1. ABOUT OPTIMIZATION

The field of optimization is concerned with the study of *maximization and minimization of mathematical functions.* Very often the arguments of (i.e., *variables* or *unknowns* in) these functions are subject to side conditions or *constraints.* By virtue of its great utility in such diverse areas as applied science, engineering, economics, finance, medicine, and statistics, optimization holds an important place in the practical world and the scientific world. Indeed, as far back as the Eighteenth Century, the famous Swiss mathematician and physicist Leonhard Euler (1707-1783) proclaimed[1] that *... nothing at all takes place in the Universe in which some rule of maximum or minimum does not appear.* The subject is so pervasive that we even find some optimization terms in our everyday language.

Optimization is a large subject; it cannot adequately be treated in the short amount time available in one quarter of an academic year. In this course, we shall restrict our attention mainly to some aspects of nonlinear programming and discuss linear programming as a special case. Among the many topics that will *not* be covered in this course are integer programming, network programming, and stochastic programming.

As a discipline, optimization is often called ***mathematical programming***. The latter name tends to be used in conjunction with finite-dimensional optimization problems, which in fact are what we shall be studying here. The word "programming" should not be confused with computer programming which in fact it antedates. As originally used, the term refers to the timing and magnitude of actions to be carried out so as to achieve a goal in the best possible way.

General form of a mathematical programming problem

The class of mathematical programming problems considered in this course can all be expressed in the form

$$\text{(P)} \qquad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X \end{array}$$

We call f the ***objective function*** and X the ***feasible region*** of (P). We assume that the feasible region is a subset of R^n, and f is a real-valued function on X. In particular, this means that for each x in the feasible region, the corresponding objective function value is a

[1]See Leonhardo Eulero, *Methodus Inviendi Lineas Curvas Maximi Minimive Proprietate Gaudentes*, Lausanne & Geneva, 1744, p. 245.

well-defined real number, $f(x)$. The decision variable x may be a vector $x = (x_1, \ldots, x_n)$ or a scalar (when $n = 1$).

A problem (P) in which $X = R^n$ is said to be ***unconstrained***. The study of ***unconstrained optimization*** has a long history and continues to be of interest. When X is a proper subset of R^n, we say that (P) is a ***constrained optimization problem***. In most cases, the set feasible region X is specified through a system of inequalities or equations—or both. Thus, X is often given as the set of all solutions of the system

$$c_i(x) \leq 0 \quad i \in \mathcal{I}$$
$$c_i(x) = 0 \quad i \in \mathcal{E}$$

Note that $\mathcal{I}$ and $\mathcal{E}$ are index sets. These conditions imposed on the decision variable x are called ***constraints***, and the c_i are ***constraint functions***. In a constrained optimization problem, either one of the sets $\mathcal{I}$ and $\mathcal{E}$ may be empty. For that matter, both may be empty if x is still required to belong to a proper subset of R^n. As implied earlier, the optimization problems considered in this course will not involve integer constraints on the individual variables x_j $(j = 1, \ldots, n)$ even though such restrictions are of great importance in many practical problems.

If the feasible region of a constrained optimization problem is empty, the problem is said to be ***infeasible***; otherwise it is ***feasible***.

There are many ways to categorize optimization problems. One of the most straightforward is in accordance with properties of the objective function and constraint functions (if any). Thus, when f and all the c_i are ***affine functions***, that is of the form

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n + a_{n+1},$$

we have what is called a ***linear programming problem***. Linear programming problems (or ***linear programs*** as they are also called) are of great importance in the field, partly because there are so many real-world problems that are naturally of this type or can be approximated by them. When f or any of the constraint functions is nonlinear (or not affine), a constrained optimization problem (P) is called a ***nonlinear programming problem*** (or ***nonlinear program***).

In some cases, constrained optimization problems are classified according to their structure. This is especially true of linear programs. Since there are many types of nonlinear functions, there is a wide range of possibilities among nonlinear programs. Among the most simplest—yet most important—of nonlinear programming problems is the ***quadratic programming problem***: the optimization of a quadratic function subject to affine constraints.

For information on the historical development of mathematical programming, especially that of the last half century, see

- J.K. Lenstra, A.H.G. Rinnooy Kan, and A. Schrijver, *History of Mathematical Programming*, North-Holland, Amsterdam, 1991.
- G.B. Dantzig and M.N. Thapa, *Linear Programming 1: Introduction*, Springer, New York, 1997. [See especially pp. xxi–xxxii.]

SELECTED BOOKS ON OPTIMIZATION

[1] J. Abadie, ed., *Nonlinear Programming*, North-Holland, Amsterdam, 1967.

[2] J. Abadie, ed., *Integer and Nonlinear Programming*, North-Holland, Amsterdam, 1970.

[3] J-P. Aubin, *Optima and Equilibria, Second Edition*, Springer-Verlag, Berlin, 1998.

[4] M. Avriel, *Nonlinear Programming*, Prentice-Hall, Englewood Cliffs, New Jersey, 1976.

[5] M.S. Bazaraa, H.D. Sherali and C.M. Shetty, *Nonlinear Programming*, second edition, John Wiley & Sons, New York, 1993.

[6] E.M.L. Beale, *Mathematical Programming in Practice*, Sir Isaac Pitman & Sons, Ltd., London, 1968.

[7] D.P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, Massachusetts, 1995.

[8] D.M. Bertsimas and J.N. Tsitsiklis, *Linear Optimization*, Athena Scientific, Belmont, Massachusetts, 1997.

[9] E. Blum and W. Oettli, *Mathematische Optimierung*, Springer-Verlag, Berlin, 1975. [In German.]

[10] J.M. Borwein and A.S. Lewis, *Convex Analysis and Nonlinear Optimization*, Springer-Verlag, New York, 2000.

[11] J. Bracken and G.P. McCormick, *Selected Applications of Nonlinear Programming*, John Wiley & Sons, New York, 1968.

[12] G. Bradley, A. Hax, and T.L. Magnanti, *Applied Mathematical Programming*, Addison-Wesley, Reading, Massachusetts, 1977.

[13] N. Cameron, *Introduction to Linear and Convex Programming*, Cambridge University Press, Cambridge, 1985.

[14] J. Céa, *Optimisation*, Dunod, Paris, 1971. [In French.]

[15] P.G. Ciarlet, *Introduction to Numerical Linear Algebra and Optimization*, Cambridge University Press, Cambridge, 1989. [Translated from the French.]

[16] F. Clarke, *Optimization and Nonsmooth Analysis*, Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[17] R.W. Cottle, J.S. Pang, and R.E. Stone, *The Linear Complementarity Problem*, Academic Press, Boston, 1992.

[18] G.B. Dantzig and M.N. Thapa, *Linear Programming 1: Introduction*, Springer-Verlag, New York, 1997.

[19] G.B. Dantzig and M.N. Thapa, *Linear Programming 2: Theory and Extensions*, Springer-Verlag, New York, 2003.

[20] G.B. Dantzig and A.F. Veinott, Jr., eds., *Mathematics of the Decision Sciences, Part 1*, American Mathematical Society, Providence, RI, 1968.

[21] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976. [Translated from the French.]

[22] A.V. Fiacco and G.P. McCormick, *Nonlinear Programming: Sequential Unconstrained Optimization Techniques*, John Wiley & Sons, New York, 1968.

[23] R. Fletcher, *Practical Methods of Optimization, Volume 1: Unconstrained Optimization*, John Wiley & Sons, Chichester, 1980.

[24] R. Fletcher, *Practical Methods of Optimization, Volume 2: Constrained Optimization*, John Wiley & Sons, Chichester, 1980.

[25] R. Fourer, D.M. Gay, and B.W. Kernighan, *AMPL*, Scientific Press, South San Francisco, California, 1993.

[26] J. Franklin, *Methods of Mathematical Economics*, Springer-Verlag, New York, 1980.

[27] D. Gale, *The Theory of Linear Economic Models*, McGraw-Hill Book Company, Inc., New York, 1960.

[28] C.B. Garcia and W.I. Zangwill, *Pathways to Solutions, Fixed Points, and Equilibria*, Prentice-Hall, Engelwood Cliffs, New Jersey, 1981.

[29] S.I. Gass and C.M. Harris, eds., *Encyclopedia of Operations Research and Management Science*, Kluwer Academic Publishers, Boston.

[30] P.E. Gill, W. Murray and M.H. Wright, *Practical Optimization*, Academic Press, London, 1981.

[31] P.E. Gill, W. Murray and M.H. Wright, *Numerical Linear Algebra and Optimization*, Addison-Wesley, Redwood City, California, 1991.

[32] E.G. Golstein, *Dualitätstheorie in der nichtlinearen ihre Anwendungen*, Akademie-Verlag, Berlin, 1975. [German translation of the original Russian.]

[33] H. Hancock, *Theory of Maxima and Minima*, Dover, New York, 1960. [Reproduction of the first edition of the work published by Ginn and Company, 1917.]

[34] R. Korn and E. Korn, *Option Pricing and Portfolio Optimization*, American Mathematical Society, Providence, R.I., 2001. [Translated from the German.]

[35] P.-J. Laurent, *Approximation et Optimisation*, Hermann, Paris, 1972. [In French.]

[36] D.G. Luenberger, *Optimization by Vector Space Methods*, John Wiley & Sons, New York, 1969.

[37] D.G. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley, Reading, Massachusetts, 1989.

[38] Z.Q. Luo, J.S. Pang and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, 1996.

[39] O.L. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New York 1969.

[40] G.P. McCormick, *Nonlinear Programming*, John Wiley & Sons, New York, 1983.

[41] M. Minoux, *Mathematical Programming*, John Wiley & Sons, New York, 1986. [Translated from the French.]

[42] J.J. Moré and S.J. Wright, *Optimization Software Guide*, Society for Industrial and Applied Mathematics, Philadelphia, 1994.

[43] K.G. Murty, *Linear Programming*, John Wiley & Sons, New York, 1983.

[44] K.G. Murty, *Linear Complementarity, Linear and Nonlinear Programming*, Heldermann-Verlag, Berlin, 1988.

[45] G.L. Nemhauser, A.H.G. Rinnooy Kan, and M.J. Todd (eds.), *Optimization*, North-Holland, Amsterdam, 1989.

[46] H. Nikaido, *Convex Structures and Economic Theory*, Academic Press, New York, 1968.

[47] J. Nocedal and S.J. Wright, *Numerical Optimization*, Springer, New York, 1999.

[48] A.L. Peressini, F.E. Sullivan, and J.J. Uhl, Jr., *The Mathematics of Nonlinear Programming*, Springer-Verlag, New York, 1993.

[49] E.L. Polak, *Optimization: Algorithms and Consistent Approximations*, Springer-Verlag, New York, 1997.

[50] B.T. Polyak, *Introduction to Optimization*, Optimization New York, 1987.

[51] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1970.

[52] R.T. Rockafellar, *Conjugate Duality and Optimization*, Society for Industrial and Applied Mathematics, Philadelphia, 1974.

[53] R.T. Rockafellar and R.J-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.

[54] A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley & Sons, Chichester, 1986.

[55] J.F. Shapiro, *Mathematical Programming*, John Wiley & Sons, New York, 1979.

[56] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimensions I*, Springer-Verlag, New York, 1970.

[57] R.K. Sundaram, *A First Course in Optimization Theory*, Cambridge University Press, Cambridge, 1996.

[58] T. Terlaky (ed.), *Interior Point Methods of Mathematical Programming*, Kluwer Academic Publishers, Dordrecht, 1996.

[59] J. van Tiel, *Convex Analysis*, John Wiley & Sons, Chicester, 1984.

[60] R. Webster, *Convexity*, Oxford University Press, Oxford 1994.

[61] S.J. Wright, *Primal-Dual Interior-Point Methods*, SIAM Publications, Philadelphia, 1997.

[62] Y. Ye, *Interior Point Algorithms*, New York, John Wiley & Sons, 1997.

[63] W.I. Zangwill, *Nonlinear Programming*, Prentice-Hall, Engelwood Cliffs, New Jersey, 1969.

[64] G.M. Ziegler, *Lectures on Polytopes*, Springer-Verlag, New York, 1995.

[65] G. Zoutendijk, *Methods of Feasible Directions*, Elsevier, Amsterdam, 1960.

2. CONVEX SETS AND FUNCTIONS

2.1 Real n-space; Euclidean n-space

In these pages, R denotes the field of real numbers (the real number system) and R^n denotes real n-space, the n-dimensional vector space of all n-tuples $(x_1, \ldots, x_n)$ with $x_i \in R$ for $i = 1, \ldots, n$. The elements of R^n are called ***vectors***[2] or ***points***. The latter term is preferred in geometric discussions. For all $x, y \in R^n$, the ***inner product*** of x and y is

$$x^{\mathrm{T}}y = \sum_{i=1}^{n} x_i y_i.$$

It is sometimes convenient to use the notation $\langle x, y \rangle$ for the inner product of x and y. For $x \in R^n$, let

$$\|x\| = (x^{\mathrm{T}}x)^{1/2}.$$

This function is called the ***Euclidean norm (metric)***.[3] When "equipped" with this norm (also known as the 2-norm), R^n becomes ***Euclidean n-space***, E^n, where, of course, Euclidean geometry holds. The Euclidean norm gives rise to the ***Euclidean metric*** defined as

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

This function is used to measure the ***Euclidean distance*** between x and y. As a metric, d has the following properties:

(i) $d(x, y) \geq 0$ with equality if and only if $x = y$ (*definiteness*);

(ii) $d(x, y) = d(y, x)$ (*symmetry*);

(iii) $d(x, z) \leq d(x, y) + d(y, z)$ (*triangle inequality*).

Note that $d(x, 0) = \|x\|$. This is sometimes called the ***length***[4] of x. There are many other ways to measure distance, but the one above is the most widely used, and—as stated above—is what turns R^n into Euclidean n-space.

---

[2]Ordinarily, we treat vectors $x \in R^n$ as columns (or $n \times 1$ matrices).

[3]Notice that as with *all* norms, the property $\|\lambda x\| = |\lambda| \cdot \|x\|$ holds for all x and all $\lambda \in R$.

[4]But note that some authors apply the term "length" to $x \in R^n$, they mean n, the number of components (or coordinates) of x. See, for example Nash and Sofer, page 17, line –9.

2.2 Convex sets

The following are familiar definitions.

Definition. A subset S of R^n is ***affine*** if

$$[\, x, y \in S \text{ and } \alpha \in R\,] \Longrightarrow \alpha x + (1-\alpha)y \in S.$$

When x and y are distinct points in R^n and $\alpha \in R$, the set of all points $z = \alpha x + (1-\alpha)y$ is the ***line*** determined by x and y.

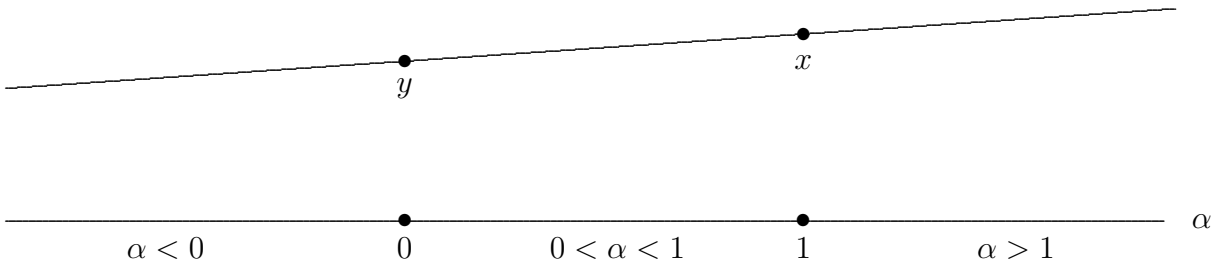


Figure 2.1

Definition. A subset S of R^n is ***convex*** if for all $x, y \in S$ and all $0 < \alpha < 1$ it follows that $\alpha x + (1-\alpha)y \in S$. The vector $z = \alpha x + (1-\alpha)y$ is called a ***convex combination*** of x and y.

The concept of the convex combination of two points can be generalized to any finite number of points. Thus, if $x^1, \ldots, x^m$ are m points in R^n and if $\alpha_1, \ldots, \alpha_m$ are nonnegative scalars that sum to 1, then a point of the form

$$x = \alpha_1 x^1 + \cdots + \alpha_m x^m$$

is called a ***convex combination*** of $x^1, \ldots, x^m$.

By default or convention, the ***empty set***, $\emptyset$, is convex. Examples of convex sets arise in linear programming and many other places within the field of optimization. Sets of the form

$$\{x : Ax \geq b\} \quad \text{and} \quad \{y : y = Ax, x \geq 0\}$$

are typical examples. So are those of the form

$$\{x : x^{\mathrm{T}} D x \leq \kappa\} \quad \kappa > 0,\ D \text{ positive definite.}$$

Convex sets have the interesting and useful property that intersections of convex sets are again convex.

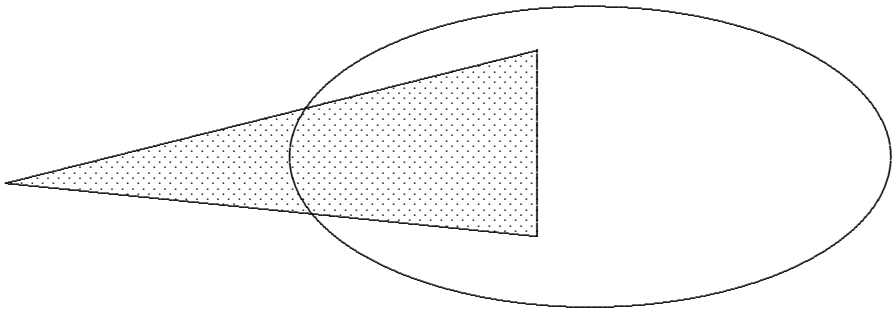


Figure 2.2

Definition. A ***linear variety*** V in R^n is a translate of a linear subspace, that is

$V = \{a\} + T := \{v \in R^n : v = a + x, \quad x \in T\}$ $\quad a \in R^n$ (fixed) and T a linear subspace.

Definition. If S is a nonempty subset of R^n, the ***carrying plane*** or ***affine hull*** of S is the linear variety $V(S)$ of least dimension containing S. The ***dimension*** of S is the dimension of $V(S)$.

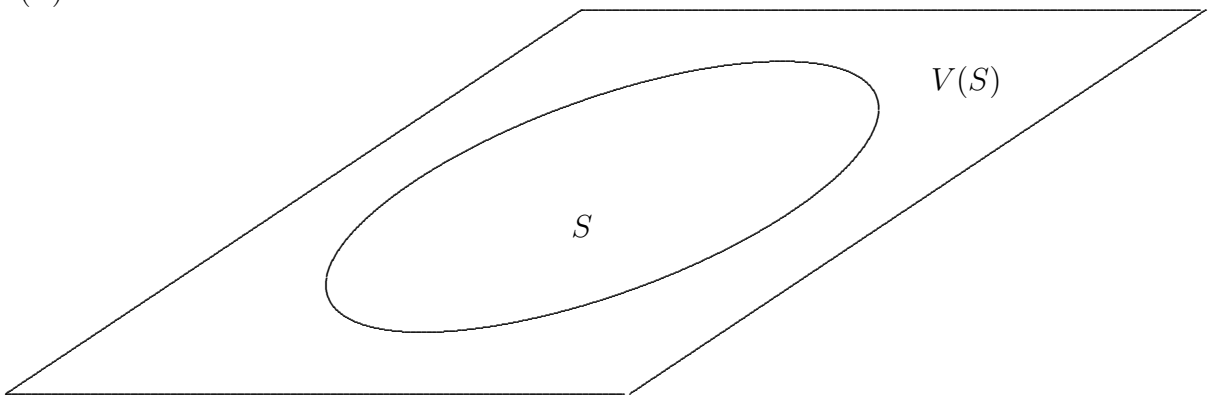


Figure 2.3

Let $B_\delta(\bar{x})$ denote the ***open ball*** of radius δ centered at the point $\bar{x}$. This means

$$B_\delta(\bar{x}) = \{x : \|x - \bar{x}\| < \delta\}.$$

The ***interior*** of a set S is the set

$$\operatorname{int} S = \{x \in S : B_\delta(x) \subset S \text{ for some } \delta > 0\}.$$

Note that $\operatorname{int} S$ may be empty, as in the case of a line segment in 2-dimensional space, but if $\operatorname{int} S \neq \emptyset$, then we say that S is ***solid.*** A point $\bar{x}$ belongs to the ***boundary*** of S if every open ball centered at $\bar{x}$ meets both S and its complement. The boundary of S is denoted $\operatorname{bdy} S$ or sometimes ∂S.

Recall that a set $S \subset R^n$ is said to be ***open*** if for each point $\bar{x} \in S$ there is a $\delta > 0$ such that $B_\delta(\bar{x}) \subset S$. A set $S \subset R^n$ is said to be ***closed*** if its complement $R^n \setminus S$ is open.

We say $\bar{x} \in S$ is a ***relative interior point***, symbolically rel int S, if there exists a $\delta > 0$ such that $(B_\delta(\bar{x}) \cap V(S)) \subset S$. The set $B_\delta(\bar{x}) \cap V(S)$ is called a ***relative neighborhood*** of $\bar{x}$. If $\bar{x} \in R^n$ and every relative neighborhood of $\bar{x}$ contains an element of S and an element of

$$R^n \setminus S = \{x \in R^n : x \notin S\},$$

then $\bar{x} \in$ rel bdy S, the ***relative boundary*** of S. When $V(S) = R^n$ we drop the term "relative". In this case, if int $S \neq \emptyset$, we say that S is ***solid***.

Remark. Compare the above with the discussion in Nash and Sofer, page 16. The development there reflects authors' desire to avoid using the terms ***relative*** interior and boundary. This leads to the kind of difficulties that arise in Exercises 7–9 on page 20.

Definition. If $S \subset R^n$, the ***convex hull*** of S is the intersection of all convex sets containing S. The convex hull of S will be denoted co (S). For all S, co(S) is convex.

This definition is an ***exterior*** characterization of the convex hull of a set. The convex hull of the set S has an ***interior*** characterization as well. Indeed, the convex hull of S can be shown to equal to the set of all convex combinations of finitely many points belonging to S.

Definition. The points $x^0, x^1, \ldots, x^m$ are said to be in ***general position*** if the vectors $x^1 - x^0, \ldots, x^m - x^0$ are linearly independent. (Another way to express the fact that the points $x^0, x^1, \ldots, x^m$ are in general position is to say that dim $V(\{x^0, x^1, \ldots, x^m\}) = m$.)

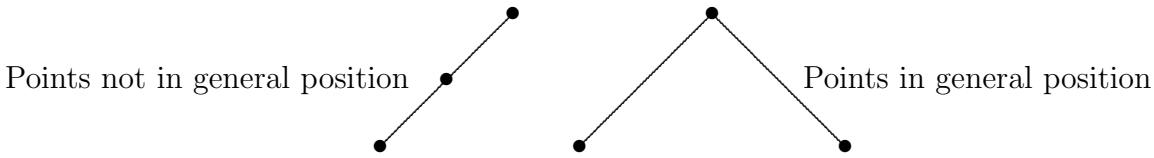


Figure 2.4

The convex hull of a finite set of points is called a ***polytope***. An ***m-simplex*** is the convex hull of $m+1$ points in general position. Hence an m-simplex is a particular kind of polytope.

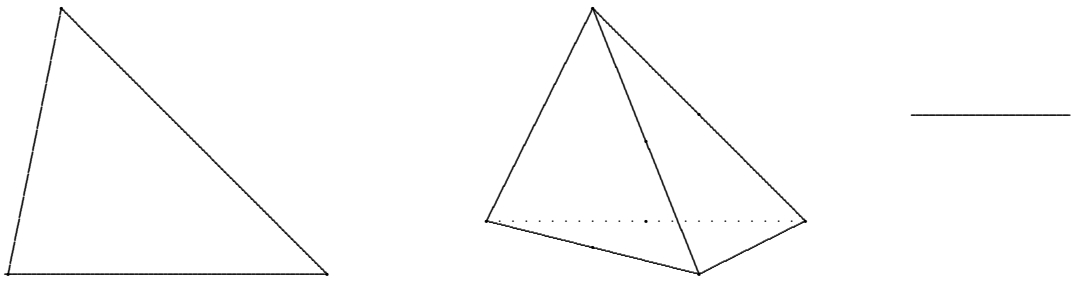


Figure 2.5

Lemma. If S is an m-simplex in R^n, then rel int $S \neq \emptyset$.

Proof. Let $S = \text{co}\,(\{x^0, x^1, \ldots, x^m\})$ and define

$$\bar{x} = \frac{1}{m+1} \sum_{i=0}^{m} x^i.$$

(The point $\bar{x}$ is the ***centroid*** of S.) Clearly $\bar{x} \in V(S)$. Let $V^i = V(\{x^0, \ldots, x^{i-1}, x^{i+1}, \ldots, x^m\})$ and let $\delta_i = \min_{x \in V^i} \|x - \bar{x}\|$. (See Figure 2.6 below.) Then $\delta_i > 0$ for all i. Take $\delta = \min_{0 \leq i \leq m} \delta_i$. Then we have $(B_\delta(\bar{x}) \cap V(S)) \subset S$. □

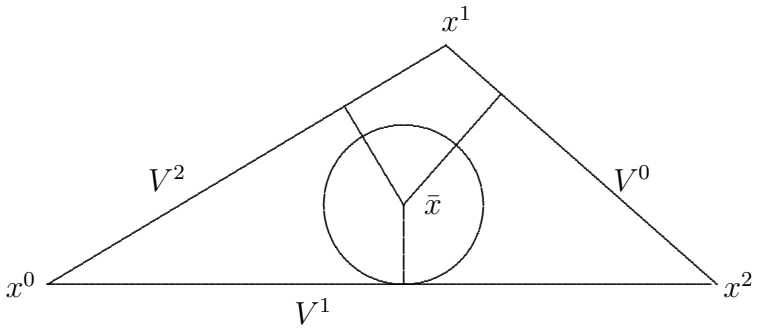


Figure 2.6

Theorem. Every convex set C of dimension $m \geq 1$ has a nonempty relative interior.

Proof. By hypothesis, dim $C \geq 1$, so there exist $m+1$ points $x^0, x^1, \ldots, x^m$ in C such that co $(\{x^0, x^1, \ldots, x^m\})$ is an m-simplex. Since co $(\{x^0, x^1, \ldots, x^m\}) \subset C$ (by convexity), and rel int co $(\{x^0, x^1, \ldots, x^m\}) \neq \emptyset$, it follows that rel int $C \neq \emptyset$. □

Theorem. If C is convex, $x \in$ rel int C, $y \in C$ and $x \neq y$, then for all $\alpha \in [0, 1)$,

$$z = (1 - \alpha)x + \alpha y \in \text{rel int}\, C.$$

Proof. We may assume that $\alpha > 0$. Since $x \in$ rel int C, there exists a number $\delta > 0$ such that $(B_\delta(x) \cap V(C)) \subset C$. Note that $z - y = (1 - \alpha)(x - y)$, so

$$0 < \frac{\|z - y\|}{\|x - y\|} = 1 - \alpha < 1.$$

Let $\epsilon = \delta(1 - \alpha)$. Then $(B_\epsilon(z) \cap V(C)) \subset C$. Indeed, let $w \in B_\epsilon(z) \cap V(C)$ be arbitrary. Define $u = x + \frac{1}{1-\alpha}(w - z)$.

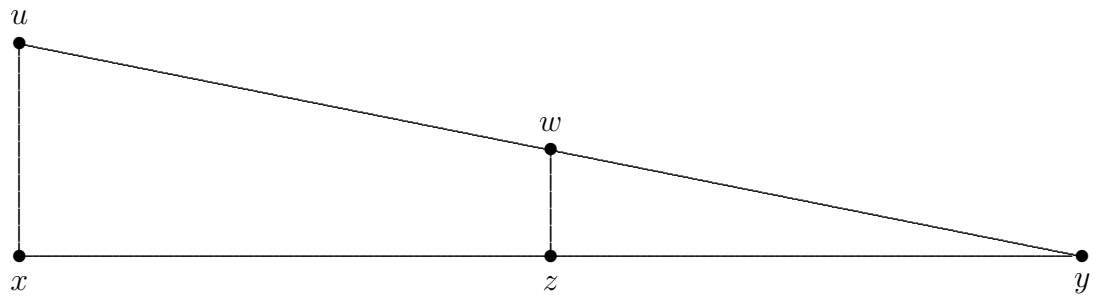


Figure 2.7

Then

$$\|u - x\| = \frac{1}{1-\alpha}\|w - z\| < \frac{1}{1-\alpha}\delta(1-\alpha) = \delta.$$

Hence $u \in (B_\delta(x) \cap V(C)) \subset C$ since $x \in \text{rel int}\, C$. Now (by definition of u)

$$w = (1-\alpha)u + z - (1-\alpha)x = (1-\alpha)u + \alpha y \in C.$$

Hence $z \in \text{rel int}\, C$. □

One implication of this theorem is that the relative interior of a convex set is itself a convex set.

Theorem. A polytope $\text{co}\,(\{x^1, \ldots, x^m\})$ is a compact set.

Proof. This is a forthcoming exercise. □

We come now to an interior representation of a convex set.

Lemma. A set $C \subset R^n$ is convex if and only if it contains every convex combination of every finite subset of its elements.

Proof. If C contains every convex combination of every finite subset of its elements, then it does so for sets of cardinality two, and hence C must be convex.

Conversely, suppose C is convex. Let $x^1, \ldots, x^m$ be arbitrary elements of C and consider an arbitrary convex combination $x = \sum_{i=1}^m \alpha_i x^i$ of these points. In the present situation, we may assume that $0 < \alpha_i < 1$ for $i = 1, \ldots, m$. The case of $m = 2$ is covered by the definition of convexity. Inductively, assume that any convex combination of $m - 1 \geq 2$ points of C belongs to C. To prove this for m points, we may write

$$x = \alpha_1 x^1 + \cdots + \alpha_m x^m = \alpha_1 x^1 + (1-\alpha_1)\left[\frac{\alpha_2}{1-\alpha_1}x^2 + \cdots + \frac{\alpha_m}{1-\alpha_1}x^m\right].$$

By the inductive hypothesis, the term in square brackets belongs to C because $x^2, \ldots, x^m \in C$ and

$$\frac{\alpha_2}{1-\alpha_1} + \cdots + \frac{\alpha_m}{1-\alpha_1} = 1, \quad \frac{\alpha_i}{1-\alpha_1} \geq 0 \quad i = 2, \ldots, m.$$

It is now clear that x has been expressed as a convex combination of *two* points of C and hence must belong to C. $\square$

The following result tells us that if S is an arbitrary subset of R^n, the external representation of $\text{co}(S)$ is the same as its internal representation (via convex combinations of finite sets of points in S).

Theorem. If $S \subset R^n$, then $\text{co}\,(S)$ is the set T of all convex combinations of finitely many points in S.

Proof. Let $C = \text{co}\,(S)$. The object is to show that $C = T$. Now, C contains S and all convex combinations of finite sets of points belonging to S. That is, $T \subset C$. Since $S \subset T$, the proof will be complete once we show that T is a convex set.

Let x and y belong to T. Then

$$x = \textstyle\sum_{i=1}^m \alpha_i x^i, \quad \sum_{i=1}^m \alpha_i = 1, \quad \alpha_i \geq 0 \quad i = 1, \ldots, m$$

$$y = \textstyle\sum_{j=1}^p \mu_j y^j, \quad \sum_{j=1}^p \mu_j = 1, \quad \mu_j \geq 0 \quad j = 1, \ldots, p$$

where all the points x^i and y^j belongs to S. Let θ be a number between 0 and 1 and consider the convex combination

$$z = \theta x + (1-\theta)y = \sum_{i=1}^m \theta\alpha_i x^i + \sum_{j=1}^p (1-\theta)\mu_j y^j.$$

Defining

$$\theta_i = \theta\alpha_i, \qquad z^i = x^i, \quad i = 1, \ldots, m,$$

$$\theta_{m+j} = (1-\theta)\mu_j, \quad z^{m+j} = y^j, \quad j = 1, \ldots, p,$$

we may write $z = \sum_{k=1}^{m+p} \theta_k z^k$ where, clearly, for all $k = 1, \ldots, m+p$

$$z^k \in S, \quad \theta_k \geq 0, \quad \text{and} \quad \sum_{k=1}^{m+p} \theta_k = 1.$$

This shows that the point z belongs to T which therefore must be convex. $\square$

Note that if $x \in R^n$ and $x^0, x^1, \ldots, x^m \in R^n$, then $x \in \text{co}\{x^0, x^1, \ldots, x^m\}$ if and only if there exists nonnegative scalars $\lambda_0, \lambda_1, \ldots, \lambda_m$ such that

$$\begin{bmatrix} x^0 \\ 1 \end{bmatrix} \lambda_0 + \begin{bmatrix} x^1 \\ 1 \end{bmatrix} \lambda_1 + \cdots + \begin{bmatrix} x^m \\ 1 \end{bmatrix} \lambda_m = \begin{bmatrix} x \\ 1 \end{bmatrix}.$$

Remark. The following theorem rests in part on a classical result called the ***Cauchy-Schwartz inequality*** which says that for all $a, b \in R^n$

$$|a^{\mathrm{T}} b| \leq \|a\| \cdot \|b\|$$

with equality if and only if a and b are linearly dependent.

Theorem. Let C be a nonempty closed convex subset of R^n and let $y \in R^n \setminus C$. Then there exists a unique point $\bar{x} \in C$ such that $\|\bar{x} - y\| \leq \|x - y\|$ for all $x \in C$. [Furthermore, $\bar{x}$ is the minimizing point (i.e., the point of C closest to y) if and only if $(y - \bar{x})^{\mathrm{T}}(x - \bar{x}) \leq 0$ for all $x \in C$.]

Proof. This theorem is concerned with minimizing the Euclidean distance from y to C. Note that this is equivalent to minimizing the square of the Euclidean distance from y to C.

Let $x^* \in C$ be arbitrary. Then $\{\, x \in C : \|x - y\| \leq \|x^* - y\| \,\}$ is compact. It follows that $\|x - y\|^2$ has a minimum over this set and hence over C. The minimizer $\bar{x}$ belongs to C (by definition) and is unique. (This follows from the Cauchy-Schwarz inequality.) The remaining condition is a consequence of the ***gradient inequality*** for differentiable convex functions which is given later. $\square$

The point $\bar{x}$ in the above theorem is called the ***projection*** of y onto C.

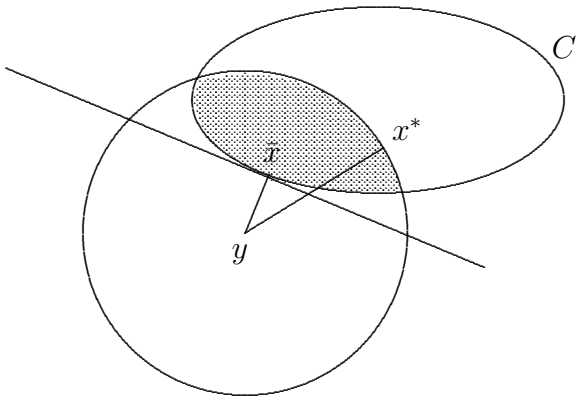
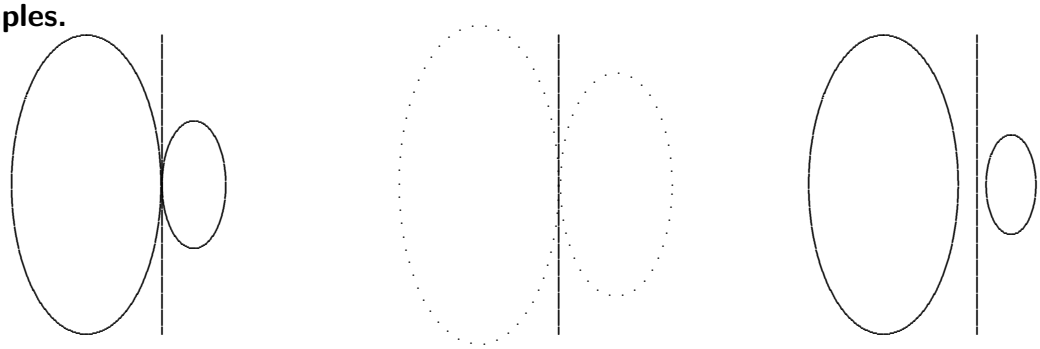


Figure 2.8

Separation and support

Definition. Let $H = \{ x : p^{\mathrm{T}}x = \alpha \}$ be a hyperplane in R^n. (Hence $p \neq 0$.) Let S_1 and S_2 be two nonempty subsets of R^n. Then H ***separates*** S_1 and S_2 if $p^{\mathrm{T}}x \geq \alpha$ for all $x \in S_1$ and $p^{\mathrm{T}}x \leq \alpha$ for all $x \in S_2$. ***Proper separation*** requires that $S_1 \cup S_2 \not\subset H$. (Notice that this would still allow *one* of the sets to lie in H.) The sets S_1 and S_2 are ***strictly separated*** by H if $p^{\mathrm{T}}x > \alpha$ for all $x \in S_1$ and $p^{\mathrm{T}}x < \alpha$ for all $x \in S_2$. (Notice that this does not prevent points of S_1 and S_2 from becoming arbitrarily close to H.) The sets S_1 and S_2 are ***strongly separated*** by H if there exists a number $\epsilon > 0$ such that $p^{\mathrm{T}}x > \alpha + \epsilon$ for all $x \in S_1$ and $p^{\mathrm{T}}x < \alpha - \epsilon$ for all $x \in S_2$.

Examples.



Properly separated sets Strictly separated sets Strongly separated sets

Figure 2.9

Notice how it is possible to place a ***slab*** (the region between two parallel hyperplanes (or lines in this case)) between strongly separated convex sets.

Theorem. If C is a nonempty closed convex set and $y \notin C$, then there exists a vector p and a real number α such that $p^{\mathrm{T}}y > \alpha$ and $p^{\mathrm{T}}x \leq \alpha$ for all $x \in C$.

Proof. Take $p = y - \bar{x}$ and $\alpha = \sup\{ p^{\mathrm{T}}x : x \in C \}$ as in the previous theorem. □

Corollary. If C is a closed convex proper subset of R^n, then C equals the intersection of all closed halfspaces containing C.

Proof. This is obvious. □

Definition. Let S be a nonempty subset of R^n, and let $\bar{x} \in \operatorname{bdy} S$. Then the hyperplane $H = \{ x : p^{\mathrm{T}}(x - \bar{x}) = 0 \}$ is a ***supporting hyperplane*** at $\bar{x}$ if $S \subset H^+$ or $S \subset H^-$ where

$$H^+ = \{ x : p^{\mathrm{T}}(x - \bar{x}) \geq 0 \},$$

$$H^- = \{ x : p^{\mathrm{T}}(x - \bar{x}) \leq 0 \}.$$

Proper support requires that $S \not\subset H$.

