a) - f(a)\big]\right\} \\ &= \big[f(x+a) - f(a)\big] + \big[f(y+a) - f(a)\big] \\ &= g(x) + g(y). \end{aligned}$$

This shows that g is linear. To see that g does not depend on the choice of a, let a and b belong to A and let x belong to the linear subspace $M = A - a = A - b$. We want to show that $f(x+a) - f(a) = f(x+b) - f(b)$.

Now $\frac{1}{2}x$ also belongs to the linear subspace M and $\frac{1}{2}a + \frac{1}{2}b$ belongs to the affine subspace A. Recall that for any point $z \in M$ and $y \in A$, the point $z + y$ belongs to A. In particular,

$$\tfrac{1}{2}x + \tfrac{1}{2}a + \tfrac{1}{2}b \in A, \quad x + a \in A, \quad x + b \in A.$$

Now

$$\begin{aligned}
f(\tfrac{1}{2}x + \tfrac{1}{2}a + \tfrac{1}{2}b) &= f(\tfrac{1}{2}x + \tfrac{1}{2}b + \tfrac{1}{2}a)\\
f\big(\tfrac{1}{2}[x + a] + \tfrac{1}{2}b\big) &= f(\tfrac{1}{2}[x + b] + \tfrac{1}{2}a)\\
\tfrac{1}{2}f(x + a) + \tfrac{1}{2}f(b) &= \tfrac{1}{2}f(x + b) + \tfrac{1}{2}f(a)\\
f(x + a) + f(b) &= f(x + b) + f(a)\\
f(x + a) - f(a) &= f(x + b) - f(b)
\end{aligned}$$

which shows that the linear function g is independent of the choice of a or b in A.

The last paragraph follows by taking $a = 0$ when $A = X$. ∎

Exercise 127 (p. 67)

1. The sum of two convex sets is convex.

   Let $S = A + B = \{x + y : x \in A,\ y \in B\}$. Let u, v belong to S where $u = x_1 + y_1$, $v = x_2 + y_2$, each $x_i \in A$, $y_i \in B$. Then

   $$\lambda u + (1-\lambda)v = \lambda(x_1 + y_1) + (1-\lambda)(x_2 + y_2) = \underbrace{\big(\lambda x_1 + (1-\lambda)y_1\big)}_{\in A} + \underbrace{\big(\lambda x_2 + (1-\lambda)y_2\big)}_{\in B} \in S.$$

2. Scalar multiples of convex sets are convex.

   Obvious, as $\lambda\alpha x + (1-\lambda)\alpha y = \alpha\big(\lambda x + (1-\lambda)y\big)$.

3. A set C is convex if and only if

   $$\alpha C + \beta C = (\alpha + \beta)C$$

   for all nonnegative scalars α and β.

   ($\Longleftarrow$) Assume C is convex, and let $\alpha, \beta \geqslant 0$ be given. If $\alpha = \beta = 0$, the claim reduces to $\{0\} = \{0\}$, which is true. So assume $\alpha + \beta > 0$.

   First we show that $(\alpha+\beta)C \subset \alpha C + \beta C$. This uses no convexity. Let y belong to $(\alpha+\beta)C$. Then $y = (\alpha + \beta)x$, where $x \in C$. But $(\alpha + \beta)x = \alpha x + \beta x \in \alpha C + \beta C$.

   Next we show $\alpha C + \beta C \subset (\alpha + \beta)C$. Let y belong to $\alpha C + \beta C$, that is, $y = \alpha u + \beta v$, where $u, v \in C$. Dividing by the nonzero quantity $\alpha + \beta$ gives

   $$\frac{1}{\alpha + \beta}y = \frac{\alpha}{\alpha + \beta}u + \frac{\beta}{\alpha + \beta}v \in C,$$

   where the set membership follows from the convexity of C. Multiplying by $\alpha + \beta$ gives $y \in (\alpha + \beta)C$.

   ($\Longrightarrow$) This is easy: Assume $\alpha C + \beta C = (\alpha + \beta)C$ for all nonnegative scalars α and β. Let $\lambda \in [0, 1]$, and let $x, y \in C$. Then

   $$\lambda x + (1-\lambda)y \in \lambda C + (1-\lambda)C = \big(\lambda + (1-\lambda)\big)C = C.$$

4. The intersection of an arbitrary family of convex sets is convex.

   Let A be a set and for each $\alpha \in A$, let C_α be a convex set. Let $x, y \in \bigcap_\alpha C_\alpha$. Since each C_α is convex, $\lambda x + (1-\lambda)y \in C_\alpha$ too. But then $\lambda x + (1-\lambda)y \in \bigcap_\alpha C_\alpha$.

5. A convex set C contains every convex combination of its members.

   The proof is by induction. The definition of convexity is that the set includes every convex combination of two of its members. So assume that C includes every convex combination of n or fewer members, and let $x = \sum_{i=1}^{n+1} \lambda_i x_i$ be a convex combination of $n+1$ members. If any $\lambda_i = 0$, then we have a combination of no more thn n members, and so are finished. Therefore we may assume every $\lambda_i > 0$. Let $\alpha = \sum_{i=1}^{n} \lambda_i$. Then

   $$\sum_{i=1}^{n+1} \lambda_i x_i = \alpha \sum_{i=1}^{n} (\lambda_i/\alpha) x_i + (1-\alpha) x_{n+1}.$$

   But $y = \sum_{i=1}^{n} (\lambda_i/\alpha) x_i$ is a convex combination, and so belongs to C by the induction hypothesis. Thus $x = \alpha y + (1-\alpha) x_{n+1}$ is a convex combination that belongs to C.

6. The convex hull of a set A is the smallest (with respect to inclusion) convex set that includes A.

   Recall that

   $$\operatorname{co} A = \Big\{ \sum_{i=1}^{m} \lambda_i x_i : \text{where } m \geqslant 1 \text{ and each } x_i \in A, \lambda_i \geq 0, \text{ and } \sum_{i=1}^{m} \lambda_i = 1 \Big\}.$$

   To show that this is the smallest convex set that includes A, we need to show three things. (i) $\operatorname{co} A$ includes A. (ii) $\operatorname{co} A$ is convex. (iii) If C is convex and includes A, then C also includes $\operatorname{co} A$. Note that this also proves that $\operatorname{co} A$ is the intersection of all convex sets that include A.

