Nonlinear Programming

THIRD EDITION

Dimitri P. Bertsekas

Massachusetts Institute of Technology

WWW site for book information and orders

http://www.athenasc.com

Athena Scientific, Belmont, Massachusetts

APPENDIX A:
Mathematical Background

In this appendix, we collect definitions, notational conventions, and several results from linear algebra and real analysis that are used extensively in nonlinear programming. Only a few proofs are given. Additional proofs can be found in Appendix A of the book by Bertsekas and Tsitsiklis [BeT89], which provides a similar but more extended summary of linear algebra and analysis. Related and additional material can be found in the books by Hoffman and Kunze [HoK71], Lancaster and Tismenetsky [LaT85], and Strang [Str76] (linear algebra), and the books by Ash [Ash72], Ortega and Rheinboldt [OrR70], and Rudin [Rud76] (real analysis).

Set Notation

If X is a set and x is an element of X, we write $x \in X$. A set can be specified in the form $X = \{x \mid x \text{ satisfies } P\}$, as the set of all elements satisfying property P. The union of two sets X_1 and X_2 is denoted by $X_1 \cup X_2$, and their intersection by $X_1 \cap X_2$. The symbols $\exists$ and $\forall$ have the meanings "there exists" and "for all," respectively. The empty set is denoted by $\emptyset$.

The set of real numbers (also referred to as scalars) is denoted by $\Re$. The set $\Re$ augmented with $+\infty$ and $-\infty$ is called the *set of extended real numbers*. We write $-\infty < x < \infty$ for all real numbers x, and $-\infty \leq x \leq \infty$ for all extended real numbers x. We denote by $[a, b]$ the set of (possibly extended) real numbers x satisfying $a \leq x \leq b$. A rounded, instead of square, bracket denotes strict inequality in the definition. Thus $(a, b]$, $[a, b)$, and (a, b) denote the set of all x satisfying $a < x \leq b$, $a \leq x < b$, and $a < x < b$, respectively. Furthermore, we use the natural extensions of the rules of arithmetic: $x \cdot 0 = 0$ for every extended real number x, $x \cdot \infty = \infty$ if $x > 0$, $x \cdot \infty = -\infty$ if $x < 0$, and $x + \infty = \infty$ and $x - \infty = -\infty$ for every scalar x. The expression $\infty - \infty$ is meaningless and is never allowed to occur.

Inf and Sup Notation

The *supremum* of a nonempty set X of scalars, denoted by $\sup X$, is defined as the smallest scalar y such that $y \geq x$ for all $x \in X$. If no such scalar y exists, we say that the supremum of X is ∞. Similarly, the *infimum* of X, denoted by $\inf X$, is defined as the largest scalar y such that $y \leq x$ for all $x \in X$, and is equal to $-\infty$ if no such scalar y exists. For the empty set, we use the convention

$$\sup \emptyset = -\infty, \qquad \inf \emptyset = \infty.$$

If $\sup X$ is equal to a scalar $\bar{x}$ that belongs to the set X, we say that $\bar{x}$ is the *maximum point* of X and we write $\bar{x} = \max X$. Similarly, if $\inf X$ is equal to a scalar $\bar{x}$ that belongs to the set X, we say that $\bar{x}$ is the *minimum point* of X and we write $\bar{x} = \min X$. Thus, when we write $\max X$ (or $\min X$) in place of $\sup X$ (or $\inf X$, respectively), we do so just for emphasis: we indicate that it is either evident, or it is known through earlier analysis, or it is about to be shown that the maximum (or minimum, respectively) of the set X is attained at one of its points.

Function Notation

If f is a function, we use the notation $f : X \mapsto Y$ to indicate the fact that f is defined on a nonempty set X (its *domain*) and takes values in a set Y (its *range*). Thus when using the notation $f : X \mapsto Y$, we implicitly assume that X is nonempty. If $f : X \mapsto Y$ is a function, and U and V are subsets of X and Y, respectively, the set $\big\{f(x) \mid x \in U\big\}$ is called the *image* or *forward image of* U *under* f, and the set $\big\{x \in X \mid f(x) \in V\big\}$ is called the *inverse image of* V *under* f.

A.1 VECTORS AND MATRICES

We denote by $\Re^n$ the set of n-dimensional real vectors. For any $x \in \Re^n$, we use x_i to indicate its ith *coordinate*, also called its ith *component*, and we also write $x = (x_1, \ldots, n)$.

Vectors in $\Re^n$ will be viewed as column vectors, unless the contrary is explicitly stated. For any $x \in \Re^n$, x' denotes the n-dimensional row vector that has the same components as x, arranged in the same order. The *inner product* of two vectors $x, y \in \Re^n$ is defined by $x'y = \sum_{i=1}^n x_i y_i$. Two vectors $x, y \in \Re^n$ satisfying $x'y = 0$ are called *orthogonal*.

If x is a vector in $\Re^n$, the notations $x > 0$ and $x \geq 0$ indicate that all components of x are positive and nonnegative, respectively. For any two vectors x and y, the notation $x > y$ means that $x - y > 0$. The notations $x \geq y$, $x < y$, etc., are to be interpreted accordingly.

If X is a set and λ is a scalar, we denote by λX the set $\{\lambda x \mid x \in X\}$. If X_1 and X_2 are two subsets of $\Re^n$, we denote by $X_1 + X_2$ the set

$$\{x_1 + x_2 \mid x_1 \in X_1,\, x_2 \in X_2\},$$

which is referred to as the *vector sum of* X_1 *and* X_2. We use a similar notation for the sum of any finite number of subsets. In the case where one of the subsets consists of a single vector $\bar{x}$, we simplify this notation as follows:

$$\bar{x} + X = \{\bar{x} + x \mid x \in X\}.$$

We also denote by $X_1 - X_2$ the set

$$\{x_1 - x_2 \mid x_1 \in X_1,\, x_2 \in X_2\}.$$

Given sets $X_i \subset \Re^{n_i}$, $i = 1, \ldots, m$, the *Cartesian product* of the X_i, denoted by $X_1 \times \cdots \times X_m$, is the set

$$\big\{(x_1, \ldots, x_m) \mid x_i \in X_i,\, i = 1, \ldots, m\big\},$$

which is a subset of $\Re^{n_1 + \cdots + n_m}$.

Subspaces and Linear Independence

A nonempty subset S of $\Re^n$ is called a *subspace* if $ax + by \in S$ for every $x, y \in S$ and every $a, b \in \Re$. An *affine set* or *linear manifold* in $\Re^n$ is a translated subspace, i.e., a set X of the form $X = \bar{x} + S = \{\bar{x} + x \mid x \in S\}$, where $\bar{x}$ is a vector in $\Re^n$ and S is a subspace of $\Re^n$, called the *subspace parallel to* X. Note that there can be only one subspace S associated with an affine set in this manner. [To see this, let $X = x + S$ and $X = \bar{x} + \bar{S}$ be two representations of the affine set X. Then, we must have $x = \bar{x} + \bar{s}$ for some $\bar{s} \in \bar{S}$ (since $x \in X$), so that $X = \bar{x} + \bar{s} + S$. Since we also have $X = \bar{x} + \bar{S}$, it follows that $S = \bar{S} - \bar{s} = \bar{S}$.] The *span* of a finite collection $\{x_1, \ldots, x_m\}$ of elements of $\Re^n$ is the subspace consisting of all vectors y of the form $y = \sum_{k=1}^m \alpha_k x_k$, where each α_k is a scalar.

The vectors $x_1, \ldots, x_m \in \Re^n$ are called *linearly independent* if there exists no set of scalars $\alpha_1, \ldots, \alpha_m$, at least one of which is nonzero, such that $\sum_{k=1}^m \alpha_k x_k = 0$. An equivalent definition is that $x_1 \neq 0$, and for every $k > 1$, the vector x_k does not belong to the span of $x_1, \ldots, x_{k-1}$.

If S is a subspace of $\Re^n$ containing at least one nonzero vector, a *basis* for S is a collection of vectors that are linearly independent and whose span is equal to S. Every basis of a given subspace has the same number of vectors. This number is called the *dimension* of S. By convention, the subspace $\{0\}$ is said to have dimension zero. The *dimension of an affine set*

$\bar{x} + S$ is the dimension of the corresponding subspace S. Every subspace of nonzero dimension has a basis that is orthogonal (i.e., any pair of distinct vectors from the basis is orthogonal).

Given any set X, the set of vectors that are orthogonal to all elements of X is a subspace denoted by $X^\perp$:

$$X^\perp = \{y \mid y'x = 0,\ \forall\ x \in X\}.$$

If S is a subspace, $S^\perp$ is called the *orthogonal complement* of S. Any vector x can be uniquely decomposed as the sum of a vector from S and a vector from $S^\perp$. Furthermore, we have $(S^\perp)^\perp = S$.

Matrices

For any matrix A, we use A_{ij}, $[A]_{ij}$, or a_{ij} to denote its ijth element. The *transpose* of A, denoted by A', is defined by $[A']_{ij} = a_{ji}$. For two matrices A and B of compatible dimensions, we have $(AB)' = B'A'$.

If X is a subset of $\Re^n$ and A is an $m \times n$ matrix, then the *image of* X *under* A is denoted by AX (or $A \cdot X$ if this enhances notational clarity):

$$AX = \{Ax \mid x \in X\}.$$

extrem If Y is a subset of $\Re^m$, the *inverse image of* Y *under* A is denoted by $A^{-1}Y$ or $A^{-1} \cdot Y$:

$$A^{-1}Y = \{x \mid Ax \in Y\}.$$

If X and Y are subspaces, then AX and $A^{-1}Y$ are also subspaces.

Let A be a square matrix. We say that A is *symmetric* if $A' = A$. We say that A is *diagonal* if $[A]_{ij} = 0$ when $i \neq j$. We say that A is *lower triangular* if $[A]_{ij} = 0$ when $i < j$, and *upper triangular* if $[A]_{ij} = 0$ when $i > j$. We denote by I the identity matrix (the diagonal matrix whose diagonal elements are 1). We denote the *determinant* of A by $\det(A)$.

Let A be an $m \times n$ matrix. The *range space* of A, denoted by $R(A)$, is the set of all $y \in \Re^m$ such that $y = Ax$ for some $x \in \Re^n$. The *nullspace* of A, denoted by $N(A)$, is the set of all $x \in \Re^n$ such that $Ax = 0$. It is seen that $R(A)$ and $N(A)$ are subspaces. The *rank* of A is the dimension of $R(A)$. The rank of A is equal to the maximal number of linearly independent columns of A, and is also equal to the maximal number of linearly independent rows of A. The matrix A and its transpose A' have the same rank. We say that A has *full rank*, if its rank is equal to $\min\{m, n\}$. This is true if and only if either all the rows of A are linearly independent, or all the columns of A are linearly independent.

The range space of an $m \times n$ matrix A is equal to the orthogonal complement of the nullspace of its transpose, i.e.,

$$R(A) = N(A')^\perp.$$

Another way to state this result is that given vectors $a_1, \ldots, a_n \in \Re^m$ (the columns of A) and a vector $x \in \Re^m$, we have $x'y = 0$ for all y such that $a_i'y = 0$ for all i if and only if

$$x = \lambda_1 a_1 + \cdots + \lambda_n a_n$$

for some scalars $\lambda_1, \ldots, \lambda_n$ [compare with Farkas' Lemma (Prop. B.15 in Appendix B)].

A function $f : \Re^n \mapsto \Re$ is said to be *affine* if it has the form $f(x) = a'x + b$ for some $a \in \Re^n$ and $b \in \Re$. Similarly, a function $f : \Re^n \mapsto \Re^m$ is said to be *affine* if it has the form $f(x) = Ax + b$ for some $m \times n$ matrix A and some $b \in \Re^m$. If $b = 0$, f is said to be a *linear function* or *linear transformation*. Sometimes, with slight abuse of terminology, an equation or inequality involving a linear function, such as $a'x = b$ or $a'x \leq b$, is referred to as a *linear equation or inequality*, respectively.

A.2 NORMS, SEQUENCES, LIMITS, AND CONTINUITY

Definition A.1: A *norm* $\|\cdot\|$ on $\Re^n$ is a mapping that assigns a scalar $\|x\|$ to every $x \in \Re^n$ and that has the following properties:

(a) $\|x\| \geq 0$ for all $x \in \Re^n$.