Theorem. If C is a nonempty convex set in R^n and $\bar{x} \in \operatorname{bdy} C$, then there exists a hyperplane that supports C at $\bar{x}$, i.e., there exists a vector $p \neq 0$ such that

$$p^{\mathrm{T}}(x - \bar{x}) \leq 0 \quad \text{for all } x \in \operatorname{cl} C.$$

Proof. This is clear. □

2.3 Convex functions

Definition. Let $C \subset R^n$ be a nonempty convex set. Then $f : C \longrightarrow R$ is ***convex*** (on C) if for all $x, y \in C$ and all $\alpha \in (0, 1)$

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y).$$

If strict inequality holds whenever $x \neq y$, then f is said to be ***strictly convex.*** The negative of a (strictly) convex function is called a ***(strictly) concave function.***

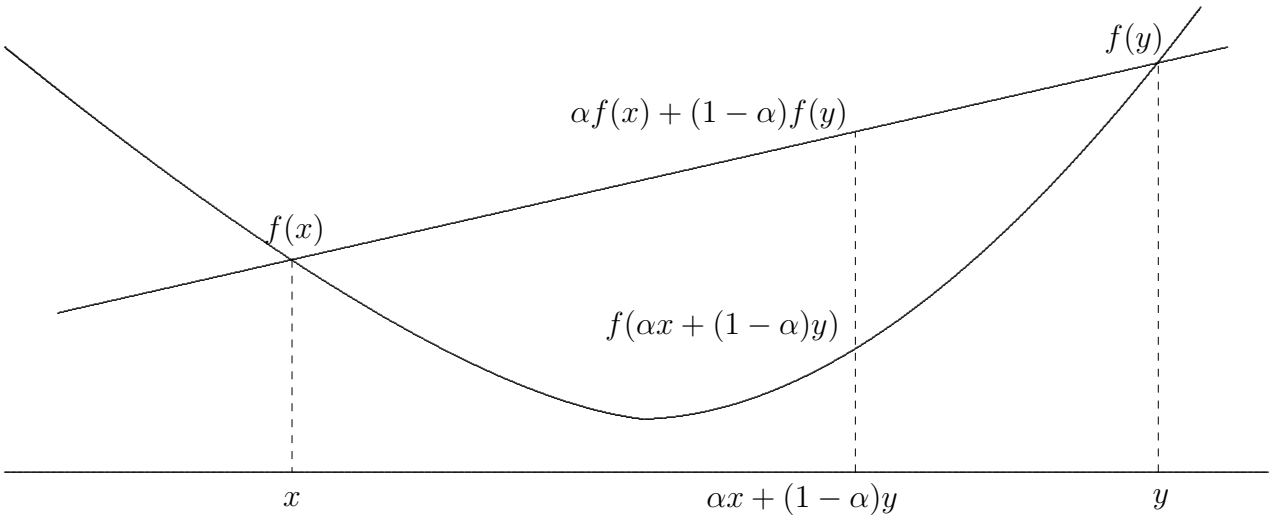


Figure 2.10

Convex functions are of interest in the context of optimization theory for several reasons. They arise frequently and have many significant properties, among which is the fact that a local minimum of a convex function (on a convex domain) is a global minimum. This makes it possible to use local conditions to test for optimality.

Examples.

1. Linear functions are both convex and concave.

2. Positive semidefinite quadratic forms, i.e., functions $x^{\mathrm{T}}Ax$ such that $x^{\mathrm{T}}Ax \geq 0$ for all x, are convex.

3. Positively-weighted sums of convex functions are convex.

Proposition. If $f : C \longrightarrow R$ is convex, then for all $\alpha \in R$ the ***level set*** $\{\, x \in C : f(x) \leq \alpha \,\}$ is convex.

Proof. This is an immediate consequence of the definition. □

Remark. The converse of this statement is not true. There are nonconvex functions whose level sets are all convex.

Theorem. Let C be a nonempty convex set in R^n, and let f be differentiable on C. Then f is convex on C if and only if for all $x, y \in C$:

$$f(y) \geq f(x) + (\nabla f(x))^{\mathrm{T}}(y - x). \quad \text{(This is called the } \textbf{\textit{gradient inequality}}.\text{)}$$

Proof. Suppose f is convex on C. Let x and y be arbitrary elements of C and let $\alpha \in (0, 1)$. Then

$$\begin{aligned} f(y) &\geq \frac{f[(1-\alpha)x + \alpha y] - (1-\alpha)f(x)}{\alpha} \\ &= f(x) + \frac{f[x + \alpha(y - x)] - f(x)}{\alpha}. \end{aligned}$$

Let $\alpha \longrightarrow 0^+$; then the right-hand side becomes $f(x) + (\nabla f(x))^{\mathrm{T}}(y - x)$.

Conversely, let x and y be distinct points in C and let $\alpha \in (0, 1)$. If the gradient inequality holds, we have

$$\begin{aligned} f(x) &\geq f[(1-\alpha)x + \alpha y] + (\nabla f[(1-\alpha)x + \alpha y])^{\mathrm{T}}\alpha(x - y), \\ f(y) &\geq f[(1-\alpha)x + \alpha y] + (\nabla f[(1-\alpha)x + \alpha y])^{\mathrm{T}}(1-\alpha)(y - x). \end{aligned}$$

Multiplying these inequalities by $(1 - \alpha)$ and α, respectively, and then adding the resulting inequalities, we obtain

$$(1-\alpha)f(x) + \alpha f(y) \geq f[(1-\alpha)x + \alpha y]$$

which proves that f is convex on C. $\square$

Remark. Let f be differentiable on the convex set C. Then f is strictly convex on C if and only if strict inequality holds in the gradient inequality for all pairs of distinct points x and y in C.

Example. This result can be used to prove that the univariate function $f(x) = \frac{1}{x}$ is strictly convex when C is the positive real line. We need to prove that for all distinct real numbers $x, y > 0$, we have

$$\frac{1}{y} > \frac{1}{x} - \frac{1}{x^2}(y - x).$$

This follows from the observation that for distinct $x, y > 0$,

$$\frac{1}{yx^2}(y - x)^2 > 0.$$

Proposition. Let f be a differentiable convex function on C. Then solving the problem

$$\text{minimize } f(x) \text{ subject to } x \in C \tag{1}$$

is equivalent to the ***variational inequality problem***

$$\text{Find } \bar{x} \in C \text{ such that } (\nabla f(\bar{x}))^{\mathrm{T}}(y - \bar{x}) \geq 0 \quad \text{for all } y \in C. \tag{2}$$

Proof. If $\bar{x}$ solves (2), then for all $y \in C$, the gradient inequality gives

$$f(y) - f(\bar{x}) \geq (\nabla f(\bar{x}))^{\mathrm{T}}(y - \bar{x}) \geq 0.$$

Conversely, if $\bar{x}$ solves (1), then $\bar{x}$ must also solve (2), for suppose there exists $\tilde{y} \in C$ such that

$$(\nabla f(\bar{x}))^{\mathrm{T}}(\tilde{y} - \bar{x}) < 0.$$

This says that f *decreases* when its argument moves away from $\bar{x}$ in the direction $\tilde{y} - \bar{x}$. In particular, there exists a real number $\theta \in (0, 1)$ such that

$$f[(1 - \theta)\bar{x} + \theta\tilde{y}] = f[\bar{x} + \theta(\tilde{y} - \bar{x})] < f(\bar{x}).$$

But this contradicts the assumption that $\bar{x}$ minimizes f on C. $\square$

Theorem. Let f be twice continuously differentiable on the open convex set C. Then f is convex on C if and only if its Hessian matrix $\nabla^2 f(x)$ is positive semidefinite for all $x \in C$.

Proof. Let f be convex on C. Let $x \in C$ and $d \in R^n$ be arbitrary. For α sufficiently small $x + \alpha d \in C$ because C is open. Since f is twice continuously differentiable,

$$f(x + \alpha d) - f(x) - \alpha(\nabla f(x))^{\mathrm{T}} d = \frac{1}{2}\alpha^2 d^{\mathrm{T}} \nabla^2 f(x) d + \alpha^2 \beta(x, x + \alpha d)\|d\|$$

where $\lim_{\alpha \longrightarrow 0} \beta(x, x+\alpha d) = 0$. When α is sufficiently small, the left-hand side is nonnegative by the gradient inequality. Hence $\frac{1}{2} d^{\mathrm{T}} \nabla^2 f(x) d \geq 0$.

Conversely, suppose $\nabla^2 f(x)$ is positive semidefinite for all $x \in C$. Then with $x, y \in C$, Taylor's theorem implies that there exists $\alpha \in (0, 1)$ such that for $z = (1 - \alpha)x + \alpha y$:

$$f(y) - f(x) - (\nabla f(x))^{\mathrm{T}}(y - x) = \tfrac{1}{2}(y - x)^{\mathrm{T}} \nabla^2 f(z)(y - x).$$

The right-hand side is nonnegative, hence the gradient inequality holds for all $x, y \in C$. □

Remark. If $\nabla^2 f(x)$ is positive definite for all $x \in C$, then f is strictly convex on C. But the converse is false, as shown by the function $f(x) = x^4$ with domain $C = R$. This function is strictly convex and twice continuously differentiable on R, but $f''(0) = 0$ which is not positive definite.

References

J.M. Borwein and A.S. Lewis, *Convex Analysis and Nonlinear Optimization* Springer-Verlag, New York, 2000.

H.G. Eggleston [1958], *Convexity*, Cambridge University Press, Cambridge.

H. Guggenheimer [1977], *Applicable Geometry: Global and Local Convexity*, Robert E. Krieger Publishing Company, Huntington, New York.

D.G. Luenberger [1969], *Optimization by Vector Space Methods*, John Wiley & Sons, New York.

O.L. Mangasarian [1969], *Nonlinear Programming*, McGraw-Hill, New York. [Reprinted in softcover by SIAM Publications.]

R.T. Rockafellar [1970], *Convex Analysis*, Princeton University Press, Princeton, New Jersey.

A. Schrijver [1986], *Theory of Linear and Integer Programming*, Chichester: John Wiley & Sons.

D.M.Y. Sommerville [1958], *An Introduction to the Geometry of N Dimensions*, Dover, New York. [This is a reproduction of the 1929 edition, published in London by Metheun & Co., Ltd.]

J. Stoer and C. Witzgall [1970], *Convexity and Optimization in Finite Dimensions I*, Springer-Verlag, New York.

R. Webster [1994]. *Convexity*, Oxford University Press, Oxford.

2. CONVEX SETS AND FUNCTIONS (CONTINUED)

2.3 Convex functions (continued)

In Handout No. 2 we covered first- and second-order characterizations of convexity for suitably differentiable functions on convex sets. It should be pointed out, however, that not all convex functions are differentiable. A simple example of a nondifferentiable convex function is

$$f(x) = |x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x \leq 0 \end{cases}$$

the graph of which is shown in Figure 2.11.

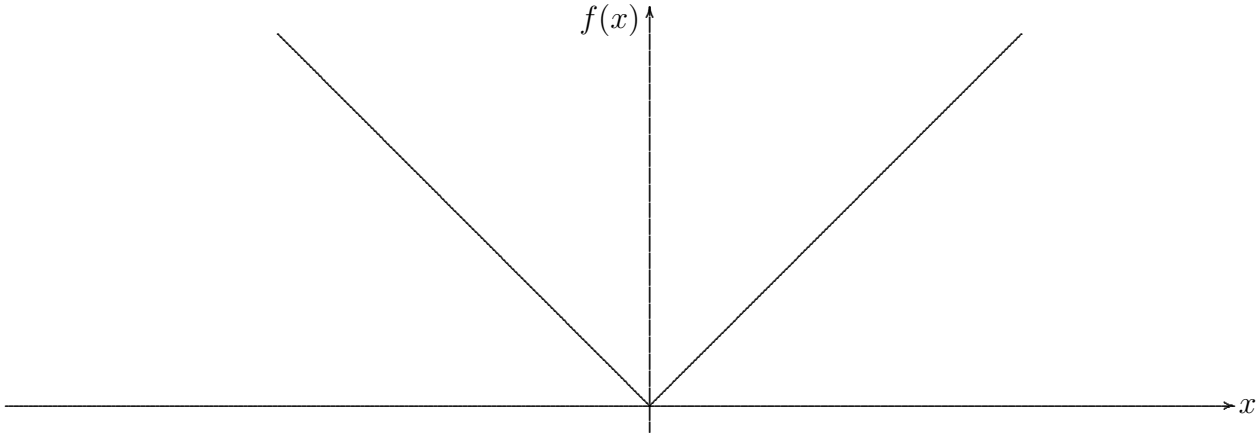


Figure 2.11

We can put this example into a more general framework. Note that for every $x \in R$, we have $|x| = \max\{x, -x\}$. That is to say, $f(x) = |x|$ is the (pointwise) maximum of two linear functions: $f_1(x) = x$ and $f_2(x) = -x$. More generally, now, we may consider a function that is the pointwise maximum of a finite set of linear functions (all of which have the same domain, R), say

$$F(x) = \max\{F_1(x), \ldots, F_m(x)\}.$$

Such a function would be convex as well as ***piecewise linear***. A piecewise linear function $\varphi(x)$ of a single variable, $x \in R$ (or perhaps an interval $[\ell, u] \subset R$), is one for which there is a set of ***breakpoints*** $x_1 < x_2 < \cdots < x_r$ in the domain of the function such that on the

closed interval defined by two consecutive breakpoints, the function value $\varphi(x)$ is given by an affine function (a constant times x plus a constant). A univariate piecewise linear function will be ***continuous*** if the function values on neighboring subintervals agree at their common breakpoint. This is illustrated in Figure 2.12 which exhibits two ***nonconvex*** piecewise linear functions.

Not continuous Continuous

Figure 2.12

More will be said later about functions defined in this way. Right now, though, it will be helpful to bring up an alternate way of looking at convex functions. This requires the introduction of a new term.

Definition. Let f be a real-valued function on $S \subset R^n$. The ***epigraph*** of f is the set

$$\operatorname{epi} f = \{(x, \mu) \in R^{n+1} : x \in S,\ \mu \geq f(x)\}.$$

This is just the set of points in R^{n+1} that lie on or above the graph of f.

Theorem. Let $C \subset R^n$ be a nonempty convex set. Then $f : C \longrightarrow R$ is convex function if and only if epi f is a convex subset of R^{n+1}.

Proof. This is an exercise. □

This theorem links convex functions to convex sets; it can be very useful in proving the convexity of certain functions. Here is one such instance.

Example. Let $\langle a^i, x\rangle + b_i = \sum_{j=1}^n a_{ij}x_j + b_i$ for $i = 1, \ldots, m$ be a given finite set of affine functions. Then

$$f(x) = \max_{1 \leq i \leq m} \{\langle a^i, x\rangle + b_i\}$$

is a convex function. The theorem makes it easy to see why this must be so. Indeed the epigraph of f is just the intersection of the halfspaces associated with the inequalities $\mu \geq \langle a^i, x\rangle + b_i$ $(i = 1, \ldots, m)$, each of which a convex set. Hence the epigraph of f is convex.

We have seen that a convex function need not be differentiable. As a matter of fact, it is not even necessary for a convex function to be continuous. This observation is illustrated by

the function $f : [0,1] \longrightarrow R$ shown below:

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

It should be noticed, however, that the function in this example is continuous on the interior of its domain. This is a consequence of a general result. We first prove two lemmas.

Lemma (Jensen's inequality). If $f : C \longrightarrow R$ is convex, then

$$f(\lambda_1 x^1 + \cdots + \lambda_m x^m) \leq \lambda_1 f(x^1) + \cdots + \lambda_m f(x^m)$$

for any $x^1, \ldots, x^m \in C$ and any $\lambda_1 \geq 0, \ldots, \lambda_m \geq 0$ such that $\lambda_1 + \cdots + \lambda_m = 1$.

Proof. For $m = 2$ the inequality is just the definition of convexity. Arguing inductively, we now assume $m > 2$ and that the inequality holds for $m - 1$ points. The rest of the proof is just a matter of clever writing (as in the lemma on page 6 of Handout No. 2), namely

$$x = \lambda_1 x^1 + \cdots + \lambda_m x^m = \lambda_1 x^1 + (1 - \lambda_1) \left[\frac{\lambda_2}{1 - \lambda_1} x^2 + \cdots + \frac{\lambda_m}{1 - \lambda_1} x^m \right],$$

and then using the convexity of f together with the inductive hypothesis. □

Lemma. Let f be a convex function on the nonempty convex set C. Then f is bounded above on every nonempty compact convex set K contained in the relative interior of C.

Proof. First of all, note that by what we have already proved (see Handout 2, page 5) the set C has a nonempty relative interior. By the Heine-Borel Theorem[5], we may "cover" K by a finite collection of simplices each contained in C. By the finiteness of this covering, it suffices to prove that f is bounded above on a simplex. Suppose the simplex is the convex hull of the points $x^1, \ldots, x^r$. Let

$$\mu = \max\{f(x^1), \ldots, f(x^r)\}.$$

Every point in the simplex is of the form $x = \sum_{i=1}^r \lambda_i x^i$ where $\sum_{i=1}^r \lambda_i = 1$ and $\lambda_i \geq 0,\ i = 1, \ldots, r$. We can then write

$$f(x) = f\left(\sum_{i=1}^r \lambda_i x^i \right) \leq \sum_{i=1}^r \lambda_i f(x^i) \leq \sum_{i=1}^r \lambda_i \mu = \mu$$

[5]In some treatments of real analysis, a compact set is defined as one for which every "open cover" has a finite subcover. (See Royden [1968, p.157] and Rudin [1964, p.32], for example.) The Heine-Borel Theorem then asserts the equivalence of this definition with the one we customarily use in these notes.

This shows that f is bounded above on the simplex. Since the given compact convex set is covered by finitely many simplices, there is a finite upper bound on f. □

We shall use this lemma in proving the following theorem.

Theorem. Every convex function is continuous on the relative interior of its domain.

Proof. Let C be the domain of the convex function f and let x^0 be a point in the relative interior of C. By a simple translation and a minor redefinition of f we may assume that $x^0 = 0$ and $f(0) = 0$. Now let $N(0, \lambda) \subset C$ denote a relatively open neighborhood of x^0. (This means that $N(0, \lambda)$ is the intersection of an open ball of radius λ and center x^0 with the carrying plane of C.) For an arbitrary $\varepsilon \in (0, 1]$, consider $x \in N(0, \lambda)$ such that $\|x\| \leq \varepsilon\lambda$. We can then write

$$f(x) = f\left[(1-\varepsilon)0 + \varepsilon\left(\frac{1}{\varepsilon}x\right)\right] \leq \varepsilon f\left(\frac{1}{\varepsilon}x\right) \leq \varepsilon\mu$$

for some $\mu > 0$. Furthermore,

$$\begin{aligned} 0 = f(0) &= f\left[\frac{1}{1+\varepsilon}x + \left(\frac{\varepsilon}{1+\varepsilon}\right)\left(\frac{1}{\varepsilon}\right)(-x)\right] \\ &\leq \left(\frac{1}{1+\varepsilon}\right) f(x) + \left(\frac{\varepsilon}{1+\varepsilon}\right) f\left[\frac{1}{\varepsilon}(-x)\right]. \end{aligned}$$

Hence $f(x) \geq -\varepsilon\mu$. Together, these inequalities imply $|f(x)| \leq \varepsilon\mu$. This proves the continuity of f at x^0. □

A consequence of this theorem is that if a convex function fails to be continuous, then the points of discontinuity must lie on the relative boundary of the function's domain.

Some univariate convex functions[6]
The second-order criteria for convexity can be used to prove the following:

1. $f(x) = e^{\alpha x}$ (for all real α);
2. $f(x) = x^p$ if $x \geq 0$ and $1 \leq p < \infty$;
3. $f(x) = -x^p$ if $x \geq 0$ and $0 \leq p \leq 1$;
4. $f(x) = x^p$ if $x > 0$ and $-\infty < p \leq 0$;
5. $f(x) = -\log x$ if $x > 0$.

With regard to the last of these, one can use Jensen's inequality to demonstrate the inequality between the arithmetic mean and the geometric mean. The reasoning goes like this. Let

[6]Based on Rockafellar [1970, Section 4].

$x_1, \ldots, x_m$ be a set of positive real numbers and let $x = \lambda_1 x_1 + \cdots + \lambda_m x_m$ be an arbitrary convex combination of these m scalars. Using the convexity of $f(x) = -\log x$ and Jensen's inequality, we get

$$-\log(\lambda_1 x_1 + \cdots + \lambda_m x_m) \leq -\lambda_1 \log x_1 - \cdots - \lambda_m \log x_m.$$

Multiplying through this inequality by -1 and taking exponentials of both sides yields

$$\lambda_1 x_1 + \cdots + \lambda_m x_m \geq x_1^{\lambda_1} \cdots x_m^{\lambda_m}.$$

In particular, when $\lambda_1 = \cdots = \lambda_m = \frac{1}{m}$ we get

$$\tfrac{1}{m}(x_1 + \cdots + x_m) \geq (x_1 \cdots x_m)^{1/m}.$$

This is the ***arithmetic-mean/geometric-mean inequality.***

Some general facts about (multivariate) convex functions

From the second-order characterization of convex functions, it follows that a positive semidefinite ***quadratic form*** $f(x) = x^{\mathrm{T}}Qx$ is convex on R^n. Note that the Hessian matrix of f is just a constant matrix: Q. It follows at once that a ***quadratic function***, i.e., one of the form

$$f(x) = \tfrac{1}{2}x^{\mathrm{T}}Qx + c^{\mathrm{T}}x + \kappa,$$

is convex if and only if Q is a positive semidefinite matrix.

Remark. In most linear algebra books, discussions of positive semidefiniteness include a symmetry assumption. Thus, in talking about the positive semidefiniteness of a quadratic form, $x^{\mathrm{T}}Qx$, or of a matrix, Q, it is assumed that $Q = Q^{\mathrm{T}}$. Notice that the definition $x^{\mathrm{T}}Qx \geq 0$ for all x does not require a symmetry assumption on Q. For instance

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \geq 0 \quad \text{for all } x_1 \text{ and } x_2.$$

But when the values of the quadratic form are essentially the issue, it is not restrictive to assume that Q is symmetric because for any square matrix Q, we have the identity:

$$x^{\mathrm{T}}Qx \equiv \tfrac{1}{2}x^{\mathrm{T}}(Q + Q^{\mathrm{T}})x.$$

and the matrix $\frac{1}{2}(Q+Q^{\mathrm{T}})$ is symmetric. It is called the ***symmetric part*** of Q. One motivation for the symmetry assumption is that any real symmetric matrix has real ***eigenvalues***, and a symmetric matrix is positive semidefinite[7] if and only if its eigenvalues are all nonnegative. This is a useful characterization of positive semidefiniteness. There are others as well. One

[7] One says that a square matrix Q is positive semidefinite if the associated quadratic form $x^{\mathrm{T}}Qx$ is positive semidefinite (and vice versa).

of these is that a symmetric matrix is positive semidefinite if and only if each of its principal minors is nonnegative. Still another is that a symmetric matrix is positive semidefinite if and only if it equals AA^{T} for some matrix A. This sort of representation is called a ***matrix factorization***. Matrix factorizations are briefly discussed in Nash and Sofer [1996, Appendix A.6]. For much more, see Golub and Van Loan [1983]. Two particular matrix factorizations are noteworthy. One is called the LDL^{T} factorization. In it, the matrix L is lower triangular and D is diagonal. Any symmetric matrix possesses such a factorization. But note that when the diagonal elements of D are nonnegative, it is possible to take their square roots. Thus, if $D = \operatorname{diag}(d_{11}, \ldots, d_{nn})$ and all the d_{ii} are nonnegative, then we can define $\hat{d}_{ii} = \sqrt{d_{ii}}$ for $i = 1, \ldots, n$ and obtain $LDL^{\mathrm{T}} = \hat{L}\hat{L}^{\mathrm{T}}$ where $\hat{L} = L\hat{D}$ and $\hat{D} = \operatorname{diag}(\hat{d}_{11}, \ldots, \hat{d}_{nn})$. The expression $Q = \hat{L}\hat{L}^{\mathrm{T}}$ is called the ***Cholesky factorization***[8] of Q. Another criterion for positive semidefiniteness of a symmetric matrix is the nonnegativity of all ***principal minors*** (i.e., the determinants of all principal submatrices, the submatrices formed by deleting a set of rows and the same set of columns). For example, the principal minors of $Q = Q^{\mathrm{T}} \in R^{3\times 3}$ are $\det[\ \ ] = 1,\ q_{11},\ q_{22},\ q_{33}$,

$$\det \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix}, \quad \det \begin{bmatrix} q_{11} & q_{13} \\ q_{31} & q_{33} \end{bmatrix}, \quad \det \begin{bmatrix} q_{22} & q_{23} \\ q_{32} & q_{33} \end{bmatrix}, \quad \text{and} \quad \det Q.$$

References

H.G. Eggleston [1958], *Convexity*, Cambridge University Press, Cambridge.

G.H. Golub and C.F. Van Loan [1983], *Matrix Computations*, North Oxford Academic, Oxford.

S. Karlin [1959], *Mathematical Methods and Theory in Games, Programming, and Economics, Volume 1*, Addison-Wesley, Reading, Mass.

D.G. Luenberger [1969], *Optimization by Vector Space Methods*, John Wiley & Sons, New York.

O.L. Mangasarian [1969], *Nonlinear Programming*, McGraw-Hill, New York. [Reprinted in softcover by SIAM Publicationns.]

S.G. Nash and A. Sofer [1996], *Linear and Nonlinear Programming*, McGraw-Hill, New York.

R.T. Rockafellar [1970], *Convex Analysis*, Princeton University Press, Princeton, New Jersey.

H.L. Royden [1968], *Real Analysis*, Macmillan, New York.

W. Rudin [1964], *Principles of Mathematical Analysis*, McGraw-Hill, New York.

J. Stoer and C. Witzgall [1970], *Convexity and Optimization in Finite Dimensions I*, Springer-Verlag, New York.

R. Webster [1994]. *Convexity*, Oxford University Press, Oxford.

[8]The Cholesky factorization is named for André Louis Cholesky (1875-1918), a French geodesist.

2.4 Polyhedral convex sets

Within the family of convex sets in R^n, the ones that occur most frequently and are the most tractable are the ***polyhedral convex sets***. Such sets arise as the feasible regions of linearly constrained optimization problems. This class includes linear and quadratic programming problems and many others.

Recall that a ***hyperplane*** is the solution set of a (nontrivial) linear equation. Thus, if $p \neq 0$ is an n-vector, and α is a scalar (possibly zero), then the linear equation $p^{\mathrm{T}}x = \alpha$ defines a ***hyperplane***

$$H = \{x \in R^n : p^{\mathrm{T}}x = \alpha\}.$$

Such sets are clearly convex.

A hyperplane is either a linear subspace (as when $\alpha = 0$) or else is a translate of a subspace. Suppose for example that $\alpha \neq 0$ in the definition of H above. Let us define

$$H_0 = \{x \in R^n : p^{\mathrm{T}}x = 0\}.$$

Then H_0 is a linear subspace (of dimension $n-1$). Now notice that the point

$$x^* = \frac{\alpha}{p^{\mathrm{T}}p}p$$

belongs to H. Furthermore, for arbitrary $x \in H_0$ the point $x + x^* \in H$. Conversely, if $y \in H$, then $x = y - x^* \in H_0$. This discussion shows that H is a ***translate*** of H_0. In particular,

$$H = H_0 + \{x^*\}.$$

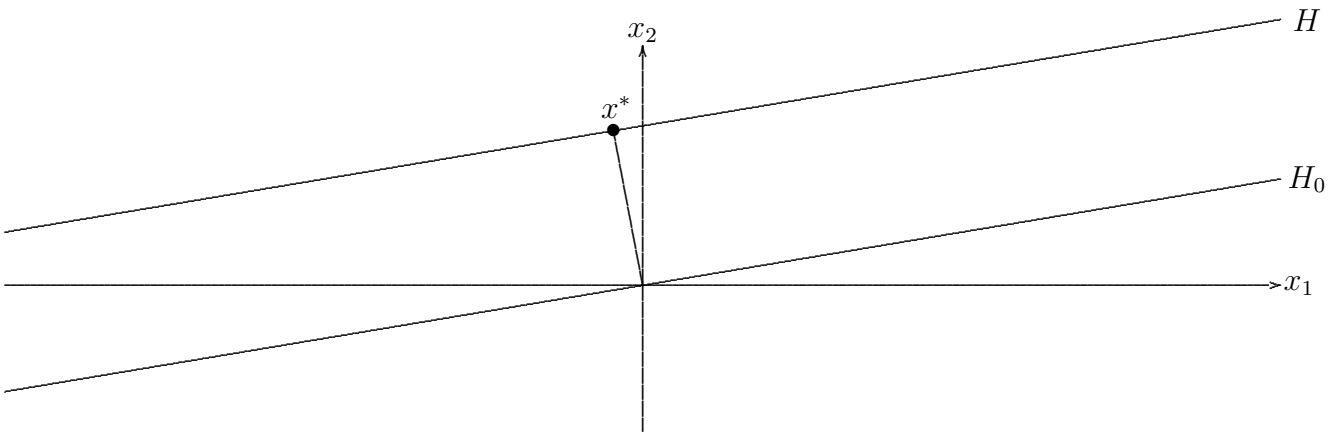


Figure 2.13

This representation is not unique. The point x^* used here was chosen as the projection of the origin on H. Notice that when $\|p\| = 1$, the scalar α equals the distance from the origin in R^n to the hyperplane H.

Associated with every hyperplane $H = \{x : p^{\mathrm{T}}x = \alpha\}$ is a pair of linear inequalities:

$$p^{\mathrm{T}}x \leq \alpha \quad \text{and} \quad p^{\mathrm{T}}x \geq \alpha.$$

The solution sets of these linear inequalities are called ***halfspaces***; we denote them (somewhat arbitrarily)

$$H^+ = \{x : p^{\mathrm{T}}x \leq \alpha\},$$

$$H^- = \{x : p^{\mathrm{T}}x \geq \alpha\}.$$

Note that $H^+ \cup H^- = R^n$, so the name halfspace seems appropriate. Notice also that every hyperplane is the intersection of two halfspaces. In particular

$$H = H^+ \cap H^-.$$

In algebraic terms, this says that the linear equation $p^{\mathrm{T}}x = \alpha$ is equivalent to (i.e., has the same solutions as) the pair of linear inequalities

$$p^{\mathrm{T}}x \leq \alpha \quad \text{and} \quad p^{\mathrm{T}}x \geq \alpha.$$

The significance of this observation is shown in the following definition.

Definition. A ***polyhedron*** (or ***polyhedral convex set***) is the intersection of a finite set of linear inequalities.

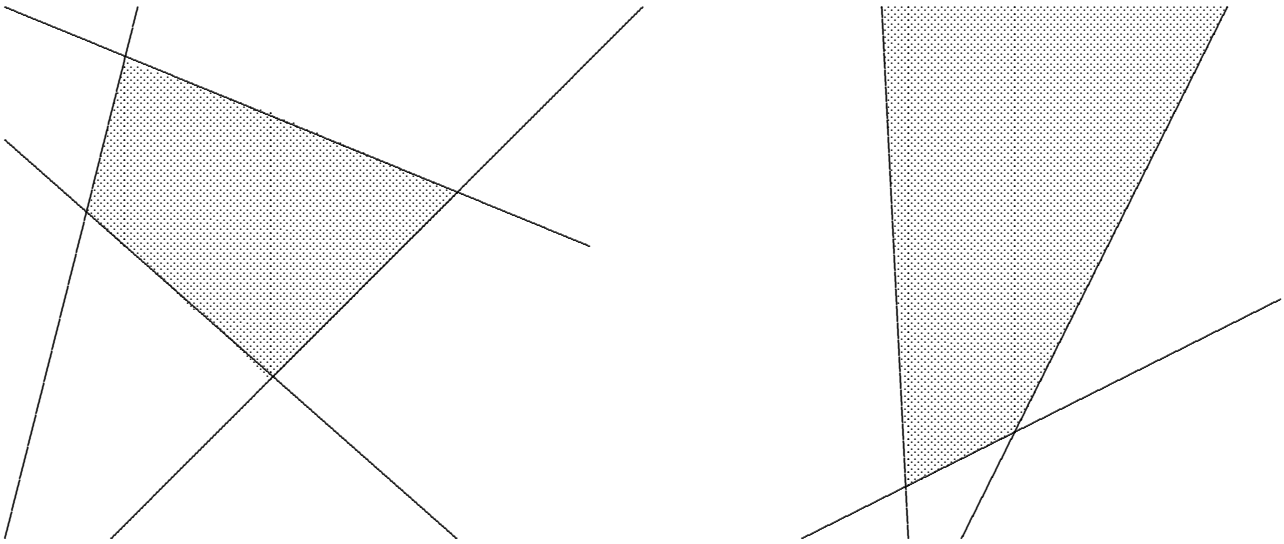


Figure 2.14

Remark. The kind of linear inequalities we have in mind in this definition are of the "weak" type ("$\leq$" or "$\geq$") rather than the "strict" type ("$<$" or "$>$"). Notice that H^+ and H^- are ***closed sets*** since their complements are open sets.

One sees immediately that an inequality such as $q^{\mathrm{T}}x \leq \beta$ is equivalent to one of the form $p^{\mathrm{T}}x \geq \alpha$ where $p = -q$ and $\alpha = -\beta$. For this reason, it is not restrictive to say that a polyhedron is the solution set of a finite system of linear inequalities such as

$$p_i^{\mathrm{T}}x \geq \alpha_i,\ i = 1, \ldots, m.$$

This ***system*** of linear inequalities can be represented in matrix form as

$$P^{\mathrm{T}}x \geq a$$

where $P^{\mathrm{T}} \in R^{m \times n}$ (that is, P^{T} is a matrix of m rows and n columns) and $a \in R^m$.

Proposition 1. Every polyhedron is a closed convex set.

Proof. Let S be a polyhedron, say

$$S = \{x : p_i^{\mathrm{T}}x \geq a_i\ i = 1, \ldots, m\}$$

Then S is the intersection of m closed convex sets (halfspaces). Since the intersection of closed sets is closed and the intersection of convex sets is convex, it follows that S is closed and convex. (It may, however, be empty.) $\square$

There is a slightly more direct way to prove the convexity assertion. If S contains less than 2 points, then it is clearly convex. Assume that S contains at least two distinct points x and y. For an arbitrary $\theta \in (0, 1)$, let $z = \theta x + (1 - \theta)y$. Then we have

$$P^{\mathrm{T}}[\theta x + (1 - \theta)y] = \theta P^{\mathrm{T}}x + (1 - \theta)P^{\mathrm{T}}y \geq \theta a + (1 - \theta)a = a.$$

This again shows that S is convex.

Simple examples (such as line segments and halfspaces) show that a polyhedron may or may not be bounded. In the literature, bounded polyhedra are called ***polytopes***. Recall that in Handout No. 2 we defined a polytope to be the convex hull of a finite set of points. Some further development is needed before we can justify the use of the term "polytope" for a bounded polyhedron. This issue is related to the matter of internal and external representations of convex sets. In fact, our definition of polyhedron is an external one.

2.4.1 Extreme points

For the moment let S denote any convex set, not necessarily a polyhedron.

Definition. A point $\hat{x} \in S$ is an ***extreme point*** of S if it cannot be expressed as a convex combination of two other distinct points of S.

A convex set need not have any extreme points at all. Certainly an open ball (or any relatively open convex set) has no extreme points. On the other hand, a convex set (such as a closed ball) may have infinitely many extreme points. Some convex sets such as polyhedra have finitely many extreme points, but a convex set with finitely many extreme points is not necessarily a polyhedron.

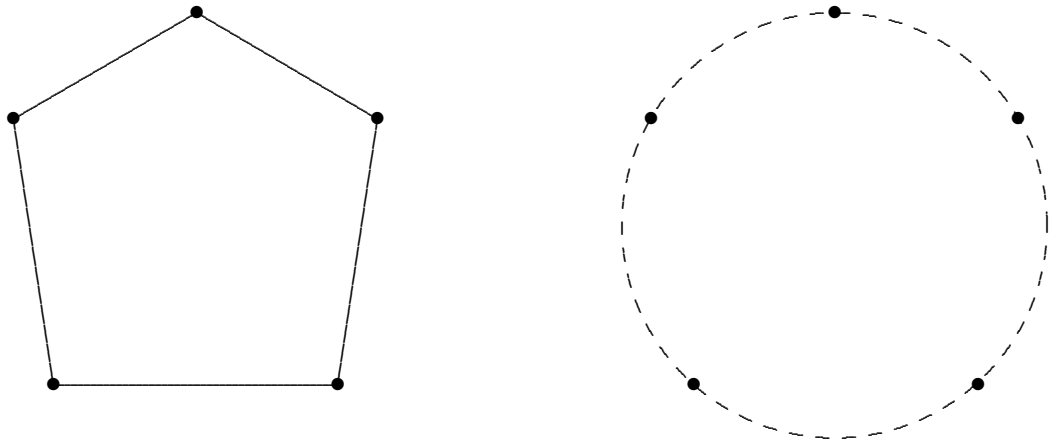


Figure 2.15

One thing that makes extreme points interesting is the fact that certain kinds of optimization problems have extreme point solutions. The most notable example of this is the ***linear programming problem*** which in so-called ***standard form*** is

$$\begin{array}{ll} \text{minimize} & c^{\mathrm{T}}x \\ \text{subject to} & Ax = b \\ & x \geq 0 \end{array}$$

Notice that the ***feasible region***

$$S = \{x : Ax = b,\ x \geq 0\}$$

is mathematically equivalent to the intersection of the solution sets of a finite collection of linear inequalities and hence is polyhedral. We shall not express the problem that way just now because it is important to bring out the connection between the extreme points of S and certain types of nonnegative solutions of the system $Ax = b$. To accomplish this (and many other things), it will be helpful to introduce a little notation.

Let A be a real $m \times n$ matrix. The elements of A are ordinarily denoted a_{ij} where $i = 1, \ldots, m$ and $j = 1, \ldots, n$. The matrix A is just a rectangular array having m rows and n columns. It is very useful to be able to refer to these rows and columns in a precise way, and that is the purpose of the notational convention described here. For $i = 1, \ldots, m$ let $A_{i\bullet}$ denote the i-th

row of A. In like manner, for $j = 1, \ldots, n$ let $A_{\bullet j}$ denote the j-th column of A. These ideas can be generalized. Let $\sigma \subset \{1, \ldots, m\}$ and $\tau \subset \{1, \ldots, n\}$. Then $A_{\sigma\tau}$ is the ***submatrix*** $[a_{ij}]$ of A such that $i \in \sigma$ and $j \in \tau$. Then, for example, $A_{\bullet j}$ corresponds to the case where $\sigma = \{1, \ldots, m\}$ and $\tau = \{j\}$.