   (i) is easy: Just take $m = 1$, so $\lambda_1 = 1$, and (iii) follows from the definition of convexity. So it remains to show that $\operatorname{co} A$ is actually convex. This too is easy. Let $x = \sum_{i=1}^{n} \alpha_i x_i$ and $y = \sum_{j=1}^{m} \gamma_j y_j$ belong to $\operatorname{co} A$, where each $x_i, y_j \in A$, etc. Let $0 \leqslant \lambda \leqslant 1$. Then

   $$\lambda x + (1-\lambda) y = \sum_{i=1}^{n} \lambda \alpha_i x_i + \sum_{j=1}^{m} (1-\lambda) \gamma_j y_j.$$

   Since each $\lambda\alpha_i, (1-\lambda)\gamma_j \geqslant 0$ and their sum is $\lambda + 1 - \lambda = 1$, and each $x_i, y_j \in A$, we see that $\lambda x + (1-\lambda) y$ belongs to $\operatorname{co} A$, so it is convex.

7. The interior and closure of a convex set are also convex.

   Let C be convex, and let C° denote its interior, and $\overline{C}$ its closure. Recall that C° is open, and indeed is the largest open set included in C. Let $0 < \lambda < 1$, and observe that $\lambda C^\circ + (1-\lambda) C^\circ$ is open.[3] Since C is convex, $\lambda C^\circ + (1-\lambda) C^\circ \subset C$. Since C° is the largest open set included in C we must have$\lambda C^\circ + (1-\lambda) C^\circ \subset C^\circ$. But this shows that C° is convex.

   To see that $\overline{C}$ is convex, let $x, y \in \overline{C}$, $x_n \to x$, $y_n \to y$, each $x_n, y_n \in C$, and let $0 \leqslant \lambda \leqslant 1$. Then $\lambda x_n + (1-\lambda) y_n \in C$, and $\lambda x_n + (1-\lambda) y_n \to \lambda x + (1-\lambda) y$, so $\lambda x + (1-\lambda) y \in \overline{C}$.

Exercise 135 (p. 70)

1. The sum of concave functions is concave.

   $$\begin{aligned} (f+g)\big(\lambda x + (1-\lambda) y\big) &= f\big(\lambda x + (1-\lambda) y\big) + g\big(\lambda x + (1-\lambda) y\big) \\ &\leqslant \lambda f(x) + (1-\lambda) f(y) + \lambda g(x) + (1-\lambda) g(y) \\ &= \lambda (f+g)(x) + (1-\lambda)(f+g)(y). \end{aligned}$$

[3] Indeed if G is open, then $G + A$ is open for any set A, for if $x = u + v$ with $u \in G$ and $v \in A$, then $G + v$ is an open neighborhood of x included in $G + A$. Also if G is open and $\lambda \neq 0$, then λG is open, for if $B_\varepsilon(x)$ is included in G, then $B_{|\lambda|\varepsilon}(x)$ is included in λG.

2. A nonnegative multiple of a concave function is concave.

   If $f\big(\lambda x + (1-\lambda)y\big) \geqslant \lambda f(x) + (1-\lambda)f(y)$, since multiplying by positive α preserves inequalities, we also have $\alpha f\big(\lambda x + (1-\lambda)y\big) \geqslant \lambda\alpha f(x) + (1-\lambda)\alpha f(y)$ for $\alpha > 0$. The case $\alpha = 0$ is obviously true.

3. The pointwise limit of a sequence of concave functions is concave.

   By concavity $f_n\big(\lambda x + (1-\lambda)y\big) \geqslant \lambda f_n(x) + (1-\lambda)f_n(y)$ for each n, so by elementary properties of limits of real sequences, $\lim_n f_n\big(\lambda x + (1-\lambda)y\big) \geqslant \lambda \lim_n f_n(x) + (1-\lambda)\lim_n f_n(y)$.

4. The pointwise infimum of a family of concave functions is concave.

   Let $g = \inf_\alpha f_\alpha$. Then

$$\begin{aligned}
\lambda g(x) + (1-\lambda)g(y) &= \inf_\alpha \lambda f_\alpha(x) + \inf_\beta (1-\lambda) f_\beta(y) \\
&\geqslant \inf_\gamma \lambda f_\gamma(x) + (1-\lambda) f_\gamma(y) \quad \text{(inf over smaller set)} \\
&\geqslant \inf_\gamma f_\gamma\big(\lambda x + (1-\lambda)y\big) \quad \text{(by concavity of } f_\gamma\text{)} \\
&= g\big(\lambda x + (1-\lambda)y\big).
\end{aligned}$$

5. A function is both concave and convex if and only if it is affine.

   This is just Exercise 133.

Exercise 159 (p. 78)

Let A and B be disjoint nonempty convex subsets of $\boldsymbol{R}^{\mathrm{n}}$ and suppose nonzero p in $\boldsymbol{R}^{\mathrm{n}}$ properly separates A and B with $p \cdot A \geqslant p \cdot B$.

1. If A is a linear subspace, then p annihilates A. That is, $p \cdot x = 0$ for every x in A.
2. If A is a cone, then $p \cdot x \geqslant 0$ for every x in A.
3. If B is a cone, then $p \cdot x \leqslant 0$ for every x in B.
4. If A includes a set of the form $x + \boldsymbol{R}^{\mathrm{n}}_{++}$, then $p > 0$.
5. If B includes a set of the form $x - \boldsymbol{R}^{\mathrm{n}}_{++}$, then $p > 0$.