(b) $\|cx\| = |c| \cdot \|x\|$ for every $c \in \Re$ and every $x \in \Re^n$.

(c) $\|x\| = 0$ if and only if $x = 0$.

(d) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \Re^n$.

The *Euclidean norm* is defined by

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

The space $\Re^n$, equipped with this norm, is called a *Euclidean space*. We will use the Euclidean norm almost exclusively in this book. In particular, *in the absence of a clear indication to the contrary, $\|\cdot\|$ will denote the Euclidean norm*. Two important results for the Euclidean norm are:

Proposition A.1: (Pythagorean Theorem) If x and y are orthogonal then

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2.$$

Proposition A.2: (Schwarz inequality) For any two vectors x and y, we have

$$|x'y| \le \|x\| \cdot \|y\|,$$

with equality holding if and only if $x = \alpha y$ for some scalar α.

Two other important norms are the *maximum norm* $\|\cdot\|_\infty$ (also called *sup-norm* or ℓ_∞*-norm*), defined by

$$\|x\|_\infty = \max_i |x_i|,$$

and the ℓ_1*-norm* $\|\cdot\|_1$, defined by

$$\|x\|_1 = \sum_{i=1}^n |x_i|.$$

Sequences

We use both subscripts and superscripts in sequence notation. Generally, we prefer subscripts, but we use superscripts whenever we need to reserve the subscript notation for indexing components of vectors and functions. The meaning of the subscripts and superscripts should be clear from the context in which they are used.

A sequence $\{x_k \mid k = 1, 2, \ldots\}$ (or $\{x_k\}$ for short) of scalars is said to *converge* if there exists a scalar x such that for every $\epsilon > 0$ we have $|x_k - x| < \epsilon$ for every k greater than some integer K (that depends on ϵ). The scalar x is said to be the *limit* of $\{x_k\}$, and the sequence $\{x_k\}$ is said to *converge to* x; symbolically, $x_k \to x$ or $\lim_{k\to\infty} x_k = x$. If for every scalar b there exists some K (that depends on b) such that $x_k \ge b$ for all $k \ge K$, we write $x_k \to \infty$ and $\lim_{k\to\infty} x_k = \infty$. Similarly, if for every scalar b there exists some integer K such that $x_k \le b$ for all $k \ge K$, we write $x_k \to -\infty$ and $\lim_{k\to\infty} x_k = -\infty$. Note, however, that implicit in any of the statements "$\{x_k\}$ converges" or "the limit of $\{x_k\}$ exists" or "$\{x_k\}$ has a limit" is that the limit of $\{x_k\}$ is a scalar. A scalar sequence $\{x_k\}$ is called a *Cauchy sequence* if for every $\epsilon > 0$, there exists some integer K (depending on ϵ) such that $|x_k - x_m| < \epsilon$ for all $k \ge K$ and $m \ge K$.

A scalar sequence $\{x_k\}$ is said to be *bounded above* (respectively, *below*) if there exists some scalar b such that $x_k \le b$ (respectively, $x_k \ge b$) for all k. It is said to be *bounded* if it is bounded above and bounded below. The sequence $\{x_k\}$ is said to be monotonically *nonincreasing* (respectively, *nondecreasing*) if $x_{k+1} \le x_k$ (respectively, $x_{k+1} \ge x_k$) for all k. If $x_k \to x$ and $\{x_k\}$ is monotonically nonincreasing (nondecreasing), we also use the notation $x_k \downarrow x$ ($x_k \uparrow x$, respectively).

Proposition A.3: Every bounded and monotonically nonincreasing or nondecreasing scalar sequence converges.

Note that a monotonically nondecreasing sequence $\{x_k\}$ is either bounded, in which case it converges to some scalar x by the above proposition, or else it is unbounded, in which case $x_k \to \infty$. Similarly, a monotonically nonincreasing sequence $\{x_k\}$ is either bounded and converges, or it is unbounded, in which case $x_k \to -\infty$.

Given a scalar sequence $\{x_k\}$, let

$$y_m = \sup\{x_k \mid k \geq m\}, \qquad z_m = \inf\{x_k \mid k \geq m\}.$$

The sequences $\{y_m\}$ and $\{z_m\}$ are nonincreasing and nondecreasing, respectively, and therefore have a limit whenever $\{x_k\}$ is bounded above or is bounded below, respectively (Prop. A.3). The limit of y_m is denoted by $\limsup_{k\to\infty} x_k$, and is referred to as the *upper limit* of $\{x_k\}$. The limit of z_m is denoted by $\liminf_{k\to\infty} x_k$, and is referred to as the *lower limit* of $\{x_k\}$. If $\{x_k\}$ is unbounded above, we write $\limsup_{k\to\infty} x_k = \infty$, and if it is unbounded below, we write $\liminf_{k\to\infty} x_k = -\infty$.

Proposition A.4: Let $\{x_k\}$ and $\{y_k\}$ be scalar sequences.

(a) We have

$$\inf\{x_k \mid k \geq 0\} \leq \liminf_{k\to\infty} x_k \leq \limsup_{k\to\infty} x_k \leq \sup\{x_k \mid k \geq 0\}.$$

(b) $\{x_k\}$ converges if and only if

$$-\infty < \liminf_{k\to\infty} x_k = \limsup_{k\to\infty} x_k < \infty.$$

Furthermore, if $\{x_k\}$ converges, its limit is equal to the common scalar value of $\liminf_{k\to\infty} x_k$ and $\limsup_{k\to\infty} x_k$.

(c) If $x_k \leq y_k$ for all k, then

$$\liminf_{k\to\infty} x_k \leq \liminf_{k\to\infty} y_k, \qquad \limsup_{k\to\infty} x_k \leq \limsup_{k\to\infty} y_k.$$

(d) We have

$$\liminf_{k\to\infty} x_k + \liminf_{k\to\infty} y_k \leq \liminf_{k\to\infty}(x_k + y_k),$$

$$\limsup_{k\to\infty} x_k + \limsup_{k\to\infty} y_k \geq \limsup_{k\to\infty}(x_k + y_k).$$

A sequence $\{x_k\}$ of vectors in $\Re^n$ is said to converge to some $x \in \Re^n$ if the ith component of x_k converges to the ith component of x for every i. We use the notations $x_k \to x$ and $\lim_{k\to\infty} x_k = x$ to indicate convergence for vector sequences as well. The sequence $\{x_k\}$ is called bounded (or Cauchy) if each of its corresponding coordinate sequences is bounded (or Cauchy, respectively). It can be seen that $\{x_k\}$ is bounded if and only if there exists a scalar c such that $\|x_k\| \leq c$ for all k. An infinite subset of a sequence $\{x_k\}$ is called a *subsequence* of $\{x_k\}$. A subsequence can itself be viewed as a sequence, and can be represented as a set $\{x_k \mid k \in \mathcal{K}\}$, where $\mathcal{K}$ is an infinite subset of positive integers (the notation $\{x_k\}_\mathcal{K}$ will also be used).

Definition A.2: We say that a vector $x \in \Re^n$ is a *limit point of a sequence* $\{x_k\}$ in $\Re^n$ if there exists a subsequence of $\{x_k\}$ that converges to x.

Proposition A.5:

(a) A bounded sequence of vectors in $\Re^n$ converges if and only if it has a unique limit point.

(b) A sequence in $\Re^n$ converges if and only if it is a Cauchy sequence.

(c) Every bounded sequence in $\Re^n$ has at least one limit point.

(d) Let $\{x_k\}$ be a scalar sequence. If $\limsup_{k\to\infty} x_k$ ($\liminf_{k\to\infty} x_k$) is finite, then it is the largest (respectively, smallest) limit point of $\{x_k\}$.

$o(\cdot)$ Notation

For a positive integer p and a function $h : \Re^n \mapsto \Re^m$ we write

$$h(x) = o\big(\|x\|^p\big)$$

if

$$\lim_{k\to\infty} \frac{h(x_k)}{\|x_k\|^p} = 0,$$

for all sequences $\{x_k\}$ such that $x_k \to 0$ and $x_k \neq 0$ for all k.

Closed and Open Sets

We say that x is a *closure point* or *limit point* of a subset X of $\Re^n$ if there exists a sequence $\{x_k\} \subset X$ that converges to x. The *closure* of X, denoted $\mathrm{cl}(X)$, is the set of all closure points of X.

Definition A.3: A subset X of $\Re^n$ is called *closed* if it is equal to its closure. It is called *open* if its complement, $\{x \mid x \notin X\}$, is closed. It is called *bounded* if there exists a scalar c such that $\|x\| \leq c$ for all $x \in X$. It is called *compact* if it is closed and bounded. A *neighborhood* of a vector x is an open set containing x. If $X \subset \Re^n$ and $x \in X$, we say that x is an *interior* point of X if there exists a neighborhood of x that is contained in X. A vector $x \in X$ which is not an interior point of X is said to be a *boundary* point of X. The set of all boundary points of X is called the *boundary* of X.

For any norm $\|\cdot\|$ in $\Re^n$, $\epsilon > 0$, and $x^* \in \Re^n$, consider the sets

$$\big\{x \mid \|x - x^*\| < \epsilon\big\}, \qquad \big\{x \mid \|x - x^*\| \leq \epsilon\big\}.$$

The first set is open and is called an *open sphere* centered at x^*, while the second set is closed and is called a *closed sphere* centered at x^*. Sometimes the terms *open ball* and *closed ball* are used, respectively.

Proposition A.6:

(a) The union of finitely many closed sets is closed.

(b) The intersection of closed sets is closed.

(c) The union of open sets is open.

(d) The intersection of finitely many open sets is open.

(e) A set is open if and only if all of its elements are interior points.

(f) Every subspace of $\Re^n$ is closed.

(g) A subset of $\Re^n$ is compact if and only if it is closed and bounded.

Continuity

Let $f : X \mapsto \Re^m$ be a function, where X is a subset of $\Re^n$, and let x be a vector in X. If there exists a vector $y \in \Re^m$ such that the sequence $\big\{f(x_k)\big\}$ converges to y for every sequence $\{x_k\} \subset X$ such that $\lim_{k\to\infty} x_k = x$, we write $\lim_{z\to x} f(z) = y$. If there exists a vector $y \in \Re^m$ such that the sequence $\{f(x_k)\}$ converges to y for every sequence $\{x_k\} \subset X$ such that $\lim_{k\to\infty} x_k = x$ and $x_k \leq x$ (respectively, $x_k \geq x$) for all k, we write $\lim_{z\uparrow x} f(z) = y$ [respectively, $\lim_{z\downarrow x} f(z)$].

Definition A.4: Let X be a subset of $\Re^n$.

(a) A function $f : X \mapsto \Re^m$ is called *continuous* at a vector $x \in X$ if $\lim_{z\to x} f(z) = f(x)$.

(b) A function $f : X \mapsto \Re^m$ is called *right-continuous* (respectively, *left-continuous*) at a vector $x \in X$ if $\lim_{z\downarrow x} f(z) = f(x)$ [respectively, $\lim_{z\uparrow x} f(z) = f(x)$].

(c) A real-valued function $f : X \mapsto \Re$ is called *upper semicontinuous* (respectively, *lower semicontinuous*) at a vector $x \in X$ if $f(x) \geq \limsup_{k\to\infty} f(x_k)$ [respectively, $f(x) \leq \liminf_{k\to\infty} f(x_k)$] for every sequence $\{x_k\} \subset X$ that converges to x.

(d) A function $f : X \mapsto \Re$ is called *coercive* if for every sequence $\{x_k\} \subset X$ such that $\|x_k\| \to \infty$, we have $\lim_{k\to\infty} f(x_k) = \infty$.

If $f : X \mapsto \Re^m$ is continuous at every vector in a subset of its domain X, we say that f *is continuous over that subset*. If $f : X \mapsto \Re^m$ is continuous at every vector in its domain X, we say that f *is continuous*. We say that f is *Lipschitz continuous* if $\|f(x) - f(y)\| \leq L\|x - y\|$ for some scalar L and all $x, y \in X$. We also say that $f : X \mapsto \Re$ is *coercive over a subset* of its domain X if for every sequence $\{x_k\}$ from that subset such that $\|x_k\| \to \infty$, we have $\lim_{k\to\infty} f(x_k) = \infty$. If f is coercive over X, we simply say that f is coercive.