Example. If we let

$$A = \begin{bmatrix} 4 & -1 & 2 & 5 \\ -3 & 0 & -1 & 9 \\ 6 & 1 & 4 & -5 \end{bmatrix},$$

then we have

$$A_{2\bullet} = \begin{bmatrix} -3 & 0 & -1 & 9 \end{bmatrix} \quad \text{and} \quad A_{\bullet 3} = \begin{bmatrix} 2 \\ -1 \\ 4 \end{bmatrix}.$$

In terms of this notation, we can write the system of equations $Ax = b$ as

$$\sum_{j=1}^{n} A_{\bullet j} x_j = b.$$

The feasibility condition for the linear programming problem asks for the (right-hand side) vector b to be expressed as a nonnegative linear combination of the columns of A.

Definition. A vector $\bar{x} \in S = \{x : Ax = b,\ x \geq 0\}$ gives a ***representation*** of the vector b as a nonnegative linear combination of columns of A. The representation $\bar{x}$ ***uses*** column j if and only if $\bar{x}_j > 0$.

Definition. The ***support*** of a vector $\bar{x}$ is the set of indices (subscripts) j such that $\bar{x}_j \neq 0$. The support of $\bar{x}$ is denoted supp $\bar{x}$.

This is a special case of a more general concept.[9]

Theorem 1. Let $A \in R^{m \times n}$ be a nonzero matrix and let $b \in R^m$ be a nonzero vector. If $S = \{x : Ax = b,\ x \geq 0\} \neq \emptyset$, then there exists an element of S that uses only linearly independent columns of A in representing b.

Proof. Let $x^0 \in S$ be given. The hypothesis $b \neq 0$ implies that $x^0 \neq 0$. In other words, x^0 has a nonempty support, say τ. If $A_{\bullet\tau}$ has linearly independent columns, we are done. If the columns of $A_{\bullet\tau}$ are linearly dependent, then there exists a nonzero vector $\tilde{x}$ whose support is a subset of τ such that $A_{\bullet\tau}\tilde{x}_\tau = 0$. Without loss of generality, we may assume that $\tilde{x}_\tau$ has at least one positive component. (Otherwise, use $-\tilde{x}_\tau$ which does have a positive component.)

[9] In general, the *support* of a numerical function is the subset of its domain at which the function value is nonzero. Vectors and matrices can be regarded as functions defined on their index sets. For example, if $x \in R^n$, then x is a function defined on $\{1, \ldots, n\}$ and taking values in R. In particular, for $j = 1, \ldots, n$, we have $x(j) = x_j$.

Then for sufficiently small $\lambda > 0$ the vector $x^0_\tau - \lambda\tilde{x}_\tau \geq 0$ and for some positive value of λ the vector $x^0_\tau - \lambda\tilde{x}_\tau$ will satisfy $A_{\bullet\tau}[x^0_\tau - \lambda\tilde{x}_\tau] = b$ and have a smaller support than x^0 does. This new vector can be used to replace x^0, and the process can be repeated. Eventually one obtains a vector that uses only linearly independent columns of A in representing b. □

Definition. Let $\hat{x}$ be a solution of the system

$$Ax = b,\ x \geq 0.$$

If $\tau = \operatorname{supp}\hat{x}$, and $A_{\bullet\tau}$ has linearly independent columns, then $\hat{x}$ is called a ***basic solution*** of the system.

Theorem 2. Every basic solution of the system $Ax = b,\ x \geq 0$ is an extreme point of the set of solutions of the system.

Proof. Let $\hat{x}$ be a basic solution of the system $Ax = b,\ x \geq 0$ and let $\tau = \operatorname{supp}\hat{x}$. If $\hat{x}$ is not an extreme point of the set $S = \{x : Ax = b,\ x \geq 0\}$, then it lies on the open line segment between two distinct points of S. Let these points be $\tilde{x}$ and $\tilde{y}$. For some $\alpha \in (0, 1)$ we have $\hat{x} = \alpha\tilde{x} + (1 - \alpha)\tilde{y}$. It follows from this that

$$\operatorname{supp}\tilde{x} \subset \tau \quad \text{and} \quad \operatorname{supp}\tilde{y} \subset \tau.$$

Since $\hat{x}$, $\tilde{x}$, $\tilde{y} \in S$, we have

$$A_{\bullet\tau}\hat{x}_\tau = A_{\bullet\tau}\tilde{x}_\tau = A_{\bullet\tau}\tilde{y}_\tau = b.$$

This in turn implies that

$$A_{\bullet\tau}(\hat{x}_\tau - \tilde{x}_\tau) = A_{\bullet\tau}(\hat{x}_\tau - \tilde{y}_\tau) = 0.$$

Since $A_{\bullet\tau}$ has linearly independent columns, the above equations imply that $\hat{x} = \tilde{x} = \tilde{y}$ which is a contradiction. □

Remark. It is clear that a polyhedral set can have at most finitely many extreme points.

2.4.2 Cones

The set of all nonnegative scalar multiples of a nonzero vector is called a ***ray***. If $d \neq 0$, then the ray ***generated*** by d is the set

$$< d > = \{x : x = \lambda d,\ \lambda \geq 0\}.$$

A translate of a ray, that is, a set of the form

$$\{\bar{x}\} + < d >,$$

is called a ***halfline***.

Definition. Let C be a nonempty subset of a linear space. If C has the property that it contains all nonnegative multiples of its elements, then C is called a ***cone***.

Examples. The following are some cones:

- The entire space or any subspace thereof.
- The zero vector.
- A ray generated by a nonzero vector.
- The set of all nonnegative-valued functions (for example positive semidefinite quadratic forms).
- The solution set of the linear inequality system $Ax \geq 0$. These are called ***polyhedral cones***. In particular, the ***nonnegative orthant***,

  $$R^n_+ = \{x \in R^n : x \geq 0\},$$

  is a polyhedral cone.
- The ***polar*** of a cone C in R^n is defined to be

  $$C^* = \{y : y^{\mathrm{T}}x \leq 0 \text{ for all } x \in C\};$$

  it is a (closed convex) cone.

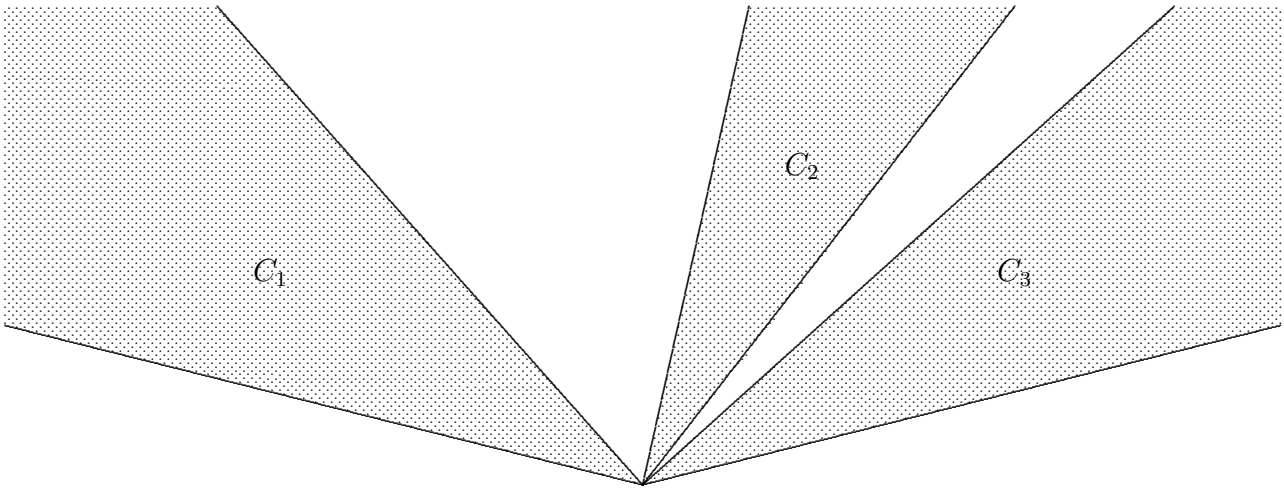


Figure 2.16

Remark. Some cones are convex, others are not. Any ***polyhedral cone*** i.e., the solution set of a homogeneous linear inequality system such as $Ax \geq 0$ is convex. The set of all nonnegative multiples of the elements of a nonempty convex set is a convex cone. The cone generated by all nonnegative multiples of points in the union of two ***disjoint*** sets may or may not be convex. For instance, let one of the sets be the epigraph of the convex function

e^{-x} where $x \in R_+$. Let the other be the ray $< (1,0) >$. The cone of nonnegative multiples of points in the union of these two sets is R^2_+, which is convex. But consider the example illustrated in Figure 2.16. Each of the cones C_1, C_2, C_3 is convex, and each can be considered as the set of all nonnegative multiples of a convex set (such as a line segment cutting across the cone). The union of any two of these three cones is not convex.

Proposition 2. Let $a^1, a^2, \ldots, a^n$ be vectors in R^m. Then the set

$$< a^1 > + < a^2 > + \cdots + < a^n > = \{y : y = Ax,\ x \geq 0\}$$

is a convex cone denoted $\operatorname{pos} A$ where $a^j = A_{\bullet j}$.

Proof. This is obvious. □

Cone of this form are said to be ***finitely generated***. Sometimes they are called ***finite cones***, but they are not finite sets unless they are just the zero vector.

The following is an important result as we shall see later.

Theorem 3. Every finite cone is a closed set.

Proof. Let C be a finite cone. If C is just $\{0\}$, then it is trivially closed, so we assume that $C \neq \{0\}$. Thus we have

$$C = \{y : y = Ax,\ x \geq 0\}$$

where $A \in R^{m \times n}$ is a nonzero matrix. Let $\{y^k\}$ be a sequence of points in C converging to y^*. The object is to show that $y^* \in C$. To do so, we prove that there exists a vector $x^* \geq 0$ such that $y^* = Ax^*$.

Now if $y^* = 0$, we may take $x^* = 0$ and we're done. So we assume $y^* \neq 0$. For each y^k in the sequence, let x^k be a nonnegative vector such that $y^k = Ax^k$ and the columns of A corresponding to the support of x^k are linearly independent. There must be a subsequence of $\{y^k\}$ converging to y^* such that the corresponding $\{x^k\}$ all have the same support, say τ. To make the notation simple, we may assume that the entire sequence has this property. Suppose the index set τ contains r elements (i.e., $A_{\bullet\tau}$ has r columns). Then there exists an index set σ such that the submatrix $A_{\sigma\tau}$ is nonsingular. It then follows that

$$x^k_\tau = A^{-1}_{\sigma\tau} y^k_\sigma \longrightarrow A^{-1}_{\sigma\tau} y^*_\sigma,$$

and it is clear that

$$x^*_\tau := A^{-1}_{\sigma\tau} y^*_\sigma \geq 0.$$

If $j \notin \tau$, let $x^k_j = 0$. Then $x^* \geq 0$ and $y^* = Ax^*$, so $y^* \in C$. □

Proposition 3. Any linear subspace L of R^n is a finite cone.

Proof. Let $\{a^1, a^2, \ldots, a^r\}$ be a basis for L. Then

$$L = < a^1 > + < a^2 > + \cdots + < a^r > + < -a^1 > + < -a^2 > + \cdots + < -a^r >,$$

so L is a finite cone. □

Notice that in the following theorem, the set C can be specified entirely by linear inequalities.

Lemma. Let $A \in R^{m\times n}$ be given. A polyhedral cone of the form

$$C = \{x : Ax = 0,\ x \geq 0\}$$

is a finite cone.

Proof. We first dispense with two trivial cases. First, if C consists of only the zero vector, then $C = < 0 >$. The second case is where $A = 0$. Then $C = R^n_+$ and that is clearly the finite cone generated by the unit vectors $e^1, \ldots, e^n \in R^n$. Hereafter, we assume that both C and A have nonzero elements.[10] Now a suitable positive multiple of any nonzero element of C belongs to

$$S = \{x : Ax = 0,\ e^{\mathrm{T}}x = 1,\ x \geq 0\}$$

where $e = (1, \ldots, 1)^{\mathrm{T}} \in R^n$. Notice that the polyhedral set S is properly contained in C. If $n = 2$, our assumptions above lead us to the consideration of only one possibility: that the rank and nullity of A equal 1. Then $C = < p >$ where $p > 0$, and S is a single point, which is an extreme point of S by default.

As an inductive hypothesis, assume that when A has at most $n - 1$ columns, then every point of S is expressible as a convex combination of extreme points of S. Since S has only finitely many extreme points in all, the inductive hypothesis amounts to saying that S is a polytope. Now suppose A has n columns. If the support of a point $\bar{x} \in S$ has cardinality less than n (i.e., fewer than n elements), then the inductive hypothesis applies. If $\bar{x} > 0$, define

$$\bar{A} = [\bar{A}_{\bullet 1}\ \bar{A}_{\bullet 2}\ \cdots\ \bar{A}_{\bullet n}] = \left[\left[\begin{array}{c} A_{\bullet 1} \\ 1 \end{array}\right]\ \left[\begin{array}{c} A_{\bullet 2} \\ 1 \end{array}\right]\ \cdots\ \left[\begin{array}{c} A_{\bullet n} \\ 1 \end{array}\right]\right]$$

If the columns of $\bar{A}$ are linearly independent, then $\bar{x}$ is an extreme point. If the columns of $\bar{A}$ are linearly dependent, then by Theorem 1, there exists an extreme point x^* of S. Under the present assumptions, x^* has at least one zero component, and since $e^{\mathrm{T}}x^* = e^{\mathrm{T}}\bar{x} = 1$, it follows that $x^*_k > \bar{x}_k$ for some k. Now define the scalar

$$\lambda = \max_i \frac{x^*_i}{\bar{x}_i} > 1.$$

It is easy to see that

$$\tilde{x} := \frac{1}{\lambda - 1}(\lambda\bar{x} - x^*) \in S.$$

But $\tilde{x}$ has at most $n - 1$ positive components and hence is a convex combination of extreme points of S. From the definition of $\tilde{x}$, it follows that

$$\bar{x} = \frac{1}{\lambda}x^* + \left(1 - \frac{1}{\lambda}\right)\tilde{x}.$$

[10] This implies $n > 1$.

Thus, $\bar{x}$ is a convex combination of extreme points of S. Since S has only finitely many extreme points and every element of C is a nonnegative linear combination of these extreme points, it follows that C is a finite cone. □

This lemma can be expressed as follows.

Corollary 1. If L is a linear subspace of R^n, then $L \cap R^n_+$ is a finite cone.

Proof. Any linear subspace of R^n is the nullspace of some matrix A. Hence

$$L \cap R^n_+ = \{x : Ax = 0,\ x \geq 0\} = C,$$

and we have just shown that C is a finite cone. □

We come now to an important result. Recall that $A \in R^{m \times n}$.

Theorem 4 (Minkowski). Every polyhedral convex cone

$$C = \{x : Ax \geq 0\}$$

is finite.

Proof. Let $L = \{y : y = Ax,\ x \in R^n\}$. (This is just the column space of A.) Then L is a subspace of R^m, and by the lemma (or by Corollary 1) above, we have

$$L \cap R^m_+ = < y^1 > + \cdots + < y^r >$$

for some vectors $y^1, \ldots, y^r \in L \cap R^m_+$. For each $i = 1, \ldots, r$ we have $y^i = Ax^i$. If $x \in C$, then $Ax \in L \cap R^m_+$, so

$$Ax = \sum_{i=1}^{r} \lambda_i y^i = \sum_{i=1}^{r} \lambda_i (Ax^i)$$

where $\lambda_i \geq 0$ for $i = 1, \ldots, r$. Thus,

$$A\left(x - \sum_{i=1}^{r} \lambda_i x^i\right) = 0.$$

This equation means $x - \sum_{i=1}^{r} \lambda_i x^i$ belongs to the nullspace of A which is a finite cone. Hence there exist vectors $\hat{x}^1, \ldots, \hat{x}^s$ and nonnegative scalars $\mu_1, \ldots, \mu_s$ such that

$$x - \sum_{i=1}^{r} \lambda_i x^i = \sum_{j=1}^{s} \mu_j \hat{x}^j.$$

Rewriting this equation, we have

$$x = \sum_{i=1}^{r} \lambda_i x^i + \sum_{j=1}^{s} \mu_j \hat{x}^j$$

in which all the scalar coefficients are nonnegative. Thus we conclude that

$$C \subset < x^1 > + \cdots + < x^r > + < \hat{x}^1 > + \cdots + < \hat{x}^s > .$$

The reverse inclusion is straightforward hence we have equality which shows that C is a finite cone. □

Minkowski's theorem says that every polyhedral cone is finite. The next theorem will complete the two-fold representation of convex cones, but before we come to that, we need the following proposition.

Proposition 4. Let C be a nonempty cone in R^n. Then $C = C^{**}$ (the polar of its polar) if and only if C is closed and convex.

Proof. Recall that the polar of a cone is always closed and convex. Since C^{**} is the polar of the cone C^*, it is closed and convex. Accordingly, if $C = C^{**}$, then C must be closed and convex because C^{**} is so.

For the converse, we first observe that $C \subset C^{**}$. Indeed, for all $x \in C$ we have $y^{\mathrm{T}}x \leq 0$ for all $y \in C^*$. But $y^{\mathrm{T}}x = x^{\mathrm{T}}y$. This implies that $x \in C^{**}$. Since x was arbitrary, we have $C \subset C^{**}$, as claimed. Next we claim that the set $C^{**} \setminus C$ is empty. (In otherwords, there are no vectors in C^{**} that do not belong to C.) Otherwise, there must exist a vector $\tilde{z} \in C^{**}$ such that $\tilde{z} \notin C$. Since $\tilde{z} \in C^{**}$, we know (from the definition of the polar) that $\tilde{z}^{\mathrm{T}}y \leq 0$ for all $y \in C^*$. Since C is assumed to be closed and convex, there exists a hyperplane that separates $\tilde{z}$ from C. Because the latter is a cone—and therefore contains the zero vector—it follows that there exists a vector $\tilde{y}$ such that

$$\tilde{y}^{\mathrm{T}}x \leq 0 \text{ for all } x \in C \tag{1}$$

and

$$\tilde{y}^{\mathrm{T}}\tilde{z} > 0. \tag{2}$$

Equation (1) implies $\tilde{y} \in C^*$. But then equation (2) contradicts the assumption that $\tilde{z} \in C^{**}$. Hence $C^{**} \setminus C$ is empty as asserted. □

Theorem 5 (Weyl). Every finite cone is polyhedral.

Proof. Consider the finite cone

$$C = \sum_{i=1}^{r} < a^i >.$$

Now consider the polar cone

$$C^* = \{y : y^{\mathrm{T}}a^i \leq 0, \quad i = 1, \ldots, r\}.$$

It is clear that C^* is a polyhedral cone. By Minkowski's theorem, the polar cone C^* is finitely generated. Let

$$C^* = \sum_{j=1}^{r} < b^j >.$$

Now the polar cone C^{**} of C^* is polyhedral. Because C is closed we have $C = C^{**}$. These observations show that C is polyhedral. □

As an application of Weyl's theorem, we can prove something that seems geometrically obvious but still requires a proof.

Corollary 2. Every polytope is a polyhedral convex set.

Proof. Let $S = \text{co}\{x^1, \ldots, x^r\}$. Then we have

$$x \in S \Longleftrightarrow \begin{bmatrix} x \\ 1 \end{bmatrix} = \sum_{i=1}^{r} \lambda_i \begin{bmatrix} x^i \\ 1 \end{bmatrix} \qquad \lambda_i \geq 0 \qquad i = 1, \ldots, r.$$

Let C be the finite cone generated by the vectors

$$\begin{bmatrix} x^1 \\ 1 \end{bmatrix}, \ldots, \begin{bmatrix} x^r \\ 1 \end{bmatrix}.$$

By Weyl's theorem, there exists a matrix $A = [A_{\bullet 1} \ \ldots, \ A_{\bullet n}, \ A_{\bullet\, n+1}]$ such that

$$\begin{bmatrix} x \\ \xi \end{bmatrix} \in C \Longleftrightarrow A_{\bullet 1} x_1 + \cdots + A_{\bullet n} x_n + A_{\bullet\, n+1} \xi \geq 0.$$

Now take

$$A = [A_{\bullet 1} \ \cdots \ A_{\bullet n}] \quad \text{and} \quad b = -A_{\bullet\, n+1}.$$

Then

$$\begin{bmatrix} x \\ \xi \end{bmatrix} \in C \Longleftrightarrow Ax \geq b\xi.$$

Since

$$x \in S \Longrightarrow \begin{bmatrix} x \\ 1 \end{bmatrix} \in C,$$

it follows that $Ax \geq b$ for all $x \in S$. Conversely, if $Ax \geq b$, then $\begin{bmatrix} x \\ 1 \end{bmatrix} \in C$, and this says that x is a convex combination of $x^1, \ldots, x^r$. That is, $x \in S$. This shows

$$S = \text{co}\{x^1, \ldots, x^r\} = \{x : Ax \geq b\}$$

so that S is polyhedral. □

2.4.3 The structure of polyhedra

The definition of a polyhedral set as the intersection of finitely many halfspaces is an external one. There is another way to generate polyhedral sets as we shall see below.

Relative to the polyhedral set $S = \{x : Ax \geq b\}$ we define the polyhedral cone given by the linear inequalities

$$Ax - \xi b \geq 0, \quad \xi \geq 0.$$

To bring out the similarity with the formulation used earlier, we write

$$Y = \left\{ \begin{bmatrix} x \\ \xi \end{bmatrix} : \begin{bmatrix} A & -b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ \xi \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}.$$

On one hand, if $\bar{x} \in S$, then $\bar{y} = \begin{bmatrix} \bar{x} \\ 1 \end{bmatrix} \in Y$. On the other hand, if Y contains any point $\bar{y} = \begin{bmatrix} \bar{x} \\ \xi \end{bmatrix}$ such that $\xi > 0$, then $\frac{1}{\xi}\bar{x} \in S$ and $\begin{bmatrix} \frac{1}{\xi}\bar{x} \\ 1 \end{bmatrix} \in Y$. By Minkowski's theorem, Y is a finite cone, say

$$Y = < y^1 > + \cdots + < y^t > .$$

If S is nonempty, then Y contains some points $y = \begin{bmatrix} x \\ \xi \end{bmatrix}$ for which $\xi > 0$. Hence at least one of the generators of Y must have a positive $(n+1)$-st component. Let the generators $y^1, \ldots, y^t$ be separated into two subsets: those with $\xi > 0$ and those with $\xi = 0$. By positively scaling and reordering the $y^k,\ k = 1, \ldots, t$, we may assume they are

$$\begin{bmatrix} x^1 \\ 1 \end{bmatrix}, \ldots, \begin{bmatrix} x^r \\ 1 \end{bmatrix}, \begin{bmatrix} x^{r+1} \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} x^t \\ 0 \end{bmatrix}.$$

It could happen that $r = t$, however. Thus, $y = \begin{bmatrix} x \\ \xi \end{bmatrix} \in Y$ means

$$\begin{bmatrix} x \\ \xi \end{bmatrix} = \sum_{i=1}^{r} \lambda_i \begin{bmatrix} x^i \\ 1 \end{bmatrix} + \sum_{i=r+1}^{t} \lambda_i \begin{bmatrix} x^i \\ 0 \end{bmatrix}$$

where $\lambda_i \geq 0,\ i = 1, \ldots, t$. For every $x \in S$, the point $\begin{bmatrix} x \\ 1 \end{bmatrix} \in Y$. The significance of the equation above for the last component is that $\lambda_1 + \cdots + \lambda_r = 1$. Thus it follows that

$$\sum_{i=1}^{r} \lambda_i x^i \in \text{co}\{x^1, \ldots, x^r\} \quad \text{and} \quad \sum_{i=r+1}^{t} \lambda_i x^i \in \text{pos}\{x^{r+1}, \ldots, x^t\}.$$

This proves the following important resolution theorem.

Theorem 6 (Motzkin; Goldman). If $S = \{x : Ax \geq b\}$ is nonempty, it is the sum of a polytope and a finite cone. □

This theorem has a useful corollary.

Corollary 3. If the polyhedral set $S = \{x : Ax \geq b\}$ is nonempty, then it is bounded if and only if the cone $C = \{x : Ax \geq 0\}$ contains only the zero vector.

Proof. If C contains a nonzero vector, say d, then $x + \alpha d \in S$ for any $\alpha \geq 0$. This implies that S cannot be bounded. In plainer language: If S is bounded, $C = \{0\}$. Now assume

$C = \{0\}$. In the proof of the resolution theorem, the generators y^k of Y with $\xi = 0$ (i.e., $y^{r+1}, \ldots, y^t$) give generators $x^{r+1}, \ldots, x^t$ for C. But if $C = \{0\}$, these generators are zero, so S is a polytope, and hence is bounded. □

Extreme-point optima

The resolution theorem can be used to prove an important fact about optimal solutions of linear programs. We shall discuss this result in terms of linear programs in ***standard form***. This is not restrictive because a feasible region described by the constraints

$$Ax = b, \quad x \geq 0$$

is also expressible as the set of vectors satisfying

$$Ax \geq b, \quad -Ax \geq -b, \quad \text{and} \quad x \geq 0,$$

consequently, the Motzkin-Goldman resolution theorem applies to the feasible region of a linear programming problem in standard form.

Theorem. If the linear program

$$\text{(P)} \quad \begin{array}{ll} \text{minimize} & c^{\mathrm{T}}x \\ \text{subject to} & Ax = b \\ & x \geq 0 \end{array}$$

has an optimal solution, then there exists an extreme point of the feasible region of (P) that is optimal.

Proof. Let S denote the feasible region of (P), and let $\bar{x} \in S$ be an optimal solution of this linear program. Thus, S is a nonempty polyhedral set. As such, it is the sum of a polytope and a finite cone. If $c^{\mathrm{T}}\tilde{x} < 0$ for any element $\tilde{x}$ of the finite cone, then the objective function in (P) has no lower bound on S, so there cannot exist an optimal solution such as $\bar{x}$. Now letting

$$\bar{x} = \sum_{i=1}^{r} \lambda_i x^i + \sum_{i=r+1}^{t} \lambda_i x^i,$$

we can see that

$$c^{\mathrm{T}} \left(\sum_{i=r+1}^{t} \lambda_i x^i \right) = 0.$$

Indeed, we have just shown that this quantity must be nonnegative. If it is positive, then

$$c^{\mathrm{T}}\bar{x} = c^{\mathrm{T}} \left(\sum_{i=1}^{r} \lambda_i x^i + \sum_{i=r+1}^{t} \lambda_i x^i \right) > c^{\mathrm{T}} \left(\sum_{i=1}^{r} \lambda_i x^i \right).$$

Since

$$\sum_{i=1}^{r} \lambda_i x^i \in S,$$

we have a contradiction regarding the optimality of $\bar{x}$. We may now assume that all the points $x^1, \ldots, x^r$ are in fact extreme points of S. (Otherwise remove any points that are convex combinations of the other points in this set.) Let

$$x^* = \arg\min\{c^{\mathrm{T}}x^i : i = 1, \ldots, r\}.$$

Then

$$c^{\mathrm{T}}\bar{x} = c^{\mathrm{T}}\left(\sum_{i=r+1}^{t} \lambda_i x^i\right) \geq c^{\mathrm{T}}x^* \geq c^{\mathrm{T}}\bar{x}.$$

Hence x^* is an optimal extreme-point solution of (P). □

References

D. Gale [1960], *The Theory of Linear Economic Models*, McGraw-Hill Book Company, Inc., New York.

A.J. Goldman [1956], Resolution and separation theorems for polyhedral convex sets, in (H.W. Kuhn and A.W. Tucker, eds.) *Linear Inequalities and Related Systems*, Annals of Mathematics Study 38, Princeton University Press, Princeton, N.J.

O.L. Mangasarian [1969], *Nonlinear Programming*, McGraw-Hill, New York. [Reprinted in softcover by SIAM Publications.]

T.S. Motzkin [1936], *Beiträge zur Theorie det linear Ungleichungen*, Azriel, Jerusalem. [Doctoral dissertation, University of Basel, 1933.]

K.G. Murty [1983], *Linear Programming*, John Wiley & Sons, New York.

A. Schrijver [1986], *Theory of Linear and Integer Programming*, John Wiley & Sons, Chichester.

J. Stoer and C. Witzgall [1970], *Convexity and Optimization in Finite Dimensions I*, Springer-Verlag, New York.

G.M. Ziegler [1995], *Lectures on Polytopes*, Springer-Verlag, New York.

2.5 Alternative theorems

What can you say when a system of linear equations $Ax = b$ has no solution? One thing you can say is that the rows of the matrix A must be linearly dependent. This, in turn, implies that there exists a nonzero vector y such that $y^{\mathrm{T}}A = 0$. But more can be said, as the following lemma reveals.

Lemma. If $A \in R^{m\times n}$ and $b \in R^m$, then exactly one of the following systems

$$Ax = b, \tag{i}$$

$$y^{\mathrm{T}}A = 0^{\mathrm{T}}, \quad y^{\mathrm{T}}b > 0 \tag{ii}$$

has a solution.

Proof. It is impossible for both systems—(i) and (ii)—to have a solution, for otherwise, there exist x and y such that

$$0 = y^{\mathrm{T}}Ax = y^{\mathrm{T}}b > 0$$

which is a contradiction. Now suppose (i) has no solution. From the fact[11] that $R^m = \mathcal{R}(A) + \mathcal{N}(A^{\mathrm{T}})$, we know there exist vectors x and u such that

$$b = Ax + Z^{\mathrm{T}}u$$

where Z^{T} is a basis for the nullspace of A^{T}. Now $y := Z^{\mathrm{T}}u \neq 0$ (otherwise system (i) has a solution). Thus, we have $y^{\mathrm{T}}A = 0^{\mathrm{T}}$ and

$$y^{\mathrm{T}}b = u^{\mathrm{T}}ZAx + u^{\mathrm{T}}Z\,Z^{\mathrm{T}}u = 0 + y^{\mathrm{T}}y > 0,$$

which is to say that (ii) has a solution. □

This lemma will play an important role when we study nonlinear programming. For now, it illustrates a class of results known as ***theorems of the alternative.*** More simply, these are called ***alternative theorems*** or ***transposition theorems***.

[11] Courses like Mathematics 113 make a big point of the fact that given $A \in R^{m\times n}$, the space R^m can be expressed as the range of A plus the nullspace of A^{T}.

As an extremely important application of the inequality theory developed in previous sections, we have other alternative theorem involving pairs of inequality systems[12] that use the same data, but in different ways. What these theorems have in common is the assertion that *exactly one* of the linear inequality systems in the given pair has a solution.

Theorem 3 (Handout 4) says that for any real matrix A, the set $\operatorname{pos} A = \{Ax : x \geq 0\}$ is closed. The following is probably most famous of all alternative theorems.

Theorem 1 (Farkas). If $A \in R^{m\times n}$ and $b \in R^m$, then exactly one of the following systems

$$Ax = b, \quad x \geq 0, \tag{1}$$

$$y^{\mathrm{T}}A \leq 0^{\mathrm{T}}, \quad y^{\mathrm{T}}b > 0 \tag{2}$$

has a solution.

Proof. First, suppose systems (1) and (2) both have solutions, say $\tilde{x}$ and $\tilde{y}$, respectively. Then we have

$$0 < \tilde{y}^{\mathrm{T}}b = \tilde{y}^{\mathrm{T}}(A\tilde{x}) = (\tilde{y}^{\mathrm{T}}A)\tilde{x} \leq 0.$$

This is a contradiction. Hence at most one of the systems (1) and (2) can have a solution. To complete the proof, if is enough to show that if (1) has no solution, then (2) does have a solution.

If (1) has no solution, then $b \notin \operatorname{pos} A$. By the separation theorem (Handout No. 2, middle of page 9) there is a nonzero vector y and a scalar α such that

$$y^{\mathrm{T}}(Ax) \leq \alpha \quad \text{for all } x \geq 0 \quad \text{and} \quad y^{\mathrm{T}}b > \alpha.$$

Letting $x = 0$, we see that $\alpha \geq 0$. This implies $y^{\mathrm{T}}b > 0$. Since $\operatorname{pos} A$ is a cone, we may assume that in fact $y^{\mathrm{T}}z \leq 0$ for all $z \in \operatorname{pos} A$. (Otherwise, some positive multiple of z would be larger than $y^{\mathrm{T}}b$ which is impossible.) Now, among the elements of $\operatorname{pos} A$ are the columns of A itself. (To see this, consider $A_{\bullet j}$ for some fixed j. Take $x_j = 1$ and $x_k = 0$ for $k \neq j$. Then $z = Ax = A_{\bullet j} \in \operatorname{pos} A$.) Thus, we may assume $\alpha = 0$. This shows that y is a solution of the system (2). □

Remark. For some reason, Theorem 1 is known in the literature as "Farkas's lemma." The theorem appeared in a paper by the Hungarian mathematician J. Farkas (Theorie der einfachen Ungleichungen, *Journal für die reine und angewandte Mathematik* 124 (1902), 1-27) where it was identified as a *Grundsatz*, or principle, not as a *Hilfsatz*, or lemma. Reluctantly, we follow the well established tradition.

Farkas's lemma can be used to establish many other alternative theorems. Our presentation is not exhaustive but should be extensive enough to make the point that Farkas's lemma is a useful theoretical tool. It should be noticed here that in proving alternative theorems, our approach is first to show that not both systems can have solutions and then to cast one of the two systems as an equivalent system having the form (1). The application of Farkas's lemma then implies that some alternative system does have a solution, and this alternative system is equivalent to the member of the pair for which we want to demonstrate the existence of a solution.

[12] By "inequality systems" we mean systems of linear equations and/or linear inequalities that involve at least some linear inequalities.

Theorem 2. If $A \in R^{m\times n}$ and $b \in R^m$, then exactly one of the following systems

$$Ax \geq b, \quad x \geq 0, \tag{3}$$

$$y \geq 0, \quad y^{\mathrm{T}}A \leq 0^{\mathrm{T}}, \quad y^{\mathrm{T}}b > 0 \tag{4}$$

has a solution.

Proof. It is impossible for both (3) and (4) to have solutions, for otherwise there would be $\tilde{x}$ and $\tilde{y}$ such that

$$0 < \tilde{y}^{\mathrm{T}}b \leq \tilde{y}^{\mathrm{T}}(A\tilde{x}) = (\tilde{y}^{\mathrm{T}}A)\tilde{x} \leq 0$$

which is a contradiction.

It is sufficient to prove that if (3) has no solution, then (4) does. Now if (3) has no solution, then the system

$$Ax - Iu = b, \quad x \geq 0,\ u \geq 0 \tag{3$'$}$$

has no solution. Note that (3′) can be regarded as an instance of a system like (1), that is, a system of linear equations in nonnegative variables. Now if (3′) has no solution, then there exists a vector y such that

$$y^{\mathrm{T}}[A,\ -I] \leq (0^{\mathrm{T}},\ 0^{\mathrm{T}}), \quad y^{\mathrm{T}}b > 0. \tag{4$'$}$$

Any solution of (4′) is a solution of (4) and vice versa. This completes the proof. □

As another application of Farkas's lemma we consider a pair of *homogeneous* systems. Notice that the system (5) below has strict inequalities, whereas system (6) involves a nonzero, nonnegative vector.

Theorem 3 (Gordan). If $A \in R^{m\times n}$, then exactly one of the following systems

$$Ax > 0, \tag{5}$$

$$y \geq 0, \quad y \neq 0, \quad y^{\mathrm{T}}A = 0^{\mathrm{T}} \tag{6}$$

has a solution.

Proof. It is impossible for both (5) and (6) to have solutions, for otherwise there would be $\tilde{x}$ and $\tilde{y}$ such that

$$0 < \tilde{y}^{\mathrm{T}}(A\tilde{x}) = (\tilde{y}^{\mathrm{T}}A)\tilde{x} = 0$$

which is a contradiction.

Assume that (5) has no solution. Then neither does the system $Ax \geq e$ where e is the vector of ones in R^m. Now this is a system like (3) except for the fact that the variables are not sign restricted. To take care of that we can substitute the difference of two nonnegative vectors for x. Thus, we let $x = x' - x''$ where $x' \geq 0$ and $x'' \geq 0$. Then the system under consideration can be written as

$$Ax' - Ax'' \geq e, \quad x' \geq 0,\ x'' \geq 0. \tag{5$'$}$$

Hence if (5) has no solution, then neither does (5′). By Theorem 2, there exists a vector y such that

$$y \geq 0, \quad y^{\mathrm{T}}[A, -A] \leq (0^{\mathrm{T}}, 0^{\mathrm{T}}), \quad y^{\mathrm{T}}e > 0 \tag{6'}$$

It follows from (6′) that y satisfies the conditions of (6). □

We now turn to another application of alternative theorems.

2.6 Duality in linear programming

With every linear programming problem there is another linear programming problem called its ***dual***. Relative to this second linear program (LP), the original one is called the ***primal problem***. These problems are intimately related as we shall see.

As an example, consider the linear programming problem in standard form

$$\begin{aligned} \text{minimize} \quad & c^{\mathrm{T}}x \\ \text{subject to} \quad & Ax = b \\ & x \geq 0 \end{aligned}$$

The ***dual*** of this problem is *defined* to be

$$\begin{aligned} \text{maximize} \quad & b^{\mathrm{T}}y \\ \text{subject to} \quad & A^{\mathrm{T}}y \leq c \end{aligned}$$

Notice the following important relationships between this primal/dual pair.

1. The primal is a minimization problem whereas the dual is a maximization problem.
2. The main constraints of the primal are linear equations in nonnegative variables, whereas the constraints of the dual are linear inequalities (of the $\leq$ type) and the variables have no explicit sign restriction on them. Such variables are said to be ***free***.
3. When $A \in R^{m \times n}$, there are n primal variables and n dual constraints, and there are m linear equations in the primal and m free variables in the dual. Thus, the variables of each problem are in one-to-one correspondence with the constraints of the other problem.

To illustrate the intimate connection between an LP and its dual, we begin by considering the following simple relationship.

Proposition 1 (Weak duality). Let $\bar{x}$ be an arbitrary feasible solution of the linear program in standard form

$$\begin{aligned} \text{minimize} \quad & c^{\mathrm{T}}x \\ \text{subject to} \quad & Ax = b \\ & x \geq 0, \end{aligned}$$

and let $\bar{y}$ be an arbitrary feasible solution of the corresponding dual problem

$$\begin{aligned} \text{maximize} \quad & b^{\mathrm{T}}y \\ \text{subject to} \quad & A^{\mathrm{T}}y \leq c. \end{aligned}$$

Then

$$b^{\mathrm{T}}\bar{y} \leq c^{\mathrm{T}}\bar{x}.$$

Proof. Using the feasibility of the two vectors $\bar{x}$ and $\bar{y}$, we obtain

$$b^{\mathrm{T}}\bar{y} = (A\bar{x})^{\mathrm{T}}\bar{y} = \bar{x}^{\mathrm{T}}A^{\mathrm{T}}\bar{y} \leq \bar{x}^{\mathrm{T}}c = c^{\mathrm{T}}\bar{x}$$

which gives the assertion. □

This ***weak duality inequality*** has an important consequence.