Proof: The proofs of all these are more or less the same, so I shall just prove (4). Since p is nonzero by hypothesis, it suffice to show that $p \geqq 0$. Suppose by way of contradiction that $p_i < 0$ for some i. Note that $te^i + \varepsilon\mathbf{1}$ belongs to $\boldsymbol{R}^{\mathrm{m}}_{++}$ for every $t, \varepsilon > 0$. Now $p \cdot (x + te^i + \varepsilon\mathbf{1}) = p \cdot x + tp_i + \varepsilon p \cdot \mathbf{1}$. By letting $t \to \infty$ and $\varepsilon \downarrow 0$ we see that $p \cdot x + tp_i + \varepsilon p \cdot \mathbf{1} \downarrow -\infty$, which contradicts $p \cdot (x + te^i + \varepsilon\mathbf{1}) \geqslant p \cdot y$ for any y in B. Therefore $p > 0$. ∎

Exercise 200 (p. 93)

Prove Lemma 199: Let f be a real-valued function defined on some interval I of $\boldsymbol{R}$. If f is concave, then for every $x < y < z$ in I,

$$\frac{f(y) - f(x)}{y - x} \geqslant \frac{f(z) - f(x)}{z - x} \geqslant \frac{f(z) - f(y)}{z - y}.$$

Conversely, if one of the (three) inequalities is satisfied for every $x < y < z$ in I, then f is concave.

Equivalently,

$\dfrac{f(z) - f(x)}{z - x}$ is decreasing in both x and z over $\{(x, z) : x < z\}$ if and only f is concave.

Proof: ($\Longrightarrow$) Assume f is concave. Since $x < y < z$ we can write y as a convex combination of x and z, namely

$$y = \frac{z-y}{z-x}x + \frac{y-x}{z-x}z.$$

By concavity

$$f(y) \geqslant \frac{z-y}{z-x}f(x) + \frac{y-x}{z-x}f(z). \tag{$*$}$$

Subtracting $f(x)$ from both sides gives

$$f(y) - f(x) \geqslant \frac{x-y}{z-x}f(x) + \frac{y-x}{z-x}f(z).$$

Dividing by $y - x > 0$ gives

$$\frac{f(y)-f(x)}{y-x} \geqslant \frac{-1}{z-x}f(x) + \frac{1}{z-x}f(z) = \frac{f(z)-f(x)}{z-x}.$$

Similarly, subtracting $f(z)$ from both sides of $(*)$ gives

$$f(y) - f(z) \geqslant \frac{z-y}{z-x}f(x) + \frac{y-z}{z-x}f(z).$$

Dividing by $y - z < 0$ gives

$$\frac{f(z)-f(y)}{z-y} = \frac{f(y)-f(z)}{y-z} \leqslant \frac{-1}{z-x}f(x) + \frac{1}{z-x}f(z) = \frac{f(z)-f(x)}{z-x}.$$

Combining these inequalities completes this part of the proof.

($\Longleftarrow$) It suffices to show that $f\big(\alpha x + (1-\alpha)z\big) \geqslant \alpha f(x) + (1-\alpha)f(z)$ whenever $x < z$ belong to I and $0 < \alpha < 1$. (The cases $x = z$, or $\alpha = 0$, or $\alpha = 1$ take care of themselves.) Define $y = \alpha x + (1-\alpha)z$, so $x < y < z$, and note that $\alpha = \dfrac{z-y}{z-x}$ and $1-\alpha = \dfrac{y-x}{z-x}$. There are three cases to consider:

Case 1. The outer inequality

$$\frac{f(y)-f(x)}{y-x} \geqslant \frac{f(z)-f(y)}{z-y}$$

is satisfied. Multiplying by $y - x = (1-\alpha)(z-x) > 0$, we have

$$f(y) - f(x) \geqslant (1-\alpha)\frac{z-x}{z-y}\big(f(z)-f(y)\big).$$

Multiplying by $\dfrac{z-x}{z-y} = \alpha > 0$ gives

$$\alpha\big(f(y)-f(x)\big) \geqslant (1-\alpha)\big(f(z)-f(y)\big)$$

and regrouping we get

$$f(y) \geqslant \alpha f(x) + (1-\alpha)f(z).$$

Case 2. The left hand inequality

$$\frac{f(y)-f(x)}{y-x} \geqslant \frac{f(z)-f(x)}{z-x}$$

is satisfied. Multiplying by $y - x = (1-\alpha)(z-x) > 0$ gives

$$f(y) - f(x) \geqslant (1-\alpha)\big(f(z)-f(x)\big)$$

and regrouping we get

$$f(y) \geqslant \alpha f(x) + (1-\alpha)f(z).$$

Case 3. The right hand inequality

$$\frac{f(z)-f(x)}{z-x} \geqslant \frac{f(z)-f(y)}{z-y}$$

is satisfied. If you don't trust me by now, you should be able to figure this one out, but here it is anyhow. Multiply by $z-y = \alpha(z-x)$ to get

$$\alpha\big(f(z)-f(x)\big) \geqslant f(z)-f(y),$$

and rearrange to get

$$f(y) \geqslant \alpha f(x) + (1-\alpha)f(z).$$

∎

Exercise 203 (p. 94)

Prove Corollary 202.

Proof: Recall that

$$\Delta^2_{v,w} f(x) = f(x+w+v) - f(x+w) - \big(f(x+v) - f(x)\big).$$

There are four cases.

Case 1: $v, w > 0$. By Corollary 201 with $x_1 = x$, $y_1 = x+v$, $x_2 = x+w$, and $y_2 = x+w+v$, we have

$$\frac{f(x+v)-f(x)}{v} \geqslant \frac{f(x+w+v)-f(x+w)}{v},$$

so

$$\frac{f(x+w+v)-f(x+w)}{v} - \frac{f(x+v)-f(x)}{v} \leqslant 0,$$

so multiplying by $v^2 w > 0$ gives the desired conclusion.

Case 2: $v < 0$, $w > 0$. Use Corollary 201 with $x_1 = x+v$, $y_1 = x$, $x_2 = x+w+v$, and $y_2 = x+w$ to get

$$\frac{f(x)-f(x+v)}{-v} \geqslant \frac{f(x+w)-f(x+w+v)}{-v}$$

so

$$\frac{f(x+w)-f(x+w+v)}{-v} - \frac{f(x)-f(x+v)}{-v} \leqslant 0,$$

and rearrange.