Proposition A.7:

(a) Any vector norm on $\Re^n$ is a continuous function.

(b) Let $f : \Re^m \mapsto \Re^p$ and $g : \Re^n \mapsto \Re^m$ be continuous functions. The composition $f \cdot g : \Re^n \mapsto \Re^p$, defined by $(f \cdot g)(x) = f\big(g(x)\big)$, is a continuous function.

(c) Let $f : \Re^n \mapsto \Re^m$ be continuous, and let Y be an open (respectively, closed) subset of $\Re^m$. Then the inverse image of Y, $\big\{x \in \Re^n \mid f(x) \in Y\big\}$, is open (respectively, closed).

(d) Let $f : \Re^n \mapsto \Re^m$ be continuous, and let X be a compact subset of $\Re^n$. Then the image of X, $\big\{f(x) \mid x \in X\big\}$, is compact.

(e) Let X be a closed subset of $\Re^n$ and let $f : X \mapsto \Re$ be lower semicontinuous at all points of X. Then the level set $\big\{x \in X \mid f(x) \leq \gamma\big\}$ is closed for all $\gamma \in \Re$.

If X is a nonempty subset of $\Re^n$ and f is a real-valued function whose domain contains X, we say that a vector $x^* \in X$ is a *minimum of f over X* if $f(x^*) = \inf_{x \in X} f(x)$. We also call x^* a *minimizing point* or a *minimizer* or a *minimum* of f over X. Alternatively, we say that f *attains a minimum over X at x^**, and we indicate this by writing

$$x^* \in \arg\min_{x \in X} f(x).$$

If x^* is known to be the unique minimizer of f over X, by slight abuse of notation, we also write

$$x^* = \arg\min_{x \in X} f(x).$$

We use similar notation for maxima. An important property of compactness in connection with optimization problems is the following theorem, which provides conditions for existence of solutions of optimization problems.

Proposition A.8: (Weierstrass' Theorem) Let X be a nonempty subset of $\Re^n$ and let $f : X \mapsto \Re$ be lower semicontinuous at all points of X. Assume that one of the following three conditions holds:

(1) X is compact.

(2) X is closed and f is coercive.

(3) There exists a scalar γ such that the level set

$$\big\{x \in X \mid f(x) \leq \gamma\big\}$$

is nonempty and compact.

Then, the set of minima of f over X is nonempty and compact.

Proof: Assume condition (1). Let $\{z_k\} \subset X$ be a sequence such that

$$\lim_{k \to \infty} f(z_k) = \inf_{z \in X} f(z).$$

Since X is bounded, this sequence has at least one limit point x [Prop. A.5(c)]. Since X is closed, x belongs to X, while the lower semicontinuity of f implies that $f(x) \leq \lim_{k \to \infty} f(z_k) = \inf_{z \in X} f(z)$. Therefore, we must have $f(x) = \inf_{z\in X} f(z)$. The set of all minima of f over X is the level set $\big\{x \in X \mid f(x) \le \inf_{z\in X} f(z)\big\}$, which is closed by the lower semicontinuity of f [Prop. A.7(e)], and hence compact since X is bounded.

Assume condition (2). Consider a sequence $\{z_k\}$ as in the proof of part (a). Since f is coercive, $\{z_k\}$ must be bounded and the proof proceeds like the proof of part (a).

Assume condition (3). If the given γ is equal to $\inf_{z\in X} f(z)$, the set of minima of f over X is $\big\{x \in X \mid f(x) \le \gamma\big\}$, and since by assumption this set is nonempty, we are done. If $\gamma > \inf_{z\in X} f(z)$, consider a sequence $\{z_k\}$ as in the proof of part (a). Then, for all k sufficiently large, z_k must belong to the set $\big\{x \in X \mid f(x) \le \gamma\big\}$. Since this set is compact, $\{z_k\}$ must be bounded and the proof proceeds like the proof of part (a). **Q.E.D.**

Note that with appropriate adjustments, the above proposition applies to the existence of maxima of f over X. In particular, if f is upper semicontinuous at all points of X and X is compact, there exists a vector $y \in X$ such that $f(y) = \sup_{z\in X} f(z)$. Note also that under additional convexity assumptions on X and f, there is a more refined theory of existence of optimal solutions, whereby the boundedness assumptions underlying Weierstrass' Theorem are replaced by alternative conditions involving directions of recession (see [BNO03], Section 2.3, [Ber09], Section 3.2).

With an application of Weierstrass' Theorem, we obtain the following *norm equivalence property in* $\Re^n$, which shows that if a sequence converges with respect to one norm, it converges with respect to all other norms.

Proposition A.9: For any two norms $\|\cdot\|$ and $\|\cdot\|'$ on $\Re^n$, there exists some positive constant $c \in \Re$ such that $\|x\| \le c\|x\|'$ for all $x \in \Re^n$.

Proof: Let a be the minimum of $\|x\|'$ over the set of all $x \in \Re^n$ such that $\|x\| = 1$. The latter set is closed and bounded and, therefore, the minimum is attained at some $\tilde{x}$ (Prop. A.8) that must be nonzero since $\|\tilde{x}\| = 1$. For any $x \in \Re^n$, $x \ne 0$, the $\|\cdot\|$ norm of $x/\|x\|$ is equal to 1. Therefore,

$$0 < a = \|\tilde{x}\|' \le \left\|\frac{x}{\|x\|}\right\|' = \frac{\|x\|'}{\|x\|}, \qquad \forall\, x \ne 0,$$

which proves the desired result with $c = 1/a$. **Q.E.D.**

As a corollary, we obtain the following.

Proposition A.10: If a subset of $\Re^n$ is open (respectively, closed, bounded, or compact) for some norm, it is open (respectively, closed, bounded, or compact), for all other norms.

Matrix Norms

A norm $\|\cdot\|$ on the set of $n \times n$ matrices is a real-valued mapping that has the same properties as vector norms do when the matrix is viewed as an element of $\Re^{n^2}$. The norm of an $n \times n$ matrix A is denoted by $\|A\|$.

We are mainly interested in *induced norms*, which are constructed as follows. Given any vector norm $\|\cdot\|$, the corresponding induced matrix norm, also denoted by $\|\cdot\|$, is defined by

$$\|A\| = \max_{\{x\in\Re^n \mid \|x\|=1\}} \|Ax\|. \tag{A.1}$$

The set over which the maximization takes place above is closed [Prop. A.7(c)] and bounded, while the function being maximized is continuous [Prop. A.7(b)]. Therefore, by Weiestrass' theorem (Prop. A.8) the maximum is attained. It is easily verified that for any vector norm, Eq. (A.1) defines a matrix norm having all the required properties.

Note that by the Schwarz inequality (Prop. A.2), we have

$$\|A\| = \max_{\|x\|=1} \|Ax\| = \max_{\|y\|=\|x\|=1} |y'Ax|.$$

By reversing the roles of x and y in the above relation and by using the equality $y'Ax = x'A'y$, it follows that

$$\|A\| = \|A'\|. \tag{A.2}$$

A.3 SQUARE MATRICES AND EIGENVALUES

Definition A.5: A square matrix A is called *singular* if its determinant is zero. Otherwise it is called *nonsingular* or *invertible*.

Proposition A.11:

(a) Let A be an $n \times n$ matrix. The following are equivalent:

(i) The matrix A is nonsingular.

(ii) The matrix A' is nonsingular.

(iii) For every nonzero $x \in \Re^n$, we have $Ax \neq 0$.

(iv) For every $y \in \Re^n$, there exists a unique $x \in \Re^n$ such that $Ax = y$.

(v) There exists an $n \times n$ matrix B such that $AB = I = BA$.

(vi) The columns of A are linearly independent.

(vii) The rows of A are linearly independent.

(b) Assuming that A is nonsingular, there is a unique matrix B satisfying $AB = I = BA$, which is called the *inverse* of A and is denoted by A^{-1}.

(c) For any two square invertible matrices A and B of the same dimensions, we have $(AB)^{-1} = B^{-1}A^{-1}$.

Let A and B be square matrices, and let C be a matrix of appropriate dimension. Then we have

$$(A + CBC')^{-1} = A^{-1} - A^{-1}C(B^{-1} + C'A^{-1}C)^{-1}C'A^{-1},$$

provided all the inverses appearing above exist. For a proof, multiply the right-hand side by $A + CBC'$ and show that the product is the identity.

Another useful formula provides the inverse of the partitioned matrix

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}.$$

There holds

$$M^{-1} = \begin{bmatrix} Q & -QBD^{-1} \\ -D^{-1}CQ & D^{-1} + D^{-1}CQBD^{-1} \end{bmatrix},$$

where

$$Q = (A - BD^{-1}C)^{-1},$$

provided all the inverses appearing above exist. For a proof, multiply M with the given expression for M^{-1} and verify that the product is the identity.

Definition A.6: The *characteristic polynomial* ϕ of an $n \times n$ matrix A is defined by $\phi(\lambda) = \det(\lambda I - A)$, where I is the identity matrix of the same size as A. The n (possibly repeated and complex) roots of ϕ are called the *eigenvalues* of A. A vector x (with possibly complex coordinates) such that $Ax = \lambda x$, where λ is an eigenvalue of A, is called an *eigenvector* of A associated with λ.

Proposition A.12: Let A be a square matrix.

(a) A complex number λ is an eigenvalue of A if and only if there exists a nonzero eigenvector associated with λ.

(b) A is singular if and only if it has an eigenvalue that is equal to zero.

Note that the only use of complex numbers in this book is in relation to eigenvalues and eigenvectors. All other matrices or vectors are implicitly assumed to have real components.

Proposition A.13: Let A be an $n \times n$ matrix.

(a) The eigenvalues of a triangular matrix are equal to its diagonal entries.

(b) If S is a nonsingular matrix and $B = SAS^{-1}$, then the eigenvalues of A and B coincide.

(c) The eigenvalues of $cI + A$ are equal to $c + \lambda_1, \ldots, c + \lambda_n$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A.

(d) The eigenvalues of A^k are equal to $\lambda_1^k, \ldots, \lambda_n^k$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A.

(e) If A is nonsingular, then the eigenvalues of A^{-1} are the reciprocals of the eigenvalues of A.

(f) The eigenvalues of A and A' coincide.

Definition A.7: The *spectral radius* $\rho(A)$ of a square matrix A is defined as the maximum of the magnitudes of the eigenvalues of A.

It can be shown that the roots of a polynomial depend continuously on the coefficients of the polynomial. For this reason, the eigenvalues of a square matrix A depend continuously on A, and we obtain the following.

Proposition A.14: The eigenvalues of a square matrix A depend continuously on the elements of A. In particular, $\rho(A)$ is a continuous function of A.

The next two propositions are fundamental for the convergence theory of linear iterative methods.

Proposition A.15: For any induced matrix norm $\|\cdot\|$ and any square matrix A we have

$$\lim_{k\to\infty} \|A^k\|^{1/k} = \rho(A) \leq \|A\|.$$

Furthermore, given any $\epsilon > 0$, there exists an induced matrix norm $\|\cdot\|$ such that

$$\|A\| = \rho(A) + \epsilon.$$

Proposition A.16: Let A be a square matrix. We have

$$\lim_{k\to\infty} A^k = 0$$

if and only if $\rho(A) < 1$.

A corollary of the above proposition is that the iteration $x_{k+1} = Ax_k$ converges to 0 for every initial condition x_0 if and only if $\rho(A) < 1$. From this it also follows that if $\rho(A) < 1$, the iteration $x_{k+1} = Ax_k + b$ converges to the vector $x^* = (I - A)^{-1}b$ for every initial condition x_0 and every vector b. To see this, note that the iteration $x_{k+1} = Ax_k + b$ can equivalently be written as $y_{k+1} = Ay_k$, where $y_k = x_k - x^*$.

A.4 SYMMETRIC AND POSITIVE DEFINITE MATRICES

Symmetric matrices have several special properties, particularly with respect to their eigenvalues and eigenvectors. In this section, $\|\cdot\|$ denotes the Euclidean norm throughout.

Proposition A.17: Let A be a symmetric $n \times n$ matrix. Then:

(a) The eigenvalues of A are real.

(b) The matrix A has a set of n mutually orthogonal, real, and nonzero eigenvectors $x_1, \ldots, x_n$.