Corollary 1. Let $\bar{x}$ and $\bar{y}$ be feasible solutions of the linear program in standard form and its dual, respectively. If $b^{\mathrm{T}}\bar{y} = c^{\mathrm{T}}\bar{x}$, then $\bar{x}$ and $\bar{y}$ are optimal solutions of their respective linear programs.

Proof. Suppose $\bar{y}$ is not optimal for the dual problem. Then there exists a dual feasible vector $\tilde{y}$ such that $b^{\mathrm{T}}\tilde{y} > b^{\mathrm{T}}\bar{y} = c^{\mathrm{T}}\bar{x}$. This contradicts Proposition 1. Hence $\bar{y}$ is optimal for the dual problem. In like manner we can show that $\bar{x}$ is optimal for the primal problem. □

One place where an alternative theorem can be put to use is in proving the ***strong duality theorem***.

Theorem 1 (Strong duality). If $\bar{x}$ is an optimal solution of the linear program

$$\begin{aligned} \text{minimize} \quad & c^{\mathrm{T}}x \\ \text{subject to} \quad & Ax = b \\ & x \geq 0, \end{aligned}$$

then there exists an optimal solution $\bar{y}$ of the corresponding dual linear program and moreover

$$b^{\mathrm{T}}\bar{y} = c^{\mathrm{T}}\bar{x}.$$

Proof. It will suffice to prove the existence of a solution to the linear inequality system

$$\begin{aligned} A^{\mathrm{T}}y &\leq c \\ -b^{\mathrm{T}}y &\leq -c^{\mathrm{T}}\bar{x} \end{aligned}$$

since any solution of the latter system would be a dual feasible solution yielding a dual objective value equal to the optimal objective value of the primal. Now this system can be converted to an equivalent system of linear equations in nonnegative variables. That system is

$$\begin{bmatrix} A^{\mathrm{T}} \\ -b^{\mathrm{T}} \end{bmatrix} y' - \begin{bmatrix} A^{\mathrm{T}} \\ -b^{\mathrm{T}} \end{bmatrix} y'' + \begin{bmatrix} I & 0 \\ 0^{\mathrm{T}} & 1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} c \\ -c^{\mathrm{T}}\bar{x} \end{bmatrix}$$

$$y', y'', u, v \geq 0.$$

Now if this system has no solution, then by Farkas's lemma, there exists a scalar $\tilde{\xi}$ and a vector $\tilde{x}$ satisfying the follow conditions:

$$A\tilde{x} = \tilde{\xi} b$$
$$\begin{bmatrix} \tilde{x} \\ \tilde{\xi} \end{bmatrix} \leq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$
$$c^{\mathrm{T}}\tilde{x} > \tilde{\xi} c^{\mathrm{T}}\bar{x}$$

Defining $\hat{x} = -\tilde{x}$ and $\hat{\xi} = -\tilde{\xi}$ we obtain a solution to the system

$$A\hat{x} = \hat{\xi} b$$
$$\begin{bmatrix} \hat{x} \\ \hat{\xi} \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$
$$c^{\mathrm{T}}\hat{x} < \hat{\xi} c^{\mathrm{T}}\bar{x}$$

Now we note that $\hat{\xi} > 0$. Otherwise, $\hat{\xi} = 0$, implying that $\bar{x} + \lambda\hat{x}$ is primal feasible for all $\lambda \geq 0$. Since $c^{\mathrm{T}}(\bar{x} + \lambda\hat{x}) \longrightarrow -\infty$ as $\lambda \longrightarrow +\infty$, we see that the primal has no optimal solution, but instead has an unbounded objective function. Therefore $\hat{\xi} > 0$. Finally, we obtain the contradiction that $(1/\hat{\xi})\hat{x}$ is a primal feasible vector for which the primal objective function value is less than the assumed optimal value $c^{\mathrm{T}}\bar{x}$. □

Remark. In effect, strong duality theorem amounts to the converse of Corollary 1. Thus, a vector $\bar{x}$ that is feasible for the LP (in standard form) is an optimal solution of that problem if and only if there exists a vector $\bar{y}$ that is feasible for the dual problem and $c^{\mathrm{T}}\bar{x} = b^{\mathrm{T}}\bar{y}$. This is an example of an *optimality criterion*, a set of necessary and sufficient conditions for a vector to be optimal for the linear programming problem. Here is another.

Corollary 2 (Complementary slackness conditions). A vector $\bar{x}$ is optimal for the LP (above) if and only if there exists a vector $\bar{y}$ such that

$$A\bar{x} = b \tag{7}$$
$$A^{\mathrm{T}}\bar{y} \leq c \tag{8}$$
$$\bar{x}^{\mathrm{T}}(c - A^{\mathrm{T}}\bar{y}) = 0 \tag{9}$$
$$\bar{x} \geq 0. \tag{10}$$

Proof. Conditions (7), (8), and (10) assert the feasibility of $\bar{x}$ and $\bar{y}$ in the primal and dual problems, respectively; along with (9), these conditions are equivalent to the equality of the two objective function values. □

Looking closely at (9), we see a statement to the effect that the scalar product of two nonnegative vectors, $\bar{x}$ and $c - A^{\mathrm{T}}\bar{y}$ is zero. Such a thing holds if and only if

$$\bar{x}_j(c - A^{\mathrm{T}}\bar{y})_j = 0, \quad j = 1, \ldots, n.$$

This means that for all $j = 1, \ldots, n$, at most one of the two factors $\bar{x}_j$ and $(c - A^{\mathrm{T}}\bar{y})_j$ can be positive. Thus, if the jth primal variable $\bar{x}_j$ is positive, the jth dual constraint $(A_{\bullet j})^{\mathrm{T}}\bar{y} \leq c_j$ must hold with equality. By the same token, if the jth dual constraint is *slack* (i.e., does not hold as equality), then the jth decision variable must be zero. These are known as ***complementary slackness conditions***; they come up frequently in optimization work.

References

D. Bertsimas and J.N. Tsitsiklis [1997], *Introduction to Linear Optimization*, Athena Scientific, Belmont, Mass.

D. Gale [1960], *The Theory of Linear Economic Models*, McGraw-Hill Book Company, Inc., New York.

H.W. Kuhn and A.W. Tucker, eds. [1956], *Linear Inequalities and Related Systems*, Princeton University Press, Princeton, N.J.

O.L. Mangasarian [1969], *Nonlinear Programming*, McGraw-Hill, New York. [Reprinted in softcover by SIAM Publications.]

A. Schrijver [1986], *Theory of Linear and Integer Programming*, John Wiley & Sons, Chichester.

J. Stoer and C. Witzgall [1970], *Convexity and Optimization in Finite Dimensions I*, Springer-Verlag, New York.

3. OPTIMALITY CONDITIONS

3.1 Generalities

As defined in these notes, the linear programming problem in standard form is

$$\begin{aligned} \text{minimize} \quad & c^{\mathrm{T}}x \\ \text{subject to} \quad & Ax = b \\ & x \geq 0. \end{aligned}$$

Corollary 2 at the end of Section 2.6 (Handout No. 5, page 7), states the so-called ***complementary slackness conditions***

$$\begin{aligned} A\bar{x} &= b \\ A^{\mathrm{T}}\bar{y} &\leq c \\ \bar{x}^{\mathrm{T}}(c - A^{\mathrm{T}}\bar{y}) &= 0 \\ \bar{x} &\geq 0. \end{aligned}$$

These are conditions which—for some $\bar{y}$—must hold if $\bar{x}$ is an optimal solution of the given LP. In other words, these are ***necessary*** conditions of optimality for the stated linear programming problem. The vector $\bar{y}$ is, in fact, an optimal solution of the dual of the given (primal) linear programming problem. These complementary slackness conditions are also ***sufficient*** conditions. That is to say, if $\bar{x}$ and $\bar{y}$ satisfy them, then, $\bar{x}$ is optimal for the primal problem (and $\bar{y}$ is optimal for the dual problem). Making minor modifications, we can develop corresponding optimality conditions for other forms of the linear programming problem. For the time being, we may regard the treatment of optimality conditions in linear programming as complete.

In these notes, we discuss an important optimization-theoretic question: How does one recognize an optimal solution to a ***nonlinear*** programming problem?

In all forms of mathematical programming, a ***feasible solution*** of a given problem is a vector that satisfies the constraints of the problem. If the problem has the form

$$\text{(P)} \qquad \text{minimize } f(x) \quad \text{subject to} \quad x \in S,$$

a ***global minimizer*** is a vector $\bar{x}$ such that

$$\bar{x} \in S \quad \text{and} \quad f(\bar{x}) \leq f(x) \quad \text{for all} \ \ x \in S.$$

Finding a global minimizer is normally the goal of any minimization problem, but sometimes one has to settle for a ***local minimizer***, that is, a vector $\bar{x}$ such that

$$\bar{x} \in S \quad \text{and} \quad f(\bar{x}) \leq f(x) \quad \text{for all} \ \ x \in S \cap N(\bar{x})$$

where $N(\bar{x})$ is a ***neighborhood*** of $\bar{x}$. Typically, $N(\bar{x})$ is just some open ball $B_\delta(\bar{x})$ centered at $\bar{x}$ and having what is deemed to be a suitably small radius, $\delta > 0$.

The value of the objective function f at a global minimizer or a local minimizer is also of interest. Accordingly, we can speak of $f(\bar{x})$ as the ***global minimum value*** or a ***local minimum value***, according to whether $\bar{x}$ is a global minimizer or a local minimizer, respectively.

In many instances, nonlinear programming problems are specified by functions that are ***differentiable*** or even ***continuously differentiable*** over the feasible region. Sometimes the functions are ***twice continuously differentiable***. The theory distinguishes these two cases and develops ***first-order optimality conditions*** and ***second-order optimality conditions***. As the names suggest, first-order optimality conditions involve derivatives of order no higher than one, whereas second-order optimality conditions involve derivatives no higher than two.

To put this in a familiar context, consider a differentiable function f of one variable defined on an open set S. If $\bar{x}$ is a local minimizer of f, then $f'(\bar{x}) = 0$. This is a first-order necessary condition. As is well known, this condition is not, in general, sufficient. It does not distinguish between local minimizers, local maximizers or points of inflection. However, if the function is twice differentiable and (in addition to the first-order condition) the second-order condition $f''(\bar{x}) > 0$ is satisfied, then $\bar{x}$ is a local minimizer.

The theory of optimality conditions developed in this chapter will be of a more general nature. In addition to being applicable to the multivariate case, it will pertain to ***constrained optimization problems***. The types of constraints considered will first be equations, then inequalities, and finally combinations of the two. We shall begin by dealing with first-order optimality conditions and then turn to second-order conditions.

3.2 Classical first-order conditions

In this section we restrict our attention to first-order optimality conditions.

3.2.1 Descent directions

In many cases, the functions involved in the specification of a nonlinear program are ***differentiable***. When this is so, there are usually conditions involving (partial) derivatives that must hold. The following is a familiar simple case from multivariate differential calculus.

Proposition 1. Let $U \subset R^n$ be open and suppose $f : U \to R$ is differentiable at the point $\bar{x} \in U$. If $\bar{x}$ is a local minimizer of f on U, then[13]

$$\nabla f(\bar{x}) = \left(\frac{\partial f(\bar{x})}{\partial x_1}, \ldots, \frac{\partial f(\bar{x})}{\partial x_n} \right) = 0. \qquad \square$$

It should be noted that this is a first-order ***necessary*** condition of local optimality. This vanishing of the gradient vector must occur when the hypotheses of the proposition are fulfilled. The conditions are not sufficient, however. That is, a vector that makes the gradient vector vanish need not be a local minimizer. Plenty of illustrations of this fact are available with differentiable functions of a single variable. In that case the point in question could be a local maximizer or a point of inflection. As a two-variable example one thinks immediately of the function $f(x_1, x_2) = x_1^2 - x_2^2$. The gradient of this function vanishes (i.e., equals the zero vector) at the origin, but the origin is not a local minimizer (or maximizer) but rather a ***saddlepoint***.

Let $U \subset R^n$ and let $f : U \to R$ be a differentiable function on U. If $\bar{x} \in U$ and there exists a vector p such that

$$p^{\mathrm{T}} \nabla f(\bar{x}) < 0,$$

then there exists a scalar $\bar{\tau} > 0$ such that

$$f(\bar{x} + \tau p) < f(\bar{x}) \quad \text{for all} \quad \tau \in (0, \bar{\tau}).$$

The vector p (above) is called a ***descent direction*** at $\bar{x}$.

Recall from multivariate calculus that if $\nabla f(\bar{x}) \neq 0$, then $\nabla f(\bar{x})$ is the direction of ***steepest ascent*** at $\bar{x}$. (This follows from the Cauchy-Schwarz inequality.) It then follows that $-\nabla f(\bar{x})$ is the direction of ***steepest descent*** at $\bar{x}$.

[13] Remember that we consider all vectors to be columns unless they are transposed to become row vectors. In particular, when we write $\nabla f(\bar{x}) = \left(\frac{\partial f(\bar{x})}{\partial x_1}, \ldots, \frac{\partial f(\bar{x})}{\partial x_n} \right)$, we are actually referring to a ***column vector*** even though it is written horizontally as if it were a row. The use of parentheses around the components of the vector is intended to indicate that we are referring to a column vector.

3.2.2 The method of Lagrange multipliers

Let us consider the classical equality-constrained problem

$$\begin{array}{lll} & \text{minimize} & f(x) \\ \text{(P)} & \text{subject to} & c_i(x) = 0 \quad i \in \mathcal{E} \\ & & x \in R^n \end{array}$$

Let the index set $\mathcal{E} = \{1, \ldots, m\}$. Collectively, the functions c_i above can be thought of as the components of a ***mapping***

$$c(x) = \begin{bmatrix} c_1(x) \\ \vdots \\ c_m(x) \end{bmatrix}$$

from R^n to R^m. Thus $x = (x_1, \ldots, x_n) \mapsto c(x) = (c_1(x), \ldots, c_m(x))$.

Suppose the functions $f, c_1, \ldots, c_m$ are differentiable on R^n. Accordingly, each of these functions has a gradient at every point $x \in R^n$ so we can form the ***Jacobian matrix*** of the mapping c. Our definition of the Jacobian matrix will be

$$\left[\frac{\partial c_i(x)}{\partial x_j}\right].$$

We are defining the Jacobian of c to be the matrix whose rows are the ***transposes*** of the gradients of functions $c_1, \ldots, c_m$.

The notation used for the Jacobian of the mapping c in the mathematical (programming) literature is far from standard.[14] The one used here will be

$$\nabla c(x) = \left[\frac{\partial c_i(x)}{\partial x_j}\right].$$

It is helpful to remember that when the mapping $c(x)$ is the linear transformation Ax, the Jacobian matrix is just $\nabla c(x) = A$. Notice that in this special case, the Jacobian is a ***constant***, whereas in general it is mapping from R^n to $R^{m \times n}$.

For what we are going to consider below, it makes sense to add the restriction that $m \leq n$. This is because we are going to need a linear independence assumption on the rows of the Jacobian matrix. Since ∇c is an $m \times n$ matrix, its rows can be linearly independent only if $m \leq n$. A vector $\bar{x}$ at which $\nabla c(\bar{x})$ has linearly independent rows is sometimes called ***regular point*** of the mapping c. (See, for example, Bertsekas [1995].)

[14] In fact, it is chaotic.

The following is another result from multivariate differential calculus. It gives necessary conditions of local optimality for the equality-constrained problem (P). It should be remembered that the objective function f and the constraint functions c_i are assumed to be continuously differentiable at the local minimizer.

Theorem 1 (Lagrange). Let $\bar{x}$ be a local minimizer of (P). If the functions f and $c_1, \ldots, c_m$ are continuously differentiable at $\bar{x}$ and the Jacobian matrix $\nabla c(\bar{x})$ has rank m, then there exist scalars $\bar{y}_1, \ldots, \bar{y}_m$ such that

$$\nabla f(\bar{x}) - \sum_{i=1}^{m} \bar{y}_i \nabla c_i(\bar{x}) = 0. \qquad \square$$

Recall that the numbers $\bar{y}_1, \ldots, \bar{y}_m$ are called ***Lagrange multipliers***; the function

$$L(x, y) = f(x) - \sum_{i=1}^{m} y_i c_i(x)$$

is called the ***Lagrangian function***, or simply ***Lagrangian***, for (P). For any such problem[15], one can always form the Lagrangian function. In some circumstances, however, the assumptions made in the theorem might not hold, and hence the theorem's conclusion might not be valid.

The theorem stated above says, in effect, that when $\bar{x}$ is a local minimizer for (P), and the Jacobian matrix has full rank, then there exists a vector $\bar{y} \in R^m$ such that the pair $(\bar{x}, \bar{y})$ is a ***stationary point*** of the associated Lagrangian function. To appreciate this, it is helpful to define the ***partial gradients***

$$\nabla_x L = \left(\frac{\partial L}{\partial x_1}, \ldots, \frac{\partial L}{\partial x_n} \right).$$

$$\nabla_y L = \left(\frac{\partial L}{\partial y_1}, \ldots, \frac{\partial L}{\partial y_m} \right).$$

Then, first of all, it is clear that $\nabla L(x, y) = (\nabla_x L(x, y), \nabla_y L(x, y))$. Moreover,

$$\nabla_x L(x, y) = \nabla f(x) - \sum_{i=1}^{m} y_i \nabla c_i(x)$$

$$\nabla_y L(x, y) = -c(x)$$

3.2.3 The need for a regularity condition

[15]There is a slightly subtle point here. The Lagrangian function for (P) is given by the functions used to represent (P). If the functions are changed, the Lagrangian will change correspondingly. Even if the feasible region is not altered by the change in its representation, the properties of the functions used may be different, possibly invalidating certain assumptions.

In the preceding theorem, the assumption that $\nabla c(\bar{x})$ has full rank is an example of a ***regularity condition.*** Lagrange's theorem is not valid unless this regularity condition holds.

Example 1. Consider the problem

$$\begin{array}{ll} \text{minimize} & x_1 \\ \text{subject to} & x_1^2 + (x_2 - 1)^2 - 1 = 0 \\ & x_1^2 + (x_2 + 1)^2 - 1 = 0 \end{array}$$

Notice that this problem has exactly one feasible point: $\bar{x} = (0, 0)$, which must therefore be optimal. In this case we have

$$\nabla f(\bar{x}) = (1, 0)$$

$$\nabla c_1(\bar{x}) = (0, -2)$$

$$\nabla c_2(\bar{x}) = (0, 2)$$

Under these circumstances, there cannot exist Lagrange multipliers $\bar{y}$ satisfying the condition $\nabla_x L(\bar{x}, \bar{y}) = 0$.

The explanation for what happens in Example 1 is very simple. Indeed, the conclusion of Lagrange's Theorem says that $\nabla_x L(\bar{x}, \bar{y}) = 0$. This equation can be written in the form

$$\nabla f(\bar{x}) = \sum_{i=1}^{m} \bar{y}_i \nabla c_i(\bar{x}),$$

an interpretation of which is the statement that $\nabla f(\bar{x})$ belongs to the column space of the matrix $(\nabla c(\bar{x}))^{\mathrm{T}}$. When the latter is rank deficient (as in the example), there will be vectors in R^n that do not belong to this column space, namely all nonzero vectors in the orthogonal complement of the column space of the transposed Jacobian matrix, $(\nabla c(\bar{x}))^{\mathrm{T}}$. Since any vector in the orthogonal complement can be regarded as the gradient of an affine function, it follows that when the Jacobian matrix of c at a local minimizer has linearly dependent rows, there will always be functions f for which $\nabla_x L(\bar{x}, \bar{y}) \neq 0$, that is, where the conclusion of Lagrange's Theorem does not hold.

This material should be somewhat familiar from your multivariate calculus course. If not, you might wish to brush up on it.[16]

3.3 First-order conditions for inequality-constrained problems

Let us now consider the inequality-constrained problem

$$\text{(P)} \qquad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & c_i(x) \geq 0 \quad i \in \mathcal{I} \end{array}$$

[16] See, for example, Fleming [1977, p. 161].

In this case we shall let $\mathcal{I} = \{1, \ldots, m\}$ and use the functions $c_1, \ldots, c_m$ as coordinates of the mapping $g : R^n \to R^m$. If, at some vector $\bar{x}$, we have $c_i(\bar{x}) = 0$, then the i-th constraint is said to be ***active*** or ***binding*** at $\bar{x}$. Relative to this problem, we define the (possibly empty) set

$$\mathcal{A}(\bar{x}) := \{i \in \mathcal{I} : c_i(\bar{x}) = 0\}.$$

3.3.1 The KKT constraint qualification (regularity condition)

Let $\bar{x}$ be a feasible point for the inequality-constrained problem (above) in which all the functions are differentiable. Assume $\mathcal{A}(\bar{x}) \neq \emptyset$. We say the ***Karush-Kuhn-Tucker constraint qualification*** is satisfied at $\bar{x}$ if for every nonzero solution v of the inequality system

$$v^{\mathrm{T}} \nabla c_i(\bar{x}) \geq 0 \quad \text{for all} \quad i \in \mathcal{A}(\bar{x})$$

there exists a differentiable curve

$$\gamma : [0, 1] \to R^n$$

whose image is contained in the feasible region such that

$$\gamma(0) = \bar{x}, \quad \gamma'(0) = \kappa v \quad \text{for some } \kappa > 0.$$

Theorem 2 (Karush [1939]; Kuhn & Tucker [1951]). If $\bar{x}$ is a local minimizer for (P), and the (KKT) constraint qualification is satisfied at $\bar{x}$, then there exist numbers $\bar{y}_1, \ldots, \bar{y}_m$ such that

$$\nabla f(\bar{x}) - \sum_{i=1}^{m} \bar{y}_i \nabla c_i(\bar{x}) = 0$$
$$\left.\begin{array}{c} \bar{y}_i \geq 0 \\ \bar{y}_i c_i(\bar{x}) = 0 \end{array}\right\} \; i = 1, \ldots, m$$

Proof. If $\mathcal{A}(\bar{x}) = \emptyset$, take all the $\bar{y}_i = 0$. If $\mathcal{A}(\bar{x}) \neq \emptyset$, consider the linear inequality system:

$$v^{\mathrm{T}} \nabla f(\bar{x}) < 0$$
$$v^{\mathrm{T}} \nabla c_i(\bar{x}) \geq 0 \quad \text{for all} \quad i \in \mathcal{A}(\bar{x})$$

Notice that any solution v of this system must not be a zero vector. Because of the constraint qualification, this system cannot have a solution because $\bar{x}$ is a local minimizer. By Farkas's Lemma, there exist scalars $\bar{y}_i \geq 0$ for all $i \in \mathcal{A}(\bar{x})$ such that

$$\nabla f(\bar{x}) - \sum_{i \in A(\bar{x})} \bar{y}_i \nabla c_i(\bar{x}) = 0.$$

For $i \notin \mathcal{A}(\bar{x})$, take $\bar{y}_i = 0$. This does it. □

The first-order necessary conditions of local optimality (in the theorem above) are called the ***Karush-Kuhn-Tucker conditions.*** The vector $\bar{x}$ is called a ***KKT stationary point***, and $(\bar{x}, \bar{y})$ is called a ***KKT pair.*** Thus, to say that $\bar{x}$ is a KKT stationary point means that there exists a vector $\bar{y}$ such that $(\bar{x}, \bar{y})$ is a KKT pair, i.e., satisfies the KKT (first-order necessary) conditions of local optimality.

Example 2. The need for a constraint qualification is illustrated by the following problem.

$$\begin{array}{ll} \text{minimize} & -x_1 \\ \text{subject to} & (1-x_1)^3 - x_2 \geq 0 \\ & x_1 \geq 0 \\ & x_2 \geq 0 \end{array} .$$

The feasible region of this problem is the compact (but nonconvex) subset of the first quadrant lying between the nonnegative axes and the cubic curve. Here we see that $\bar{x} = (1, 0)$ is the one and only global minimizer. The corresponding index set of active constraints is $\mathcal{A}(\bar{x}) = \{1, 3\}$. Denoting the objective function by f and the three constraints in order as c_1, c_2, c_3, we find that

$$\nabla f(\bar{x}) = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \qquad \nabla c_1(\bar{x}) = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \qquad \nabla c_3(\bar{x}) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

As in the previous example, this reveals that the stationarity condition cannot hold.

Definition. Let S be a nonempty subset of R^n, and let $\bar{x} \in \operatorname{cl} S$ be given. The ***cone of feasible directions*** at $\bar{x}$ is the set

$$\mathcal{D} := \{d \in R^n : d \neq 0,\ \bar{x} + \lambda d \in S \text{ for all } \lambda \in (0, \delta) \text{ for some } \delta > 0\}.$$

Note that the cone of feasible directions is not a cone in the strict sense of the word because it cannot contain the zero vector which is ordinarily required in the definition of a cone. This caveat pertains to the following term.

Definition. Let the function f be differentiable at the point $\bar{x} \in R^n$. If $\nabla f(\bar{x}) \neq 0$, the ***cone of descent directions*** at $\bar{x}$ is the open halfplane

$$\mathcal{F}_0 := \{d \in R^n : d^{\mathrm{T}} \nabla f(\bar{x}) < 0\}.$$

If S is the feasible region of a minimization problem in which f is the objective function, then $\mathcal{F}_0 \cap \mathcal{D} = \emptyset$ can be regarded as a ***geometric condition*** for $\bar{x} \in \operatorname{cl} S$ to be a local minimizer.

Now consider a nonlinear programming problem

$$\begin{array}{lll} & \text{minimize} & f(x) \\ \text{(P)} & \text{subject to} & c_i(x) \geq 0 \quad i = 1, \ldots, m \\ & & x \in X \quad \text{nonempty, open} \end{array}$$

Define the set

$$\mathcal{G}_0 := \{d : d^{\mathrm{T}} \nabla c_i(\bar{x}) > 0 \quad \text{for all } i \in \mathcal{A}(\bar{x})\}.$$

Then the statement

$$\mathcal{F}_0 \cap \mathcal{G}_0 = \emptyset$$

can be regarded as a necessary condition of local optimality for (P).

Example 3. Consider the optimization problem

$$\begin{array}{lll} & \text{minimize} & (x_1 - 1)^2 + (x_2 - 1)^2 \\ & \text{subject to} & (1 - x_1 - x_2)^3 \geq 0 \\ \text{(P)} & & x_1 \geq 0 \\ & & x_2 \geq 0 \end{array}$$

This problem has a unique optimal solution: $\bar{x} = (\frac{1}{2}, \frac{1}{2})$. The feasible region of (P) is the same as that with constraints

$$x_1 + x_2 \leq 1, \quad x_1 \geq 0, \quad x_2 \geq 0.$$

Let $\tilde{x} = (\tilde{x}_1, \tilde{x}_2)$ be any point satisfying $\tilde{x}_1 + \tilde{x}_2 = 1$. With $c_1(x) = (1 - x_1 - x_2)^3$, we have

$$\nabla c_1(\tilde{x}) = 3(1 - \tilde{x}_1 - \tilde{x}_2)^2 \begin{bmatrix} -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

so $\mathcal{G}_0 = \emptyset$ and hence $\mathcal{F}_0 \cap \mathcal{G}_0 = \emptyset$.

This example illustrates the fact that the necessary condition of optimality given above can be satisfied by infinitely many ***nonoptimal*** points in the feasible region as well as by the optimal solution. The example also gives an instance of a (feasible) set that can be represented by linear constraints as well as nonlinear constraints. It must be conceded, however, that this example is a bit of a museum piece. Such cases are not normally encountered in practice.

3.3.2 Fritz John's Theorem

Theorem 3 (F. John [1948]). If $\bar{x}$ is a local minimizer of the optimization problem

$$\begin{array}{lll} & \text{minimize} & f(x) \\ \text{(P)} & \text{subject to} & c_i(x) \geq 0 \quad i = 1, \ldots, m \\ & & x \in X \quad \text{nonempty, open} \end{array}$$

in which the functions f and c_i $(i = 1, \ldots, m)$ are differentiable on X, then there exists a set of nonnegative scalars $\lambda_0, \lambda_1, \ldots, \lambda_m$ not all of which are zero such that

$$\lambda_i c_i(\bar{x}) = 0, \quad i = 1, \ldots, m,$$

and

$$\lambda_0 \nabla f(\bar{x}) - \sum_{i=1}^{m} \lambda_i \nabla c_i(\bar{x}) = 0.$$

Proof. We may assume that $\mathcal{A}(\bar{x}) \neq \emptyset$. Since $\bar{x}$ is a local minimizer of (P), the system

$$\begin{aligned} v^{\mathrm{T}} \nabla f(\bar{x}) &< 0 \\ v^{\mathrm{T}} \nabla c_i(\bar{x}) &> 0 \quad \text{for all } i \in \mathcal{A}(\bar{x}) \end{aligned}$$

has no solution. Accordingly, by Gordan's Theorem (Handout No. 5, page 3), there exist nonnegative scalars λ_0, λ_i $(i \in \mathcal{A}(\bar{x}))$ not all of which are zero such that

$$\lambda_0 \nabla f(\bar{x}) - \sum_{i \in \mathcal{A}(\bar{x})} \lambda_i \nabla c_i(\bar{x}) = 0.$$

If $i \in \{1, \ldots, m\} \setminus \mathcal{A}(\bar{x})$, define $\lambda_i = 0$. This completes the proof. □

Remark. Notice that this theorem requires no constraint qualification, yet we know that a constraint qualification is required. So what's the catch? The catch is that the scalar λ_0 might equal zero. Notice, though, that if $\lambda_0 > 0$, then the conclusion of the Karush-Kuhn-Tucker theorem holds with the multipliers $\lambda_1/\lambda_0, \ldots, \lambda_m/\lambda_0$. Notice also that if $\lambda_0 = 0$, then the vectors $\nabla c_i(\bar{x})$ must be (positively) linearly dependent. Hence any condition that rules out such linear dependence will imply $\lambda_0 > 0$.

3.3.3 Variables with sign restrictions

The preceding discussion applies to problems with sign restricted variables. Consider, for instance, the problem

$$\text{(P)} \qquad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & c(x) \geq 0 \\ & x \geq 0 \end{array}$$

Corollary 1. If $\bar{x}$ is a local minimizer of (P) and the constraint qualification is satisfied at $\bar{x}$, then

there exist numbers $\bar{y}_1, \ldots, \bar{y}_m$ such that

$$\left.\begin{array}{c} \nabla f(\bar{x}) - \sum_{i=1}^{m} \bar{y}_i \nabla c_i(\bar{x}) \geq 0 \\ \bar{y}_i \geq 0 \\ \bar{y}_i c_i(\bar{x}) = 0 \end{array}\right\} \; i = 1, \ldots, m \qquad .$$

$$\bar{x}^{\mathrm{T}} \left[\nabla f(\bar{x}) - \sum_{i=1}^{m} \bar{y}_i \nabla c_i(\bar{x}) \right] = 0. \qquad \square$$

3.3.4 Some sufficient conditions for optimality[17]

Let us consider again the inequality-constrained problem

$$\text{(P)} \qquad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & c_i(x) \geq 0 \quad i = 1, \ldots, m \end{array}$$

Theorem 4. If f and $-c_1, \ldots, -c_m$ are differentiable convex functions, then the first-order (KKT) optimality conditions are sufficient for the global optimality of a feasible vector in the inequality-constrained problem, (P).

Proof. Let $(\bar{x}, \bar{y})$ be a KKT pair for (P) in which $\bar{x}$ is a feasible vector. Consider the Lagrangian function $L(x, y) = f(x) - y^{\mathrm{T}} c(x)$ associated with (P). Let x be feasible and let y be nonnegative. Then, by our hypotheses, L is a convex, differentiable function of x. Hence by the gradient inequality applied to L

$$L(x, \bar{y}) \geq L(\bar{x}, \bar{y}) + (x - \bar{x})^{\mathrm{T}} \nabla_x L(\bar{x}, \bar{y}) \quad \text{for all feasible } x.$$

More explicitly,

$$f(x) - \bar{y}^{\mathrm{T}} c(x) \geq f(\bar{x}) - \bar{y}^{\mathrm{T}} c(\bar{x}) + (x - \bar{x})^{\mathrm{T}} [\nabla f(\bar{x}) - \bar{y}^{\mathrm{T}} \nabla c(\bar{x})].$$

Hence,

$$f(x) \geq f(\bar{x}) + \bar{y}^{\mathrm{T}} c(x) \geq f(\bar{x}).$$

This proves that $\bar{x}$ is a global minimizer for (P). $\square$

Remark. Notice that the preceding theorem does not mention the constraint qualification (CQ). It is simply *given* that $(\bar{x}, \bar{y})$ is a KKT pair, and this being the case, no CQ assumption is required. The convexity assumption is a big one, however. This assumption can be weakened, slightly, as we shall see later on.

[17] Theorem 4 below states the frequently invoked sufficient conditions for optimality in (P). These can be generalized by using the concepts of *quasiconvexity* and *pseudoconvexity* which we do not have time to cover in this course. For details, see Mangasarian [1969, Chapters 9 and 10].

3.4 Problems with equality constraints

We now consider first-order optimality conditions for problems having both inequality and equality constraints. These can be denoted[18]

$$\text{(P)} \qquad \begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & c_i(x) \geq 0 & i \in \mathcal{I} \\ & h_i(x) = 0 & i \in \mathcal{E} \end{array}$$

Typically, we take

$$\mathcal{I} = \{1, \ldots, m\} \quad \text{and} \quad \mathcal{E} = \{1, \ldots, \ell\}.$$

Our aim in this discussion is to establish analogues of the Fritz John and Karush-Kuhn-Tucker first-order necessary conditions of optimality in (P). For any feasible point $\bar{x}$ of (P) we have the sets

$$\mathcal{A}(\bar{x}) = \{i \in \mathcal{I} : c_i(\bar{x}) = 0\}$$

$$\mathcal{F}_0 = \{d \in R^n : d^{\mathrm{T}} \nabla f(\bar{x})\} < 0\}$$

$$\mathcal{G}_0 = \{d : d^{\mathrm{T}} \nabla c_i(\bar{x}) > 0 \quad \text{for all } i \in \mathcal{A}(\bar{x})\}.$$

To these sets, we add

$$\mathcal{H}_0 = \{d : d^{\mathrm{T}} \nabla h_i(\bar{x}) = 0 \quad \text{for all } i \in \mathcal{E}\}.$$

If h were an affine mapping, it would be intuitively clear that $\mathcal{F}_0 \cap \mathcal{G}_0 \cap \mathcal{H}_0 = \emptyset$ would be a necessary condition of local optimality of $\bar{x}$, for an element of this set would be a feasible descent direction. The object of the discussion below[19] is to show that if $\bar{x}$ is a regular point with respect to h, the set $\mathcal{F}_0 \cap \mathcal{G}_0 \cap \mathcal{H}_0$ is empty even if h is not affine.

3.4.1 A necessary condition of optimality

Theorem 5. Let $\bar{x}$ be a local minimizer for (P). If the functions c_i are continuous at $\bar{x}$ for all $i \notin \mathcal{A}(\bar{x})$, the functions c_i are differentiable at $\bar{x}$ for all $i \in \mathcal{A}(\bar{x})$, the functions h_i are continuously differentiable at $\bar{x}$ for all $i \in \mathcal{E}$, and the rows of $\nabla h(\bar{x})$ are linearly independent, then

$$\mathcal{F}_0 \cap \mathcal{G}_0 \cap \mathcal{H}_0 = \emptyset.$$

Proof. Suppose there is a vector $v \in \mathcal{F}_0 \cap \mathcal{G}_0 \cap \mathcal{H}_0$. The object is to show that there is a curve γ in the feasible region of (P) with that starts at $\bar{x}$ and has a positive multiple of v as its tangent at that point. This will give a contradiction to the local optimality of $\bar{x}$.

[18] Notational practice varies on how the functions involved in the equality constraints of (P) are represented. Some authors (see, e.g., Gill, Murray, and Wright [1981]) use the same letter for both the equality and inequality constraints and distinguish the two groups only by their index sets. Others use different letters for the functions in addition to different index set names for their subscripts. We follow the latter notational system because it seems to facilitate the discussion somewhat.

[19] This material is adapted from Bazaraa, Sherali and Shetty [1993].

For $\theta \geq 0$ consider the differential equation and boundary condition

$$\frac{d}{d\theta}\gamma(\theta) = P(\theta)v, \quad \gamma(0) = \bar{x}.$$

In this differential equation, $P(\theta)$ denotes a projection matrix into the null space of $\nabla h[\gamma(\theta)]$. The existence of a solution to this differential equation follows from the fact that h is continuously differentiable at $\bar{x}$ and $\nabla h(\bar{x})$ has full rank. The matrix $P(\cdot)$ is continuous in θ and has the property $\gamma(\theta) \to \bar{x}$ as $\theta \to 0^+$.

Now if $\theta > 0$ is sufficiently small, we have

$$\frac{d}{d\theta}c_i(\gamma(\theta)) = (\nabla c_i(\gamma(\theta)))^{\mathrm{T}} P(\theta)v \quad \text{for all } i \in \mathcal{A}(\bar{x}).$$

By definition, v is in the null space of $\nabla h(\bar{x})$ so that when $\theta = 0$ the equation $P(0)v = v$ is satisfied. Next we have

$$\frac{d}{d\theta} c_i(\gamma(0)) = v^{\mathrm{T}} \nabla c_i(\bar{x}) > 0$$

for all $i \in \mathcal{A}(\bar{x})$. This means $c_i(\gamma(\theta)) > 0$ for $\theta > 0$ and sufficiently small. Indeed, for sufficiently positive θ, the curve satisfies $c(\gamma(\theta)) > 0$.

Next we need to show that when $\theta > 0$ is sufficiently small, the curve satisfies $h_i(\gamma(\theta)) = 0$ for all $i \in \mathcal{E}$. Now the mean value theorem implies that

$$h_i(\gamma(\theta)) = h_i(\gamma(0)) + \theta \frac{d}{d\theta} h_i(\gamma(\bar{\theta})) = \theta \frac{d}{d\theta} h_i(\gamma(\bar{\theta}))$$

where $\bar{\theta} \in (0, \theta)$. By the chain rule, we obtain

$$\frac{d}{d\theta} h_i(\gamma(\bar{\theta})) = \nabla h_i(\gamma(\bar{\theta}))^{\mathrm{T}} P(\bar{\theta}) v.$$

This implies that $P(\bar{\theta})v$ lies in the null space of $\nabla h_i(\gamma(\bar{\theta}))$, so we have

$$\frac{d}{d\theta} h_i(\gamma(\bar{\theta})) = 0.$$

It now follows that $h_i(\gamma(\theta)) = 0$ for all $i \in \mathcal{E}$ if $\gamma > 0$ is sufficiently small.