Case 3: $v > 0$, $w < 0$. Use $\Delta^2_{v,w} f(x) = \Delta^2_{w,v} f(x)$ and interchange the roles of v and w in case (2).

Case 4: $v, w < 0$. Use Corollary 201 with $x_1 = x+w+v$, $y_1 = x+w$, $x_2 = x+v$, and $y_2 = x$,

$$\frac{f(x+w)-f(x+w+v)}{-v} \geqslant \frac{f(x)-f(x+v)}{-v},$$

or

$$\frac{f(x+w+v)-f(x+w)}{v} \geqslant \frac{f(x+v)-f(x)}{v},$$

so

$$\frac{f(x+w+v)-f(x+w)}{v} - \frac{f(x+v)-f(x)}{v} \geqslant 0,$$

and multiply by $v^2 w < 0$. ∎

Bibliography

[1] S. N. Afriat. 1971. Theory of maxima and the method of Lagrange. *SIAM Journal of Applied Mathematics* 20:343–357. http://www.jstor.org/stable/2099955

[2] A. C. Aitken. 1954. *Determinants and matrices*, 8th. ed. New York: Interscience.

[3] C. D. Aliprantis and K. C. Border. 2006. *Infinite dimensional analysis: A hitchhiker's guide*, 3d. ed. Berlin: Springer–Verlag. 5, 14, 17, 24, 59, 81, 82, 86, 94, 114, 115, 116

[4] C. D. Aliprantis and O. Burkinshaw. 1990. *Problems in real analysis*. Boston: Academic Press. 31

[5] E. J. Anderson and P. Nash. 1987. *Linear programming in infinite dimensional spaces*. New York: John Wiley and Sons.

[6] T. M. Apostol. 1957. *Mathematical analysis: A modern approach to advanced calculus*. Addison-Wesley series in mathematics. Reading, Massachusetts: Addison Wesley. 11, 25, 28, 30, 31, 47, 50, 55, 56, 65

[7] ———— . 1967. *Calculus*, 2d. ed., volume 1. Waltham, Massachusetts: Blaisdell. 3, 25, 26, 27, 28, 29, 31, 32, 46, 47, 154

[8] ———— . 1969. *Calculus*, 2d. ed., volume 2. Waltham, Massachusetts: Blaisdell. 25, 46, 47, 55, 65, 162

[9] K. J. Arrow and A. C. Enthoven. 1961. Quasi-concave programming. *Econometrica* 29(4):779–800. http://www.jstor.org/stable/1911819 153

[10] K. J. Arrow and L. Hurwicz. 1956. Reduction of constrained maxima to saddle-point problems. In J. Neyman, ed., *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry*, pages 1–20. Berkeley and Los Angeles: University of California Press. http://projecteuclid.org/euclid.bsmsp/1200511853

[11] K. J. Arrow, L. Hurwicz, and H. Uzawa, eds. 1958. *Studies in linear and non-linear programming*. Number 2 in Stanford Mathematical Studies in the Social Sciences. Stanford, California: Stanford University Press. 182, 185

[12] ———— . 1961. Constraint qualifications in maximization problems. *Naval Research Logistics Quarterly* 8(2):175–191. DOI: 10.1002/nav.3800080206 149

[13] J.-P. Aubin. 1979. *Mathematical methods of game and economic theory*. Number 7 in Studies in mathematics and its applications. New York: North Holland. 36

[14] ———— . 1984. *L'analyse non linéaire et ses motivations économiques*. Paris: Masson. 106

[15] J.-P. Aubin and I. Ekeland. 1984. *Applied nonlinear analysis*. Pure and Applied Mathematics: A Wiley-Interscience Series of Texts, Monographs, and Tracts. Mineola, New York: John Wiley & Sons. Reprint of the 1984 edition by John Wiley and Sons. 36

[16] A. Bachem and W. Kern. 1992. *Linear programming duality: An introduction to oriented matroids*. Berlin: Springer–Verlag.

[17] L. M. Benveniste and J. A. Scheinkman. 1979. On the differentiability of the value function in dynamic models of economics. *Econometrica* 47:727–732. http://www.jstor.org/stable/1910417

[18] C. Berge. 1959. *Espaces topologiques et fonctions multivoques*. Paris: Dunod. 21, 22

[19] ——— . 1997. *Topological spaces*. Mineola, New York: Dover. Originally published in French by Dunod, Paris, 1962 as *Espaces topologiques, fonctions multivoques*. Reprint of the English translation by E. M. Patterson, originally published by Oliver and Boyd, Edinburgh and London, 1963.

[20] A. Berman. 1973. *Cones, matrices and mathematical programming*. Number 79 in Lecture Notes in Economics and Mathematical Systems. Berlin: Springer–Verlag.

[21] D. P. Bertsekas. 1976. *Dynamic programming and stochastic control*. Number 125 in Mathematics in Science and Engineering. New York: Academic Press.

[22] ——— . 1999. *Nonlinear programming*, 2d. ed. Belmont, Massachusetts: Athena Scientific.

[23] D. P. Bertsekas and S. E. Shreve. 1978. *Stochastic optimal control: The discrete time case*. Number 139 in Mathematics in Science and Engineering. New York: Academic Press. http://hdl.handle.net/1721.1/4852

[24] D. Blackwell. 1965. Discounted dynamic programming. *Annals of Mathematical Statistics* 36:226–235. http://www.jstor.org/stable/2238089

[25] ——— . 1967. Positive dynamic programming. In L. M. Le Cam and J. Neyman, eds., *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 3: Physical Sciences*, pages 415–418. Berkeley and Los Angeles: University of California Press. http://projecteuclid.org/euclid.bsmsp/1200513001

[26] G. A. Bliss. 1938. Normality and abnormality in the calculus of variations. *Transactions of the American Mathematical Society* 43(3):365–376. http://www.jstor.org/stable/1990066

[27] ——— . 1946. *Lectures on the calculus of variations*. Chicago: University of Chicago Press.