(c) Suppose that the eigenvectors in part (b) have been normalized so that $\|x_i\| = 1$ for each i. Then

$$A = \sum_{i=1}^{n} \lambda_i x_i x_i',$$

where λ_i is the eigenvalue corresponding to x_i.

Proposition A.18: Let A be a symmetric $n \times n$ matrix, let $\lambda_1 \leq \cdots \leq \lambda_n$ be its (real) eigenvalues, and let $x_1, \ldots, x_n$ be associated orthogonal eigenvectors, normalized so that $\|x_i\| = 1$ for all i. Then:

(a) $\|A\| = \rho(A) = \max\big\{|\lambda_1|, |\lambda_n|\big\}$, where $\|\cdot\|$ is the matrix norm induced by the Euclidean norm.

(b) $\lambda_1\|y\|^2 \leq y'Ay \leq \lambda_n\|y\|^2$ for all $y \in \Re^n$.

(c) (*Courant-Fisher Minimax Principle*) For all $i = 1, \ldots, n$, and for all i-dimensional subspaces $\overline{S}_i$ and all $(n-i+1)$-dimensional subspaces $\underline{S}_i$, there holds

$$\min_{\|y\|=1,\, y \in \underline{S}_i} y'Ay \leq \lambda_i \leq \max_{\|y\|=1,\, y \in \overline{S}_i} y'Ay.$$

Furthermore, equality on the left (right) side above is attained if $\underline{S}_i$ is the subspace spanned by $x_i, \ldots, x_n$ ($\overline{S}_i$ is the subspace spanned by $x_1, \ldots, x_i$, respectively).

(d) (*Interlocking Eigenvalues Lemma*) Let $\tilde{\lambda}_1 \leq \tilde{\lambda}_2 \leq \cdots \leq \tilde{\lambda}_n$ be the eigenvalues of $A + bb'$, where b is a vector in $\Re^n$. Then,

$$\lambda_1 \leq \tilde{\lambda}_1 \leq \lambda_2 \leq \tilde{\lambda}_2 \leq \cdots \leq \lambda_n \leq \tilde{\lambda}_n.$$

Proof: (a) We know that $\|A\| \geq \rho(A)$ (Prop. A.15), so we need to show the reverse inequality. We express an arbitrary vector $y \in \Re^n$ in the form $y = \sum_{i=1}^n \xi_i x_i$, where each ξ_i is a suitable scalar. Using the orthogonality of the vectors x_i and the Pythagorean Theorem (Prop. A.1), we obtain $\|y\|^2 = \sum_{i=1}^n |\xi_i|^2 \cdot \|x_i\|^2$. Using the Pythagorean Theorem again, we obtain

$$\|Ay\|^2 = \left\|\sum_{i=1}^n \lambda_i \xi_i x_i\right\|^2 = \sum_{i=1}^n |\lambda_i|^2 \cdot |\xi_i|^2 \cdot \|x_i\|^2 \leq \rho^2(A)\|y\|^2.$$

Since this is true for every y, we obtain $\|A\| \leq \rho(A)$ and the desired result follows.

(b) As in part (a), we express the generic $y \in \Re^n$ as $y = \sum_{i=1}^n \xi_i x_i$. We have, using the orthogonality of the vectors x_i, $i = 1, \ldots, n$, and the fact $\|x_i\| = 1$,

$$y'Ay = \sum_{i=1}^n \lambda_i |\xi_i|^2 \|x_i\|^2 = \sum_{i=1}^n \lambda_i |\xi_i|^2$$

and

$$\|y\|^2 = \sum_{i=1}^n |\xi_i|^2 \|x_i\|^2 = \sum_{i=1}^n |\xi_i|^2.$$

These two relations prove the desired result.

(c) Let $\underline{X}_i$ be the subspace spanned by $x_1, \ldots, x_i$. The subspaces $\underline{X}_i$ and $\underline{S}_i$ must have a common vector x_0 with $\|x_0\| = 1$, since the sum of their dimensions is $n+1$ [if there was no common nonzero vector, we could take sets of basis vectors for $\underline{X}_i$ and S_i (a total of $n+1$ in number), which would have to be linearly independent, yielding a contradiction]. The vector x_0 can be expressed as a linear combination $x_0 = \sum_{j=1}^{i} \xi_j x_j$, and since $\|x_0\| = 1$ and $\|x_i\| = 1$ for all $i = 1, \ldots, n$, we must have

$$\sum_{j=1}^{i} \xi_j^2 = 1.$$

We also have using the expression

$$A = \sum_{j=1}^{n} \lambda_j x_j x_j'$$

[cf. Prop. A.17(c)],

$$x_0' A x_0 = \sum_{j=1}^{i} \lambda_j \xi_j^2 \le \lambda_i \left(\sum_{j=1}^{i} \xi_j^2 \right).$$

Combining the last two relations, we obtain $x_0' A x_0 \le \lambda_i$, which proves the left-hand side of the desired inequality. The right-hand side is proved similarly. Furthermore, we have $x_i' A x_i = \lambda_i$, so equality is attained as in the final assertion.

(d) From part (c) we have

$$\lambda_i = \max_{\underline{S}_i} \min_{\|y\|=1,\, y \in \underline{S}_i} y' A y \le \max_{\underline{S}_i} \min_{\|y\|=1,\, y \in \underline{S}_i} y'(A + bb')y \le \tilde{\lambda}_i,$$

so that $\lambda_i \le \tilde{\lambda}_i$ for all i. Furthermore, from part (c), for some $(n-i+1)$-dimensional subspace $\underline{\tilde{S}}_i$ we have

$$\tilde{\lambda}_i = \min_{\|y\|=1,\, y \in \underline{\tilde{S}}_i} y'(A + bb')y.$$

Using this relation and the left-hand side of the inequality of part (c), applied to the subspace $\{y \mid y \in \underline{\tilde{S}}_i,\, b'y = 0\}$, whose dimension is at least $(n-i)$, we obtain

$$\tilde{\lambda}_i \le \min_{\|y\|=1,\, y \in \underline{\tilde{S}}_i,\, b'y=0} y'(A + bb')y = \min_{\|y\|=1,\, y \in \underline{\tilde{S}}_i,\, b'y=0} y' A y \le \lambda_{i+1},$$

and the proof is complete. **Q.E.D.**

Proposition A.19: Let A be a square matrix, and let $\|\cdot\|$ be the matrix norm induced by the Euclidean norm. Then:

(a) If A is symmetric, then $\|A^k\| = \|A\|^k$ for any positive integer k.

(b) $\|A\|^2 = \|A'A\| = \|AA'\|$.

(c) If A is symmetric and nonsingular, then $\|A^{-1}\|$ is equal to the reciprocal of the smallest of the absolute values of the eigenvalues of A.

Proof: (a) If A is symmetric then A^k is symmetric. Using Prop. A.18(a), we have $\|A^k\| = \rho(A^k)$. Using Prop. A.13(d), we obtain $\rho(A^k) = \rho(A)^k$, which is equal to $\|A\|^k$ by Prop. A.18(a).

(b) For any vector x such that $\|x\| = 1$, we have, using the Schwarz inequality (Prop. A.2),

$$\|Ax\|^2 = x'A'Ax \leq \|x\| \cdot \|A'Ax\| \leq \|x\| \cdot \|A'A\| \cdot \|x\| = \|A'A\|.$$

Thus, $\|A\|^2 \leq \|A'A\|$. On the other hand,

$$\|A'A\| = \max_{\|y\|=\|x\|=1} |y'A'Ax| \leq \max_{\|y\|=\|x\|=1} \|Ay\| \cdot \|Ax\| = \|A\|^2.$$

Therefore, $\|A\|^2 = \|A'A\|$. The equality $\|A\|^2 = \|AA'\|$ is obtained by replacing A by A' and using Eq. (A.2).

(c) This follows by combining Prop. A.13(e) with Prop. A.18(a). **Q.E.D.**

Definition A.8: A symmetric $n\times n$ matrix A is called *positive definite* if $x'Ax > 0$ for all $x \in \Re^n$, $x \neq 0$. It is called *nonnegative definite* or *positive semidefinite* if $x'Ax \geq 0$ for all $x \in \Re^n$.

Throughout this book, the notion of positive and negative definiteness applies exclusively to symmetric matrices. Thus *whenever we say that a matrix is positive or negative (semi)definite, we implicitly assume that the matrix is symmetric*.

Proposition A.20:

(a) For any $m \times n$ matrix A, the matrix $A'A$ is symmetric and nonnegative definite. The matrix $A'A$ is positive definite if and only if A has rank n. In particular, if $m = n$, $A'A$ is positive definite if and only if A is nonsingular.

(b) A square symmetric matrix is nonnegative definite (respectively, positive definite) if and only if all of its eigenvalues are nonnegative (respectively, positive).

(c) The inverse of a symmetric positive definite matrix is symmetric and positive definite.

Proof: (a) Symmetry is obvious. For any vector $x \in \Re^n$, we have $x'A'Ax = \|Ax\|^2 \geq 0$, which establishes nonnegative definiteness. Positive definiteness is obtained if and only if the inequality is strict for every $x \neq 0$, which is the case if and only if $Ax \neq 0$ for every $x \neq 0$. This is equivalent to A having rank n.

(b) Let λ and x be an eigenvalue and a corresponding real nonzero eigenvector of a symmetric nonnegative definite matrix A. Then $0 \leq x'Ax = \lambda x'x = \lambda\|x\|^2$, which proves that $\lambda \geq 0$. For the converse result, let y be an arbitrary vector in $\Re^n$. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A, assumed to be nonnegative, and let $x_1, \ldots, x_n$ be a corresponding set of nonzero, real, and orthogonal eigenvectors. Let us express y in the form $y = \sum_{i=1}^n \xi_i x_i$. Then $y'Ay = (\sum_{i=1}^n \xi_i x_i)'(\sum_{i=1}^n \xi_i \lambda_i x_i)$. From the orthogonality of the eigenvectors, the latter expression is equal to $\sum_{i=1}^n \xi_i^2 \lambda_i \|x_i\|^2 \geq 0$, which proves that A is nonnegative definite. The proof for the case of positive definite matrices is similar.

(c) The eigenvalues of A^{-1} are the reciprocal of the eigenvalues of A [Prop. A.13(e)], so the result follows using part (b). **Q.E.D.**

Proposition A.21: Let A be a square symmetric nonnegative definite matrix.

(a) There exists a symmetric matrix Q with the property $Q^2 = A$. Such a matrix is called a *symmetric square root* of A and is denoted by $A^{1/2}$.

(b) A symmetric square root $A^{1/2}$ is invertible if and only if A is invertible. Its inverse is denoted by $A^{-1/2}$.

(c) There holds $A^{-1/2}A^{-1/2} = A^{-1}$.

(d) There holds $AA^{1/2} = A^{1/2}A$.

Proof: (a) Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A and let $x_1, \ldots, x_n$ be corresponding nonzero, real, and orthogonal eigenvectors normalized so that $\|x_k\| = 1$ for each k. We let

$$A^{1/2} = \sum_{k=1}^{n} \lambda_k^{1/2} x_k x_k',$$

where $\lambda_k^{1/2}$ is the nonnegative square root of λ_k. We then have

$$A^{1/2}A^{1/2} = \sum_{i=1}^{n}\sum_{k=1}^{n} \lambda_i^{1/2}\lambda_k^{1/2} x_i x_i' x_k x_k' = \sum_{k=1}^{n} \lambda_k x_k x_k' = A.$$

Here the second equality follows from the orthogonality of distinct eigenvectors; the last equality follows from Prop. A.17(c). We now notice that each one of the matrices $x_k x_k'$ is symmetric, so $A^{1/2}$ is also symmetric.

(b) This follows from the fact that the eigenvalues of A are the squares of the eigenvalues of $A^{1/2}$ [Prop. A.13(d)].

(c) We have $(A^{-1/2}A^{-1/2})A = A^{-1/2}(A^{-1/2}A^{1/2})A^{1/2} = A^{-1/2}IA^{1/2} = I$.