Using similar reasoning, we can show that

$$\frac{d}{d\theta} f(\gamma(0)) = \nabla f(\bar{x})^{\mathrm{T}} v < 0$$

whence $f(\gamma(\theta)) < f(\bar{x})$ for all sufficiently small $\theta > 0$. Assembling these facts, we see that a contradiction has been obtained. Hence $\mathcal{F}_0 \cap \mathcal{G}_0 \cap \mathcal{H}_0 = \emptyset$ as asserted. □

3.4.2 The Fritz John Theorem

Theorem 6. Let $\bar{x}$ be a local minimizer for (P). If the functions c_i are continuous at $\bar{x}$ for all $i \notin \mathcal{A}(\bar{x})$, the functions c_i are differentiable at $\bar{x}$ for all $i \in \mathcal{A}(\bar{x})$, and the functions h_i are continuously differentiable at $\bar{x}$ for all $i \in \mathcal{E}$, then there exist multipliers $\lambda_0, \lambda_1, \ldots, \lambda_m, \mu_1, \ldots, \mu_\ell$ not all zero such that

$$\lambda_0 \nabla f(\bar{x}) - \sum_{i=1}^{m} \lambda_i \nabla c_i(\bar{x}) - \sum_{i=1}^{\ell} \mu_i \nabla h_i(\bar{x}) = 0$$

$$\lambda_i c_i(\bar{x}) = 0 \qquad \text{for all } i = 1, \ldots, m$$

$$\lambda_0, \lambda_i \geq 0 \qquad \text{for all } i = 1, \ldots, m$$

Proof. First note that if the rows of the Jacobian matrix $\nabla h(\bar{x})$ are linearly dependent, then taking $\lambda_0, \lambda_1, \ldots, \lambda_m = 0$, we can satisfy the required conditions, albeit trivially. The case where the rows of the Jacobian matrix are linearly independent is more interesting. In this case, we can

invoke the conclusion of the previous theorem. The existence of the desired multipliers follows from an alternative theorem, rather like that of Gordan. We leave the details as an exercise. □

3.4.3 The KKT Theorem again

Theorem 7. Let $\bar{x}$ be a local minimizer for (P). Assume the functions c_i are differentiable at $\bar{x}$ for all $i \in \mathcal{I}$, and the functions h_i are continuously differentiable at $\bar{x}$ for all $i \in \mathcal{E}$. If all the vectors $\nabla c_i(\bar{x})$ for $i \in \mathcal{A}(\bar{x})$ and $\nabla h_i(\bar{x})$ for $i \in \mathcal{E}$ are linearly independent, then there exist (unique) multipliers $\lambda_1, \ldots, \lambda_m, \mu_1, \ldots, \mu_\ell$ such that

$$\begin{aligned} \nabla f(\bar{x}) - \sum_{i=1}^{m} \lambda_i \nabla c_i(\bar{x}) - \sum_{i=1}^{\ell} \mu_i \nabla h_i(\bar{x}) &= 0 \\ \lambda_i c_i(\bar{x}) &= 0 \qquad \text{for all } i = 1, \ldots, m \\ \lambda_i &\geq 0 \qquad \text{for all } i = 1, \ldots, m \end{aligned}$$

Proof. Since the hypotheses of this theorem are stronger[20] than those of the corresponding (Fritz John) Theorem 6, the conclusion of that theorem holds. In particular, the "Fritz John" multiplier λ_0 must be positive, for otherwise we obtain a contradiction of the linear independence assumption. By virtue of the homogeneity of the first-order optimality conditions of the Fritz John Theorem, we may assume that $\lambda_0 = 1$. Under this condition, the uniqueness of the remaining multipliers follows from the linear independence assumption. □

Remark. The linear independence assumption used in Theorem 7 can be weakened somewhat. This is accomplished in the so-called ***Mangasarian-Fromovitz constraint qualification***. See Mangasarian and Fromovitz [1967], Mangasarian [1969, p. 173] and Bazaraa, Sherali, and Shetty [1993, Exercise 5.20, p. 197].

3.5 Saddlepoint problems

Let $F : A \times B \to R$ be a given function. If $(\bar{x}, \bar{y}) \in A \times B$ and

$$F(\bar{x}, y) \leq F(\bar{x}, \bar{y}) \leq F(x, \bar{y}) \quad \text{for all} \quad (x, y) \in A \times B,$$

then $(\bar{x}, \bar{y})$ is called a ***saddlepoint*** of F on $A \times B$.

Now consider the inequality-constrained problem

$$\text{(P)} \qquad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & c(x) \geq 0 \end{array}$$

and define its associated Lagrangian function $L(x, y) = f(x) - y^{\mathrm{T}} c(x)$ on the set $A \times B = R^n \times R^m_+$. (Notice that this means that c must be a mapping from R^n into R^m.)

[20] We have assumed differentiability of all the c_i, not just those which are active at $\bar{x}$.

The following theorem relates saddlepoints of L to (global) minima for (P).

Theorem 8. If $(\bar{x}, \bar{y})$ is a saddlepoint of L (as defined above), then $\bar{x}$ solves (P).

Proof. The vector $\bar{x}$ is feasible for (P), for if $c(\bar{x})$ has a negative component, then the inequality $L(\bar{x}, y) \leq L(\bar{x}, \bar{y})$ for all $y \geq 0$ cannot hold. Moreover, $\bar{y}^{\mathrm{T}} c(\bar{x}) = 0$ since

$$y^{\mathrm{T}} c(\bar{x}) \geq \bar{y}^{\mathrm{T}} c(\bar{x}) \geq 0 \quad \text{for all} \quad y \geq 0,$$

and the value 0 is attainable. The vector $\bar{x}$ is a global minimizer, for otherwise there exists a vector $\tilde{x}$ such that $c(\tilde{x}) \geq 0$ and

$$f(\tilde{x}) - \bar{y}^{\mathrm{T}} c(\tilde{x}) \leq f(\tilde{x}) < f(\bar{x}) = f(\bar{x}) - \bar{y}^{\mathrm{T}} c(\bar{x}) \leq f(\tilde{x}) - \bar{y}^{\mathrm{T}} c(\tilde{x}).$$

This is a contradiction. □

Remarks. The condition of being a saddlepoint of the Lagrangian function L for the problem (P) is obviously very strong, for it yields *sufficient* conditions for a vector to be a global minimizer using

- no differentiability assumption,
- no regularity assumption, and
- no convexity assumption.

To obtain necessary conditions, we normally make regularity and convexity assumptions.

Example 4 (Nonexistence of a saddlepoint).

This example exhibits a nonlinear program having a globally optimal solution but no saddlepoint for the Lagrangian function. Consider the problem (P)

$$\begin{array}{ll} \text{minimize} & -x_1 \\ \text{subject to} & x_2 - x_1^2 \geq 0 \\ & -x_2 \geq 0 \end{array}$$

Then $\bar{x} = (0, 0)$ is the (unique) global minimizer. Indeed, since $0 \geq x_2 \geq x_1^2 \geq 0$ it is the *only* feasible solution. The associated Lagrangian function is

$$L(x_1, x_2, y_1, y_2) = -x_1 - y_1(x_2 - x_1^2) + y_2 x_2.$$

Note that $L(\bar{x}, y) = 0$ for all y. If there exists a saddlepoint $(\bar{x}, \bar{y})$, then

$$0 \leq -x_1 - \bar{y}_1(x_2 - x_1^2) + \bar{y}_2 x_2 = L(x, \bar{y}) \quad \text{for all } x.$$

Let $x_2 = 0$. If $x_1 > 0$ and sufficiently small, we get $L(x, \bar{y}) < 0$, whereas for $(\bar{x}, \bar{y})$ to be a saddlepoint, we need $0 = L(\bar{x}, \bar{y}) \leq L(x, \bar{y})$, so we have a contradiction.

3.6 Appendix. Is the KKT constraint qualification indispensable?

If $\bar{x}$ is a local minimizer and the KKT conditions hold at $\bar{x}$, does the KKT constraint qualification have to hold there as well? The following example[21] shows that it does not.

Example 5. Define the functions $s(t)$, and $c(t)$ of the real variable t:

$$s(t) = \begin{cases} t^4 \sin \frac{1}{t} & \text{if } t \neq 0 \\ 0 & \text{if } t = 0 \end{cases}$$

$$c(t) = \begin{cases} t^4 \cos \frac{1}{t} & \text{if } t \neq 0 \\ 0 & \text{if } t = 0 \end{cases}$$

These functions are continuously differentiable. The functions and their derivatives vanish at 0.

Now consider the nonlinear program

$$\begin{array}{lrcl} \text{minimize} & f(x) & = & x_2 \\ \text{subject to} & c_1(x) & = & s(x_1) - x_2 + x_1^2 \geq 0 \\ & c_2(x) & = & x_2 - x_1^2 - c(x_1) \geq 0 \\ & c_3(x) & = & 1 - x_1^2 \geq 0 \end{array}$$

The feasible region lies between the curves $x_2 = x_1^2 + x_1^4$ and $x_2 = x_1^2 - x_1^4$. Indeed,

$$x_1^2 - x_1^4 \;\leq\; x_1^2 + c(x_1) \;\leq\; x_2 \;\leq\; x_1^2 + s(x_1) \;\leq\; x_1^2 + x_1^4.$$

The feasible region also lies between the lines $x_1 = -1$ and $x_1 = +1$. It is easy to see that the unique optimal solution to the problem is $\bar{x} = 0$.

In this instance, we have $\mathcal{A}(\bar{x}) = \mathcal{A}(0) = \{1, 2\}$. Moreover,

$$\nabla c_1(0) = (0, -1),$$
$$\nabla c_2(0) = (0, 1).$$

[21] See Abadie [1967, p. 35].

If v is a solution of

$$v^{\mathrm{T}}\nabla c_i(0) \geq 0,\ i = 1, 2,$$

then v_1 is arbitrary and $v_2 = 0$. It can be shown that the only curve in the feasible region is identically 0. [To see this, compute

$$x_1^2 + s(x_1) - (x_1^2 + c(x_1)) = x_1^4 \left(\sin \frac{1}{x_1} - \cos \frac{1}{x_1} \right).$$

This is not always nonnegative.] The derivative of this curve is not a positive multiple of $v = (v_1, 0)$ where $v_1 \neq 0$. This means that the KKT constraint qualification does not hold at $\bar{x} = 0$.

On the other hand, the KKT conditions for this problem give:

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} - u_1 \begin{bmatrix} 0 \\ -1 \end{bmatrix} - u_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$u_1 \geq 0$$

$$u_2 \geq 0$$

This reduces to

$$u_1 \geq 0, \quad u_2 \geq 0, \quad 1 + u_1 - u_2 = 0.$$

Thus, $\bar{x} = 0$ and any $\bar{u} \geq 0$ such that $1 + \bar{u}_1 - \bar{u}_2 = 0$ (for example, $\bar{u}_1 = 1$ and $\bar{u}_2 = 2$) will satisfy the KKT conditions.

References

J. Abadie [1967], On the Kuhn-Tucker Theorem in (J. Abadie,ed.) *Nonlinear Programming*, North-Holland, Amsterdam, pp. 19-36.

M. Avriel [1976], *Nonlinear Programming*, Prentice-Hall, Englewood Cliffs, N.J.

M.S. Bazaraa, H.D. Sherali, and C.M. Shetty [1993], *Nonlinear Programming*, John Wiley & Sons, New York.

D.P. Bertsekas [1999], *Nonlinear Programming*, Athena Scientific, Belmont, Mass.

J.M. Borwein and A.S. Lewis [2000], *Convex Analysis and Nonlinear Optimization*, Springer, New York.

A.V. Fiacco and G.P. McCormick [1968], *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley & Sons, New York.

P.E. Gill, W. Murray, and M.H. Wright [1981], *Practical Optimization*, Academic Press, London.

W. Fleming [1977], *Functions of Several Variables*, Springer-Verlag, New York.

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*, Wiley-Interscience, New York.

W. Karush [1939], *Minima of functions of several variables with inequalities as side conditions*, Masters Thesis, Department of Mathematics, University of Chicago.

H.W. Kuhn and A.W. Tucker [1951], Nonlinear programming, in (J. Neyman, ed.) *Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley.

D.G. Luenberger [1989], *Linear and Nonlinear Programming*, Addison-Wesley, Reading, Mass.

O.L. Mangasarian [1969], *Nonlinear Programming*, McGraw-Hill, New York.

O.L. Mangasarian and S. Fromovitz [1967], Fritz John's necessary optimality conditions in the presence of equality and inequality constraints, *Journal of Mathematical Analysis and Applications* 17, 37-47.

S.G. Nash and A. Sofer [1996], *Linear and Nonlinear Programming*, McGraw-Hill, New York.

3.7 Second-order optimality conditions

In the case of nonconvex nonlinear optimization, satisfaction of the KKT conditions not is not enough to guarantee that a vector is a local minimizer. As in the case of unconstrained univariate optimization, one needs second-order conditions to distinguish local minimizers from other kinds of points. The results covered here are largely due to McCormick [1967]. The latter work also appears in the historically important little book of Fiacco and McCormick [1968].

Here we take up the nonlinear programming problem (P):

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & c_i(x) \geq 0 \quad i \in \mathcal{I} \\ & c_i(x) = 0 \quad i \in \mathcal{E} \\ & x \in R^n \end{array}$$

Let us assume all the functions in (P) are ***twice continuously differentiable***.

Let S denote the feasible region of (P). For $\bar{x} \in S$, we have the set $\mathcal{A}(\bar{x})$ of active constraints. Relative to these ingredients, consider the set[22]

$$T(\bar{x}) := \{z : z^{\mathrm{T}} \nabla c_i(\bar{x}) = 0, \text{ for all } i \in \mathcal{A}(\bar{x}),\ z^{\mathrm{T}} \nabla c_i(\bar{x}) = 0, \text{ for all } i \in \mathcal{E}\}.$$

Actually, $T(\bar{x})$ is a linear subspace of R^n; it is sometimes called the ***tangent space*** at $\bar{x}$.

Definition. The ***second-order constraint qualification*** holds at $\bar{x}$ if for every nonzero $z \in T(\bar{x})$ there is a twice continuously differentiable curve $\alpha : [0, 1] \to R^n$ such that

$$\alpha(0) = \bar{x}, \qquad \frac{d\alpha(0)}{d\theta} = \kappa z \quad (\kappa > 0)$$

and for all $\theta \in [0, 1]$

$$c_i(\alpha(\theta)) = 0, \text{ for all } i \in \mathcal{A}(\bar{x}), \qquad c_i(\alpha(\theta)) = 0, \text{ for all } i \in \mathcal{E}.$$

The second-order constraint qualification (SOCQ) requires that every nonzero vector in the tangent space be tangent to a twice continuously differentiable curve lying in the boundary of the constraint

[22] Recall that $\mathcal{A}(\bar{x})$ is a set of *indices*, whereas $T(\bar{x})$ is a set of vectors.

set. As in the KKT constraint qualification[23] case, the conditions are designed to facilitate the proof of a theorem on necessary conditions of local optimality.

Theorem (Second-order necessary conditions). Let $\bar{x}$ be a local minimizer of (P) and let $\bar{u}, \bar{v}$ denote vectors such that $(\bar{x}, \bar{u}, \bar{v})$ satisfies the KKT conditions of (P). If the second-order constraint qualifications holds at $\bar{x}$, then

$$z^{\mathrm{T}} \nabla_x^2 L(\bar{x}, \bar{u}, \bar{v}) z \geq 0 \qquad \text{for all } z \in T(\bar{x}).$$

Proof. The condition obviously holds when $z = 0$. Let z be an arbitrary nonzero vector in $T(\bar{x})$. Let α denote a twice continuously differentiable curve as guaranteed by the SOCQ, and let $\kappa = 1$ (scaling z, if necessary). Now define $w = \frac{d^2\alpha(0)}{d\theta^2} \in R^n$. By the SOCQ and the chain rule of differentiation, we obtain

$$\frac{d^2 c_i(\alpha(0))}{d\theta^2} = z^{\mathrm{T}} \nabla^2 c_i(\bar{x}) z + w^{\mathrm{T}} \nabla c_i(\bar{x}) = 0 \quad \text{for all } i \in \mathcal{A}(\bar{x})$$

and

$$\frac{d^2 c_i(\alpha(0))}{d\theta^2} = z^{\mathrm{T}} \nabla^2 c_i(\bar{x}) z + w^{\mathrm{T}} \nabla c_i(\bar{x}) = 0 \quad \text{for all } i \in \mathcal{E}.$$

By assumption the triple $(\bar{x}, \bar{u}, \bar{v})$ satisfies the KKT conditions, so we have

$$\nabla f(\bar{x}) - \sum_{i \in \mathcal{I}} \bar{u}_i \nabla c_i(\bar{x}) - \sum_{i \in \mathcal{E}} \bar{v}_i \nabla c_i(\bar{x}) = 0. \tag{11}$$

By (1) and the definition of $T(\bar{x})$:

$$\frac{df(\alpha(0))}{d\theta} = z^{\mathrm{T}} \nabla f(\bar{x}) = z^{\mathrm{T}} \left[\sum_{i \in \mathcal{I}} \bar{u}_i \nabla c_i(\bar{x}) + \sum_{i \in \mathcal{E}} \bar{v}_j \nabla c_i(\bar{x}) \right] = 0.$$

Since $\bar{x}$ is a local minimizer and $\frac{df(\alpha(0))}{d\theta} = 0$, we have $\frac{d^2 f(\alpha(0))}{d\theta^2} \geq 0$. This translates into

$$\frac{d^2 f(\alpha(0))}{d\theta^2} = z^{\mathrm{T}} \nabla^2 f(\bar{x}) z + w^{\mathrm{T}} \nabla f(\bar{x}) \geq 0.$$

The conclusion of the theorem now follows. □

[23] This is also known as the first-order constraint qualification or FOCQ, for short.

The SOCQ is difficult to verify in general. See Fiacco and McCormick [1968, page 26] for a proof of the following result.

Theorem. Given the nonlinear program (P), the SOCQ holds at the feasible point $\bar{x}$ if the vectors

$$\begin{aligned} &\nabla c_i(\bar{x}) \quad \text{for all } i \in \mathcal{A}(\bar{x}) \\ &\nabla c_i(\bar{x}) \quad \text{for all } i \in \mathcal{E} \end{aligned}$$

are linearly independent. □

The following two examples are given in Fiacco and McCormick [1968], pages 27–28. They demonstrate the independence of the FOCQ and the SOCQ. In other words, neither one implies the other.

Example 1. (FOCQ satisfied; SOCQ not satisfied.) Consider the nonlinear program

$$\begin{aligned} \text{minimize} \quad & x_2 \\ \text{subject to} \quad & -x_1^9 + x_2^3 \geq 0 \qquad (c_1) \\ & x_1^9 + x_2^3 \geq 0 \qquad (c_2) \\ & x_1^2 + (x_2+1)^2 - 1 \geq 0 \qquad (c_3) \end{aligned}$$

The feasible region of this nonlinear program is the nonshaded region shown in Figure 3.1. This problem has the optimal solution $\bar{x} = (0, 0)$. It is easy to show that in this example, the first-order constraint qualification is satisfied at $\bar{x}$. The elements of the tangent space $T(\bar{x})$ are of the form $z = (z_1, 0)^{\mathrm{T}}$ where $z_1 \neq 0$. However, there is no arc in S along which all the c_i vanish, hence the SOCQ is not satisfied here. It can also be shown that the second-order necessary conditions of optimality fail in this example.

Example 2. (FOCQ not satisfied; SOCQ satisfied.) Consider the constraints

$$\begin{aligned} 1 - x_1^2 - (x_2-1)^2 \geq 0 \qquad (c_1) \\ 1 - x_1^2 - (x_2+1)^2 \geq 0 \qquad (c_2) \\ x_1 \geq 0 \qquad (c_3) \end{aligned}$$

The only point satisfying these inequalities is $\bar{x} = (0, 0)$. (See Figure 3.2 on the next page.) The SOCQ is satisfied vacuously at $\bar{x}$ since $T(\bar{x}) = \{0\}$. However, the first-order constraint qualification is *not* satisfied in this instance.

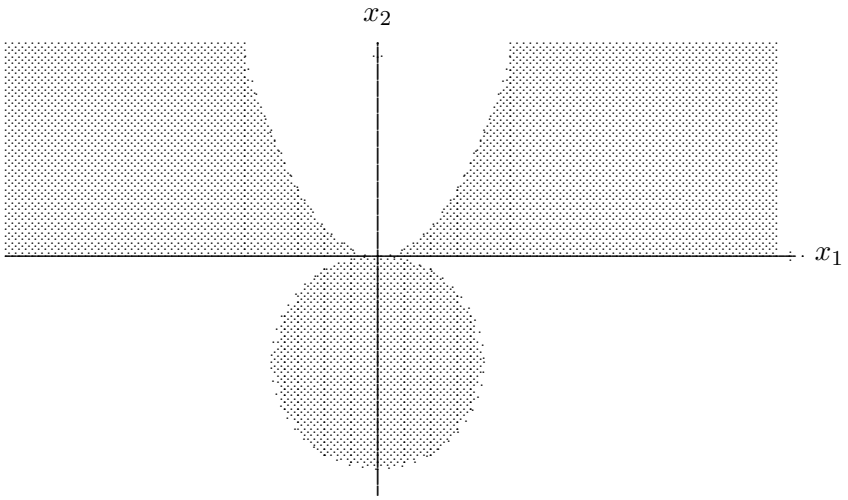


Figure 3.1

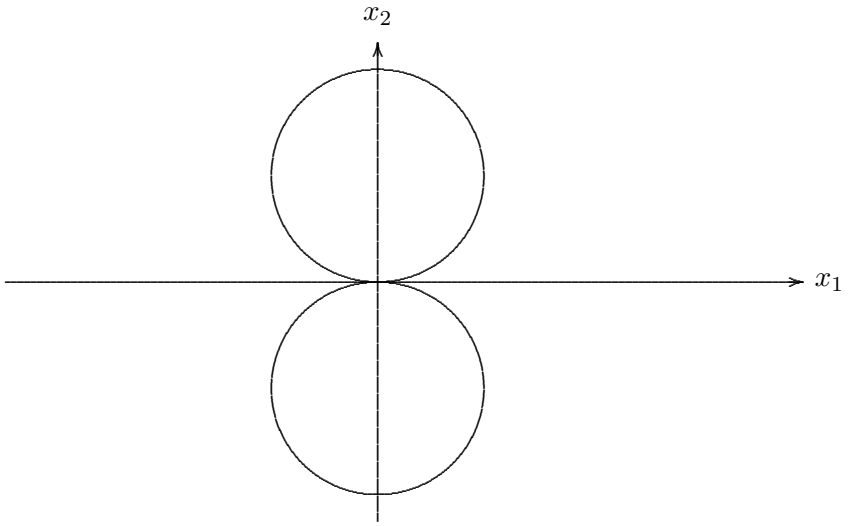


Figure 3.2

Example 3 (G.P. McCormick [1967]). Consider the optimization problem

$$\begin{aligned} \text{minimize} \quad & (x_1 - 1)^2 + x_2^2 \\ \text{subject to} \quad & -x_1 + \frac{x_2^2}{\beta} \geq 0 \qquad \beta > 0 \end{aligned}$$

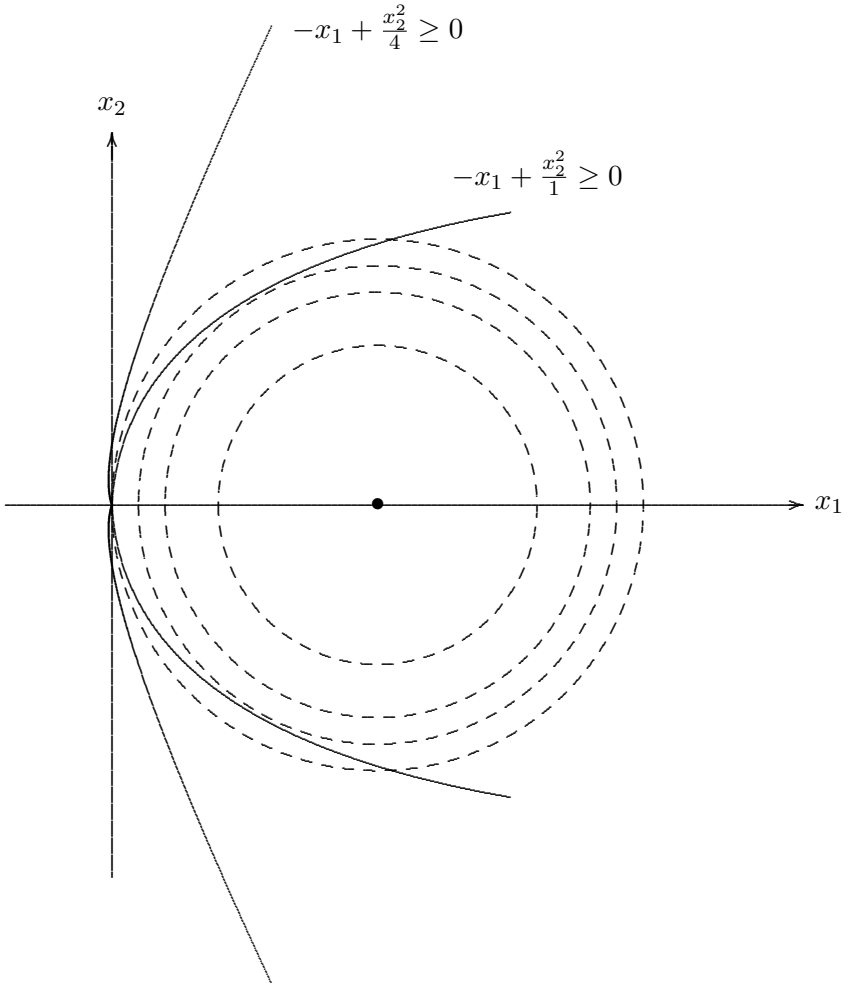


Figure 3.3

It can be shown that the first- and second-order constraint qualifications are satisfied at $\bar{x} = (0, 0)$. Question: For what values of the parameter β is this point a local minimizer?

The Lagrangian function for this problem is

$$L(x, u) = (x_1 - 1)^2 + x_2^2 - u\left[-x_1 + \frac{x_2^2}{\beta}\right].$$

For all $\beta \neq 0$, $\bar{x} = (0, 0)$ and $\bar{u} = 2$ satisfy the KKT conditions of this problem. We have $\mathcal{A}(\bar{x}) = \{1\}$ and

$$T(\bar{x}) = \{z : z \in R^2,\ z_1 = 0\}.$$

If $\bar{x}$ is to be a local minimizer, then for all $z \in T(\bar{x})$:

$$z^{\mathrm{T}} \nabla_x^2 L(\bar{x}, \bar{u}) z = \begin{bmatrix} 0 \\ z_2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 2 & 0 \\ 0 & 2 - \frac{4}{\beta} \end{bmatrix} \begin{bmatrix} 0 \\ z_2 \end{bmatrix}$$

$$= \left(2 - \frac{4}{\beta}\right) z_2^2 \geq 0.$$

This can hold only if $\beta \geq 2$. In particular, the Hessian matrix of L evaluated at $(\bar{x}, \bar{u})$ is indefinite for $\beta < 2$, so $\bar{x} = (0, 0)$ cannot be a local minimum for such values of β.

Definition. Suppose $\bar{x}$ is a feasible point for (P). Then $\bar{x}$ is an ***isolated (strict, strong) local minimizer*** if there exists a neighborhood $N(\bar{x})$ of $\bar{x}$ such that $f(\bar{x}) < f(x)$ for all $x \in S \cap N(\bar{x})$.

Theorem (Second-order sufficient conditions). Let $\bar{x} \in S$, and let $(\bar{x}, \bar{u}, \bar{v})$ satisfy the KKT conditions for (P). Then $\bar{x}$ is an isolated local minimizer if

$$z^{\mathrm{T}} \nabla_x^2 L(\bar{x}, \bar{u}, \bar{v}) z > 0$$

for all nonzero vectors $z \in R^n$ satisfying

$$\begin{aligned} z^{\mathrm{T}} \nabla c_i(\bar{x}) &= 0 \quad \text{if } \bar{u}_i > 0 \\ z^{\mathrm{T}} \nabla c_i(\bar{x}) &\geq 0 \quad \text{if } i \in \mathcal{A}(\bar{x}) \text{ and } \bar{u}_i = 0 \\ z^{\mathrm{T}} \nabla c_i(\bar{x}) &= 0 \quad \text{for all } i \in \mathcal{E} \end{aligned}$$

Proof. Suppose $\bar{x}$ is not an isolated local minimizer. Then there exists a feasible sequence $x^k \to \bar{x}$ such that $f(\bar{x}) \geq f(x^k)$ for all integers $k \geq 1$. Put $x^k = \bar{x} + \theta_k d^k$ where $\theta_k > 0$ and $\|d^k\| = 1$. Without loss of generality, we may assume $(\theta_k, d^k) \to (0, \bar{d})$ where $\|\bar{d}\| = 1$.

By the mean value theorem:

$$\begin{aligned} c_i(x^k) - c_i(\bar{x}) &= \theta_k (d^k)^{\mathrm{T}} \nabla c_i(\bar{x} + \sigma_{ik} \theta_k d^k) \geq 0 \quad i \in \mathcal{A}(\bar{x}) \\ c_i(x^k) - c_i(\bar{x}) &= \theta_k (d^k)^{\mathrm{T}} \nabla c_i(\bar{x} + \bar{\sigma}_{jk} \theta_k d^k) = 0 \quad i \in \mathcal{E} \\ f(x^k) - f(\bar{x}) &= \theta_k (d^k)^{\mathrm{T}} \nabla f(\bar{x} + \sigma_k \theta_k d^k) \leq 0 \end{aligned}$$

where $\sigma_{ik}, \bar{\sigma}_{jk}, \sigma_k \in (0, 1)$. Dividing by θ_k and taking limits, we obtain

$$\begin{aligned} \bar{d}^{\mathrm{T}} \nabla c_i(\bar{x}) &\geq 0 \quad i \in \mathcal{A}(\bar{x}) \\ \bar{d}^{\mathrm{T}} \nabla c_i(\bar{x}) &= 0 \quad i \in \mathcal{E} \\ \bar{d}^{\mathrm{T}} \nabla f(\bar{x}) &\leq 0 \end{aligned}$$

If $\bar{d}^{\mathrm{T}}\nabla c_i(\bar{x}) > 0$ for some $i \in \mathcal{A}(\bar{x})$ such that $\bar{u}_i > 0$, then the KKT conditions lead to the contradiction that $\bar{d}^{\mathrm{T}}\nabla f(\bar{x}) > 0$. Hence $\bar{u}_i > 0$ implies that $\bar{d}^{\mathrm{T}}\nabla c_i(\bar{x}) = 0$.

Next, apply Taylor's Theorem and the hypothesis about the sequence $\{x^k\}$ to obtain

$$\begin{aligned}
c_i(x^k) &= c_i(\bar{x}) + \theta_k (d^k)^{\mathrm{T}} \nabla c_i(\bar{x}) + \tfrac{1}{2}\theta_k^2 (d^k)^{\mathrm{T}} \nabla^2 c_i(\bar{x} + \tau_{ik}\theta_k d^k) d^k \geq 0 \\
c_i(x^k) &= c_i(\bar{x}) + \theta_k (d^k)^{\mathrm{T}} \nabla c_i(\bar{x}) + \tfrac{1}{2}\theta_k^2 (d^k)^{\mathrm{T}} \nabla^2 h_j(\bar{x} + \bar{\tau}_{jk}\theta_k d^k) d^k = 0 \qquad (i \in \mathcal{E}) \\
f(x^k) - f(\bar{x}) &= \theta_k (d^k)^{\mathrm{T}} \nabla f(\bar{x}) + \tfrac{1}{2}\theta_k^2 (d^k)^{\mathrm{T}} \nabla^2 f(\bar{x} + \tau_k \theta_k d^k) d^k \leq 0
\end{aligned}$$

where $\tau_{ik}, \bar{\tau}_{jk}, \tau_k \in (0,1)$. It now follows that

$$\begin{aligned}
0 \geq \theta_k (d^k)^{\mathrm{T}} &\left\{ \nabla f(\bar{x}) - \sum_{i=1}^{m} \bar{u}_i \nabla c_i(\bar{x}) - \sum_{i \in \mathcal{E}} \bar{v}_i \nabla c_i(\bar{x}) \right\} \\
&+ \frac{1}{2}\theta_k^2 (d^k)^{\mathrm{T}} \left\{ \nabla^2 f(\bar{x} + \tau_k \theta_k d^k) - \sum_{i=1}^{m} \bar{u}_i \nabla^2 c_i(\bar{x} + \tau_{ik}\theta_k d^k) \right. \\
&\left. - \sum_{i \in \mathcal{E}} \bar{v}_i \nabla^2 c_i(x)(\bar{x} + \bar{\tau}_{ik}\theta_k d^k) \right\} d^k.
\end{aligned}$$

The first term in curly brackets equals zero. Now divide by $\frac{1}{2}\theta_k^2$ and take limits to obtain a contradiction.

Definition. The set

$$S^{n-1} = \{x \in R^n : \|x\| = 1\}.$$

is called the ***unit sphere*** in R^n.

Definition. Let $(\bar{x}, \bar{u}, \bar{v})$ satisfy the KKT conditions for the nonlinear program (P) above. Then

$$\begin{aligned}
\mathcal{A}_+(\bar{x}) &= \{i \in \mathcal{A}(\bar{x}) : \bar{u}_i > 0\}, \\
\mathcal{A}_0(\bar{x}) &= \{i \in \mathcal{A}(\bar{x}) : \bar{u}_i = 0\}.
\end{aligned}$$

Constraints for which $i \in \mathcal{A}_+(\bar{x})$ are said to be ***strongly binding*** at $\bar{x}$, whereas those for which $i \in \mathcal{A}_0(\bar{x})$ are said to be ***weakly binding*** at $\bar{x}$.

Definition. Recall the definition of $T(\bar{x})$. The set $T_+(\bar{x})$ consists of the (necessarily nonzero) vectors $z \in S^{n-1}$ satisfying the homogeneous linear conditions

$$\begin{aligned}
z^{\mathrm{T}} \nabla c_i(\bar{x}) &= 0 \quad \text{for all } i \in \mathcal{A}_+(\bar{x}), \\
z^{\mathrm{T}} \nabla c_i(\bar{x}) &\geq 0 \quad \text{for all } i \in \mathcal{A}_0(\bar{x}), \\
z^{\mathrm{T}} \nabla c_i(\bar{x}) &= 0 \quad \text{for all } i \in \mathcal{E}.
\end{aligned}$$

Theorem (Second-order sufficient conditions). Let $\bar{x} \in S$, and let $(\bar{x}, \bar{u}, \bar{v})$ satisfy the KKT conditions for (P). Define the set

$$Y(\epsilon, \delta) = \{y \in S^{n-1} : \|y - z\| \leq \epsilon \text{ for some } z \in T_+(\bar{x}),\ \bar{x} + \delta_y y \in S,\ 0 < \delta_y < \delta,\ \epsilon > 0\}.$$

If there exists a set $\bar{Y} = Y(\bar{\epsilon}, \bar{\delta})$ such that for all $(y, \lambda) \in \bar{Y} \times [0, 1]$

$$y^{\mathrm{T}} \nabla_x^2 L(\bar{x} + \lambda \delta_y y, \bar{u}, \bar{v}) y \geq 0,$$

then $\bar{x}$ is a local minimizer for (P).

Proof. Suppose the theorem is false. Then there exists a sequence of points z^k converging to $\bar{x}$ such that $f(z^k) < f(\bar{x})$ for all k. Let $z^k = \bar{x} + \delta_k d^k$ where $d^k \in S^{n-1}$ and $\delta_k > 0$. Without loss of generality, we may assume the entire sequence $\{(\delta_k, d^k)\}$ converges to $(0, \bar{d})$ where $\bar{d} \in S^{n-1}$.

Now if $\bar{d}^{\mathrm{T}} \nabla c_i(\bar{x}) > 0$ for some $i \in \mathcal{A}_+(\bar{x})$, we get the same contradiction as in the previous theorem. If $\bar{d}^{\mathrm{T}} \nabla c_i(\bar{x}) = 0$ for all $i \in \mathcal{A}_+(\bar{x})$ or if $\mathcal{A}_+(\bar{x}) = \emptyset$, then (by definition) $\bar{d} \in T_+(\bar{x})$. By Taylor's expansion, we have

$$L(z^k, \bar{u}, \bar{v}) = L(\bar{x}, \bar{u}, \bar{v}) + \delta_k (d^k)^{\mathrm{T}} \nabla_x L(\bar{x}, \bar{u}, \bar{v}) + \frac{(\delta_k)^2}{2} (d^k)^{\mathrm{T}} \nabla_x^2 L(\eta^k, \bar{u}, \bar{v}) d^k$$

where $\eta^k = \bar{x} + \lambda \delta_k d^k$ and $0 \leq \lambda \leq 1$. From the assumptions, we have

$$(d^k)^{\mathrm{T}} \nabla_x^2 L(\eta^k, \bar{u}, \bar{v}) d^k < 0 \quad \text{for all } k.$$

But for k sufficiently large, $d^k \in Y(\bar{\epsilon}, \bar{\delta})$. Hence, by the hypothesis, when k is large enough,

$$(d^k)^{\mathrm{T}} \nabla_x^2 L(\eta^k, \bar{u}, \bar{v}) d^k \geq 0$$

which contradicts the inequality above. □

3.8 Second-order optimality criteria for quadratic programming

We have a theorem on second-order ***necessary conditions*** for local optimality in nonlinear programming, and we have a separate theorem on second-order ***sufficient conditions*** for an isolated local minimizer in such a problem. This is not a symmetrical state of affairs. In the case of quadratic programming, the situation is somewhat better: There is a single set of second-order conditions that are both necessary and sufficient for local optimality.

Consider the quadratic program (QP) of finding $x \in R^n$ so as to

$$\begin{aligned} &\text{minimize} && f(x) = c^{\mathrm{T}} x + \tfrac{1}{2} x^{\mathrm{T}} Q x \\ &\text{subject to} && Ax \geq b. \end{aligned}$$

Let S denote the polyhedral feasible region of (QP). Let $\bar{x} \in S$ be any point such that $\mathcal{A}(\bar{x}) \neq \emptyset$. The *nonzero* solutions of the system

$$A_{i\bullet} v \geq 0 \quad \text{for all } i \in \mathcal{A}(\bar{x}) \tag{12}$$

are called ***feasible directions*** at $\bar{x}$. Let $\mathcal{F}$ denote the collection of all solutions of (1). Note that $0 \in \mathcal{F}$ and that $\mathcal{F}$ is a polyhedral cone.[24] If $\mathcal{A}(\bar{x}) = \emptyset$, we put $\mathcal{F} = R^n$. In the literature, the set $\mathcal{F}$ is sometimes call called the ***cone of feasible directions*** at $\bar{x}$, even though it contains the zero vector which is not a genuine direction.