[28] O. Bolza. 1961. *Lectures on the calculus of variations*. New York: Dover. Reprint of the 1904 edition published by the University of Chicago Press.

[29] T. Bonnesen and W. Fenchel. 1948. *Theorie der konvexen Körper*. New York: Chelsea. Reprint of 1934 edition published by Julius Springer, number 3 in the series Ergebnisse der Mathematik.

[30] K. C. Border. 1985. *Fixed point theorems with applications to economics and game theory*. New York: Cambridge University Press. 20, 57

[31] J. M. Borwein and A. S. Lewis. 2006. *Convex analysis and nonlinear optimization: Theory and examples*, 2d. ed. Number 3 in CMS Books in Mathematics/Ouvrages de Mathèmatiques de la SMC. New York: Springer.

[32] J. M. Borwein and Q. J. Zhu. 2005. *Techniques of variational analysis*. Number 20 in CMS Books in Mathematics/Ouvrages de Mathèmatiques de la SMC. New York: Springer Science+Business Media.

[33] D. Boyd. 1972. Classical analysis, volume 1. Notes prepared for Ma 108 abc. Published occasionally since at least 1972 by the Department of Mathematics, California Institute of Technology, 253-37, Pasadena CA 91125. 33

[34] E. Burger. 1955. On extrema with side conditions. *Econometrica* 23(4):451–452. http://www.jstor.org/stable/1905352

[35] C. Carathéodory. 1982. *Calculus of variations*, 2d. ed. New York: Chelsea. This was originally published in 1935 in two volumes by B. G. Teubner in Berlin as *Variationsrechnung und Partielle Differentialgleichungen erster Ordnung*. In 1956 the first volume was edited and updated by E. Hölder. The revised work was translated by Robert B. Dean and Julius J. Brandstatter and published in two volumes as *Calculus of variations and partial differential equations of the first order* by Holden-Day in 1965–66. The Chelsea second edition combines and revises the 1967 edition. 137, 161

[36] C. Castaing and M. Valadier. 1977. *Convex analysis and measurable multifunctions*. Number 580 in Lecture Notes in Mathematics. Berlin: Springer–Verlag. 70

[37] F. H. Clarke. 1990. *Optimization and nonsmooth analysis*. Number 5 in Classics in Applied Mathematics. New York: SIAM. Reprint of 1983 edition published by Wiley.

[38] F. H. Clarke, Y. S. Ledyaev, R. J. Stern, and P. R. Wolenski. 1998. *Nonsmooth analysis and control theory*. Number 178 in Graduate Texts in Mathematics. New York: Springer–Verlag. 159

[39] G. B. Dantzig. 1963. *Linear programming and extensions*. Princeton: Princeton University Press. 126, 131

[40] G. Debreu. 1952. Definite and semidefinite quadratic forms. *Econometrica* 20(2):295–300. http://www.jstor.org/stable/1907852 164, 169

[41] ——— . 1959. *Theory of value: An axiomatic analysis of economic equilibrium*. Number 17 in Cowles Foundation Monographs. New Haven: Yale University Press. http://cowles.econ.yale.edu/P/cm/m17/m17-all.pdf 3

[42] P. A. Diamond and D. L. McFadden. 1974. Some uses of the expenditure function in public finance. *Journal of Public Economics* 3:3–21. 159

[43] J. Dieudonné. 1969. *Foundations of modern analysis*. Number 10-I in Pure and Applied Mathematics. New York: Academic Press. Volume 1 of Treatise on Analysis. 20, 25, 36, 38, 41, 47, 55, 56, 57, 61

[44] W. E. Diewert and A. D. Woodland. 1977. Frank Knight's theorem in linear programming revisited. *Econometrica* 45(2):375–398. http://www.jstor.org/stable/1911216 166

[45] R. Dorfman, P. A. Samuelson, and R. M. Solow. 1987. *Linear programming and economic analysis*. New York: Dover. Reprint of the 1958 edition published by McGraw-Hill in New York.

[46] N. Dunford and J. T. Schwartz. 1957. *Linear operators: Part I*. New York: Interscience. 81

[47] I. Ekeland and R. Temam. 1976. *Convex analysis and variational problems*. Number 1 in Studies in Mathematics and its Applications. Amsterdam: North Holland. 36

[48] M. A. El-Hodiri. 1971. *Constrained extrema: Introduction to the differentiable case with economic applications*. New York: Springer–Verlag.

[49] L. G. Epstein. 1981. Generalized duality and integrability. *Econometrica* 49(3):655–678. http://www.jstor.org/stable/1911517

[50] K. Fan, I. Glicksberg, and A. J. Hoffman. 1957. Systems of inequalities involving convex functions. *Proceedings of the American Mathematical Society* 8:617–622. http://www.jstor.org/stable/2033529 82

[51] W. Fenchel. 1953. Convex cones, sets, and functions. Lecture notes, Princeton University, Department of Mathematics. From notes taken by D. W. Blackett, Spring 1951. 86, 90, 93, 95, 97, 98, 104

[52] B. de Finetti. 1949. Sulle stratificazioni convesse. *Annali di Matematica Pura ed Applicata. Serie 4* 30(1):173–183. DOI: 10.1007/BF02415006

[53] W. H. Fleming and R. W. Rishel. 1975. *Deterministic and stochastic optimal control*. Number 1 in Applications of Mathematics. New York: Springer–Verlag.

[54] J. Franklin. 1980. *Methods of mathematical economics*. Undergraduate Texts in Mathematics. New York: Springer–Verlag. 114

[55] D. Gale. 1951. Convex polyhedral cones and linear inequalities. In Koopmans [93], chapter 17, pages 287–297. http://cowles.econ.yale.edu/P/cm/m13/m13-17.pdf

[56] ——— . 1960. A note on revealed preference. *Economica N.S.* 27(108):348–354. http://www.jstor.org/stable/2550547 112

[57] ——— . 1989. *Theory of linear economic models*. Chicago: University of Chicago Press. Reprint of the 1960 edition published by McGraw-Hill. 114, 115, 117, 126, 127, 131

[58] D. Gale and H. Nikaidô. 1965. Jacobian matrix and global univalence of mappings. *Mathematische Annalen* 159(2):81–93. DOI: 10.1007/BF01360282

[59] F. R. Gantmacher. 1959. *Matrix theory*, volume 1. New York: Chelsea. 161, 164, 165, 166

[60] M. Gerstenhaber. 1951. Theory of convex polyhedral cones. In Koopmans [93], chapter 18, pages 298–316. http://cowles.econ.yale.edu/P/cm/m13/m13-18.pdf

[61] M. A. Goberna and M. A. López. 1998. *Linear semi-infinite optimization*. Number 2 in Wiley Series in Mathematical Methods in Practice. New York: John Wiley and Sons.