(d) We have $AA^{1/2} = A^{1/2}A^{1/2}A^{1/2} = A^{1/2}A$. **Q.E.D.**

A symmetric square root of A is not unique. For example, let $A^{1/2}$ be as in the proof of Prop. A.21(a) and notice that the matrix $-A^{1/2}$ also has the property $(-A^{1/2})(-A^{1/2}) = A$. However, if A is positive definite, it can be shown that the matrix $A^{1/2}$ we have constructed is the only symmetric and positive definite square root of A.

A.5 DERIVATIVES

Let $f : \Re^n \mapsto \Re$ be some function, fix some $x \in \Re^n$, and consider the expression

$$\lim_{\alpha \to 0} \frac{f(x + \alpha e_i) - f(x)}{\alpha},$$

where e_i is the ith unit vector (all components are 0 except for the ith component which is 1). If the above limit exists, it is called the ith *partial derivative* of f at the vector x and it is denoted by $(\partial f/\partial x_i)(x)$ or $\partial f(x)/\partial x_i$ (x_i in this section will denote the ith component of the vector

x). Assuming all of these partial derivatives exist, the *gradient* of f at x is defined as the column vector

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}.$$

For any $y \in \Re^n$, we define the one-sided *directional derivative* of f in the direction y to be

$$f'(x; y) = \lim_{\alpha \downarrow 0} \frac{f(x + \alpha y) - f(x)}{\alpha},$$

provided that the limit exists.

If the directional derivative of f at a vector x exists in all directions y and $f'(x; y)$ is a linear function of y, we say that f is *differentiable* at x. This type of differentiability is also called *Gateaux differentiability*. It is seen that f is differentiable at x if and only if the gradient $\nabla f(x)$ exists and satisfies

$$\nabla f(x)' y = f'(x; y), \qquad \forall\ y \in \Re^n.$$

The function f is called *differentiable over a subset U of $\Re^n$* if it is differentiable at every $x \in U$. The function f is called *differentiable* (without qualification) if it is differentiable at all $x \in \Re^n$.

If f is differentiable over an open set U and $\nabla f(\cdot)$ is continuous at all $x \in U$, f is said to be *continuously differentiable over U*. It can then be shown that

$$\lim_{y \to 0} \frac{f(x + y) - f(x) - \nabla f(x)' y}{\|y\|} = 0, \qquad \forall\ x \in U, \tag{A.3}$$

where $\|\cdot\|$ is an arbitrary vector norm. If f is continuously differentiable over $\Re^n$, then f is also called a *smooth* function. If f is not smooth, it is called *nonsmooth*.

The preceding equation can also be used as an alternative definition of differentiability. In particular, f is called *Frechet differentiable* at x if there exists a vector g satisfying Eq. (A.3) with $\nabla f(x)$ replaced by g. If such a vector g exists, it can be seen that all the partial derivatives $(\partial f / \partial x_i)(x)$ exist and that $g = \nabla f(x)$. Frechet differentiability implies (Gateaux) differentiability but not conversely (see for example Ortega and Rheinboldt [OrR70] for a detailed discussion). In this book, when dealing with a differentiable function f, we will always assume that f is continuously differentiable over some open set [$\nabla f(\cdot)$ is a continuous function over that set], in which case f is both Gateaux and Frechet differentiable, and the distinctions made above are of no consequence.

The definitions of differentiability of f at a vector x only involve the values of f in a neighborhood of x. Thus, these definitions can be used for functions f that are not defined on all of $\Re^n$, but are defined instead in a neighborhood of the vector at which the derivative is computed. In particular, for functions $f : X \mapsto \Re$, where X is a strict subset of $\Re^n$, we use the above definition of differentiability of f at a vector x, *provided x is an interior point of the domain X*. Similarly, we use the above definition of continuous differentiability of f over a subset U, *provided U is an open subset of the domain X*. Thus any mention of continuous differentiability of a function over a subset implicitly assumes that this subset is open.

Differentiation of Vector-Valued Functions

A function $f : \Re^n \mapsto \Re^m$, with component functions $f_1, \ldots, f_m$, is called differentiable (or smooth) if each component is differentiable (or smooth, respectively). The *gradient matrix* of f, denoted $\nabla f(x)$, is the $n \times m$ matrix whose ith column is the gradient $\nabla f_i(x)$ of f_i:

$$\nabla f(x) = \Big[\nabla f_1(x) \cdots \nabla f_m(x) \Big].$$

The transpose of ∇f is called the *Jacobian* of f and is the matrix whose ijth entry is equal to the partial derivative $\partial f_i / \partial x_j$.

Now suppose that each one of the partial derivatives of a function $f : \Re^n \mapsto \Re$ is a smooth function of x. We use the notation $(\partial^2 f / \partial x_i \partial x_j)(x)$ to indicate the ith partial derivative of $\partial f / \partial x_j$ at a vector $x \in \Re^n$. The *Hessian* of f is the matrix whose ijth entry is equal to $(\partial^2 f / \partial x_i \partial x_j)(x)$, and is denoted by $\nabla^2 f(x)$. We have $(\partial^2 f / \partial x_i \partial x_j)(x) = (\partial^2 f / \partial x_j \partial x_i)(x)$ for every x, which implies that $\nabla^2 f(x)$ is symmetric.

If $f : \Re^{m+n} \mapsto \Re$ is a function of (x, y), where $x \in \Re^m$ and $y \in \Re^n$, and $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ denote the components of x and y, respectively, we write

$$\nabla_x f(x, y) = \begin{pmatrix} \frac{\partial f(x,y)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x,y)}{\partial x_m} \end{pmatrix}, \qquad \nabla_y f(x, y) = \begin{pmatrix} \frac{\partial f(x,y)}{\partial y_1} \\ \vdots \\ \frac{\partial f(x,y)}{\partial y_n} \end{pmatrix}.$$

We denote by $\nabla^2_{xx} f(x, y)$, $\nabla^2_{xy} f(x, y)$, and $\nabla^2_{yy} f(x, y)$ the matrices with components

$$\big[\nabla^2_{xx} f(x, y)\big]_{ij} = \frac{\partial^2 f(x, y)}{\partial x_i \partial x_j}, \qquad \big[\nabla^2_{xy} f(x, y)\big]_{ij} = \frac{\partial^2 f(x, y)}{\partial x_i \partial y_j},$$

$$\big[\nabla^2_{yy} f(x, y)\big]_{ij} = \frac{\partial^2 f(x, y)}{\partial y_i \partial y_j}.$$

If $f : \Re^{m+n} \mapsto \Re^r$, and $f_1, f_2, \ldots, f_r$ are the component functions of f, we write

$$\nabla_x f(x,y) = \big[\nabla_x f_1(x,y) \cdots \nabla_x f_r(x,y)\big],$$

$$\nabla_y f(x,y) = \big[\nabla_y f_1(x,y) \cdots \nabla_y f_r(x,y)\big].$$

Let $f : \Re^k \mapsto \Re^m$ and $g : \Re^m \mapsto \Re^n$ be smooth functions, and let h be their composition, i.e.,

$$h(x) = g\big(f(x)\big).$$

Then, the *chain rule* for differentiation states that

$$\nabla h(x) = \nabla f(x) \nabla g\big(f(x)\big), \qquad \forall\ x \in \Re^k.$$

Some examples of useful relations that follow from the chain rule are:

$$\nabla\big(f(Ax)\big) = A' \nabla f(Ax), \qquad \nabla^2\big(f(Ax)\big) = A' \nabla^2 f(Ax) A,$$

where A is a matrix,

$$\nabla_x\Big(f\big(h(x), y\big)\Big) = \nabla h(x) \nabla_h f\big(h(x), y\big),$$

$$\nabla_x\Big(f\big(h(x), g(x)\big)\Big) = \nabla h(x) \nabla_h f\big(h(x), g(x)\big) + \nabla g(x) \nabla_g f\big(h(x), g(x)\big).$$

Differentiation Theorems

We now state some theorems relating to differentiable functions that will be useful for our purposes.

Proposition A.22: (Mean Value Theorem) If $f : \Re \mapsto \Re$ is continuously differentiable over an open interval I, then for every $x, y \in I$, there exists some $\xi \in [x, y]$ such that

$$f(y) - f(x) = \nabla f(\xi)(y - x).$$

Proposition A.23: (Second Order Expansions) Let $f : \Re^n \mapsto \Re$ be twice continuously differentiable over an open sphere S centered at a vector x.

(a) For all y such that $x + y \in S$,

$$f(x+y) = f(x) + y'\nabla f(x) + \tfrac{1}{2}y'\left(\int_0^1 \left(\int_0^t \nabla^2 f(x+\tau y)d\tau\right) dt\right) y.$$

(b) For all y such that $x + y \in S$, there exists an $\alpha \in [0,1]$ such that

$$f(x+y) = f(x) + y'\nabla f(x) + \tfrac{1}{2}y'\nabla^2 f(x+\alpha y)y.$$

(c) For all y such that $x + y \in S$ there holds

$$f(x+y) = f(x) + y'\nabla f(x) + \tfrac{1}{2}y'\nabla^2 f(x)y + o\big(\|y\|^2\big).$$

Proposition A.24: (Descent Lemma) Let $f : \Re^n \mapsto \Re$ be continuously differentiable, and let x and y be two vectors in $\Re^n$. Suppose that

$$\big\|\nabla f(x+ty) - \nabla f(x)\big\| \le Lt\|y\|, \qquad \forall\ t \in [0,1],$$

where L is some scalar. Then

$$f(x+y) \le f(x) + y'\nabla f(x) + \frac{L}{2}\|y\|^2.$$

Proof: Let t be a scalar parameter and let $g(t) = f(x+ty)$. The chain rule yields $(dg/dt)(t) = y'\nabla f(x+ty)$. Now

$$\begin{aligned} f(x+y) - f(x) = g(1) - g(0) &= \int_0^1 \frac{dg}{dt}(t)\,dt = \int_0^1 y'\nabla f(x+ty)\,dt \\ &\le \int_0^1 y'\nabla f(x)\,dt + \left|\int_0^1 y'\big(\nabla f(x+ty) - \nabla f(x)\big)\,dt\right| \\ &\le \int_0^1 y'\nabla f(x)\,dt + \int_0^1 \|y\|\cdot\|\nabla f(x+ty) - \nabla f(x)\|dt \\ &\le y'\nabla f(x) + \|y\|\int_0^1 Lt\|y\|\,dt = y'\nabla f(x) + \frac{L}{2}\|y\|^2. \end{aligned}$$

Q.E.D.

Proposition A.25: (Implicit Function Theorem) Let $f : \Re^{n+m} \mapsto \Re^m$ be a function of $x \in \Re^n$ and $y \in \Re^m$ such that:

(1) $f(\overline{x}, \overline{y}) = 0$.

(2) f is continuous, and has a continuous and nonsingular gradient matrix $\nabla_y f(x, y)$ in an open set containing $(\overline{x}, \overline{y})$.

Then there exist open sets $S_{\overline{x}} \subset \Re^n$ and $S_{\overline{y}} \subset \Re^m$ containing $\overline{x}$ and $\overline{y}$, respectively, and a continuous function $\phi : S_{\overline{x}} \mapsto S_{\overline{y}}$ such that $\overline{y} = \phi(\overline{x})$ and $f\big(x, \phi(x)\big) = 0$ for all $x \in S_{\overline{x}}$. The function ϕ is unique in the sense that if $x \in S_{\overline{x}}$, $y \in S_{\overline{y}}$, and $f(x, y) = 0$, then $y = \phi(x)$. Furthermore, if for some integer $p > 0$, f is p times continuously differentiable the same is true for ϕ, and we have

$$\nabla \phi(x) = -\nabla_x f\big(x, \phi(x)\big)\big(\nabla_y f(x, \phi(x))\big)^{-1}, \qquad \forall\, x \in S_{\overline{x}}.$$

As a final word of caution to the reader, let us mention that one can easily get confused with gradient notation and its use in various formulas, such as for example the order of multiplication of various gradients in the chain rule and the Implicit Function Theorem. Perhaps the safest guideline to minimize errors is to remember our conventions:

(a) A vector is viewed as a column vector (an $n \times 1$ matrix).

(b) The gradient ∇f of a scalar function $f : \Re^n \mapsto \Re$ is also viewed as a column vector.

(c) The gradient matrix ∇f of a vector function $f : \Re^n \mapsto \Re^m$ with components $f_1, \ldots, f_m$ is the $n \times m$ matrix whose columns are the (column) vectors $\nabla f_1, \ldots, \nabla f_m$.