Remark. The inequality

$$\nabla f(\bar{x})^{\mathrm{T}} v \geq 0 \quad \text{for all } v \in \mathcal{F} \tag{13}$$

is just another way of stating the Karush-Kuhn-Tucker conditions. To see this, note that (2) is equivalent to saying

$$\begin{aligned} A_{i\bullet} v &\geq 0 \quad \text{for all } i \in \mathcal{A}(\bar{x}) \\ \nabla f(\bar{x})^{\mathrm{T}} v &< 0 \end{aligned} \tag{14}$$

has no solution. Farkas's lemma says that for all $i \in \mathcal{A}(\bar{x})$ there exist $\lambda_i \geq 0$ such that

$$\nabla f(\bar{x})^{\mathrm{T}} - \sum_{i \in \mathcal{A}(\bar{x})} \lambda_i A_{i\bullet} = 0.$$

Setting $\lambda_i = 0$ for all $i \notin \mathcal{A}(\bar{x})$, we see that $\bar{x}$ and $\lambda = (\lambda_1, \ldots, \lambda_m)$ satisfy the KKT conditions for (QP).

Theorem. If $\bar{x} \in S$ is a local minimizer of (QP), then the following two conditions hold:

(i) $\nabla f(\bar{x})^{\mathrm{T}} v \geq 0 \quad$ for all $v \in \mathcal{F}$;

(ii) $v^{\mathrm{T}} Q v \geq 0 \quad$ for all $v \in \mathcal{F} \cap \{\nabla f(\bar{x})\}^{\perp}$.

Proof. Since the first-order local optimality (KKT) conditions must hold at $\bar{x}$, condition (i) follows by the remark above. Accordingly, we turn to condition (ii). There are two cases.

Case 1. If $\mathcal{A}(\bar{x}) = \emptyset$, then $\bar{x}$ is an interior point of S. Since $\bar{x}$ is a local minimizer,

1. $\nabla f(\bar{x}) = 0$,
2. $\mathcal{F} = R^n = \{\nabla f(\bar{x})\}^{\perp}$,

and assertion (ii) of the theorem follows from a standard theorem of multivariate differential calculus, namely second-order necessary conditions of local optimality for a twice continuously differentiable function on an open set.

[24] Recall the "cone" $\mathcal{D}$ of feasible directions (see Handout No. 6, page 8). Thus, in the present notation, we have $\mathcal{F} = \mathcal{D} \cup \{0\}$.

Case 2. Assume $\mathcal{A}(\bar{x}) \neq \emptyset$. If v is a nonzero vector in $\mathcal{F} \cap \{\nabla f(\bar{x})\}^{\perp}$, then $\tau v = x - \bar{x}$ for some $x \in S \cap N(\bar{x})$ and $\tau > 0$. Then

$$f(x) = f(\bar{x} + \tau v) = f(\bar{x}) + \tau \nabla f(\bar{x})^{\mathrm{T}} v + \tfrac{1}{2}\tau^2 v^{\mathrm{T}} \nabla^2 f(\bar{x}) v;$$

it follows that

$$\tfrac{1}{2}\tau^2 v^{\mathrm{T}} Q v = f(x) - f(\bar{x}) \geq 0,$$

which proves (ii). □

Remark. The two ***necessary*** conditions of local optimality in (QP) are also ***sufficient*** for local optimality. The proof of this result is long and delicate, hence must be omitted.[25],

Example. Consider the (nonconvex) quadratic program

$$\begin{array}{ll} \text{minimize} & \tfrac{1}{2}x_1 - \tfrac{1}{2}x_2 - \tfrac{1}{2}x_1^2 + \tfrac{1}{2}x_2^2 \\ \text{subject to} & -2x_1 - x_2 \geq -6 \\ & x_1 - 4x_2 \geq -6 \\ & x_1 \geq 0 \\ & x_2 \geq 0 \end{array}$$

We shall examine three solutions of the KKT conditions and show that one of them is not a local minimizer (or a local maximizer) whereas two of them do correspond to local minima of the QP.

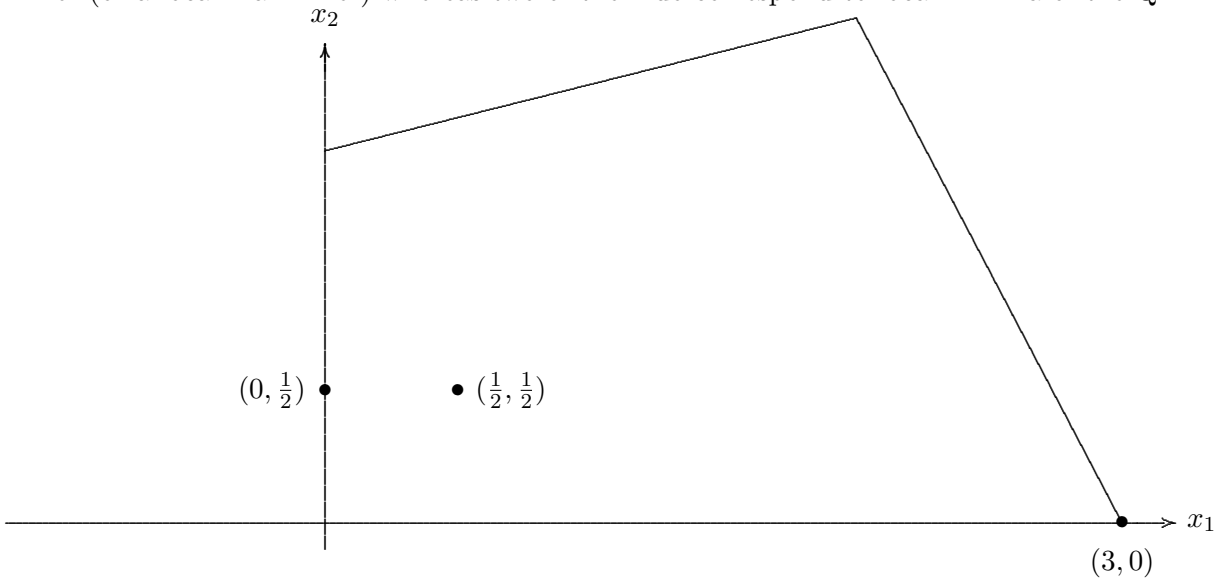


Figure 3.4

[25]For the details see Contesse [1980]. This paper, gives a more rigorous proof than the one in Majthay [1971].

First, the KKT conditions are satisfied by $(\bar{x}, \bar{y}) = (\frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0)$. We have $f(\bar{x}) = f(\frac{1}{2}, \frac{1}{2}) = 0$. For sufficiently small ϵ, the points $(\frac{1}{2} + \epsilon, \frac{1}{2})$ and $(\frac{1}{2}, \frac{1}{2} + \epsilon)$ are feasible. Now note that

$$f(\tfrac{1}{2} + \epsilon, \tfrac{1}{2}) = -\epsilon^2/2 < 0 \quad \text{and} \quad f(\tfrac{1}{2}, \tfrac{1}{2} + \epsilon) = \epsilon^2/2 > 0.$$

Thus, there are feasible directions of descent and ascent at $\bar{x}$. Hence $\bar{x}$ cannot be a local minimizer. Note that for $\bar{x} = (\frac{1}{2}, \frac{1}{2})$ we have $\mathcal{A}(\bar{x}) = \emptyset$. In this case, $\mathcal{F} = R^2$ and since $\nabla f(\frac{1}{2}, \frac{1}{2}) = (0, 0)$ we have $\{\nabla f(\bar{x})\}^\perp = R^2$. The theorem says that for $\bar{x}$ to be a local minimizer, the matrix $Q = \nabla^2 f(\bar{x})$ would have to be positive semi-definite, but it is not.

Next consider the point $\hat{x} = (0, \frac{1}{2})$. In this case, $\mathcal{A}(\hat{x}) = \{3\}$. This means that in any KKT point $(\hat{x}, \hat{y})$, we must have $\hat{y}_1 = \hat{y}_2 = \hat{y}_4 = 0$. The cone $\mathcal{F}$ at $\hat{x}$ is given by $\{v : v_1 \geq 0\}$. Since $\nabla f(\hat{x}) = (\frac{1}{2}, 0)$, it follows that

$$\nabla f(\hat{x})^{\mathrm{T}} v \geq 0 \quad \text{for all } v \in \mathcal{F}.$$

Now $\{\nabla f(\hat{x})\}^\perp = \{v : v_1 = 0\}$, and clearly we have

$$v^{\mathrm{T}} Q v = v_2^2 \geq 0 \quad \text{for all } v \in \mathcal{F} \cap \{\nabla f(\bar{x})\}^\perp.$$

Hence $\hat{x} = (0, \frac{1}{2})$ is a local minimizer. We note that $f(\hat{x}) = -1/8$.

Finally, consider $x^* = (3, 0)$. Here we have $\mathcal{A}(x^*) = \{1, 4\}$, and $\nabla f(x^*) = (-5/2, -1/2)$. The cone of feasible directions is given by

$$\begin{aligned} -2v_1 - v_2 &\geq 0 \\ v_2 &\geq 0 \end{aligned}$$

so $-2v_1 \geq v_2 \geq 0$, and $v_1 \leq 0$. With these inequalities, we can show that $\nabla f(x^*)^{\mathrm{T}} v \geq 0$ for all $v \in \mathcal{F}$, so the first-order conditions are satisfied. If $v \in \mathcal{F} \cap \{\nabla f(x^*)\}^\perp$, we have in addition $5v_1 + v_2 = 0$. This implies that for such v, $v^{\mathrm{T}} Q v = 24 v_1^2 \geq 0$. The first- and second-order sufficient conditions are satisfied at x^*, so it must be a local minimizer. Note that $f(x^*) = -3 < -1/8 = f(\hat{x})$ which means that it is a better local minimizer of (QP). In fact, it is the global minimizer.

Remark. One might wonder why the statement of the theorem involves $\{\nabla f(\bar{x})\}^\perp$. Could one just forget about it? In particular, when $\bar{x}$ is a local minimizer for (QP), is it necessary for $v^{\mathrm{T}} Q v$ to take nonnegative values on all of $\mathcal{F}$? The following example answers this question.

Example. Consider the quadratic program

$$\begin{array}{ll} \text{minimize} & f(x) = x_1 + x_2 - x_2^2 \\ \text{subject to} & x_1 \leq 1 \\ & x_2 \leq 1 \\ & x_1 \geq 0 \\ & x_2 \geq 0 \end{array} .$$

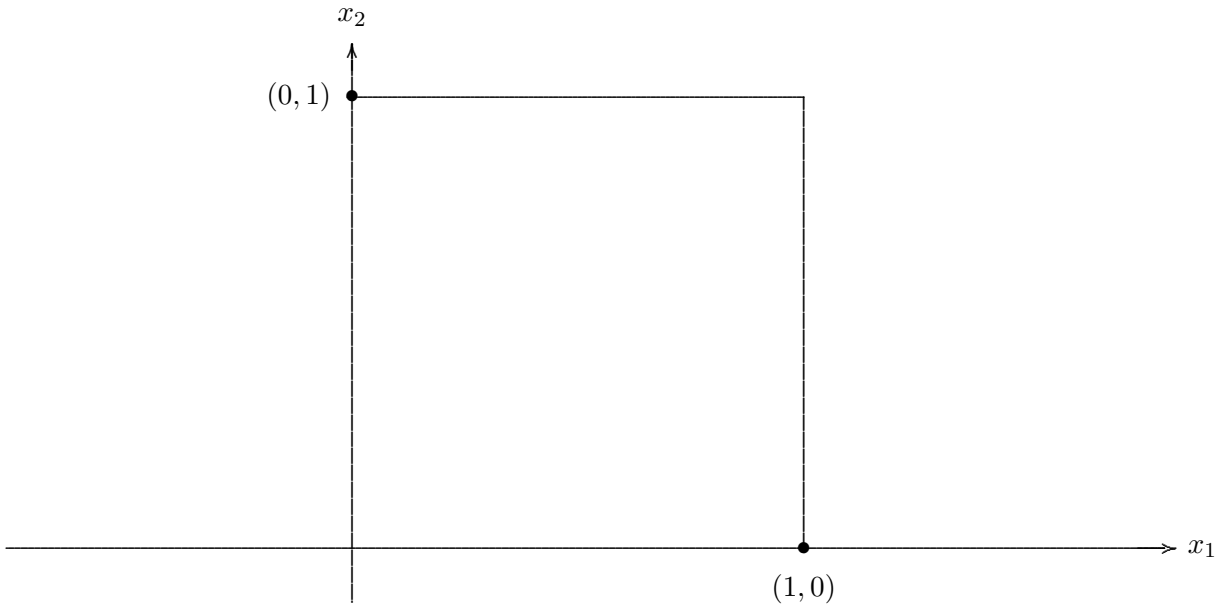


Figure 3.5

Note that in this example, we have

$$Q = \begin{bmatrix} 0 & 0 \\ 0 & -2 \end{bmatrix}.$$

It is easy to see that the feasible point $\bar{x} = (0, 1)$ is a local (in fact, global) minimizer. Indeed, $f(\bar{x}) = 0$, and $f(x) \geq 0$ for all x in the feasible region. We have

$$\mathcal{A}(\bar{x}) = \{2, 3\} \quad \text{and} \quad \nabla f(\bar{x}) = (1, -1).$$

The cone of feasible directions at this point is

$$\mathcal{F} = \{v : v_1 \geq 0 \geq v_2\}$$

and

$$\{\nabla f(\bar{x})\}^{\perp} = \{v : v_1 - v_2 = 0\}.$$

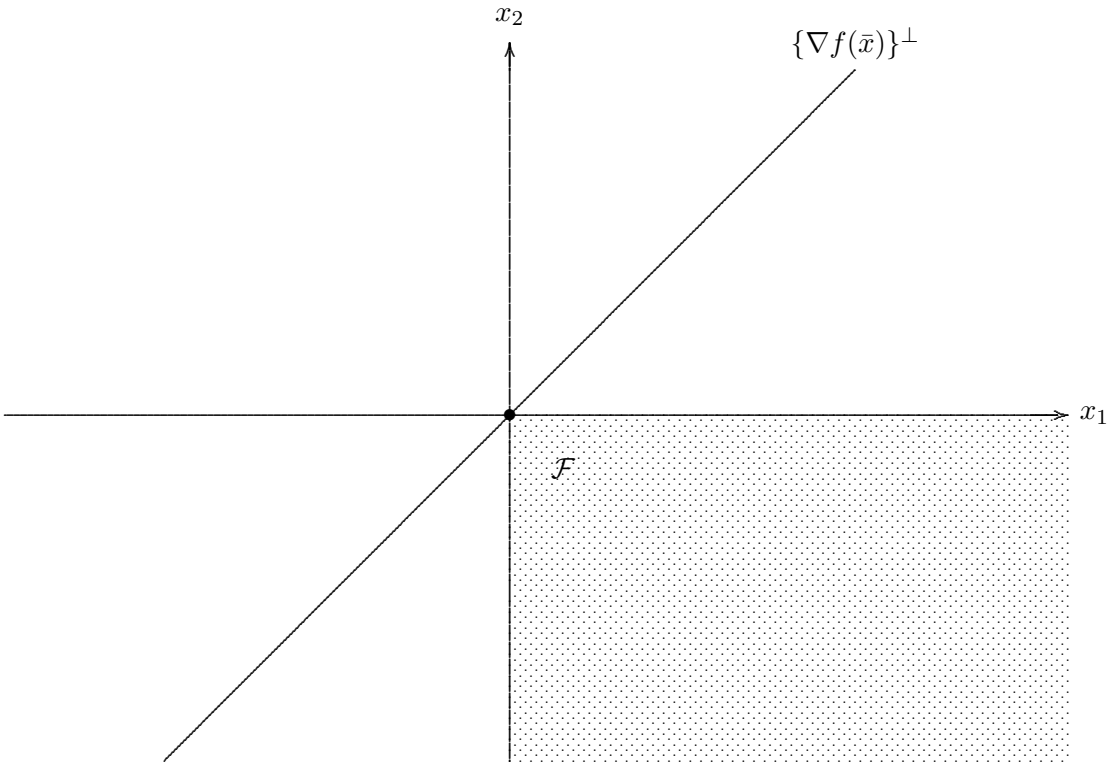


Figure 3.6

The first-order condition is satisfied at $\bar{x}$ since

$$\nabla f(\bar{x})^{\mathrm{T}} v = v_1 - v_2 \geq 0 \quad \text{for all } v \in \mathcal{F}.$$

For all $v \in R^2$, we have $v^{\mathrm{T}} Q v = -2v_2^2 \leq 0$. This takes on negative values for some elements of $\mathcal{F}$, but

$$\mathcal{F} \cap \{\nabla f(\bar{x})\}^{\perp} = \{0\},$$

so $v^{\mathrm{T}} Q v = 0$ there. In this case, (ii) is satisfied on $\mathcal{F} \cap \{\nabla f(\bar{x})\}^{\perp}$, but it is not satisfied on all of $\mathcal{F}$.

References

M. Avriel [1976], *Nonlinear Programming*, Prentice-Hall, Englewood Cliffs, N.J.

M.S. Bazaraa, H.D. Sherali, and C.M. Shetty [1993], *Nonlinear Programming*, John Wiley & Sons, New York.

D.P. Bertsekas [1995], *Nonlinear Programming*, Athena Scientific, Belmont, Mass.

J.M. Borwein and A.S. Lewis [2000], *Convex Analysis and Nonlinear Optimization*, Springer-Verlag, New York.

L. Contesse [1980], Une caractérisation complète des minima locaux en programmation quadratique, *Numerische Mathematik* 34, 315-332.

A.V. Fiacco and G.P. McCormick [1968], *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley & Sons, New York.

A. Majthay [1971], Optimality conditions for quadratic programming, *Mathematical Programming* 1, 359-365.

O.L. Mangasarian [1969], *Nonlinear Programming*, McGraw-Hill, New York.

G.P. McCormick [1967], Second order conditions for constrained minima, *SIAM J. Applied Mathematics* 15, 37-47.

S.G. Nash and A. Sofer [1996], *Linear and Nonlinear Programming*, McGraw-Hill, New York.

J. Nocedal and S.J. Wright [1999], *Numerical Optimization*, Springer, New York.

5. OPTIMIZATION ALGORITHMS

5.1 Introduction

Algorithms for linear and nonlinear programming problems tend to be iterative procedures. Starting from a given point x_0, they generate a sequence $\{x_k\}$ of *iterates* (or trial solutions).[26] The algorithms we shall study here produce these iterates according to well determined rules rather than some random selection process. The rules to be followed and the procedures that can be applied depend to a large extent on the characteristics of the problem to be solved.

5.1.1 Generalities

1. *Classes of problems.* Some of the distinctions between optimization problems stem from
   (a) differentiable versus nondifferentiable functions;
   (b) unconstrained versus constrained variables;
   (c) one-dimensional versus multi-dimensional variables.
   (c) convexity versus nonconvexity of the minimand and feasible region.
2. *Finite versus convergent iterative methods.* For some classes of optimization problems (e.g., linear and quadratic programming) there are algorithms that obtain a solution—or detect that the objective function is unbounded—in a finite number of iterations. For this reason, we call them ***finite algorithms***.[27] Most algorithms encountered in nonlinear programming are not finite, but instead are convergent—or at least they are designed to be so. Their object is to generate a sequence of trial or approximate solutions that converge to a "solution."
3. *The meaning of "solution" is needed.* What is meant by a solution may differ from one algorithm to another. In some cases, one seeks a local minimum; in some cases, one seeks

[26] Notice that we are using subscripts rather than superscripts to denote the elements of a sequence.
[27] It should be mentioneod, however, that not all algorithms for linear and quadratic programming are finite.

a global minimum; in others, one seeks a stationary point of some sort as in the method of steepest descent discussed below. In fact, there are several possibilities for defining what a solution is. Once the definition is chosen, there must be a way of testing whether or not a point (trial solution) belongs to the set of solutions.

4. *Search directions.* Typically, from a given point x_0, a nonlinear programming algorithm generates a sequence of points

$$x_{k+1} = x_k + \alpha_k p_k$$

where p_k is the ***search direction*** and α_k is the ***step size*** or ***step length***. In fact, if $\{x_k\}$ is *any* sequence of distinct vectors,

$$x_{k+1} - x_k = \alpha_k p_k$$

where, say, $\|p_k\| = 1$ and $\alpha_k > 0$. Thus, it is not very informative simply to say that the sequence $\{x_k\}$ has the property $x_{k+1} = x_k + \alpha_k p_k$. The point is that once x_k is known, then p_k is some function of x_k, and the scalar α_k may be chosen in accordance with some special rule.

5. *The general idea.* One selects a starting point x_0 and (efficiently) generates a possibly infinite sequence of trial solutions each of which is specified by the algorithm. The idea is to do this in such a way that the sequence of iterates generated by the algorithm *converges* to an element of the set of solutions of the problem. Convergence to some other sort of point is undesirable—as is failure of the sequence to converge at all.

5.1.2 Convergent sequences of real numbers

Let $\{a_k\}$ be a sequence of real numbers.[28] Then $\{a_k\}$ ***converges to*** 0 if and only if for all real numbers $\varepsilon > 0$ there exists a positive integer K such that

$$|a_k| < \varepsilon \quad \text{for all } k \geq K.$$

Let $\{x_k\}$ be a sequence of real numbers. Then $\{x_k\}$ ***converges to*** x^* if and only if $\{a_k\} = \{x_k - x^*\}$ converges to 0. For example, the sequence $\{x_k\} = \{1/k\}$ converges to 0.

5.1.3 Types of convergence

[28] Through the introduction of norms, for example the Euclidean norm, we can (and do) talk about convergent sequences of vectors.

If there exists a number $c \in [0,1)$ and an integer K such that

$$|x_{k+1} - x^*| \leq c|x_k - x^*| \quad \text{for all } k \geq K,$$

then $\{x_k\}$ converges to x^* ***q-linearly***. If there exists a number $c \in [0,1)$ and an integer K such that

$$|x_{k+1} - x^*| \leq c|x_k - x^*|^2 \quad \text{for all } k \geq K,$$

then $\{x_k\}$ converges to x^* ***q-quadraticly***. If $\{c_k\}$ is a sequence of nonnegative reals converging to 0 and

$$|x_{k+1} - x^*| \leq c_k|x_k - x^*| \quad \text{for all } k \geq K,$$

then $\{x_k\}$ converges to x^* ***q-superlinearly***.

5.2 Unconstrained minimization of smooth functions[29]

5.2.1. A generic algorithm. Let f be a smooth function on R^n. We seek $x^* \in R^n$ such that $f(x^*) \leq f(x)$ for all $x \in R^n$. Depending on the properties of f, it may be necessary to adopt a more modest goal than finding a global minimum. (For one thing, it may be impractical or even impossible to test whether a point is a global minimum or not.) We may instead have to look for a local minimum or a stationary point of the minimand, f. These considerations give rise to the different possible notions of what a solution is. Let us assume we have a way to decide whether or not any given point is a "solution."

(A1) *Test for convergence* If the termination conditions are satisfied at x_k, then it is taken (accepted) as a "solution." In practice, this may mean satisfying the desired conditions to within some tolerance.

(A2) *Compute a search direction*, say $p_k \neq 0$. This may for example be a direction in which the function value is known to decrease.

(A3) *Compute a step length*, say α_k such that

$$f(x_k + \alpha_k p_k) < f(x_k).$$

This may necessitate the use of a one-dimensional (or line) search algorithm.

[29] Although the meaning of the term "smooth function" differs from one author to another, it is generally agreed that it means at least "continuously differentiable."

(A4) *Define the new iterate* by setting

$$x_{k+1} = x_k + \alpha_k p_k$$

and return to step (A1).

5.2.2. The gradient method (steepest descent method).[30] Let f be a differentiable function and assume we can compute ∇f. We want to solve the unconstrained minimization problem

$$\min_{x \in R^n} f(x).$$

In the absence of further information, we seek a ***stationary point*** of f, that is, a point x^* at which $\nabla f(x^*) = 0$.

For this algorithm, we use the direction $p_k = -\nabla f(x_k)$ as the search direction at x_k. The number $\alpha_k \geq 0$ is chosen "appropriately," namely to satisfy

$$\alpha_k \in \arg\min f(x_k - \alpha \nabla f(x_k)).$$

Then the new iterate is defined as

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$

Now, if $\nabla f(x_k) \neq 0$, then $-\nabla f(x_k)$ is a direction of descent at x_k; in fact, it is the ***direction of steepest descent*** at x_k.

Example. Let $f(x) = c^{\mathrm{T}}x + \frac{1}{2}x^{\mathrm{T}}Qx$ where $Q \in R^{n \times n}$ is symmetric and positive definite. This implies that the eigenvalues of Q are all positive. The unique minimum x^* of $f(x)$ exists and is given by the solution of the equation

$$\nabla f(x) = c + Qx = 0,$$

or equivalently

$$Qx = -c.$$

Of course, we have $x = -Q^{-1}c$. (This is a case where we have a closed-form solution of an optimization problem.) We can also solve the equation $Qx = -c$ *directly* (i.e., by pivoting, or by matrix factorization) but that is not the point of this example. In the steepest descent method, the *iterative* scheme

$$x_{k+1} = x_k + \alpha_k p_k$$

[30] This algorithm is associated with the French mathematician, A. Cauchy (1789–1857).

becomes

$$x_{k+1} = x_k - \alpha_k(c + Qx_k)$$

by virtue of the definition, $p_k = -(c+Qx_k)$. With the search direction chosen, we need to compute the step size, α_k. To this end, we consider

$$\begin{aligned} f(x_k + \alpha p_k) &= c^{\mathrm{T}}(x_k + \alpha p_k) + \tfrac{1}{2}(x_k + \alpha p_k)^{\mathrm{T}} Q(x_k + \alpha p_k) \\ &= c^{\mathrm{T}}x_k + \alpha c^{\mathrm{T}}p_k + \tfrac{1}{2}x_k^{\mathrm{T}}Qx_k + \alpha x_k^{\mathrm{T}}Qp_k + \tfrac{1}{2}\alpha^2 p_k^{\mathrm{T}}Qp_k \end{aligned}$$

which is a strictly convex quadratic function of α. As such, its minimizer α_k is the value of α where the derivative $f'(x_k + \alpha p_k)$ vanishes, i.e., where

$$c^{\mathrm{T}}p_k + x_k^{\mathrm{T}}Qp_k + \alpha p_k^{\mathrm{T}}Qp_k^{\mathrm{T}} = 0.$$

Thus

$$\alpha_k = -\frac{(c^{\mathrm{T}} + x_k^{\mathrm{T}}Q)p_k}{p_k^{\mathrm{T}}Qp_k} = -\frac{p_k^{\mathrm{T}}p_k}{p_k^{\mathrm{T}}Qp_k}.$$

The recursion for the method of steepest descent now becomes

$$x_{k+1} = x_k - \left(\frac{p_k^{\mathrm{T}}p_k}{p_k^{\mathrm{T}}Qp_k}\right) p_k$$

where $p_k = -(c + Qx_k)$.

Convergence of the steepest descent method. The following theorem gives some conditions under which the steepest descent method will converge.

Theorem. Let $f : R^n \to R$ be given. For some given point $x_0 \in R^n$, let the level set

$$X_0 = \{x \in R^n : f(x) \leq f(x_0)\}$$

be bounded. Assume further that f is continuously differentiable on the convex hull of X_0. Let $\{x_k\}$ be the sequence of points generated by the steepest descent method initiated at x_0. Then every accumulation point[31] of $\{x_k\}$ is a stationary point of f.

Proof. Note that the assumptions imply the compactness of X_0. Since the iterates will all belong to X_0, the existence of at least one accumulation point of $\{x_k\}$ is guaranteed by the Bolzano-Weierstrass Theorem. Let $\bar{x}$ be such an accumulation point. For the sake of contradiction, assume

[31] An accumulation point of a sequence is the limit of a convergent subsequence.

$\nabla f(\bar{x}) \neq 0$. Then there exists a value $\bar{\alpha} > 0$ and a $\delta > 0$ such that $f(\bar{x} - \bar{\alpha}\nabla f(\bar{x})) + \delta = f(\bar{x})$. This means that $\bar{x} - \bar{\alpha}\nabla f(\bar{x})$ is an interior point of X_0.

For an arbitrary iterate of the sequence, say x_k, the Mean-Value Theorem implies that we can write

$$f(x_k - \bar{\alpha}\nabla f(x_k)) = f(\bar{x} - \bar{\alpha}\nabla f(\bar{x})) + (\nabla f(y_k))^{\mathrm{T}}((x_k - \bar{x}) + \bar{\alpha}(\nabla f(\bar{x}) - \nabla f(x_k)))$$

where y_k lies between $x_k - \bar{\alpha}\nabla f(x_k)$ and $\bar{x} - \bar{\alpha}\nabla f(\bar{x})$. Now as an accumulation point of $\{x_k\}$, the vector $\bar{x}$ is the limit of a subsequence of $\{x_k\}$. Denote this subsequence by $\{x_{k_t}\}$. Then

$$\{\nabla f(y_{k_t})\} \to \nabla f(\bar{x} - \bar{\alpha}\nabla f(\bar{x}))$$

and

$$\{(x_{k_t} - \bar{x}) - \bar{\alpha}(\nabla f(x_{k_t}) - \nabla f(\bar{x}))\} \to 0.$$

For sufficiently large k_t, the vector y_{k_t} is an element of the convex hull of X_0. Moreover,

$$f(x_{k_t} - \bar{\alpha}\nabla f(x_{k_t})) \leq f(\bar{x} - \bar{\alpha}\nabla f(\bar{x})) + \frac{\delta}{2} = f(\bar{x}) - \frac{\delta}{2}.$$

Let $\bar{\alpha}_{k_t}$ be the minimizer of $f(x_{k_t} - \alpha_{k_t}\nabla f(x_{k_t}))$. Since the sequence $\{f(x_{k_t})\}$ is monotonically decreasing and converges to $f(\bar{x})$, it follows that

$$f(\bar{x}) < f(x_{k_t} - \alpha_{k_t}\nabla f(x_{k_t})) \leq f(x_{k_t} - \bar{\alpha}\nabla f(x_{k_t})) \leq f(\bar{x}) - \frac{\delta}{2}$$

which is a contradiction. Hence $\nabla f(\bar{x}) = 0$. □

Remark. According to this theorem, the steepest descent method initiated at *any* point of the level set X_0 will converge to a stationary point of f. This is a nice feature; it means that (depending on the size of X_0), the starting point x_0 could be far away from the point x^* to which the sequence converges. In other words, it is not necessary to start the process in a neighborhood of the (unknown) solution.

Remark. The convergence rate of the steepest descent method applied to quadratic functions is known to be linear. Suppose Q is a symmetric positive definite matrix of order n and let its

eigenvalues be $\lambda_1 \leq \cdots \leq \lambda_n$. Obviously, the global minimizer of the quadratic form $f(x) = x^{\mathrm{T}}Qx$ is at the origin. It can be shown that when the steepest descent method is started from any nonzero point $x_0 \in R^n$, there will exist constants c_1 and c_2 such that

$$0 < c_1 \leq \frac{f(x_{k+1})}{f(x_k)} \leq c_2 \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 < 1, \qquad k = 0, 1, \ldots \ .$$

Remark. Intuitively, the slow rate of convergence of the steepest descent method can be attributed the fact that the successive search directions are perpendicular. Indeed, consider an arbitrary iterate x_k. At this point we have the search direction $p_k = -\nabla f(x_k)$. To find the next iterate x_{k+1} we minimize $f(x_k - \alpha \nabla f(x_k))$ with respect to $\alpha \geq 0$. At the minimum α_k, the derivative of this function will equal zero. Thus, we obtain $\nabla f(x_{k+1})^{\mathrm{T}} \nabla f(x_k) = 0$.

5.2.3 Newton's method

As we know, all unconstrained local minimizers of a differentiable function f are stationary points: they make the gradient ∇f vanish. Finding a solution of the stationarity condition

$$\nabla f(x) = 0$$

is a matter of solving a ***system*** of (possibly nonlinear) equations.[32] In the case of functions of a single real variable, the stationarity condition is

$$f'(x) = 0,$$

which is a single (possibly nonlinear) equation in one variable. When f is a twice continuously differentiable function, Newton's method can be a very effective way to solve such equations and hence to locate a stationary point of f.

It should be emphasized, however, that Newton's method is a procedure for solving equations. It is not necessarily a descent method, hence it cannot even be relied upon to find a local minimizer. It can, however, be modified so as to be a descent algorithm.

Newton's method for solving the equation $g(x) = 0$

Although the iterative scheme associated with Newton's method uses only first derivatives, the working hypothesis is that the function g—of which we want to find a zero—has continuous second derivatives.

[32] If f is a function of n variables, then this is a system of n equations in these n variables.

For simplicity, we start with the ***univariate*** case. Given a starting point x_0, Newton's method for solving the equation $g(x) = 0$ is to generate the sequence of iterates

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}.$$

Notice that the iteration is well defined provided that $g'(x_k) \neq 0$ at each step. The iteration stops if $g(x_k) = 0$, for then a solution of the equation has been found.

The interpretation of this iteration goes as follows. For a given iterate, x_k such that $g(x_k) \neq 0$, let $x = x_k + p$. By linearizing g at x_k we obtain

$$g(x) = g(x_k + p) \approx g(x_k) + pg'(x_k).$$

We seek a value of p at which the expression $g(x_k) + pg'(x_k) = 0$. Under the assumptions above, there will be such a value provided $g'(x_k) \neq 0$. It's a matter of where the tangent line crosses the horizontal axis. When $g'(x_k) = 0$, the tangent line is parallel to the horizontal axis, and there is no intersection. This anomaly is illustrated in Figure 5.1.

Without some further stipulations, there is no guarantee that the sequence of points defined by the iterative scheme above will converge to a zero of the function g. This can be illustrated by the case where

$$g(x) = x^{1/3}.$$

In this instance, the sequence diverges unless $x_0 = 0$, the zero of g. See Figure 5.2 below.

Theorem. If g is twice continuously differentiable and x_* is a zero of g at which $g'(x_*) \neq 0$, then provided that $|x_0 - x_*|$ is sufficiently small, the sequence generated by the iteration

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}.$$

converges quadratically to x_* with rate constant $C = |g''(x_*)/2g'(x_*)|$.

Proof. See Nash and Sofer [1996, page 40] or Luenberger [1989, page 202]. □

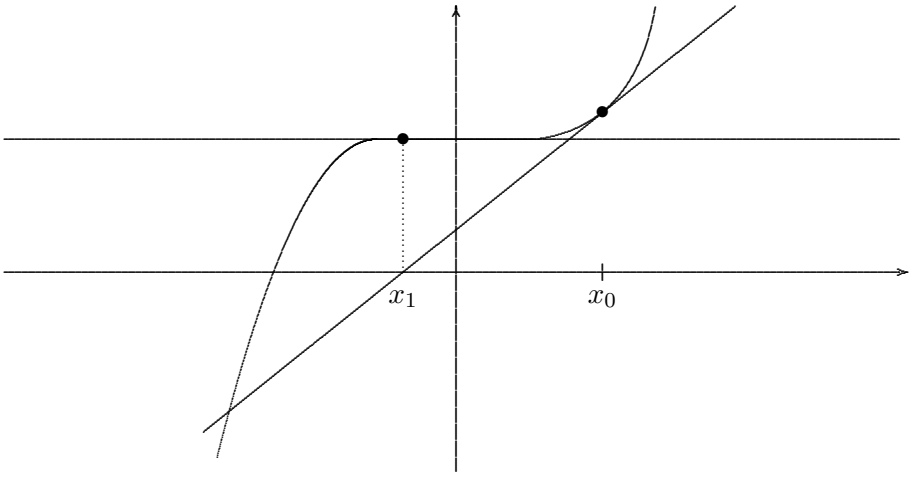


Figure 5.1

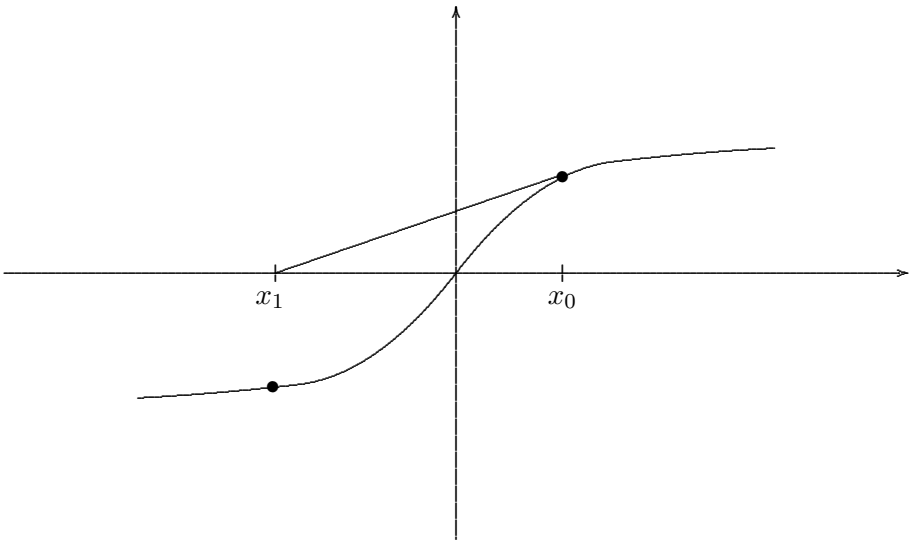


Figure 5.2

Newton's method for solving a system of equations $g(x) = 0$

Before we begin to develop the ***multivariate*** case, it will be helpful to restate our notational convention regarding the Jacobian matrix that was stated on page 4 of Handout No. 14. There we have a mapping

$$g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_\ell(x) \end{bmatrix},$$

and we define the ***Jacobian*** of g as

$$\nabla g(x) = \left[\frac{\partial g_i(x)}{\partial x_j}\right].$$

The rows of $\nabla g(x)$ are the *transposes* of the gradient vectors $\nabla g_1(x), \ldots, \nabla g_\ell(x)$. This convention is in agreement with most books on multivariate calculus.

As a method for solving a system of n equations in n variables, Newton's method is rather analogous to the one-variable case, but—as should be expected—the computational demands are much greater.

For the system $g(x) = 0$, i.e.,

$$g_i(x) = 0, \quad i = 1, \ldots, n$$

the iteration is given by

$$x_{k+1} = x_k - (\nabla g(x_k))^{-1} g(x_k).$$

This formula follows from the use of a Taylor series approximation to g at the point x_k, namely

$$g(x_k + p) \approx g(x_k) + \nabla g(x_k) p.$$

When we set the right-hand side of this equation to zero, we can solve it for p, provided that the Jacobian matrix is nonsingular. When this value of p is used, $x_{k+1} = x_k + p$ becomes our new approximation of x_*, the zero of g that we seek.

Some computational issues

When the number of variables n is very large, there can be some difficulties.

As an optimization technique, Newton's method, in its pure form, requires knowledge of (or the capacity to compute) the first and second derivatives of the minimand.[33] In an environment where individual function evaluations are expensive, this could be a drawback.

In a large-scale problem, the computation of p could also turn out to be a time-consuming task. The literal inversion of the transpose if the Jacobian matrix is not really required (no matter what the size of the problem). Instead, one solves

$$(\nabla g(x_k))p = -g(x_k)$$

using matrix factorization methods. It should be recognized that in the optimization context, $g(x)$ would be the gradient vector of the minimand f so that the Jacobian matrix of g would be the Hessian matrix of f. In the neighborhood of a minimizer, the Hessian would be positive semidefinite, and if this matrix is to be nonsingular, it must then be positive definite. The LDL^{T} factorization would be used to compute p.