[62] P. Gordan. 1873. Über die auflösung linearer Gleichungen mit reelen Coefficienten [On the solution of linear inequalities with real coefficients]. *Mathematische Annalen* 6(1):23–28. DOI: 10.1007/BF01442864

[63] E. Goursat. 1910. *Cours d'analyse mathématique*. Paris: Gauthier-Villars. 50

[64] L. M. Graves. 1956. *Theory of functions of real variables*. New York: McGraw–Hill. 38

[65] N. E. Gretsky, J. M. Ostroy, and W. R. Zame. 2002. Subdifferentiability and the duality gap. *Positivity* 6:261–274. http://www.econ.ucla.edu/zame/GOZ3.pdf

[66] G. Hadley. 1964. *Nonlinear and dynamic programming*. Reading, Massachusetts: Addison-Wesley.

[67] H. Halkin. 1972. Necessary conditions for optimal control problems with infinite horizons. Discussion Paper 7210, CORE. 56

[68] ———— . 1974. Implicit functions and optimization problems without continuous differentiability of the data. *SIAM Journal on Control* 12(2):229–236. DOI: 10.1137/0312017 53, 54, 56

[69] ———— . 1974. Necessary conditions for optimal control problems with infinite horizons. *Econometrica* 42(2):267–272. http://www.jstor.org/stable/1911976

[70] P. R. Halmos. 1974. *Measure theory*. Graduate Texts in Mathematics. New York: Springer–Verlag. Reprint of the edition published by Van Nostrand, 1950. 24

[71] H. Hancock. 1960. *Theory of maxima and minima*. New York: Dover. Reprint of 1917 edition published by Ginn–Blaisdell. 3, 50, 51, 168

[72] G. H. Hardy. 1952. *A course of pure mathematics*, 10th. ed. Cambridge: Cambridge University Press. 25, 26, 28

[73] M. R. Hestenes. 1966. *Calculus of variations and optimal control theory*. Applied Mathematics Series. New York: Wiley.

[74] J.-B. Hiriart-Urruty and C. Lemaréchal. 1993. *Convex analysis and minimization algorithms I*. Number 305 in Grundlehren der mathematischen Wissenschaften. Berlin: Springer–Verlag.

[75] ———— . 1993. *Convex analysis and minimization algorithms II*. Number 306 in Grundlehren der mathematischen Wissenschaften. Berlin: Springer–Verlag.

[76] ———— . 2001. *Fundamentals of convex analysis*. Grundlehren Text Editions. Berlin: Springer–Verlag. 71, 81, 95

[77] J. Horváth. 1966. *Topological vector spaces and distributions*, volume 1. Reading, Mass.: Addison Wesley. 31

[78] H. S. Houthakker. 1950. Revealed preference and the utility function. *Economica N.S.* 17:159–174. http://www.jstor.org/stable/2549382 112

[79] L. Hurwicz. 1958. Programming in linear spaces. In Arrow et al. [11], chapter 4, pages 38–102.

[80] L. Hurwicz and M. K. Richter. 1995. Implicit functions and diffeomorphisms without C^1. Discussion Paper 279, Center for Economic Research, University of Minnesota. 182

[81] ———— . 1997. Implicit functions and diffeomorphisms without C^1. A modified version of [80], presented at the 1997 North American Summer Meeting of the Econometric Society held at Caltech, Pasadena CA.

[82] ———— . 1997. Optimization and Lagrange multipliers: Non-C^1 constraints and "minimal" constraint qualifications. Manuscript, presented at the 1997 North American Summer Meeting of the Econometric Society held at Caltech, Pasadena CA.

[83] ———— . 2003. Implicit functions and diffeomorphisms without C^1. *Advances in Mathematical Economics* 5:65–96. 54

[84] M. D. Intriligator. 1971. *Mathematical optimization and economic theory*. Englewood Cliffs, New Jersey: Prentice-Hall.

[85] G. A. Jehle. 1991. *Advanced microeconomic theory*. Englewood Cliffs, New Jersey: Prentice-Hall. 151

[86] F. John. 1948. Extremum problems with inequalities as subsidiary conditions. In K. O. Friedrichs, O. E. Neugebauer, and J. J. Stoker, eds., *Studies and Essays: Courant Anniversary Volume*, pages 187–204. New York: Interscience.

[87] Y. Kannai. 1977. Concavifiability and construction of concave utility functions. *Journal of Mathematical Economics* 4(1):1–56. DOI: 10.1016/0304-4068(77)90015-5

[88] R. Kannan and C. K. Krueger. 1996. *Advanced analysis on the real line*. Universitext. New York: Springer. 34

[89] S. Karlin. 1987. *Mathematical methods and theory in games, programming, and economics*. New York: Dover. Reprint of the 1959 two-volume edition published by Addison–Wesley. 124

[90] W. Karush. 1939. Minima of functions of several variables with inequalities as side conditions. Master's thesis, Department of Mathematics, University of Chicago. 146

[91] D. W. Katzner. 1970. *Static demand theory*. London: Macmillan. 97

[92] R. Kihlstrom, A. Mas-Colell, and H. F. Sonnenschein. 1976. The demand theory of the weak axiom of revealed preference. *Econometrica* 44(5):971–978. http://www.jstor.org/stable/1911539 112

[93] T. C. Koopmans, ed. 1951. *Activity analysis of production and allocation: Proceedings of a conference*. Number 13 in Cowles Commission for Research in Economics Monographs. New York: John Wiley and Sons. http://cowles.econ.yale.edu/P/cm/m13/index.htm 181

[94] H. W. Kuhn. 1982. Nonlinear programming: A historical view. *ACM SIGMAP Bulletin* 31:6–18. Reprinted from SIAM–AMS Proceedings, volume IX, pp. 1–26.