With these rules in mind one can use "dimension matching" as an effective guide to writing correct formulas quickly.

A.6 CONVERGENCE THEOREMS

Many iterative algorithms can be written as

$$x_{k+1} = T\big(x_k\big), \qquad k = 0, 1, \ldots,$$

where $T : X \mapsto X$ is a mapping from a set $X \subset \Re^n$ into itself, and has the property

$$\big\|T(x) - T(y)\big\| \leq \rho \|x - y\|, \qquad \forall\, x, y \in X. \tag{A.4}$$

Here $\|\cdot\|$ is some norm, and ρ is a scalar with $0 \leq \rho < 1$. Such a mapping is called a *contraction mapping*, or simply a *contraction*. The scalar ρ is called the *contraction modulus* of T. Note that a mapping T may be a contraction for some choice of the norm $\|\cdot\|$ and fail to be a contraction under a different choice of norm.

Any vector $x^* \in X$ satisfying $T(x^*) = x^*$ is called a *fixed point* of T and the iteration $x_{k+1} = T(x_k)$ is an important algorithm for finding such a fixed point. The following is the central result regarding contraction mappings.

Proposition A.26: (Contraction Mapping Theorem) Suppose that $T : X \mapsto X$ is a contraction of modulus $\rho \in [0, 1)$ and that X is a closed subset of $\Re^n$. Then:

(a) (*Existence and Uniqueness of Fixed Point*) The mapping T has a unique fixed point $x^* \in X$.

(b) (*Convergence*) For every initial vector $x_0 \in X$, the sequence $\{x_k\}$ generated by $x_{k+1} = T(x_k)$ converges to x^*. In particular,

$$\|x_k - x^*\| \leq \rho^k \|x_0 - x^*\|, \qquad \forall\ k \geq 0.$$

Proof: (a) Fix some $x_0 \in X$ and consider the sequence $\{x_k\}$ generated by $x_{k+1} = T(x_k)$. We have, from the contraction property [cf. Eq. (A.4)],

$$\|x_{k+1} - x_k\| \leq \rho \|x_k - x_{k-1}\|,$$

for all $k \geq 1$, which implies

$$\|x_{k+1} - x_k\| \leq \rho^k \|x_1 - x_0\|, \qquad \forall\ k \geq 0.$$

It follows that for every $k \geq 0$ and $m \geq 1$, we have

$$\begin{aligned} \|x_{k+m} - x_k\| &\leq \sum_{i=1}^{m} \|x_{k+i} - x_{k+i-1}\| \\ &\leq \rho^k (1 + \rho + \cdots + \rho^{m-1}) \|x_1 - x_0\| \\ &\leq \frac{\rho^k}{1-\rho} \|x_1 - x_0\|. \end{aligned}$$

Therefore, $\{x_k\}$ is a Cauchy sequence and must converge to a limit, denoted x^* (Prop. A.5). Furthermore, since X is closed, x^* belongs to X. We have for all $k \geq 1$,

$$\big\|T(x^*) - x^*\big\| \leq \big\|T(x^*) - x_k\big\| + \|x_k - x^*\| \leq \rho \|x^* - x_{k-1}\| + \|x_k - x^*\|$$

and since x_k converges to x^*, we obtain $T(x^*) = x^*$. Therefore, the limit x^* of x_k is a fixed point of T. It is a unique fixed point because if y^* were another fixed point, we would have

$$\|x^* - y^*\| = \big\|T(x^*) - T(y^*)\big\| \le \rho\|x^* - y^*\|,$$

which implies that $x^* = y^*$.

(b) We have

$$\|x_{k'} - x^*\| = \big\|T\big(x_{k'-1}\big) - T(x^*)\big\| \le \rho\|x_{k'-1} - x^*\|,$$

for all $k' \ge 1$, so by applying this relation successively for $k' = k$, $k-1$, $\ldots$,1, we obtain the desired result. **Q.E.D.**

The type of convergence demonstrated in part (b) of the preceding proposition is referred to as *linear convergence*. More precisely, given a sequence $\{x_k\}$ that converges to some $x^* \in \Re^n$, and a continuous (error) function $e : \Re^n \mapsto \Re$ such that $e(x^*) = 0$, we say that $\big\{e(x^k)\big\}$ *converges linearly or geometrically*, if there exist $q > 0$ and $\beta \in (0,1)$ such that for all k

$$e(x^k) \le q\beta^k.$$

Typical examples of error functions that we use are $e(x) = \|x_k - x^*\|$ and $e(x) = f(x) - f(x^*)$, where f is the cost function of an optimization problem.

We note that the convergence of contraction iterations is maintained when there are additional decaying perturbations in $T(x_k)$, i.e.,

$$x_{k+1} = T(x_k) + w_k, \tag{A.5}$$

where $T : \Re^n \mapsto \Re^n$ is a contraction and $\{w_k\}$ is a sequence in $\Re^n$ such that $w_k \to 0$ (see the discussion following the subsequent Prop. A.30). A related useful fact is that when $\big\{\|w_k\|\big\}$ is linearly decaying, then the linear convergence of $\{x_k\}$ is maintained. In particular, consider the iteration (A.5), and assume that T is a contraction of modulus $\rho \in [0,1)$ and for some scalars $q > 0$ and $\sigma \in (0,1)$ we have $\|w_k\| \le q\sigma^k$, for all k. Then we claim that $\{x_k\}$ converges to x^*, the unique fixed point of T, and for every scalar γ with $\max\{\rho, \sigma\} < \gamma < 1$, there exits a scalar $p > 0$ such that

$$\|x_k - x^*\| \le p\gamma^k, \qquad \forall\ k \ge 0. \tag{A.6}$$

To see this, we note that for all k, we have

$$\|x_k - x^*\| = \big\|T(x_{k-1}) - x^* + w_{k-1}\big\| \le \big\|T(x_{k-1}) - x^*\big\| + \|w_{k-1}\|,$$

so that by using the contraction property,

$$\|x_k - x^*\| \le \rho\|x_{k-1} - x^*\| + q\sigma^{k-1}.$$

Replacing k with $k-1$, we have

$$\|x_{k-1} - x^*\| \le \rho\|x_{k-2} - x^*\| + q\sigma^{k-2},$$

and by combining the preceding two relations,

$$\|x_k - x^*\| \le \rho^2\|x_{k-2} - x^*\| + q(\sigma^{k-1} + \rho\sigma^{k-2}).$$

Proceeding similarly, we obtain for all k,

$$\begin{aligned}\|x_k - x^*\| &\le \rho^k\|x_0 - x^*\| + q(\sigma^{k-1} + \rho\sigma^{k-2} + \cdots + \rho^{k-2}\sigma + \rho^{k-1})\\ &\le \rho^k\|x_0 - x^*\| + kq\big(\max\{\rho,\,\sigma\}\big)^{k-1}\\ &\le \gamma^k\|x_0 - x^*\| + \bar q\gamma^k,\end{aligned}$$

where for a given $\gamma \in \big(\max\{\rho,\,\sigma\}, 1\big)$, $\bar q$ is such that $kq\big(\max\{\rho,\,\sigma\}\big)^{k-1} \le \bar q\gamma^k$ for all k. This shows Eq. (A.6).

In the case of a linear mapping

$$T(x) = Ax + b,$$

where A is an $n \times n$ matrix and $b \in \Re^n$, it can be shown that T is a contraction mapping with respect to some norm (but not necessarily all norms) if and only if all the eigenvalues of A lie strictly within the unit circle. For a proof, see [OrR70], or [Ber12], Example 1.5.1.

Contractions with Respect to a Weighted Maximum Norm

Given a vector $\xi = (\xi_1, \ldots, \xi_n)' \in \Re^n$, with positive components $\xi_i > 0$, the weighted maximum norm corresponding to ξ is defined by

$$\|x\|_\xi = \max_{i=1,\ldots,n} \frac{|x_i|}{\xi_i}, \qquad x \in \Re^n.$$

Consider the linear mapping

$$T(x) = Ax + b, \tag{A.7}$$

where A is an $n \times n$ matrix with components a_{ij} and b is a vector in $\Re^n$. The following proposition gives useful criteria for T to be a weighted maximum norm contraction.

Proposition A.27: Consider the mapping T of Eq. (A.7).

(a) T is a contraction with respect to $\|\cdot\|_\xi$ with modulus ρ if and only if

$$\frac{\sum_{j=1}^n |a_{ij}|\,\xi_j}{\xi_i} \le \rho, \qquad \forall\ i = 1, \ldots, n.$$

(b) Let P be a stochastic $n \times n$ matrix P (i.e., its components p_{ij} satisfy $p_{ij} \geq 0$ for all $i, j = 1, \ldots, n$, and $\sum_{j=1}^{n} p_{ij} = 1$ for all $i = 1, \ldots, n$), and assume that

$$|a_{ij}| \leq p_{ij}, \qquad \forall\ i, j = 1, \ldots, n,$$

and that for some row index $\bar{i} \in \{1, \ldots, n\}$,

$$|a_{\bar{i}j}| < p_{\bar{i}j}, \qquad \forall\ j = 1, \ldots, n.$$

Assume further that P corresponds to an irreducible Markov chain (one with a single recurrent class and no transient states) and that $\xi = (\xi_1, \ldots, \xi_n)' \in \Re^n$ is its invariant distribution, i.e.,

$$\xi_i > 0, \quad i = 1, \ldots, n, \qquad \sum_{i=1}^{n} \xi_i = 1, \qquad \xi' = \xi' P.$$

Then T is a contraction with respect to the norm $\| \cdot \|_{\xi}$.

Part (a) of the preceding proposition is given as Prop. 1.5.2(a) of [Ber12], while part (b) is given as Prop. 1 of [BeY09].

Convergence of Iterations with Delays

The following two propositions deal with iterations that involve delayed iterates.

Proposition A.28: (Iterations with Delays I) Let $\{\alpha_k\}$ be a scalar sequence such that

$$|\alpha_k| \leq \sum_{i=1}^{n} \beta_i |\alpha_{k-i}|, \qquad \forall\ k = 0, 1, \ldots,$$

where $\beta_i > 0$, $i = 1, \ldots, n$, are some scalars with $\sum_{i=1}^{n} \beta_i < 1$, and n is a positive integer. Then the sequence $\{\gamma_k\}$, where

$$\gamma_k = \max_{i=1,\ldots,n} \frac{|a_{k-i}|}{\xi_i},$$

converges to 0 linearly, where $\xi = (\xi_1, \ldots, \xi_n)'$ is the unique solution of the system of equations

$$\sum_{i=1}^{n} \xi_i = 1, \quad \xi_j = \frac{\beta_j}{\sum_{i=1}^{n} \beta_i} \xi_1 + \xi_{j+1},\ j = 1, \ldots, n-1, \quad \xi_n = \frac{\beta_n}{\sum_{i=1}^{n} \beta_i} \xi_1.$$

Proof: The given system of equations can be seen to have a unique solution by successively expressing $\xi_n, \xi_{n-1}, \ldots, \xi_2$ in terms of ξ_1, and then determining ξ_1 from the equation $\sum_{i=1}^{n} \xi_i = 1$. Furthermore, we can easily verify the equation $\xi' = \xi' P$ for ξ to be the invariant distribution of the irreducible matrix P given by

$$P = \begin{pmatrix} \beta_1/\sum_{i=1}^{n}\beta_i & \beta_2/\sum_{i=1}^{n}\beta_i & \cdots & \beta_{n-1}/\sum_{i=1}^{n}\beta_i & \beta_n/\sum_{i=1}^{n}\beta_i \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

The proof follows by using Prop. A.27(b). **Q.E.D.**

The preceding proposition can be used to show that an iteration of the form

$$\alpha_k = \gamma + \sum_{i=1}^{n} \beta_i \alpha_{k-i},$$

where γ is a scalar, and $\beta_1, \ldots, \beta_n$ are scalars satisfying $\sum_{i=1}^{n} |\beta_i| < 1$, converges to

$$\frac{\gamma}{1 - \sum_{i=1}^{n} \beta_i}.$$

The following proposition is due to [FAJ14], whose proof we follow closely. Additional related results are given in [Fey16].