Storage can also be a factor in using Newton's method, though with the storage capacity of today's computers, this issue does not loom as large as it once did.

When it is very important to keep the overall running time low, there is a question of whether the relatively small number of iterations required by Newton's method is worth the cost of the computing at each iteration. This consideration motivates the use of ***quasi-Newton methods*** which generally speaking, use a surrogate for the true Hessian matrix. Under some circumstances, it can be shown that such variants of Newton's method possess superlinear convergence to a local minimizer.

Descent

For an arbitrary twice-continuously differentiable function f, there is no reason to expect Newton's method produce a sequence of iterates that converge to a local minimizer of f. In fact, as we have seen, convergence to *anything* is not guaranteed without additional hypotheses. Inasmuch as Newton's method strives to produce a stationary of f, it is only with a satisfactory second-order (or curvature) condition that we can be sure of having—or getting to—a local minimizer.

[33] That is, the function being minimized.

If our search direction at a point, say $\bar{x}$ is

$$p = -(\nabla^2 f(\bar{x}))^{-1}\nabla f(\bar{x}),$$

then it is a descent direction only if

$$p^{\mathrm{T}}\nabla f(\bar{x}) = -(\nabla f(\bar{x}))^{\mathrm{T}}(\nabla^2 f(\bar{x}))^{-1}\nabla f(\bar{x}) < 0$$

which is to say that

$$(\nabla f(\bar{x}))^{\mathrm{T}}(\nabla^2 f(\bar{x}))^{-1}(\nabla f(\bar{x})) > 0.$$

This will hold, of course, if $\nabla^2 f(\bar{x})$ is a positive definite matrix. This is only a sufficient condition, however. It is often too strong a condition to impose when what we really want is $p^{\mathrm{T}}\nabla f(\bar{x}) < 0$.

In the case where $\nabla^2 f(\bar{x})$ is not positive definite, we can modify it by adding to it a suitable diagonal positive definite matrix E so that the sum $\nabla^2 f(\bar{x}) + E$ is positive definite. That such a thing can be done is evident from the determinantal characterization of positive definiteness (i.e., positivity of the leading principal minors of the matrix.) For example, if $A \in R^{n\times n}$ is an arbitrary symmetric matrix, then for any value of the scalar θ, the determinant of $A + \theta I$ is a polynomial of degree n. For sufficiently large θ the value of this polynomial is positive. This idea can be applied to each of the leading principal submatrices of $A + \theta I$. Hence, for θ sufficiently large, $A + \theta I$ will be positive definite. One could use $E = \theta I$ for suitably large θ. This would mean adding the same amount to each diagonal entry of A, but it is not generally necessary use such a restrictive form of perturbation.

Nash and Sofer [1996] give the following modified Newton algorithm that uses the idea sketched above to find a stationary point of f.

Modified Newton Algorithm with Line Search

Initialize the algorithm with the point x_0 and the tolerance $\varepsilon > 0$. For $k = 0, 1, \ldots$, perform the following sequence of steps.

- (Termination criterion.) If $\|\nabla f(x_k)\| < \varepsilon$, stop. Report x_k as an approximation to a stationary point, x_*.
- (Modify the Hessian.) For some diagonal positive definite matrix E, so that $\nabla^2 f(x_k) + E$ is also positive definite, compute the matrix factorization

$$LDL^{\mathrm{T}} = \nabla^2 f(x_k) + E.$$

- (Compute the search direction.) Solve the equation

$$LDL^{\mathrm{T}}p = -\nabla f(x_k).$$

  Denote the solution by p_k.

- (Compute the step length.) Do a line search to determine the step length α_k.

- (Compute the new iterate.) Compute

$$x_{k+1} = x_k + \alpha_k p_k.$$

References

M. Avriel [1976], *Nonlinear Programming*, Prentice-Hall, Engelwood Cliffs, N.J. [See Section 10.1.]

A. Ben-Tal ana A. Nemirovski [2001], *Lectures on Modern Convex Optimization*, Philadelphia: SIAM.

D.P. Bertsekas [1982], *Constrained Optimization and Lagrange Multiplier Methods*, Athena Scientific, Belmont, Mass.

D.P. Bertsekas [1999], *Nonlinear Programming*, Athena Scientific, Belmont, Mass.

D.P. Bertsekas [2003], *Convex Analysis and Optimization*, Athena Scientific, Belmont, Mass.

J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C.A. Sagastizábal [2003], *Numerical Optimization*, Springer-Verlag, Berlin.

A. Cauchy [1847], Méthode générale pour la résolution des systèms d'équations simultanées, *Comp. Rend. Acad. Sci. Paris*, pp. 536-538.

R. Fletcher [1980], *Practical Methods of Optimization, Volume 2: Constrained Optimization*, John Wiley & Sons, Chichester.

G.E. Forsythe [1968], On the asymptotic directions of the s-dimensional optimum gradient method, *Numerische Mathematik* 11, 57-76.

P.E. Gill, W. Murray, M.A. Saunders, and M.H. Wright [1989], "Constrained nonlinear programming," (Chapter III) in G.L. Nemhauser, A.H.G. Rinnooy Kan, and M.J. Todd (eds.), *Optimization*, North-Holland, Amsterdam.

P.E. Gill, W. Murray and M.H. Wright [1981], *Practical Optimization*, Academic Press, London. [See pp. 99-104.]

C. Hildreth [1954], Point estimates of concave functions, *Journal of the American Statistical Association* 49, 598-619.

D.G. Luenberger [1989], *Linear and Nonlinear Programming*, Addison-Wesley, Reading, Mass. [See Section 7.6.]

S.G. Nash and A. Sofer [1996], *Linear and Nonlinear Programming*, McGraw-Hill, New York.

Y. Nesterov and A. Nemirovski [1994], *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia.

J. Nocedal and S.J. Wright [1999], *Numerical Optimization*, Springer, New York.

J.M. Ortega and W.C. Rheinboldt [1970], *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York.

S.J. Wright [1997], *Primal-Dual Interior-Point Methods*, SIAM Publications, Philadelphia.

5.3 Linearly constrained problems

Probably the most tractable of all constrained optimization problems are those with linear constraints: either linear equations or linear inequaltiies or a mixture of both. This class of optimization problems includes linear programs, quadratic programs, and the nonlinear programs having nonquadratic objective functions. Naturally, there are an enormous number of applications for problems of this sort.

For linearly constrained problems of the form

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array}$$

where f is sufficiently differentiable and the feasible region is specified by a system of linear equations, there are modifications of the standard unconstrained optimization techniques such as steepest descent, Newton's method, quasi-Newton methods, and the method of conjugate gradients. (See Gill, Murray and Wright [1981] for some details on these methods.) For the time being, we shall be concerned with linearly constrained problems having at least some linear inequality constraints.

Algorithms for linearly constrained optimization problems can be classified according to whether or not the iterates they generate are *feasible*.[34] Our discussion will focus on primal feasible algorithms and the special issues they present. The usual (generic) algorithmic outline[35] will be in force here.

Some issues

Except for the very special case of (box) constraints that are exclusively simple bounds on the variables, i.e., constraints of the form

$$\ell \leq x \leq u, \tag{15}$$

[34]This is by no means the only way to classify these algorithms. For example, some algorithms aim to solve a problem by attacking its dual. Others are of what is called a primal-dual type.

[35]See page 3 of this handout.

the first issue that comes up is that of ***finding a feasible solution*** if one exists. As a rule, such a vector is needed in the initialization of the algorithm. The field of linear programming has a methodology for doing this, and we shall cover it here.

The next issue that comes up is that of ***retaining feasibility*** once the algorithm choses the search direction. Fortunately, the linearity of the constraints makes this relatively straightforward as will be seen below.

But there are many other issues, both theoretical and practical. With the Simplex Method of linear programming, for instance, there is the issue of ***degeneracy***. There are several ways to describe degeneracy, and we shall come to them below. The important point about degeneracy is that it can lead to a type of algorithmic behavior called ***cycling*** and thereby prevent an algorithm from making progress and terminating. There are several techniques for handling this complication, but we can give them only scant attention here due to the shortage of time. The subject of degeneracy is discussed in MS&E 310 (Linear Programming).

Over and above these matters, there are the issues of ***computational efficiency, reliability*** and ***computational complexity***. As we have seen, the problems in this class—especially those having at least some inequalities—have a decidedly combinatorial quality about them. Driven by a desire to satisfy the necessary (and sometimes sufficient) first-order optimality conditions, these methods lead to subproblems of solving of linear equations (which of course should be done efficiently). The combinatorial aspect arises out of the indeterminacy of which systems are ultimately to be solved. Recognition of this combinatorial aspect has a major impact on the field. We shall touch on this development only lightly. Both of these issues deserve serious attention.

5.4 Linear programming via the Simplex Method

The standard way to solve linear programming problems is by the ***Simplex Method*** of G.B. Dantzig. Introduced over 50 years ago, this algorithm is the one of the most widely used scientific computational procedures of all time. Our development of it will necessarily be a brief one. See for example Bertsimas and Tsitsiklis [1997], Dantzig and Thapa [1997], [2003]. Nash and Sofer [1996]. For a collection of selected publications of G.B. Dantzig, see Cottle [2003].

We shall study the Simplex Method for treating linear programs in the ***standard form***

$$\begin{array}{ll} \text{minimize} & c^{\mathrm{T}}x \\ \text{subject to} & Ax = b \\ & x \geq 0. \end{array} \tag{16}$$

As we know, any mixture of linear equations and linear inequalities can be brought to the form appearing in (2). We shall further assume that the system of linear constraints $Ax = b$ is neither vacuous nor trivial. That is, these equality constraints are really present, and the constraints of the problem are not equivalent to a system of the form (1). A linear programming problem whose constraints are only simple bounds is either unsolvable or trivially solvable.

A consequence of the Motzkin-Goldman resolution theorem

In Handout No. 4, page 13 we have the resolution theorem of Motzkin [1936] and Goldman [1956], which says that a nonempty polyhedral set is the sum of a polytope and a finite cone. This ss important structural property of polyhedral sets has the following implication for the linear programming problem.

Theorem. If the feasible region S of the linear program

$$\begin{array}{lll} & \text{minimize} & c^{\mathrm{T}}x \\ \text{(P)} & \text{subject to} & Ax \geq b \\ & & x \geq 0 \end{array}$$

has an optimal solution, then it has an optimal extreme point solution. Hence once can seek an optimal solution of (P) among the extreme points of S.

Proof. There exist vectors $x^1, \ldots, x^r$ and $x^{r+1}, \ldots, x^t$ such that $x \in S$ if and only if there exist nonnegative scalars $\lambda_1, \ldots, \lambda_t$ such that

$$\begin{bmatrix} x \\ 1 \end{bmatrix} = \sum_{i=1}^{r} \lambda_i \begin{bmatrix} x^i \\ 1 \end{bmatrix} + \sum_{i=r+1}^{t} \lambda_i \begin{bmatrix} x^i \\ 0 \end{bmatrix}.$$

We may assume that $x^1, \ldots, x^r$ are the full finite set of extreme points of S. It is clear that the objective function $c^{\mathrm{T}}x$ is bounded below on S if and only if $c^{\mathrm{T}}x^i \geq 0$ for all $i = r+1, \ldots t$ in which case, one (or more) of the points $x^1, \ldots, x^r$ must be an optimal solution of (P). □

This theorem is one way to motivate the search for an optimal solution among the extreme points of S.

What else we know

1. Extreme points of the feasible region correspond to basic feasible solutions of the constraints.
2. If the LP (2) is feasible, it has a basic feasible solution.
3. There are only finitely many extreme points of S.

4. An optimal solution (if one exists) will occur at a basic feasible solution of the constraints.

5. If the LP (2) has an optimal solution, then so does its dual; moreover, the optimal values of the primal and dual objective functions are the same.

6. If $\bar{x}$ is an optimal solution of the primal (2) and $\bar{y}$ is an optimal solution of its dual then

$$\bar{x}_j > 0 \quad \Longrightarrow \quad \bar{y}^{\mathrm{T}} A_{\bullet j} = c_j. \tag{17}$$

These facts play an important role in the Simplex Method for linear programming.

Geometric description of the Simplex Method

The geometric idea behind this method is as follows: Start at an extreme point of the (polyhedral) feasible region, S. Test it for optimality. If it does not pass the optimality test, look for an edge of the polyhedral feasible region along which objective function value decreases. (This amounts to choosing a search direction.) Move along that edge until either it is determined that the edge is unbounded, in which case the objective function is not bounded below and the procedure stops, or another extreme point is reached and the sequence of steps is repeated. Of course, some algebra is needed in order to implement these geometrically described steps.

Finding a basic feasible solution

The Simplex Method needs to be initiated at an extreme point of the feasible region, S. In some cases, one is at hand[36] or is easily produced.[37] *But when no starting point is at hand, how is one to be found?* The answer to this question comes from a remarkable idea: create an auxiliary problem of the same type which *does* have an obvious basic feasible solution as its starting point. What is remarkable about this idea is that essentially the same kind of methodology is used for solving the auxiliary problem as the original problem, the main difference being that the auxiliary problem must have a finite optimal solution. The algorithm cannot terminate with the news that the objective function is unbounded.

In the linear programming literature, the auxiliary problem mentioned above is called the ***Phase I Problem***. To formulate the Phase I Problem, we (can) begin with the LP problem in standard form (2). Multiplying each of the constraints $A_{i\bullet} x = b_i$ by -1 if necessary, we can assume that the constraints of (2) already have $b \geq 0$.

[36] This would be the case, if the problem has been solved with the same constraints and a different objective function.

[37] Sometimes the addition of slack variables takes care of this.

Let us assume that $A \in R^{m\times n}$. Identifying a basic feasible solution of a problem in standard form (2) is easy if for every $i = 1, \ldots, m$ there is a column of A, say $A_{\bullet j_i}$ such that

$$A_{\bullet j_i} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i\text{-th row}$$

Notice that $A_{\bullet j_i}$ is just the i-th column of an $m \times m$ identity matrix. Some columns of this form may already be part of A, having been created by the addition of ***slack variables*** to convert linear inequalities of the form $\leq$ to equations in nonnegative variables. To form the Phase I Problem, one adjoins new columns as needed to assure that a column of the form $A_{\bullet j_i}$ exists for each $i = 1, \ldots, m$. These adjoined columns are said to be ***artificial*** and the variables we associate with them are called ***artificial variables.*** A numerical example will help to clarify the ideas.

Example. Consider the system

$$\begin{bmatrix} 4 & -2 & 0 & 6 & 0 \\ -1 & 7 & 2 & 8 & 1 \\ 7 & 3 & 5 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 2 \\ 8 \\ 15 \end{bmatrix} \qquad x_j \geq 0,\ j = 1, \ldots, 5.$$

In this case, the matrix A already contains a column of the 3×3 identity matrix. In particular, we have $A_{\bullet 5} = I_{\bullet 2}$. We need two artificial columns to complete the set. These will look like $I_{\bullet 1}$ and $I_{\bullet 3}$. Thus, the system for the auxiliary (Phase I) problem will be

$$\begin{bmatrix} 4 & -2 & 0 & 6 & 0 & 1 & 0 \\ -1 & 7 & 2 & 8 & 1 & 0 & 0 \\ 7 & 3 & 5 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{bmatrix} = \begin{bmatrix} 2 \\ 8 \\ 15 \end{bmatrix} \qquad x_j \geq 0,\ j = 1, \ldots, 7.$$

In this system, x_6 and x_7 are artificial variables (relative to the original system). Notice that the augmented system in this case has an obvious basic feasible solution

$$(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = (0, 0, 0, 0, 8, 2, 15).$$

Now to complete the formulation of the Phase I Problem, we introduce an objective function. *The goal of the Phase I Problem is to find a basic feasible solution of the augmented system in which the artificial variables are nonbasic—and hence have value zero.* This cannot always be done. If it can be done, then the original system (containing no artificial variables) must be feasible. But since not every system is feasible, we need to have a way of detecting this outcome. The device used in formulating the Phase I Problem is to *minimize the sum of the artificial variables.* Note that the objective function of the Phase I Problem is the sum of a set of nonnegative variables. As such, it is bounded below by zero. Note also that the dual of the Phase I Problem is always feasible. For these reasons, we know that the Phase I Problem has an optimal solution (and so does its dual). If the optimal objective function value in the Phase I Problem is positive, it follows that the original system must be infeasible.

For theoretical purposes, it does no harm to assume that an artificial variable is needed for each of the m rows of the system.[38] Thus, the Phase I Problem can be formulated as follows:

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{j=n+1}^{n+m} x_j \\ \text{subject to} & \displaystyle\sum_{j=1}^{n} a_{ij}x_j + x_{n+i} = b_i, \quad i = 1, \ldots, m \\ & x_j \geq 0, \quad j = 1, \ldots, n+m \end{array} \tag{18}$$

In matrix form, the formulation above reads

$$\begin{array}{ll} \text{minimize} & e^{\mathrm{T}} x_\kappa \\ \text{subject to} & Ax + Ix_\kappa = b \\ & x \geq 0,\ x_\kappa \geq 0 \end{array} \tag{19}$$

where κ denotes the index set $\{n+1, \ldots, n+m\}$ corresponding to the artificial variables.

The Phase I Problem has the basic feasible solution $x = 0,\ x_\kappa = b$. The ***basis*** corresponding to this basic feasible solution is $[A\ I]_{\bullet\kappa}$ which is just the identity matrix.

Since the technique used to solve the Phase I Problem is the same as the Simplex Method itself, we postpone further discussion of the Phase I problem to see how the Simplex Method works.

The concept of a feasible basis

[38] Only computational efficiency is affected by introducing superfluous artificial variables.

Let us assume that we have a system of linear equations in nonnegative variables, say

$$Ax = b, \quad x \geq 0. \tag{20}$$

For this discussion, we assume that the m rows of A are ***linearly independent*** and that we know an $m \times m$ matrix of columns

$$B = A_{\bullet\beta} = [A_{\bullet j_1}\, A_{\bullet j_2}\, \cdots A_{\bullet j_m}]$$

such that

$$x_\beta = B^{-1}b \geq 0. \tag{21}$$

Notice that x_β solves the equation

$$Bx_\beta = b. \tag{22}$$

Taking $x_j = 0$ if $j \notin \beta$ then gives rise to a basic feasible solution of the system (6). The matrix B is called a ***feasible basis***; we shall call β a ***feasible basis index set***.

Testing for optimality

A feasible basis, determines a feasible point, say $\bar{x}$. (In fact, $\bar{x}$ is an extreme point of S). How do we test $\bar{x}$ for optimality? We wish to know if the optimality conditions are satisfied there. This is where matters get a little bit delicate. To simplify the discussion, we assume that all the basic components of $\bar{x}$ are positive, i.e., that $\bar{x}_\beta > 0$. When this is the case, $\bar{x}$ is a ***nondegenerate basic feasible solution.***[39] Moreover, we can say that $\bar{x}$ is an optimal solution if and only if there exists a dual feasible vector, say $\bar{y}$ such that (3) holds. Under the present circumstances, this means

$$\bar{y}^{\mathrm{T}} A_{\bullet\beta} = c_\beta^{\mathrm{T}} \tag{23}$$

Now as long as $B = A_{\bullet\beta}$ is a nonsingular $m \times m$ matrix, the equation (9) can be solved for $\bar{y}$. That is, (9) is solvable, even if $\bar{x}$ is a degenerate basic feasible solution.

Having $\bar{y}$, we want to test for optimality. To do so, we shall need to take degeneracy into account, but let's begin with the nondegenerate case.

Proposition. If B is a feasible basis for the constraints of (2), and the corresponding basic feasible solution $\bar{x}$ is nondegenerate, then $\bar{x}$ is optimal for (2) if and only if the solution $\bar{y}$ of (9) satisfies the linear inequalities

$$\bar{y}^{\mathrm{T}} A \leq c^{\mathrm{T}}. \tag{24}$$

[39] Otherwise, it is a ***degenerate basic feasible solution.***

Proof. Suppose $\bar{y}$ satisfies (9) and (10). Then together, $\bar{x}$ and $\bar{y}$ are primal and dual feasible vectors, respectively, that satisfy the complementary slackness conditions. They must both be optimal for their respective problems. This shows the sufficiency[40] of (9) and (10) for the optimality of $\bar{x}$ in (2).

As for the necessity, suppose that $\bar{x}$ is optimal for (2), $\bar{y}$ is obtained from (9), and (10) fails to hold. Then it must be the case that for some s

$$c_s - \bar{y}^{\mathrm{T}} A_{\bullet s} = \min_{j \notin \beta} \{c_j - \bar{y}^{\mathrm{T}} A_{\bullet j}\} < 0. \tag{25}$$

We shall show that, under these conditions, the value of the primal objective function over the set S can be decreased by increasing x_s. Indeed, denoting the value of the primal objective function by z, we have

$$z = \sum_{j \in \beta} c_j x_j + c_s x_s \tag{26}$$

where it is understood that all primal variables not appearing in (12) are fixed at value zero. Now we also have the relationship

$$A_{\bullet \beta} x_\beta + A_{\bullet s} x_s = b \tag{27}$$

which implicitly decribes how the basic variables x_β behave as functions of the nonbasic variable x_s. In more explicit form, this equation becomes

$$x_\beta = A_{\bullet \beta}^{-1} b - A_{\bullet \beta}^{-1} A_{\bullet s} x_s. \tag{28}$$

Substituting this expression for x_β into equation (12), we get

$$z = c_\beta^{\mathrm{T}} A_{\bullet \beta}^{-1} b + (c_s - c_\beta^{\mathrm{T}} A_{\bullet \beta}^{-1} A_{\bullet s}) x_s. \tag{29}$$

The first term on the right-hand side of (15) is the value of z at the current basic feasible solution (where $x_s = 0$). The rest of the right-hand side of (15) is a negative number (see (9) and (11)) times a nonnegative variable, x_s. When this variable is made even the least bit positive, the value of z will decrease. □

The question then becomes: how much can x_s be increased? When $x_s = 0$ (as it is in the basic solution $\bar{x}$), the value of the vector x_β is positive; in fact it is just $\bar{x}_\beta$. According to (14), it will remain so for all sufficiently small positive values of x_s,. Moreover, depending on the sign of the vector $-A_{\bullet \beta}^{-1} A_{\bullet s}$, it might remain so for all nonnegative values of x_s. Indeed, it will do so if and

[40] Notice that this sufficiency part does not make use of the nondegeneracy assumption.

only if $-A_{\bullet\beta}^{-1}A_{\bullet s} \geq 0$. If this ***sign configuration*** holds, then x_s can be increased indefinitely and the corresponding solution will remain in S, thereby making z decrease to $-\infty$. If, however, this sign configuration does not hold, there will be a limit to how much x_s can be increased before some component of x_β becomes negative and feasibility is lost. To determine this amount, we compute

$$\min_{1\leq i\leq m}\left\{\frac{(A_{\bullet\beta}^{-1}b)_i}{(A_{\bullet\beta}^{-1}A_{\bullet s})_i} : (A_{\bullet\beta}^{-1}A_{\bullet s})_i > 0\right\}. \tag{30}$$

If r belongs to the arg min of this expression, then we know that the basic variable x_{j_r} associated with the r-th equation is (among) the first to decrease to zero as x_s increases.

Changing the basis and carrying on

In the preceding calculations, we identified two indices: s and r. The index s is associated with a nonbasic *column* $A_{\bullet s}$. The index r is associated with a *row* and also with a basic column $A_{\bullet j_r}$. When the nonbasic variable x_s reaches its maximum allowable value (relative to the current basis), the basic variable x_{j_r} decreases to zero. When this occurs, we change the basis, specifically by replacing the current $A_{\bullet j_r}$ by $A_{\bullet s}$. This amounts to revising the definition of j_r, namely by setting its value to s.

Once the basis is changed, there is a new basic feasible solution. In the next iteration, the Simplex Method would test the new basic feasible solution for optimality and repeat the procedure outlined above.

The finiteness argument (assuming nondegeneracy)

If all the basic feasible solutions encountered are nondegenerate, the Simplex Method either detects unboundedness or obtains an optimal solution in a finite number of steps. The reasoning behind this is quite simple:

- The algorithm examines basic feasible solutions at each of which the objective function is uniquely determined.
- With each step from one basic feasible solution to the next, there is a strict decrease of the objective function value, hence no basic feasible solution can re-occur.
- There are only finitely many basic feasible solutions.
- Hence after finitely many steps, the algorithm either detects unboundedness (in which case it stops) or it finds an optimal basic feasible solution of the problem.

What about the degenerate case?

When a degenerate basic feasible solution is encountered, the steps of the Simplex Method can still be performed, but the finiteness argument given above breaks down. In particular, one can no longer be sure of a strict decrease in the value of the objective function. When degeneracy is present, it is possible for the Simplex Method to ***cycle***. When this happens, the sequence of feasible bases may return to one that was previously encountered. It then "cycles around this loop" without ever terminating. This sort of behavior is exhibited by the problem

$$\begin{aligned} \text{minimize} \quad & -2x_1 - 3x_2 + x_3 + 12x_4 \\ \text{subject to} \quad & -2x_1 - 9x_2 + x_3 + 9x_4 \leq 0 \\ & \tfrac{1}{3}x_1 + x_2 - \tfrac{1}{3}x_3 - 2x_4 \leq 0 \\ & x_1, \quad x_2, \quad x_3, \quad x_4 \geq 0 \end{aligned}$$

The style of argument given above is valid if it can be guaranteed that only finitely many iterations will occur between strict decreases in the objective function value. This kind of behavior is called ***stalling***, and it is commonly experienced in practice. This means, of course, that degeneracy must also be quite prevalent.

To counteract the theoretical (and occasionally practical) problem of cycling, researchers have devised ***degeneracy resolution schemes*** that keep the algorithm from stalling forever. Random

selection, perturbation (of the right-hand side), lexicographic ordering, and least-index rules are the names of some of the techniques found in the literature. In actually implementing a degeneracy resolution scheme in an LP code, one has to guard against creating intolerable computational and storage requirements.

Interpretation of dual variables

We wish to exhibit some valuable information that is available after solving a linear programming problem. For this purpose, we denote by (P) the (primal) linear programming problem in standard form:

$$\begin{array}{ll} \text{minimize} & c^{\mathrm{T}}x \\ \text{subject to} & Ax = b \\ & x \geq 0. \end{array}$$

Relative to (P), the dual problem (D) is

$$\begin{array}{ll} \text{maximize} & y^{\mathrm{T}}b \\ \text{subject to} & y^{\mathrm{T}}A \leq c^{\mathrm{T}} \end{array}$$

To ease the discussion, we assume that the matrix A is $m \times n$ and has rank m. This is not a particularly restrictive assumption.

Assume that $\bar{x}$ is an optimal basic[41] feasible solution of (P). This means that there exists an index set β such that:

1. $A_{\bullet\beta}$ is nonsingular;
2. $\bar{x}_\beta = A_{\bullet\beta}^{-1}b \geq 0$;
3. $j \in \nu = \{1, \ldots, n\} \setminus \beta \Longrightarrow \bar{x}_j = 0$.

Recall that if $\bar{x}$ is a *nondegenerate* basic feasible solution, its support has cardinality m; in these circumstances $\bar{x}_\beta > 0$. For the time being, we assume that our $\bar{x}$ is nondegenerate. Then, as assured by the Proposition on page 17 of this handout, the basis matrix $B = A_{\bullet\beta}$ yields an optimal solution of the dual problem (D), namely

[41] Note that it is sometimes possible for a *nonbasic* feasible solution to be optimal.

$$\bar{y}^{\mathrm{T}} = c_{\beta}^{\mathrm{T}} A_{\bullet\beta}^{-1}.$$

Notice that because $\bar{x}_{\nu} = 0$, we have

$$c^{\mathrm{T}}\bar{x} = c_{\beta}^{\mathrm{T}}\bar{x}_{\beta} = c_{\beta}^{\mathrm{T}}(A_{\bullet\beta}^{-1}b) = (c_{\beta}^{\mathrm{T}}A_{\bullet\beta}^{-1})b = \bar{y}^{\mathrm{T}}b.$$

Remark. Imagine a real-world situation where the elements of the vector c represent "unit production costs" and those of the vector b represent "requirements." (Note that by adopting a sign convention on the b_i, we can handle both inputs and outputs. For example, our convention could be that if item i is an input to production (a resource), we let b_i be positive, whereas if item i is an output, then we let b_i be negative. Just the opposite convention might also be used, but whatever convention is adopted, it must be consistent.) The components x_j of the decision vector x represent "activity levels". They tell us how much of the various production activities are used. Accordingly, the scalar product $c^{\mathrm{T}}\bar{x}$ is in units of ***money***. By what is called ***dimensional analysis***, this means that the quantity $\bar{y}^{\mathrm{T}}b$ should be in monetary units. But since b_i is the quantity of i either required (as an output) or available (as an input), it follows that the individual components $\bar{y}_i$ of the optimal solution to (D) must be interpreted as ***monetary units per unit of item*** i. In the jargon of fields such as economics and operations research, the values of the $\bar{y}_i$ are called ***shadow prices***. Relative to the production costs (given by c) and technology (given by A), the shadow prices are used to assess the value of resources or the cost of requirements.

An important consequence of nondegeneracy

Continuing with the preceding setup—in particular the nondegeneracy assumption—we note that

$$A_{\bullet\beta}^{-1}b > 0 \Longrightarrow A_{\bullet\beta}^{-1}b' > 0 \text{ for all } b' \text{ sufficiently close to } b.$$

This means that if the right-hand side is slightly perturbed, the previously optimal basis will remain feasible.

We also notice that changing the right-hand side vector b does not affect the ***feasibility*** of the dual problem. In particular, the vector $\bar{y}$ remains dual feasible. In fact, provided b' is close enough to b, taking $\bar{x}$ such that $\bar{x}_{\beta} = A_{\bullet\beta}^{-1}b'$ and $\bar{x}_{\nu} = 0$ yields an optimal solution for the modified problem (P$'$)

having b' as its right-hand side. Indeed, $\bar{x}$ and $\bar{y}$ satisfy the primal feasibility, dual feasibility, and complementary slackness conditions for the modified problem, and this verifies their optimality.

Let us now consider perturbations of a particular form, namely those in which only one component of b is altered. Thus, for some fixed k let

$$b' = b + \delta e_k \quad \text{where} \quad \delta \in R \quad \text{and} \quad e_k = I_{\bullet k}.$$

This has the effect of altering just the kth component of b. Assuming that $|\delta|$ is small enough to guarantee that $A_{\bullet\beta}^{-1} b' \geq 0$, we compare the optimal objective functions of the two programs, (P) and (P$'$). In fact, the ***optimal objective function values*** for (P) and (P$'$) can be written, respectively, as

$$z(b) = \bar{y}^{\mathrm{T}} b \quad \text{and} \quad z(b') = \bar{y}^{\mathrm{T}} b'.$$

From this—and the special form of b'—we conclude that

$$z(b') - z(b) = \bar{y}^{\mathrm{T}}(b' - b) = \bar{y}^{\mathrm{T}}(\delta e_k) = \delta \bar{y}_k.$$

This equation implies

$$\frac{\partial z(b)}{\partial b_k} = \bar{y}_k.$$

This says that $\bar{y}_k$ measures the marginal value of the kth right-hand side item b_k. This kind of information can be quite valuable.

Another point that is worth noting is that when the primal problem (P) has a nondegenerate optimal solution, the dual problem (D) has a *unique* optimal solution. The truth of this assertion is a simple consequence of the complementary slackness conditions:

$$\bar{x}_j (c_j - y^{\mathrm{T}} A_{\bullet j}) = 0 \quad \text{for all } j = 1, \ldots, n.$$

This and the positivity of $\bar{x}_j$ for $j \in \beta$ imply that $y^{\mathrm{T}} = c_\beta A_{\bullet\beta}^{-1}$. But this is how we defined $\bar{y}^{\mathrm{T}}$. Thus, the dual problem has only one optimal solution.

A key assumption in this discussion is that the optimal solution of the primal problem is nondegenerate. When this is not the case (the solution is degenerate), there is ambiguity in the choice of optimal solution for the dual problem. This fact makes it difficult to speak of an optimal solution to the dual as a definitive set of "prices."

References

D. Bertsimas and J.N. Tsitsiklis [1997] , *Introduction to Linear Optimization*, Athena Scientific, Belmont, Mass.

R.W. Cottle [2003], *The Basic George B. Dantzig*, Stanford: Stanford University Press.

R.W. Cottle, J.S. Pang, and R.E. Stone [1992], *The Linear Complementarity Problem*, Academic Press, Boston. [Chapter 4.]

G.B. Dantzig and M.N. Thapa [1997], *Linear Programming 1: Introduction*, Springer-Verlag, New York.

G.B. Dantzig and M.N. Thapa [2003], *Linear Programming 2: Theory and Extensions*, Springer-Verlag, New York.

P.E. Gill, W. Murray and M.H. Wright [1981], *Practical Optimization*, Academic Press, London. [See pp. 99-104.]

C. Hildreth [1954], Point estimates of concave functions, *Journal of the American Statistical Association* 49, 598-619.

D.G. Luenberger [1989], *Linear and Nonlinear Programming*, Addison-Wesley, Reading, Mass. [See Section 7.6.]

K.G. Murty [1983], *Linear Programming*, John Wiley & Sons, New York.

S.G. Nash and A. Sofer [1996], *Linear and Nonlinear Programming*, McGraw-Hill, New York.

J. Nocedal and S.J. Wright [1999], *Numerical Optimization*, Springer, New York.

A. Schrijver [1986], *Theory of Linear and Integer Programming*, John Wiley & Sons, Chichester.

S.J. Wright [1997], *Primal-Dual Interior-Point Methods*, SIAM Publications, Philadelphia.

Y. Ye [1997], *Interior Point Methods*. New York: John Wiley & Sons.

5.5 Quadratic programmming

As commonly understood, a quadratic programming problem (QP) involves the optimization (say minimization) of a quadratic function subject to linear constraints, ordinarily involving at least some linear inequalities or sign-restricted variables. This definition allows for a considerable range of problem types, several of which we shall discuss below. It does *not* allow for nonlinear constraints of any kind.[42]

About the objective function

In general, a ***quadratic function*** is the sum of three expressions: a constant, a linear form, and a quadratic form (the latter often multiplied by $\frac{1}{2}$). Such a thing looks like

$$f(x) = K + c^{\mathrm{T}}x + \tfrac{1}{2}x^{\mathrm{T}}Qx.$$

Since the additive constant K does not affect the location of an optimal solution, it can be dropped from the formulation. It is usually assumed that the matrix Q appearing in the definition of f is symmetric. This is not a restrictive assumption as long as the value of the quadratic form $x^{\mathrm{T}}Qx$ is what matters.

It should be noted that when $Q = 0$ (the zero matrix), the function $c^{\mathrm{T}}x + \frac{1}{2}x^{\mathrm{T}}Qx$ reduces to a linear function. There is a significant difference between linear and quadratic functions vis-à-vis convexity. As we know, a function is both convex and concave if and only if it is affine (linear plus a constant). Thus, a (truly) quadratic function could be convex, concave, or neither. With R^n regarded as the domain, these alternatives are equivalent to Q being positive semidefinite, negative semidefinite, or indefinite. Where quadratic functions are concerned, it is the Hessian matrix Q that determines whether a function is convex or not. If the domain of the function (e.g., the feasible region of the QP) is full-dimensional, then a quadratic function will be convex on that domain if and only if its Hessian matrix is positive semidefinite (thus making the function convex on the whole space). The general situation is that a quadratic function is convex on a convex set if and

[42] Nowadays, some people are interested in the optimization of quadratic functions subject to quadratic constraints (typically inequalities). Such problems are properly called ***quadratically constrained quadratic programs***.

only if it is convex on the affine hull (carrying plane) of that set. This means that a quadratic function f of n variables defined on a convex set S having dimension less than n could be convex on S even though its Hessian matrix is not positive semidefinite.

Example. The following case illustrates the idea just stated. Here we have $n = 2$. Let

$$f(x_1, x_2) = x_1 + x_2 - \tfrac{1}{2}x_1x_2,$$

and

$$S = \{(x_1, x_2) : x_1 + x_2 = 1,\ x_1 \geq 0,\ x_2 \geq 0\}.$$

The function f is not convex on R^2. However, since $x \in S$ implies $x_2 = 1 - x_1$, we may write the *constrained* quadratic function as

$$x_1 + (1 - x_1) - \tfrac{1}{2}x_1(1 - x_1) = 1 - \tfrac{1}{2}x_1 + \tfrac{1}{2}x_1^2,$$

which is convex on R^1.

Interesting as this example may be, one seldom gets a chance to use this sort of analysis in practice. Nevertheless, if one can analytically determine the affine hull of the feasible region, it possible to decide whether or not a given quadratic function is convex there.

Remark. This preoccupation with convex functions is pertinent to *minimization problems*. In such quadratic programs, the necessary first-order optimality conditions are also sufficient for optimality. By finding a solution of the KKT conditions, we will have found a global optimizer of the minimand. For maximization problems, we would want a *concave* objective function. And while we are at it, let us note that the minimization of a concave (truly) quadratic function can lead to local (non-global) minimizers, and, in some cases, lots of them.

Example. In R^n, the problem

$$\begin{array}{ll} \text{minimize} & -\frac{1}{2}x^{\mathrm{T}}x \\ \text{subject to} & -e \leq x \leq e \end{array}$$

gives rise to 2^n local minimizers. All of them are global minimizers. But just imagine what could happen if the feasible region were more irregular or the minimand were strictly concave but more complicated.

About the constraints

As stated above, the constraints of a quadratic programmng problem can come in many different forms, just as they can in a linear programming problem. Often the forms that are chosen as "standard" enable one to develop a particular aspect of the theory in a particularly nice way.

For our purposes, it will suffice to regard the quadratic programming problem as having the form

$$\begin{array}{ll} \text{minimize} & c^{\mathrm{T}} + \frac{1}{2}x^{\mathrm{T}}Qx \\ \text{subject to} & Ax \geq b. \end{array} \tag{31}$$

Notice that as far as the theory is concerned, this problem has free variables. Nevertheless, it is possible to include variables with explicit upper or lower bounds in such a format.

In connection with certain algorithms (particularly active set methods), it is necessary to consider ***equality-constrained quadratic optimization problems*** of the form

$$\begin{array}{ll} \text{minimize} & c^{\mathrm{T}}x + \frac{1}{2}x^{\mathrm{T}}Qx \\ \text{subject to} & \hat{A}x = \hat{b}. \end{array} \tag{32}$$

This kind of problem may arise when one has to minimize the objective of (1) over the active constraints relative to some feasible point. In that case, the matrix $\hat{A}$ and vector $\hat{b}$ in (2) would be made up of rows of A and b.

Remark. This is an opportune point at which to bring up another notational convention. In general, if f is a function with domain X and if S is a nonempty subset of X, then the ***restriction*** of f to S is the function f_S with domain S such that $f_S(x) = f(x)$ when $x \in S$.