[95] H. W. Kuhn and A. W. Tucker. 1951. Nonlinear programming. In J. Neyman, ed., *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability II, Part I*, pages 481–492. Berkeley: University of California Press. Reprinted in [119, Chapter 1, pp. 3–14]. http://projecteuclid.org/euclid.bsmsp/1200500249 146, 147

[96] K. J. Lancaster. 1987. *Mathematical economics*. Mineola, NY: Dover. Reprint of the 1968 edition published by Macmillan in New York.

[97] E. Landau. 1960. *Foundations of analysis*. New York: Chelsea. Translation of *Grundlagen der Analysis*, published in 1930. 26

[98] E. B. Leach. 1961. A note on inverse function theorems. *Proceedings of the American Mathematical Society* 12:694–697. http://www.jstor.org/stable/2034858 54

[99] ——— . 1963. On a related function theorem. *Proceedings of the American Mathematical Society* 14:687–689. http://www.jstor.org/stable/2034971 54

[100] L. H. Loomis and S. Sternberg. 1968. *Advanced calculus*. Reading, Massachusetts: Addison–Wesley. 25, 38, 41, 42, 43, 46, 47, 49, 51, 54, 56, 58

[101] R. Lucchetti. 2006. *Convexity and well-posed problems*. CMS Books in Mathematics/Ouvrages de Mathèmatiques de la SMC. New York: Springer Science+Business Media.

[102] D. G. Luenberger. 1969. *Optimization by vector space methods*. Series in Decision and Control. New York: Wiley. 36, 38

[103] ——— . 1995. *Microeconomic theory*. New York: McGraw-Hill. 3

[104] ——— . 2005. *Linear and nonlinear programming*, 2d. ed. New York: Springer Science+Business Media.

[105] O. L. Mangasarian. 1994. *Nonlinear programming*. Classics in Applied Mathematics. Philadelphia: SIAM. Reprint of the 1969 edition published by McGraw–Hill.

[106] H. B. Mann. 1943. Quadratic forms with linear constraints. *American Mathematical Monthly* 50(7):430–433. Reprinted in [119]. http://www.jstor.org/stable/2303666 169

[107] J. E. Marsden. 1974. *Elementary classical analysis*. San Francisco: W. H. Freeman and Company. 3, 25, 27, 38, 46, 47, 55, 56, 61, 62

[108] A. Mas-Colell, M. D. Whinston, and J. R. Green. 1995. *Microeconomic theory*. Oxford: Oxford University Press. 3, 86

[109] P. Milgrom and I. Segal. 2002. Envelope theorems for arbitrary choice sets. *Econometrica* 70(2):583–601. http://www.jstor.org/stable/2692283

[110] L. Montrucchio. 1987. Lipschitz continuous policy functions for strongly concave optimization problems. *Journal of Mathematical Economics* 16(3):259–273. DOI: 10.1016/0304-4068(87)90012-7

[111] J. C. Moore. 1968. Some extensions of the Kuhn–Tucker results in concave programming. In J. P. Quirk and A. M. Zarley, eds., *Papers in Quantitative Economics, 1*, pages 31–71. Lawrence, Kansas: University of Kansas Press. 129

[112] ——— . 1999. *Mathematical methods for economic theory*. Studies in Economic Theory. New York: Springer–Verlag.

[113] T. S. Motzkin. 1934. *Beiträge zur Theorie der linearen Ungleichungen*. PhD thesis, Universität Basel. 119

[114] ——— . 1951. Two consequences of the transposition theorem on linear inequalities. *Econometrica* 19(2):184–185. http://www.jstor.org/stable/1905733 119

[115] K. Murota. 2003. *Discrete convex analysis*. SIAM Monographs on Discrete Mathematics and Applications. Philadelphia: SIAM.

[116] R. B. Myerson. 1981. Optimal auction design. *Mathematics of Operations Research* 6(1):58–73. http://www.jstor.org/stable/3689266 154

[117] W. A. Neilson, T. A. Knott, and P. W. Carhart, eds. 1944. *Webster's new international dictionary of the English language*, second unabridged ed. Springfield, Massachusetts: G. & C. Merriam Company. 114, 135

[118] L. W. Neustadt. 1976. *Optimization: A theory of necessary conditions*. Princeton: Princeton University Press.

[119] P. Newman, ed. 1968. *Readings in mathematical economics I: Value theory*. Baltimore: Johns Hopkins Press. 183

[120] A. Nijenhuis. 1974. Strong derivatives and inverse mappings. *American Mathematical Monthly* 81(9):969–980. 54

[121] H. Nikaidô. 1968. *Convex structures and economic theory*. Mathematics in Science and Engineering. New York: Academic Press. 54

[122] W. Novshek. 1993. *Mathematics for economists*. Economic Theory, Econometrics, and Mathematical Economics. San Diego, California: Academic Press.

[123] J. M. Ostroy. 1993. Notes on linear programming. Notes for Economics 201C at UCLA.

[124] A. L. Peressinni, F. E. Sullivan, and J. J. Uhl, Jr. 1988. *The mathematics of nonlinear programming*. Undergraduate Texts in Mathematics. New York: Springer–Verlag.

[125] R. R. Phelps. 1993. *Convex functions, monotone operators and differentiability*, 2d. ed. Number 1364 in Lecture Notes in Mathematics. Berlin: Springer–Verlag. 90, 93

[126] J. Ponstein. 1967. Seven kinds of convexity. *SIAM Review* 9(1):115–119. http://www.jstor.org/stable/2027415

[127] ——— . 1984. Dualizing optimization problems in mathematical economics. *Journal of Mathematical Economics* 13(3):255–272. DOI: 10.1016/0304-4068(84)90033-8

[128] J. P. Quirk. 1976. *Intermediate microeconomics: Mathematical notes*. Chicago: Science Research Associates. 141, 166

[129] A. W. Roberts and D. E. Varberg. 1973. *Convex functions*. New York: Academic Press.