Proposition A.29: (Iterations with Delays II) Let $\{\alpha_k\}$ be a nonnegative sequence satisfying

$$\alpha_{k+1} \le p\alpha_k + q \max_{\max\{0,k-d\} \le \ell \le k} \alpha_\ell, \qquad \forall\ k = 0, 1, \ldots, \tag{A.8}$$

for some positive integer d and nonnegative scalars p and q such that $p + q < 1$. Then we have

$$\alpha_k \le \rho^k \alpha_0, \qquad \forall\ k = 0, 1, \ldots, \tag{A.9}$$

where $\rho = (p+q)^{\frac{1}{1+d}}$.

Proof: We first show a preliminary relation. Since $p+q<1$, we have

$$1 \leq (p+q)^{-\frac{b}{1+b}},$$

which implies that

$$\begin{aligned} p+q\rho^{-b} &= p+q(p+q)^{-\frac{b}{1+b}} \\ &\leq (p+q)(p+q)^{-\frac{b}{1+b}} \\ &= (p+q)^{\frac{1}{1+b}} \\ &= \rho. \end{aligned} \tag{A.10}$$

We now show Eq. (A.9) by induction. It clearly holds for $k=0$. Assume that it holds for all k up to some $\bar{k}$. Then

$$\alpha_k \leq \rho^k \alpha_0, \qquad \forall\ k = \max\{0, \bar{k}-b\}, \ldots, \bar{k}.$$

From this relation and Eq. (A.8), we have

$$\begin{aligned} \alpha_{\bar{k}+1} &\leq p\rho^{\bar{k}}\alpha_0 + q\left(\max_{\max\{0,\bar{k}-b\}\leq \ell \leq \bar{k}} \rho^\ell \alpha_0\right) \\ &\leq p\rho^{\bar{k}}\alpha_0 + q\rho^{\max\{0,\bar{k}-b\}}\alpha_0 \\ &\leq p\rho^{\bar{k}}\alpha_0 + q\rho^{\bar{k}-b}\alpha_0 \\ &= \big(p+q\rho^{-b}\big)\rho^{\bar{k}}\alpha_0. \end{aligned}$$

Using also Eq. (A.10), we have $\alpha_{\bar{k}+1} \leq \rho^{\bar{k}+1}\alpha_0$, and this completes the induction. **Q.E.D.**

Nonstationary Iterations

For nonstationary iterations of the form $x_{k+1} = T_k(x_k)$, where the function T_k depends on k, the ideas of the preceding propositions may apply but with modifications. The following proposition is often useful in this respect.

Proposition A.30: Let $\{\alpha_k\}$ be a nonnegative scalar sequence such that

$$\alpha_{k+1} \leq (1-\gamma_k)\alpha_k + \beta_k, \qquad \forall\ k = 0, 1, \ldots,$$

where $0 \leq \beta_k$, $0 < \gamma_k \leq 1$ for all k, and

$$\sum_{k=0}^{\infty} \gamma_k = \infty, \qquad \frac{\beta_k}{\gamma_k} \to 0.$$

Then $\alpha_k \to 0$.

Proof: We first show that given any $\epsilon > 0$, we have $\alpha_k < \epsilon$ for infinitely many k. Indeed, if this were not so, by letting $\bar{k}$ be such that $\alpha_k \geq \epsilon$ and $\beta_k/\gamma_k \leq \epsilon/2$ for all $k \geq \bar{k}$, we would have for all $k \geq \bar{k}$

$$\alpha_{k+1} \leq \alpha_k - \gamma_k \alpha_k + \beta_k \leq \alpha_k - \gamma_k \epsilon + \frac{\gamma_k \epsilon}{2} = \alpha_k - \frac{\gamma_k \epsilon}{2}.$$

Therefore, for all $m \geq \bar{k}$,

$$\alpha_{m+1} \leq \alpha_{\bar{k}} - \frac{\epsilon}{2} \sum_{k=\bar{k}}^{m} \gamma_k.$$

Since $\{\alpha_k\}$ is nonnegative and $\sum_{k=0}^{\infty} \gamma_k = \infty$, we obtain a contradiction.

Thus, given any $\epsilon > 0$, there exists $\bar{k}$ such that $\beta_k/\gamma_k < \epsilon$ for all $k \geq \bar{k}$ and $\alpha_{\bar{k}} < \epsilon$. We then have

$$\alpha_{\bar{k}+1} \leq (1 - \gamma_k)\alpha_{\bar{k}} + \beta_k < (1 - \gamma_k)\epsilon + \gamma_k \epsilon = \epsilon.$$

By repeating this argument, we obtain $\alpha_k < \epsilon$ for all $k \geq \bar{k}$. Since ϵ can be arbitrarily small, it follows that $\alpha_k \to 0$. **Q.E.D.**

As an example, consider the iteration

$$x_{k+1} = T(x_k) + w_k,$$

where $T : \Re^n \mapsto \Re^n$ is a contraction of modulus $\rho \in (0, 1)$ and $\{w_k\}$ is a sequence in $\Re^n$ such that $w_k \to 0$. Then we have

$$\|x_{k+1} - x^*\| \leq \big\|T(x_k) - x^*\big\| + \|w_k\| \leq \rho\|x_k - x^*\| + \|w_k\|,$$

and Prop. A.30 applies with $\alpha_k = \|x_k - x^*\|$, $\gamma_k = 1 - \rho$, and $\beta_k = \|w_k\|$, showing that $x_k \to x^*$.

As another example, consider a sequence of "approximate" contraction mappings $T_k : \Re^n \mapsto \Re^n$, satisfying

$$\big\|T_k(x) - T_k(y)\big\| \leq (1 - \gamma_k)\|x - y\| + \beta_k, \qquad \forall\ x, y \in \Re^n,\ k = 0, 1, \ldots,$$

where $\gamma_k \in (0, 1]$, for all k, and

$$\sum_{k=0}^{\infty} \gamma_k = \infty, \qquad \frac{\beta_k}{\gamma_k} \to 0.$$

Assume also that all the mappings T_k have a common fixed point x^*. Then

$$\|x_{k+1} - x^*\| = \big\|T_k(x_k) - T_k(x^*)\big\| \leq (1 - \gamma_k)\|x_k - x^*\| + \beta_k,$$

and from Prop. A.30, it follows that the sequence $\{x_k\}$ generated by the iteration $x_{k+1} = T_k(x_k)$ converges to x^* starting from any $x_0 \in \Re^n$.

Supermartingale Convergence

We next give a convergence theorem relating to deterministic sequences. It is a special case of a fundamental theorem, known as the *supermartingale convergence theorem*, which relates to convergence of sequences of random variables. We will not need this more general theorem in our analysis, and we refer to [Ber15a] and [WaB13] for some of its applications in incremental optimization methods with randomized order of component selection.

Proposition A.31: Let $\{Y_k\}$, $\{Z_k\}$, $\{W_k\}$, and $\{V_k\}$ be four scalar sequences such that

$$Y_{k+1} \le (1 + V_k)Y_k - Z_k + W_k, \qquad k = 0, 1, \ldots, \tag{A.11}$$

$\{Z_k\}$, $\{W_k\}$, and $\{V_k\}$ are nonnegative, and

$$\sum_{k=0}^{\infty} W_k < \infty, \qquad \sum_{k=0}^{\infty} V_k < \infty.$$

Then either $Y_k \to -\infty$, or else $\{Y_k\}$ converges to a finite value and $\sum_{k=0}^{\infty} Z_k < \infty$.

Proof: We first give the proof assuming that $V_k \equiv 0$, and then generalize. In this case, using the nonnegativity of $\{Z_k\}$, we have

$$Y_{k+1} \le Y_k + W_k.$$

By writing this relation for the index k set to $\bar{k}, \ldots, k$, where $k \ge \bar{k}$, and adding, we have

$$Y_{k+1} \le Y_{\bar{k}} + \sum_{\ell=\bar{k}}^{k} W_\ell \le Y_{\bar{k}} + \sum_{\ell=\bar{k}}^{\infty} W_\ell.$$

Since $\sum_{k=0}^{\infty} W_k < \infty$, it follows that $\{Y_k\}$ is bounded above, and by taking upper limit of the left hand side as $k \to \infty$ and lower limit of the right hand side as $\bar{k} \to \infty$, we have

$$\limsup_{k\to\infty} Y_k \le \liminf_{\bar{k}\to\infty} Y_{\bar{k}} < \infty.$$

This implies that either $Y_k \to -\infty$, or else $\{Y_k\}$ converges to a finite value. In the latter case, by writing Eq. (A.11) for the index k set to $0, \ldots, k$, and adding, we have

$$\sum_{\ell=0}^{k} Z_\ell \le Y_0 + \sum_{\ell=0}^{k} W_\ell - Y_{k+1}, \qquad \forall\ k = 0, 1, \ldots,$$

so by taking the limit as $k \to \infty$, we obtain $\sum_{\ell=0}^{\infty} Z_\ell < \infty$.

We now extend the proof to the case of a general nonnegative sequence $\{V_k\}$. We first note that

$$\log \prod_{\ell=0}^{k} (1 + V_\ell) = \sum_{\ell=0}^{k} \log(1 + V_\ell) \leq \sum_{k=0}^{\infty} V_k,$$

since we generally have $(1+a) \leq e^a$ and $\log(1+a) \leq a$ for any $a \geq 0$. Thus the assumption $\sum_{k=0}^{\infty} V_k < \infty$ implies that

$$\prod_{\ell=0}^{\infty} (1 + V_\ell) < \infty. \tag{A.12}$$

Define

$$\bar{Y}_k = Y_k \prod_{\ell=0}^{k-1} (1 + V_\ell)^{-1}, \quad \bar{Z}_k = Z_k \prod_{\ell=0}^{k} (1 + V_\ell)^{-1}, \quad \bar{W}_k = W_k \prod_{\ell=0}^{k} (1 + V_\ell)^{-1}.$$

Multiplying Eq. (A.11) with $\prod_{\ell=0}^{k} (1 + V_\ell)^{-1}$, we obtain

$$\bar{Y}_{k+1} \leq \bar{Y}_k - \bar{Z}_k + \bar{W}_k.$$

Since $\bar{W}_k \leq W_k$, the hypothesis $\sum_{k=0}^{\infty} W_k < \infty$ implies $\sum_{k=0}^{\infty} \bar{W}_k < \infty$, so from the special case of the result already shown, we have that either $\bar{Y}_k \to -\infty$ or else $\{\bar{Y}_k\}$ converges to a finite value and $\sum_{k=0}^{\infty} \bar{Z}_k < \infty$. Since

$$Y_k = \bar{Y}_k \prod_{\ell=0}^{k-1} (1 + V_\ell), \qquad Z_k = \bar{Z}_k \prod_{\ell=0}^{k} (1 + V_\ell),$$

and $\prod_{\ell=0}^{k-1} (1 + V_\ell)$ converges to a finite value by the nonnegativity of $\{V_k\}$ and Eq. (A.12), it follows that either $Y_k \to -\infty$ or else $\{Y_k\}$ converges to a finite value and $\sum_{k=0}^{\infty} Z_k < \infty$. **Q.E.D.**

Fejér Monotonicity

Supermartingale convergence theorems can be applied in a variety of contexts. One such context, the so called *Fejér monotonicity* theory, deals with iterations that "almost" decrease the distance to *every* element of some given set X^*. We may then often show that such iterations are convergent to a (unique) element of X^*. Applications of this idea arise when X^* is the set of optimal solutions of an optimization problem or the set of fixed points of a certain mapping. Examples are various gradient and subgradient projection methods with a diminishing stepsize that arise in various contexts in this book.

Proposition A.32: (Fejér Convergence Theorem) Let X^* be a nonempty subset of $\Re^n$, and let $\{x_k\} \subset \Re^n$ be a sequence satisfying for some $p > 0$ and for all k,

$$\|x_{k+1} - x^*\|^p \leq (1+\beta_k)\|x_k - x^*\|^p - \gamma_k\,\phi(x_k; x^*) + \delta_k, \qquad \forall\, x^* \in X^*,$$

where $\{\beta_k\}$, $\{\gamma_k\}$, and $\{\delta_k\}$ are nonnegative sequences satisfying

$$\sum_{k=0}^{\infty} \beta_k < \infty, \qquad \sum_{k=0}^{\infty} \gamma_k = \infty, \qquad \sum_{k=0}^{\infty} \delta_k < \infty,$$

$\phi : \Re^n \times X^* \mapsto [0, \infty)$ is some nonnegative function, and $\|\cdot\|$ is some norm. Then:

(a) The minimum distance sequence $\inf_{x^* \in X^*} \|x_k - x^*\|$ converges, and in particular, $\{x_k\}$ is bounded.