5.4.1 Equality-constrained quadratic optimization problems

We shall begin this algorithmic discussion with equality-constrained quadratic optimization problem such as (2), but to make the notation simpler, we dispense (temporarily) with the hats on A and b. Let us also make the simplifying assumption that $A \in R^{m\times n}$ and that the rows of A are *linearly independent*. Under such circumstances, there will exist a basis in A, that is, a nonsingular $m \times m$ matrix B consisting of m columns of A. Let the indices of these columns be $j_1, \ldots, j_m$. Then we could say that $B = A_{\bullet\beta}$ where $\beta = \{j_1, \ldots, j_m\}$. Let

$$\nu = \{j : 1 \leq j \leq n,\ j \notin \beta\}.$$

With these notations in place, we can write

$$A_{\bullet\nu}x_\nu + A_{\bullet\beta}x_\beta = b \tag{33}$$

from which we immediately obtain the relation

$$x_\beta = A_{\bullet\beta}^{-1}(b - A_{\bullet\nu}x_\nu). \tag{34}$$

Permuting the indices if necessary, we may assume that $x = (x_\nu,\, x_\beta)$ and then write the objective function in the form

$$f(x) = \begin{bmatrix} c_\nu \\ c_\beta \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} x_\nu \\ x_\beta \end{bmatrix} + \tfrac{1}{2} \begin{bmatrix} x_\nu \\ x_\beta \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} Q_{\nu\nu} & Q_{\nu\beta} \\ Q_{\beta\nu} & Q_{\beta\beta} \end{bmatrix} \begin{bmatrix} x_\nu \\ x_\beta \end{bmatrix} \tag{35}$$

Our goal is to minimize the quadratic function f (which is defined on $X = R^n$) for x belonging to the set S of vectors that satisfy the linear equality constraints $Ax = b$. In other words, we want to minimize f_S. Now, to do this, we can use (4) to eliminate x_β from the formula (5) by which $f(x)$ is given. This process yields

$$\begin{aligned} F(x_\nu) \;=\; & c_\beta^{\mathrm{T}} A_{\bullet\beta}^{-1} b + \tfrac{1}{2} b^{\mathrm{T}} A_{\bullet\beta}^{-\mathrm{T}} Q_{\beta\beta} A_{\bullet\beta}^{-1} b \\ & + (c_\nu^{\mathrm{T}} - c_\beta^{\mathrm{T}} A_{\bullet\beta}^{-1} A_{\bullet\nu} + b^{\mathrm{T}} A_{\bullet\beta}^{-\mathrm{T}} Q_{\beta\nu} - b^{\mathrm{T}} A_{\bullet\beta}^{-\mathrm{T}} Q_{\beta\beta} A_{\bullet\beta}^{-1} A_{\bullet\nu}) x_\nu \\ & + \tfrac{1}{2} x_\nu^{\mathrm{T}} (Q_{\nu\nu} - 2 Q_{\nu\beta} A_{\bullet\beta}^{-1} A_{\bullet\nu} + A_{\bullet\nu}^{\mathrm{T}} A_{\bullet\beta}^{-\mathrm{T}} Q_{\beta\beta} A_{\bullet\beta}^{-1} A_{\bullet\nu}) x_\nu, \end{aligned} \tag{36}$$

which is just an *unconstrained* quadratic function of $x_\nu \in R^{n-m}$. (The tiny example at the top of page 24 illustrates this process.) Equation (22) says that to evaluate f_S we can use the formula for F.

Our equality-constrained quadratic optimization problem is a matter of minimizing f_S, that is to say, F. The first-order optimality condition $\nabla F(\bar{x}_\nu) = 0$ will have to hold at a local minimizer $\bar{x}_\nu$ of F. If the Hessian matrix of F is positive semidefinite, these necessary conditions will also be sufficient for the optimality of such stationary point of F. If this Hessian matrix is positive definite, the minimizer of F will be unique.

Null space methods

We are dealing here with the set $S = \{x : Ax = b\}$ where $A \in R^{m\times n}$. A simple observation is that *any two elements of S, say* x and x', *differ by an element of the null space*[43] *of* A. Indeed,

$$Ax = b \quad \text{and} \quad Ax' = b \quad \Longrightarrow \quad A(x - x') = b - b = 0.$$

To phrase this the other way around, if we have a point $x \in S$, and we wish to obtain another point of S, then we shall have to move (from x) in a direction that belongs to the null space of S.

The null space of A is certainly a vector space, a subspace of R^n. If the rows of A are linearly independent, the null space of A will have dimension $n - m$. Now let $Z \in R^{n\times(n-m)}$ be a basis

[43] That is, the set of all vectors p such that $Ap = 0$.

matrix for the null space of A. In terms of the discussion above (where $A = [A_{\bullet\nu}\, A_{\bullet\beta}]$),

$$Z = \begin{bmatrix} I \\ -A_{\bullet\beta}^{-1} A_{\bullet\nu} \end{bmatrix} \tag{37}$$

is such a matrix. It has the prescribed order, and its columns are linearly independent. Moreover, since $AZ = 0$, it follows that $AZv = 0$ for any $v \in R^{n-m}$. But notice that it is not necessary to construct Z in this manner, i.e., with an explicit inverse of the basis matrix $A_{\bullet\beta}$.

To see why we bother with this, look back at (22).

Fortunately, any quadratic function—such as our $f(x) = c^{\mathrm{T}}x + \frac{1}{2}x^{\mathrm{T}}Qx$—satisfies the following relationship[44] for any two points x and y in its domain:

$$\begin{aligned} f(x+y) &= c^{\mathrm{T}}(x+y) + \tfrac{1}{2}(x+y)^{\mathrm{T}}Q(x+y) \\ &= c^{\mathrm{T}}x + \tfrac{1}{2}x^{\mathrm{T}}Qx + y^{\mathrm{T}}(c+Qx) + \tfrac{1}{2}y^{\mathrm{T}}Qy \\ &= f(x) + y^{\mathrm{T}}\nabla f(x) + \tfrac{1}{2}y^{\mathrm{T}}\nabla^2 f(x)y \end{aligned} \tag{38}$$

Let us now use (24) with $x = \bar{x} \in S$ and $y = Zv$ for some $v \in R^{n-m}$. We obtain

$$f(\bar{x} + Zv) = c^{\mathrm{T}}\bar{x} + \tfrac{1}{2}\bar{x}^{\mathrm{T}}Q\bar{x} + v^{\mathrm{T}}Z^{\mathrm{T}}(c + Q\bar{x}) + \tfrac{1}{2}(Zv)^{\mathrm{T}}Q(Zv) \tag{39}$$

The function values of interest can now be regarded as depending on v, so we let

$$\varphi(v) = f(\bar{x} + Zv).$$

(This is, in effect, the restriction of f to S that we saw before.) The stationarity condition for φ (that is, $\nabla\varphi(v) = 0$) gives rise to the equation

$$Z^{\mathrm{T}}QZv = -Z^{\mathrm{T}}(c + Q\bar{x}). \tag{40}$$

In solving this equation for v, we would be finding the Newton direction for φ. Relative to the equality-constrained QP, a solution v of (26) gives what is called a ***reduced Newton direction***, Zv. The matrix $Z^{\mathrm{T}}QZ$ is called the ***reduced Hessian matrix.***

5.4.2 Quadratic programs with linear inequality constraints

This discussion centers on algorithms for quadratic programming problems of the form

$$\begin{aligned} &\text{minimize} && c^{\mathrm{T}}x + \tfrac{1}{2}x^{\mathrm{T}}Qx \\ &\text{subject to} && Ax \geq b. \end{aligned} \tag{41}$$

[44] This is just saying that the Taylor series expansion of a quadratic function has just three terms.

Let us assume that $A \in R^{m \times n}$. In making this assumption, we are not declaring which of the two positive integers m and n is the larger. Indeed, because we are dealing with linear inequalities rather than equations here, it is possible for m to be (considerably) larger than n without making the system "overdetermined" as it would be if the linear inequalities were replaced by linear equations. But we may equally well have a problem in which the number n is much larger than m. The relative sizes of m and n may in fact dictate what sort of algorithm is more appropriate for the problem at hand. More will be said about this issue at a later stage in these notes.

The two classes of algorithms we shall examine are called ***active set methods*** and ***range-space methods.*** These kinds of algorithms work exclusively with feasible solutions of problems, i.e., they are ***feasible point methods.*** As we have seen, the matter of finding a feasible solution of the linear constraints $Ax \geq b$ can be handled by setting up a suitable Phase I Problem as in linear programming. *We shall assume that this aspect of the problem has already been taken care of, i.e., that a feasible solution of the constraints is known.*

For the sake of brevity, we denote the feasible region of (27) by

$$S = \{x : Ax \geq b\}.$$

The first-order (KKT) optimality conditions for a local minimizer of the problem stated in (27) are

$$c + Qx - A^{\mathrm{T}}y = 0 \tag{42}$$

$$-b + Ax \geq 0 \tag{43}$$

$$y \geq 0 \tag{44}$$

$$y^{\mathrm{T}}(-b + Ax) = 0 \tag{45}$$

In a general sense, the strategy behind the quadratic programming algorithms we shall see is to find a solution of this system. In the case where the objective function is convex, satisfying these conditions will guarantee us that the x-vector is a global minimizer for (27).

Suppose we want to know if a particular feasible vector, say $\bar{x}$, is part of a solution $(\bar{x}, \bar{y})$ to to the KKT system, (28)–(31). What can we do? Recall that for any feasible vector, say $\bar{x}$, there is a corresponding index set

$$\mathcal{A}(\bar{x}) = \{i : -b_i + A_{i\bullet}\bar{x}\} = 0.$$

If $(\bar{x}, \bar{y})$ is to be a solution of the KKT system, then for each $i \notin \mathcal{A}(\bar{x})$:

$$-b_i + A_{i\bullet}\bar{x} > 0 \quad \text{and} \quad \bar{y}_i = 0. \tag{46}$$

Given a feasible solution of the QP, this information can be of use in actually attempting to solve the KKT system, and possibly the problem. Let us define

$$\tau = \mathcal{A}(\bar{x}) \quad \text{and} \quad \hat{A} = A_{\tau\bullet}, \qquad \hat{b} = b_\tau.$$

While we are at it, we can define $\hat{y}$ as a (presently unknown) vector whose coordinates are in one-one correspondence with the $i \in \mathcal{A}(\bar{x})$. Once the coordinates of this vector are found, we can extend it to an m-vector $\bar{y}$ according to the definition

$$\bar{y}_i = \begin{cases} \hat{y}_i & \text{if } i \in \mathcal{A}(\bar{x}) \\ 0 & \text{if } i \notin \mathcal{A}(\bar{x}). \end{cases} \tag{47}$$

In aiming to satisfy the KKT system (28)–(31), we will be aided by the additional assumption that the rows of $\hat{A}$ are *linearly independent.*

Using the above notations, we can write

$$c + Q\bar{x} - \hat{A}^{\mathrm{T}}\hat{y} = 0 \tag{48}$$

$$-\hat{b} + \hat{A}\bar{x} = 0, \tag{49}$$

and with the linear independence assumption in force, we can write

$$\hat{y} = \hat{A}_r^{\mathrm{T}}(c + Q\bar{x}) \tag{50}$$

where $\hat{A}_r$ is a ***right inverse***[45] of $\hat{A}$.

Once $\hat{y}$ is obtained, it can be tested for nonnegativity. First, suppose $\hat{y} \geq 0$. Let $\bar{y}$ be the m-vector defined in (33). Then the pair $(\bar{x}, \bar{y})$ satisfies the KKT conditions (28)–(31). Whether the vector $\bar{x}$ is a local minimizer is a separate question, however. Next, suppose $\hat{y} \not\geq 0$. Then it would appear that there is more work to do. In effect, this condition says that when one of the coefficients of the reduced gradient vector (relative to the binding constraints) is negative, there is an equality-constrained descent direction.

Active set methods often include the notion of a ***working set***. This term is a potential source of confusion, not to mention rather imprecise language. When we have a point $\bar{x} \in S$ (such as the feasible region of (27) defined above), there is an associated set of indices of the linear inequalities that make up the constraints defining S. The active set corresponding to $\bar{x}$ can be thought of as the set $\mathcal{A}(\bar{x})$ defined above. This is a set of indices (subscripts). It is not the linear manifold defined by

[45] This means $\hat{A}\hat{A}_r = I$.

the corresponding system of equations, nor is it the linear inequalities themselves. The ***working set***, denoted $\mathcal{W}$, is a (possibly proper) subset of $\mathcal{A}(\bar{x})$ that is actually being used to define the search direction.

Assuming a ***search direction*** p at some feasible point x is given, the ***step size*** is partially governed by

$$\begin{aligned}\bar{\alpha} &= \max\{\alpha \geq 0 : x + \alpha p \in S\} \\ &= \min\{\frac{A_{i\bullet}x - b_i}{-A_{i\bullet}p} : A_{i\bullet}p < 0,\ i \notin \mathcal{W}\}\end{aligned}$$

In general, there is a possibility that the maximum defined above does not actually exist. (See the Remark concerning this point below.) Barring this situation, the calculation of the step size in the case of a quadratic function involves looking for the smaller of two numbers: $\bar{\alpha}$ (above) and the minimum of the quadratic $f(x + \alpha p)$ for $\alpha \geq 0$. Ordinarily, this is a strictly convex quadratic in one variable that takes its minimum value at some positive number.[46] The issue then is how this number compares to $\bar{\alpha}$.

[46] When p is the reduced Newton direction, this number is 1.

Active Set Method for (27)

0. *Initialization.* Let $x_0 \in S$ be given. Define the matrix $\bar{A}$ and the vector $\bar{b}$ corresponding to $\mathcal{A}(x_0)$. Choose $\mathcal{W} = \mathcal{A}(x_0)$ as the working set. Let Z denote a null space matrix for $\bar{A}$, and let $\bar{A}_r$ be a right inverse for $\bar{A}$. Set $k = 0$ (the iteration counter). Let $f(x) = c^{\mathrm{T}}x + \frac{1}{2}x^{\mathrm{T}}Qx$, so that $\nabla f(x) = c + Qx$ and $\nabla^2 f(x) = Q$.

1. *Test for optimality.* If $Z^{\mathrm{T}}\nabla f(x_k) = 0$, then

   (a) If $\mathcal{A}(x_k) = \emptyset$, then x_k is a stationary point of f on S. Stop.[47]

   (b) If $\mathcal{A}(x_k) \neq \emptyset$, compute the Lagrange multiplier vector

   $$\bar{y} = \bar{A}_r^{\mathrm{T}} \nabla f(x_k).$$

   (c) If $\bar{y} \geq 0$, stop. (A local stationary point has been found.)

   (d) Otherwise, drop (i.e., disregard) a single constraint index corresponding to a negative Lagrange multiplier. Revise $\mathcal{W}$, $\bar{A}$, $\bar{b}$, Z, and $\bar{A}_r$ accordingly.

2. *Find a search direction.* Compute a descent direction p relative to the constraints given by the working set.[48]

3. *Find the step length.* Compute α so that $f(x_k + \alpha p) < f(x_k)$ and $\alpha \leq \bar{\alpha}$ where $\bar{\alpha}$ is the maximum allowable step length in the direction p consistent with retaining feasibility.

4. *Update everything.* Define $x_{k+1} = x_k + \alpha p$. If a new constraint is met, add the index of this constraint to the working set and update $\mathcal{W}$, $\bar{A}$, $\bar{b}$, Z, and $\bar{A}_r$ accordingly. Replace k by $k + 1$ and return to Step 1.

Remark. The wording in Step 3 should really allow for the case where $\bar{\alpha} = \infty$. This would mean that for all $\alpha \geq 0$, $A_{i\bullet}(x_k + \alpha p) > b_i$ for all $i \notin \mathcal{A}(x_k)$. If this is the case, the objective function is unbounded below over the feasible region.

Example. Consider the quadratic program

$$\begin{aligned} \text{minimize} \quad & -x_1 - x_2 + \tfrac{1}{2}(x_1 - x_2)^2 \\ \text{subject to} \quad & x_1 + x_2 \geq 1 \\ & x_1,\ x_2 \geq 0 \end{aligned}$$

[47] In the convex case, x_k would be a global minimum.

[48] This allows for some flexibility in the method by which choice the search direction is actually computed. The reduced Newton direction is a common choice.

Starting from the interior point $x_0 = (1, 1)$, the search direction would be $p = (1, 1)$, and we could move in this direction indefinitely, all the while decreasing the objective function value.

Range space method for (27)

Consider a quadratic program of the form (27) in which there are far more variables than constraints, i.e., we have $m \ll n$. Suppose Q is an easily invertible matrix. A particular example of that would be the case where Q is positive definite and diagonal, in which case (27) would have a unique solution provided $S \neq \emptyset$. Writing the optimality conditions (28)–(31), we see that (28) implies

$$x = -Q^{-1}(c - A^{\mathrm{T}}y). \tag{51}$$

This equation permits us to substitute the right-hand side for x in the rest of the first-order conditions. They become

$$-AQ^{-1}(c - A^{\mathrm{T}}y) \geq b \tag{52}$$

$$y \geq 0 \tag{53}$$

$$y^{\mathrm{T}}[-AQ^{-1}c + AQ^{-1}A^{\mathrm{T}}y - b] = 0 \tag{54}$$

In the case where Q is positive definite, the conditions (38)–(40) can be regarded as the first-order optimality conditions of the quadratic program

$$\begin{array}{ll} \text{minimize} & -(AQ^{-1}c + b)^{\mathrm{T}}y + \frac{1}{2}y^{\mathrm{T}}AQ^{-1}A^{\mathrm{T}}y \\ \text{subject to} & y \geq 0 \end{array} \tag{55}$$

One salient feature of the quadratic program (41) is that it is a problem in m variables, namely the Lagrange multipliers, $y_1, \ldots, y_m$. The Hessian matrix $AQ^{-1}A^{\mathrm{T}}$ is at least positive semidefinite, and if the rows of A are linearly independent, then it is positive definite. If (41) has no optimal solution, then (27) must be infeasible. If (41) is a strictly convex QP, its solution will be unique. Then x can be unambiguously obtained from equation (37). But what if $AQ^{-1}A^{\mathrm{T}}$ is only positive semidefinite? This would allow for (41) to have alternate optima. But this has no adverse effect on the matter of finding x, which after all has to be unique when Q is positive definite.

This type of approach to solving (27) is called a ***range space method***. It is also called a ***dualization technique.*** Perhaps the earliest occurrence of this approach in the literature is to be found in Hildreth [1954].

5.4.3 Two numerical examples[49]

Example 1. The following is a strictly convex quadratic program with a compact feasible region containing the vector $\boldsymbol{x_0} = (0, 0)$.

$$\begin{array}{lrcrcl} \text{minimize} & -2x_1 - 6x_2 + \frac{1}{2}(2x_1^2 - 4x_1x_2 + 4x_2^2) & & & & \\ \text{subject to} & -x_1 & - & x_2 & \geq & -2 \\ & x_1 & - & 2x_2 & \geq & -2 \\ & x_1 & & & \geq & 0 \\ & & & x_2 & \geq & 0 \end{array}$$

In this discussion, we include the nonnegativity constraints with the other linear inequalities. In all, then, we have $m = 4$ linear inequality constraints for which the data are

$$A = \begin{bmatrix} -1 & -1 \\ 1 & -2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} -2 \\ -2 \\ 0 \\ 0 \end{bmatrix}, \quad c = \begin{bmatrix} -2 \\ -6 \end{bmatrix}, \quad Q = \begin{bmatrix} 2 & -2 \\ -2 & 4 \end{bmatrix}$$

The objective function $f(x) = c^{\mathrm{T}}x + \frac{1}{2}x^{\mathrm{T}}Qx$ has the gradient $\nabla f(x) = c + Qx$. Just for the record, we note that the *unconstrained* minimum of f is $-Q^{-1}c = (5, 4)$ which does *not* satisfy the constraints of the problem, hence we use an algorithm to find the solution of the constrained optimization problem.

At the starting point $\boldsymbol{x_0} = (0, 0)$, the third and fourth constraints are active, so we have $\mathcal{A}(\boldsymbol{x_0}) = \{3, 4\} = \mathcal{W}$. Then we have

$$\bar{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \bar{A}_r = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Since $\bar{A}$ is a nonsingular matrix, its null space is $\{0\}$, so for the null space matrix we take

$$Z = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

At this point, we have $Z^{\mathrm{T}}\nabla f(\boldsymbol{x_0}) = 0$, and $\mathcal{A}(\boldsymbol{x_0}) \neq \emptyset$, so we compute

$$\bar{y} = \bar{A}_r^{\mathrm{T}}\nabla f(\boldsymbol{x_0}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -2 \\ -6 \end{bmatrix} = \begin{bmatrix} -2 \\ -6 \end{bmatrix}.$$

[49] Because of the need to distinguish iteration counters from ordinary subscripts referring to individual variables, we resort here to using boldface notation for iterates and their counters.

Since this is not a nonnegative vector, we are going to drop one of the constraints from the working set, namely constraint 4. (This will allow x_2 to become positive.) Accordingly, we revise some definitions as follows:

$$\mathcal{W} = \{3\}, \quad \bar{A} = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} 0 \end{bmatrix}, \quad Z = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \bar{A}_r = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

What all this says it that we are presently enforcing the constraint $x_1 = 0$. In this simple situation, there can't be much mystery about what the search direction will be, but we compute it anyway. The reduced Newton direction vector is

$$p = -Z(Z^{\mathrm{T}}QZ)^{-1}Z^{\mathrm{T}}\nabla f(\boldsymbol{x_0}) = \begin{bmatrix} 0 \\ 3/2 \end{bmatrix}.$$

The idea now is to move from $\boldsymbol{x_0}$ to $\boldsymbol{x_0} + \alpha p$ for some $\alpha > 0$. The value $\alpha = 1$ corresponds to the global minimizer of f restricted to the line $x_1 = 0$. It is easy to check that this point is infeasible. We need a smaller α. We find it by enforcing the feasibility condition $A(0 + \alpha p) \geq b$:

$$\begin{bmatrix} -1 & -1 \\ 1 & -2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ (3/2)\alpha \end{bmatrix} \geq \begin{bmatrix} -2 \\ -2 \\ 0 \\ 0 \end{bmatrix}$$

which turns out to imply that $\alpha \leq 2/3$. Thus, the step length at the current iterate is $2/3$. We define the next iterate to be $\boldsymbol{x_1} = \boldsymbol{x_0} + (2/3)p = (0, 1)^{\mathrm{T}}$. For this iterate, we have $\mathcal{A}(\boldsymbol{x_1}) = \{2, 3\}$, and the other corresponding data are

$$\mathcal{W} = \{2, 3\}, \quad \bar{A} = \begin{bmatrix} 1 & -2 \\ 1 & 0 \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} -2 \\ 0 \end{bmatrix}, \quad Z = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \bar{A}_r = \tfrac{1}{2}\begin{bmatrix} 0 & 2 \\ -1 & 1 \end{bmatrix}.$$

At the beginning of the next iteration, we have $Z^{\mathrm{T}}\nabla f(\boldsymbol{x_1}) = 0$ and $\mathcal{A}(\boldsymbol{x_1}) \neq \emptyset$, so we define the Lagrange multiplier vector

$$\bar{y} = \tfrac{1}{2}\begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} -4 \\ -2 \end{bmatrix} = \begin{bmatrix} 1 \\ -5 \end{bmatrix}.$$

Since this vector is not nonnegative, we decide to drop constraint 3 from the working set. Geometrically, this means we are going to enforce constraint 2 but allow x_1 to become positive. Updating, we get

$$\mathcal{W} = \{2\}, \quad \bar{A} = \begin{bmatrix} 1 & -2 \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} -2 \end{bmatrix}, \quad Z = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \bar{A}_r = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Again we compute a reduced Newton search direction vector. It is given by the formula

$$p = -Z(Z^{\mathrm{T}}QZ)^{-1}Z^{\mathrm{T}}\nabla f(\boldsymbol{x_1}) = \begin{bmatrix} 5 \\ 5/2 \end{bmatrix}.$$

To compute search length, we look at the system

$$\begin{bmatrix} -1 & -1 \\ 1 & -2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 5\alpha \\ 1+(5/2)\alpha \end{bmatrix} \geq \begin{bmatrix} -2 \\ -2 \\ 0 \\ 0 \end{bmatrix}$$

The maximum value of α for which this holds is $2/15$. From this we conclude that the step size is $\alpha = \min\{1, 2/15\} = 2/15$. Using this, we define a new iterate $\boldsymbol{x_2} = \boldsymbol{x_1} + (2/15) \cdot p = (2/3\,, 4/3)$.

Relative to the new iterate $\boldsymbol{x_2}$ we have $\mathcal{A}(\boldsymbol{x_2}) = \{1,\, 2\}$. We then have

$$\mathcal{W} = \{1,\, 2\}, \quad \bar{A} = \begin{bmatrix} -1 & -1 \\ 1 & -2 \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} -2 \\ -2 \end{bmatrix}, \quad Z = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \bar{A}_r = \tfrac{1}{3}\begin{bmatrix} -2 & 1 \\ -1 & -1 \end{bmatrix}.$$

At this point we have $Z^{\mathrm{T}}\nabla f(\boldsymbol{x_2}) = 0$ and $\mathcal{A}(\boldsymbol{x_2}) \neq \emptyset$. Thus, we need to compute

$$\bar{y} = \bar{A}_r^{\mathrm{T}}\nabla f(\boldsymbol{x_2}) = \tfrac{1}{3}\begin{bmatrix} -2 & -1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} -10/3 \\ -2 \end{bmatrix} = \begin{bmatrix} 26/9 \\ -4/9 \end{bmatrix}.$$

We drop the constraint corresponding to the negative component in $\bar{y}$, namely 2. Thus we get a new data set

$$\mathcal{W} = \{1\}, \quad \bar{A} = \begin{bmatrix} -1 & -1 \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} -2 \end{bmatrix}, \quad Z = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \quad \bar{A}_r = \begin{bmatrix} -1 \\ 0 \end{bmatrix}.$$

The corresponding reduced Newton direction is

$$p = -Z(Z^{\mathrm{T}}QZ)^{-1}Z^{\mathrm{T}}\nabla f(\boldsymbol{x_2}) = \begin{bmatrix} 2/15 \\ -2/15 \end{bmatrix}.$$

We see that $\boldsymbol{x_2} + \alpha p$ is feasible for $0 \leq \alpha \leq 10$. But since 10 is greater than 1, the latter is our step length. Thus, we obtain the new iterate $\boldsymbol{x_3} = \boldsymbol{x_2} + 1 \cdot p = (4/5\,, 6/5)$.

At $\boldsymbol{x_3}$ we have $\mathcal{A}(\boldsymbol{x_3}) = \{1\}$. With

$$\mathcal{W} = \{1\}, \quad \bar{A} = \begin{bmatrix} -1 & -1 \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} -2 \end{bmatrix}, \quad Z = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \quad \bar{A}_r = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

we find that the reduced gradient is zero and the corresponding $\bar{y} = 14/5 > 0$. This tells us that $\boldsymbol{x_3}$ is a local minimizer. Since the objective function is strictly convex on R^2, $\boldsymbol{x_3}$ is the unique global minimizer for this little quadratic program.

It can be checked that $\boldsymbol{x_3} = (4/5, 6/5)$ and $y = (14/5, 0, 0, 0)$ satisfy the KKT conditions for this problem.

Remark. Something to notice about this problem is that although the iterates all lie on edges of the feasible region, the unique optimal solution is not at one of its extreme points. This illustrates a difference between linear and quadratic programming.

Example 2. This example has the same constraints as Example 1, but its objective function is strictly concave instead of convex. In such a case, we expect to find the local minimizers at extreme points of the feasible region.

$$\begin{array}{lrcrcr} \text{minimize} & -2x_1 - 6x_2 - \frac{1}{2}(2x_1^2 - 4x_1x_2 + 4x_2^2) & & & & \\ \text{subject to} & -x_1 & - & x_2 & \geq & -2 \\ & x_1 & - & 2x_2 & \geq & -2 \\ & x_1 & & & \geq & 0 \\ & & & x_2 & \geq & 0 \end{array}$$

Notice that if the constraints are ignored, this objective function has a global maximizer at $(-5, -4)$ and has no local minimizers. Let us see what the Active Set Method will do with this problem when we start it at $\boldsymbol{x_0} = (0, 0)$.

Before we apply the algorithm, let us note that $\boldsymbol{x_0}$ is not a local minimum. Indeed, the necessary condition[50] $\nabla f(\boldsymbol{x_0})^{\mathrm{T}} v \geq 0$ for all $v \in \mathcal{F}$ does not hold there. In fact, at $\boldsymbol{x_0}$, every feasible direction is a descent direction.

This time, the problem data are

$$A = \begin{bmatrix} -1 & -1 \\ 1 & -2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} -2 \\ -2 \\ 0 \\ 0 \end{bmatrix}, \quad c = \begin{bmatrix} -2 \\ -6 \end{bmatrix}, \quad Q = \begin{bmatrix} -2 & 2 \\ 2 & -4 \end{bmatrix}$$

At the starting point $\boldsymbol{x_0} = (0, 0)$, the third and fourth constraints are active, so we have $\mathcal{A}(\boldsymbol{x_0}) = \{3, 4\} = \mathcal{W}$. Then we have

$$\bar{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \bar{A}_r = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Because the matrix $\bar{A}$ is nonsingular, its null space is $\{0\}$, so for the null space matrix we take

$$Z = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

[50] See the theorem given on page 2 of Handout No. 17.

At this point, we have $Z^{\mathrm{T}}\nabla f(\boldsymbol{x_0}) = 0$, and $\mathcal{A}(\boldsymbol{x_0}) \neq \emptyset$, so we compute

$$\bar{y} = \bar{A}_r^{\mathrm{T}}\nabla f(\boldsymbol{x_0}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} -2 \\ -6 \end{bmatrix} = \begin{bmatrix} -2 \\ -6 \end{bmatrix}.$$

Continuing as in Example 1, we drop constraint 4 from the working set. The new definitions of interest are

$$\mathcal{W} = \{3\}, \quad \bar{A} = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} 0 \end{bmatrix}, \quad Z = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \bar{A}_r = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

As before, we seek a feasible descent direction. Computing the reduced Newton direction is not appropriate in this case as it does not yield a feasible direction of descent. Instead, we use the direction given by projecting $-\nabla f(\boldsymbol{x_0})$ onto the null space of $\bar{A}$. This yields[51]

$$p = -[I - \bar{A}^{\mathrm{T}}(\bar{A}\bar{A}^{\mathrm{T}})^{-1}\bar{A}]\nabla f(\boldsymbol{x_0}) = \begin{bmatrix} 0 \\ 6 \end{bmatrix}.$$

The step size computation leads to $\alpha = 1/6$ and hence we arrive at the point $\boldsymbol{x_1} = (0, 1)$. The corresponding data are

$$\mathcal{W} = \{2, 3\}, \quad \bar{A} = \begin{bmatrix} 1 & -2 \\ 1 & 0 \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} -2 \\ 0 \end{bmatrix}, \quad Z = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \bar{A}_r = \tfrac{1}{2}\begin{bmatrix} 0 & 2 \\ -1 & 1 \end{bmatrix}.$$

Here, too, we have $Z^{\mathrm{T}}\nabla f(\boldsymbol{x_1}) = 0$ and $\mathcal{A}(\boldsymbol{x_1}) \neq \emptyset$, so we define the Lagrange multiplier vector

$$\bar{y} = \tfrac{1}{2}\begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ -10 \end{bmatrix} = \begin{bmatrix} 5 \\ -5 \end{bmatrix}.$$

This vector is not nonnegative. We drop the constraint corresponding to the negative component of $\bar{y}$. This allows us to make x_1 positive. The new working set is $\mathcal{W} = \{2\}$, and we now have

$$\bar{A} = [1 \ -2], \quad \bar{b} = [-2], \quad Z = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \bar{A}_r = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Taking the search direction to be the negative of the projected gradient, we get $p = (4/5,\ 1/5)$. The next task is to find the step length. This turns out to be $1/6$, and the new point reached is $\boldsymbol{x_2} = (2/3,\ 4/3)$.

At $\boldsymbol{x_2}$ we have $\mathcal{A}(\boldsymbol{x_2}) = \mathcal{W} = \{1,\ 2\}$ and

$$\bar{A} = \begin{bmatrix} -1 & -1 \\ 1 & -2 \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} -2 \\ -2 \end{bmatrix}, \quad Z = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \bar{A}_r = \tfrac{1}{3}\begin{bmatrix} -2 & 1 \\ -1 & -1 \end{bmatrix}.$$

[51] For a discussion of this projection process, see Nash and Sofer, page 59.

Performing the optimality test, we find that we need to compute the Lagrange multipliers. We get

$$\bar{y} = \bar{A}_r^{\mathrm{T}} \nabla f(\boldsymbol{x_2}) = \begin{bmatrix} 34/9 \\ 28/9 \end{bmatrix}.$$

Since this vector is positive, we stop. We have found a local stationary point.

Is $\boldsymbol{x_2}$ a local minimum? Let's check it with the second-order optimality criterion. In this case we need to verify two conditions:

1. $\nabla f(\boldsymbol{x_2})^{\mathrm{T}} v \geq 0$ for all $v \in \mathcal{F}$;
2. $v^{\mathrm{T}} Q v \geq 0$ for all $v \in \mathcal{F} \cap \{\nabla f(\boldsymbol{x_2})\}^{\perp}$.

The set $\mathcal{F}$ is the cone of feasible directions at $\boldsymbol{x_2}$. It is given as the solution set of the homogeneous linear inequality system

$$-v_1 - \ v_2 \geq 0 \tag{56}$$

$$v_1 - 2v_2 \geq 0 \tag{57}$$

It follows (by Farkas's theorem) from the nonnegativity of the vector $\bar{y}$ that the first condition of the second-order optimality criterion holds. What about the second? Note that (42) and (43) imply $v_2 \leq 0$. The vectors $v \in \{\nabla f(\boldsymbol{x_2})\}^{\perp}$ all satisfy the equation

$$v_1 + 15v_2 = 0. \tag{58}$$

But together (42) and (44) imply $v_2 \geq 0$. Consequently $v_1 = v_2 = 0$, and therefore the other second-order optimality condition holds. In short, we have found a local minimizer.

The following figure pertains to these two examples. As we have seen, the solution found in Example 1 is $(4/5, 6/5)$ which lies on the boundary of the feasible region, but not at an extreme point. The local minimizer found in Example 2 is at the extreme point $(2/3, 4/3)$.

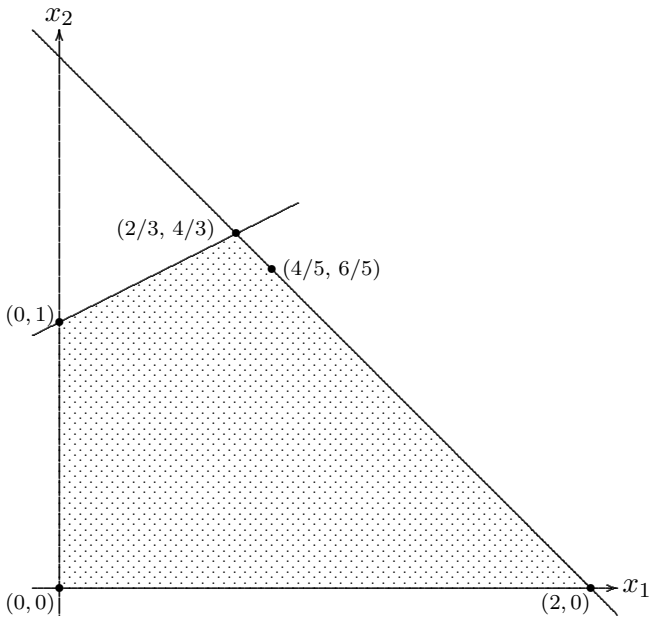


Figure 5.3

References

M. Avriel [1976], *Nonlinear Programming*, Prentice-Hall, Engelwood Cliffs, N.J. [See Section 10.1.]

A. Ben-Tal and A. Nemirovski [2001], *Lectures on Modern Convex Optimization*, MPS/SIAM, Philadelphia.

D.P. Bertsekas [1982], *Constrained Optimization and Lagrange Multiplier Methods*, Athena Scientific, Belmont, Mass.

D.P. Bertsekas [1999], *Nonlinear Programming*, Athena Scientific, Belmont, Mass.

D.P. Bertsekas [2003], *Convex Analysis and Optimization*, Athena Scientific, Belmont, Mass.

D. Bertsimas and J.N. Tsitsiklis [1997] , *Introduction to Linear Optimization*, Athena Scientific, Belmont, Mass.

J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C.A. Sagastizábal [2003], *Numerical Optimization*, Springer-Verlag, Berlin.

A. Cauchy [1847], Méthode générale pour la résolution des systèms d'équations simultanées, *Comp. Rend. Acad. Sci. Paris*, pp. 536-538.

R.W. Cottle [2003], *The Basic George B. Dantzig*, Stanford: Stanford University Press.

R.W. Cottle, J.S. Pang, and R.E. Stone [1992], *The Linear Complementarity Problem*, Academic Press, Boston. [Chapter 4.]

G.B. Dantzig and M.N. Thapa [1997], *Linear Programming 1: Introduction*, Springer-Verlag, New York.

G.B. Dantzig and M.N. Thapa [2003], *Linear Programming 2: Theory and Extensions*, Springer-Verlag, New York.

R. Fletcher [1980], *Practical Methods of Optimization, Volume 2: Constrained Optimization*, John Wiley & Sons, Chichester.

G.E. Forsythe [1968], On the asymptotic directions of the s-dimensional optimum gradient method, *Numerische Mathematik* 11, 57-76.

P.E. Gill, W. Murray, M.A. Saunders, and M.H. Wright [1989], "Constrained nonlinear programming," (Chapter III) in G.L. Nemhauser, A.H.G. Rinnooy Kan, and M.J. Todd (eds.), *Optimization*, North-Holland, Amsterdam.

P.E. Gill, W. Murray and M.H. Wright [1981], *Practical Optimization*, Academic Press, London. [See pp. 99-104.]

C. Hildreth [1954], Point estimates of concave functions, *Journal of the American Statistical Association* 49, 598-619.

D.G. Luenberger [1989], *Linear and Nonlinear Programming*, Addison-Wesley, Reading, Mass. [See Section 7.6.]

K.G. Murty [1983], *Linear Programming*, John Wiley & Sons, New York.

S.G. Nash and A. Sofer [1996], *Linear and Nonlinear Programming*, McGraw-Hill, New York.

Y. Nesterov and A. Nemirovski [1994], *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia.

J. Nocedal and S.J. Wright [1999], *Numerical Optimization*, Springer, New York.

J.M. Ortega and W.C. Rheinboldt [1970], *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York.

A. Schrijver [1986], *Theory of Linear and Integer Programming*, John Wiley & Sons, Chichester.

S.J. Wright [1997], *Primal-Dual Interior-Point Methods*, SIAM Publications, Philadelphia.