[130] R. T. Rockafellar. 1970. *Convex analysis*. Number 28 in Princeton Mathematical Series. Princeton: Princeton University Press. 70, 71, 74, 75, 76, 77, 81, 87, 88, 90, 98, 99, 100, 104, 108, 109

[131] R. T. Rockafellar and R. J.-B. Wets. 2004. *Variational analysis*, second corrected printing. ed. Number 317 in Grundlehren der mathematischen Wissenschaften. Berlin: Springer–Verlag.

[132] H. L. Royden. 1988. *Real analysis*, 3d. ed. New York: Macmillan. 24, 34, 93

[133] W. Rudin. 1973. *Functional analysis*. New York: McGraw Hill. 81

[134] ———— . 1976. *Principles of mathematical analysis*, 3d. ed. International Series in Pure and Applied Mathematics. New York: McGraw Hill. 5, 7, 25, 47, 55, 56

[135] T. Saijo. 1983. Differentiability of the cost functions is equivalent to strict quasiconcavity of the production functions. *Economics Letters* 12(2):135–139. DOI: 10.1016/0165-1765(83)90124-6

[136] P. A. Samuelson. 1965. *Foundations of economic analysis*. New York: Athenaeum. Reprint of the 1947 edition published by Harvard University Press. 65, 112, 154, 166, 169

[137] N. Schofield. 1984. *Mathematical methods in economics*. New York: New York University Press.

[138] R. J. Serfling. 1980. *Approximation theorems of mathematical statistics*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley. 28

[139] R. W. Shephard. 1981. *Cost and production functions*. Number 194 in Lecture Notes in Economics and Mathematical Systems. Berlin: Springer–Verlag. Reprint of the 1953 edition published by Princeton University Press. 151

[140] N. Z. Shor. 1985. *Minimization methods for non-differentiable functions*. Number 3 in Springer Series in Computational Mathematics. Berlin: Springer–Verlag. Translated from the Russian by K. C. Kiwiel and A. Ruszczyński.

[141] E. Silberberg. 1971. The Le Chatelier principle as a corollary to a generalized envelope theorem. *Journal of Economic Theory* 3(2):146–155. DOI: 10.1016/0022-0531(71)90012-3 159

[142] ———— . 1974. A revision of comparative statics methodology in economics, or how to do comparative statics on the back of an envelope. *Journal of Economic Theory* 7(2):159–172. DOI: 10.1016/0022-0531(74)90104-5 159

[143] ———— . 1978. *The structure of economics: A mathematical analysis*. New York: McGraw-Hill.

[144] M. L. Slater. 1950. Lagrange multipliers revisited: A contribution to non-linear programming. Discussion Paper Math. 403, Cowles Commission. Reissued as Cowles Foundation Discussion Paper #80 in 1959. http://cowles.econ.yale.edu/P/cd/d00b/d0080.pdf 124

[145] M. Spivak. 1965. *Calculus on manifolds*. Mathematics Monograph Series. New York: Benjamin. 25, 38, 45, 46, 47, 55, 56

[146] K. Sydsaeter. 1974. Letter to the editor on some frequently occurring errors in the economic literature concerning problems of maxima and minima. *Journal of Economic Theory* 9(4):464–466. DOI: 10.1016/0022-0531(74)90046-5 30, 31, 50, 136

[147] A. Takayama. 1993. *Analytical methods in economics*. Ann Arbor: University of Michigan Press.

[148] P. R. Thie. 1988. *An introduction to linear programming and game theory*, 2d. ed. New York: Wiley.

[149] D. M. Topkis. 1998. *Supermodularity and complementarity*. Princeton: Princeton University Press. 26

[150] A. H. Turunen-Red and A. D. Woodland. 1999. On economic applications of the Kuhn–Fourier theorem. In M. H. Wooders, ed., *Topics in Mathematical Economics and Game Theory: Essays in Honor of Robert J. Aumann*, Fields Institute Communications, pages 257–276. Providence, RI: American Mathematical Society.

[151] H. Uzawa. 1958. The Kuhn–Tucker conditions in concave programming. In Arrow et al. [11], chapter 3, pages 32–37. 129

[152] ———— . 1958. Preference and rational choice in the theory of consumption. In Arrow et al. [11], chapter 9, pages 129–148. 112

[153] ———— . 1960. Market mechanisms and mathematical programming. *Econometrica* 28(4):872–881. http://www.jstor.org/stable/1907569

[154] ———— . 1964. Duality principles in the theory of cost and production. *International Economic Review* 5:216–220. http://www.jstor.org/stable/2525564

[155] F. Valentine. 1964. *Convex sets*. New York: McGraw-Hill.

[156] H. R. Varian. 1992. *Microeconomic analysis*, 3d. ed. New York: W. W. Norton & Co. 3

[157] J. Viner. 1932. Cost curves and supply curves. *Zeitschrift für Nationalökonomie* 3(1):23–46. DOI: 10.1007/BF01316299 154

[158] J. von Neumann and O. Morgenstern. 1944. *The theory of games and economic behavior*. Princeton: Princeton University Press. Reprint. New York: John Wiley & Sons, 1967. 81, 117

[159] J. Warga. 1978. An implicit function theorem without differentiability. *Proceedings of the American Mathematical Society* 69(1):65–69. http://www.jstor.org/stable/2043190.pdf 55

[160] A. Wilansky. 1998. *Topology for analysis*. Mineola, NY: Dover. Unabridged republication of the work originally published by Ginn and Company, Waltham, MA in 1970. 5

[161] S. Willard. 1970. *General topology*. Reading, Massachusetts: Addison Wesley. 5

[162] M. E. Yaari. 1977. A note on separability and quasiconcavity. *Econometrica* 45:1183–1186. http://www.jstor.org/stable/1914066

[163] Zajaczkowski. 1867. *Annals of the Scientific Society of Cracow* 12. 168