(b) If $\{x_k\}$ has a limit point $\bar{x}$ that belongs to X^*, then the entire sequence $\{x_k\}$ converges to $\bar{x}$.

(c) Suppose that for some $x^* \in X^*$, $\phi(\cdot; x^*)$ is lower semicontinuous and satisfies

$$\phi(x; x^*) = 0 \qquad \text{if and only if} \qquad x \in X^*. \tag{A.13}$$

Then $\{x_k\}$ converges to a point in X^*.

Proof: (a) Let $\{\epsilon_k\}$ be a positive sequence such that $\sum_{k=0}^{\infty}(1+\beta_k)\epsilon_k < \infty$, and let x_k^* be a point of X^* such that

$$\|x_k - x_k^*\|^p \leq \inf_{x^* \in X^*} \|x_k - x^*\|^p + \epsilon_k.$$

Then since ϕ is nonnegative, we have for all k,

$$\inf_{x^* \in X^*} \|x_{k+1} - x^*\|^p \leq \|x_{k+1} - x_k^*\|^p \leq (1+\beta_k)\|x_k - x_k^*\|^p + \delta_k,$$

and by combining the last two relations, we obtain

$$\inf_{x^* \in X^*} \|x_{k+1} - x^*\|^p \leq (1+\beta_k) \inf_{x^* \in X^*} \|x_k - x^*\|^p + (1+\beta_k)\epsilon_k + \delta_k.$$

The result follows by applying Prop. A.31 with

$$Y_k = \inf_{x^* \in X^*} \|x_k - x^*\|^p, \quad Z_k = 0, \quad W_k = (1+\beta_k)\epsilon_k + \delta_k, \quad V_k = \beta_k.$$

(b) Following the argument of the proof of Prop. A.31, define for all k,

$$\bar{Y}_k = \|x_k - \bar{x}\|^p \prod_{\ell=0}^{k-1} (1 + \beta_\ell)^{-1}, \qquad \bar{\delta}_k = \delta_k \prod_{\ell=0}^{k} (1 + \beta_\ell)^{-1}.$$

Then from our hypotheses, we have $\sum_{k=0}^{\infty} \bar{\delta}_k < \infty$ and

$$\bar{Y}_{k+1} \leq \bar{Y}_k + \bar{\delta}_k, \qquad \forall\ k = 0, 1, \ldots, \tag{A.14}$$

while $\{\bar{Y}_k\}$ has a limit point at 0, since $\bar{x}$ is a limit point of $\{x_k\}$. For any $\epsilon > 0$, let $\bar{k}$ be such that

$$\bar{Y}_{\bar{k}} \leq \epsilon, \qquad \sum_{\ell=\bar{k}}^{\infty} \bar{\delta}_\ell \leq \epsilon,$$

so that by adding Eq. (A.14), we obtain for all $k > \bar{k}$,

$$\bar{Y}_k \leq \bar{Y}_{\bar{k}} + \sum_{\ell=\bar{k}}^{\infty} \bar{\delta}_\ell \leq 2\epsilon.$$

Since ϵ is arbitrarily small, it follows that $\bar{Y}_k \to 0$. We now note that as in Eq. (A.12),

$$\prod_{\ell=0}^{\infty} (1 + \beta_\ell)^{-1} < \infty,$$

so that $\bar{Y}_k \to 0$ implies that $\|x_k - \bar{x}\|^p \to 0$, and hence $x_k \to \bar{x}$.

(c) From Prop. A.31, it follows that

$$\sum_{k=0}^{\infty} \gamma_k\, \phi(x_k; x^*) < \infty.$$

Thus $\lim_{k \to \infty,\, k \in \mathcal{K}} \phi(x_k; x^*) = 0$ for some subsequence $\{x_k\}_{\mathcal{K}}$. By part (a), $\{x_k\}$ is bounded, so the subsequence $\{x_k\}_{\mathcal{K}}$ has a limit point $\bar{x}$, and by the lower semicontinuity of $\phi(\cdot; x^*)$, we must have

$$\phi(\bar{x}; x^*) \leq \lim_{k \to \infty,\, k \in \mathcal{K}} \phi(x_k; x^*) = 0,$$

which in view of the nonnegativity of ϕ, implies that $\phi(\bar{x}; x^*) = 0$. Using the hypothesis (A.13), it follows that $\bar{x} \in X^*$, so by part (b), the entire sequence $\{x_k\}$ converges to $\bar{x}$. **Q.E.D.**

APPENDIX B: Convex Analysis

Convexity is central in nonlinear programming, and has a rich mathematical theory. In this appendix, we selectively collect the definitions, notational conventions, and results that we will need. For detailed textbook accounts of convex analysis and its connections with optimization, see Rockafellar [Roc70], Ekeland and Teman [EkT76], Hiriart-Urruty and Lemarechal [HiL93], Rockafellar and Wets [RoW98], Borwein and Lewis [BoL00], Bonnans and Shapiro [BoS00], Zalinescu [Zal02], Auslender and Teboulle [AuT03], Bertsekas, Nedić, and Ozdaglar [BNO03], and Bertsekas [Ber09].

A discussion of generalized notions of convexity, including quasiconvexity and pseudoconvexity, and their applications in optimization can be found in the books by Avriel [Avr76], Bazaraa, Sherali, and Shetty [BSS93], Mangasarian [Man69], and the references quoted therein.

The author's convex optimization theory textbook [Ber09] is consistent with the notation and content of this appendix, but develops the subject in much greater depth and detail. Proofs of the results quoted are generally given in this textbook, and on some occasions, in the author's convex optimization algorithms textbook [Ber15a]. In a few cases of important convex optimization-related results, a proof is included here.

B.1 CONVEX SETS AND FUNCTIONS

A subset C of $\Re^n$ is called *convex* if

$$\alpha x + (1-\alpha)y \in C, \qquad \forall\ x, y \in C,\ \forall\ \alpha \in [0,1]. \tag{B.1}$$

The following proposition provides some means for verifying convexity of a set.

Proposition B.1:

(a) For any collection $\{C_i \mid i \in I\}$ of convex sets, the set intersection $\cap_{i\in I} C_i$ is convex.

(b) The vector sum of two convex sets C_1 and C_2 is convex.

(c) The image of a convex set under a linear transformation is convex.

(d) If C is a convex set and $f : C \mapsto \Re$ is a convex function, the level sets $\{x \in C \mid f(x) \le \alpha\}$ and $\{x \in C \mid f(x) < \alpha\}$ are convex for all scalars α.

Proof: See Prop. 1.1.1 and Section 1.1.1 of [Ber09]. **Q.E.D.**

Let C be a convex subset of $\Re^n$. A function $f : C \mapsto \Re$ is called *convex* if

$$f\big(\alpha x+(1-\alpha)y\big) \le \alpha f(x)+(1-\alpha)f(y), \qquad \forall\, x, y \in C,\ \forall\, \alpha \in [0,1]. \quad \text{(B.2)}$$

The function f is called *concave* if $-f$ is convex. The function f is called *strictly convex* if the above inequality is strict for all $x, y \in C$ with $x \neq y$, and all $\alpha \in (0,1)$. For a function $f : \Re^n \mapsto \Re$, we also say that f is *convex over the convex set* C if Eq. (B.2) holds.

We occasionally deal with functions $f : C \mapsto [-\infty, \infty]$ that can take infinite values. The *epigraph* of such a function f is the subset of $\Re^{n+1}$ given by

$$\text{epi}(f) = \big\{(x, w) \mid x \in C,\ w \in \Re,\ f(x) \le w\big\}.$$

We say that $f : C \mapsto (-\infty, \infty]$ is convex if C is convex and epi(f) is a convex set. Note that a function $f : C \mapsto (-\infty, \infty]$ is convex if Eq. (B.2) holds (here the rules of arithmetic are extended to include $\infty + \infty = \infty$, $0 \cdot \infty = 0$, and $\alpha \cdot \infty = \infty$, for all $\alpha > 0$).

The *effective domain* of f is the set

$$\text{dom}(f) = \big\{x \in C \mid f(x) < \infty\big\},$$

which is convex if f is convex. The function f is called *closed* if epi(f) is a closed set, and it is called *proper* if dom(f) is nonempty and $f(x) > -\infty$ for all $x \in C$.

By restricting the definition of a convex function to its effective domain we can avoid calculations with ∞, and we will often do this. However, in some analyses it is more economical to use convex functions that can take the value of infinity.

A useful property, obtained by repeated application of the definition of convexity [cf. Eq. (B.2)], is that if $x_1, \ldots, x_m \in C$, $\alpha_1, \ldots, \alpha_m \ge 0$, and

$\sum_{i=1}^{m} \alpha_i = 1$, then

$$f\left(\sum_{i=1}^{m} \alpha_i x_i\right) \le \sum_{i=1}^{m} \alpha_i f(x_i).$$

This is a special case of *Jensen's inequality* and can be used to prove a number of interesting inequalities in applied mathematics and probability theory.

The following proposition provides some means for recognizing convex functions.

Proposition B.2:

(a) A linear function is convex.

(b) Any vector norm is convex.

(c) The weighted sum of convex functions, with positive weights, is convex.

(d) If I is an index set, C is a convex subset of $\Re^n$, and $f_i : C \mapsto (-\infty, \infty]$ is convex for each $i \in I$, then the function $h : C \mapsto (-\infty, \infty]$ defined by

$$h(x) = \sup_{i \in I} f_i(x)$$

is also convex.

(e) If $F : \Re^{n+m} \mapsto \Re$ is a convex function of the pair (x, z) where $x \in \Re^n$, $z \in \Re^m$, and Z is a convex set such that $\inf_{Z \in Z} F(x, z) > -\infty$ for all $x \in \Re^n$, then the function $f : \Re^n \to \Re$ defined by

$$f(x) = \inf_{Z \in Z} F(x, z), \qquad \forall\, x \in \Re^n,$$

is convex.

Proof: For parts (a)-(d), see Props. 1.1.4-1.1.6 and Section 1.1.3 of [Ber09]. For part (e), see Prop. 3.3.1 of [Ber09]. **Q.E.D.**

Characterizations of Differentiable Convex Functions

For differentiable functions, there is an alternative characterization of convexity, given in the following proposition, parts (a) and (b) of which are classical. Part (c) is given as Theorem 2.1.5 in Nesterov's book [Nes04], but the proof that (iv) implies (i) given there is flawed.

Proposition B.3: (First Derivative Characterizations) Let C be a convex subset of $\Re^n$ and let $f : \Re^n \mapsto \Re$ be differentiable over $\Re^n$.

(a) f is convex over C if and only if

$$f(z) \geq f(x) + (z - x)'\nabla f(x), \qquad \forall\ x, z \in C.$$

(b) f is strictly convex over C if and only if the above inequality is strict whenever $x \neq z$.

(c) Let f be convex. For a scalar $L > 0$ the following five properties are equivalent:

(i) $\big\|\nabla f(x) - \nabla f(y)\big\| \leq L\,\|x - y\|$, for all $x, y \in \Re^n$.

(ii) $f(x) + \nabla f(x)'(y - x) + \frac{1}{2L}\big\|\nabla f(x) - \nabla f(y)\big\|^2 \leq f(y)$, for all $x, y \in \Re^n$.

(iii) $\big(\nabla f(x) - \nabla f(y)\big)'(x - y) \geq \frac{1}{L}\big\|\nabla f(x) - \nabla f(y)\big\|^2$, for all $x, y \in \Re^n$.

(iv) $f(y) \leq f(x) + \nabla f(x)'(y - x) + \frac{L}{2}\|y - x\|^2$, for all $x, y \in \Re^n$.

(v) $\big(\nabla f(x) - \nabla f(y)\big)'(x - y) \leq L\|x - y\|^2$, for all $x, y \in \Re^n$.

Proof: For parts (a) and (b), see Prop. 1.1.7 and Section 1.1.4 of [Ber09]. For part (c), see [Ber15a], Exercise 6.1 (with solution included). **Q.E.D.**

For twice differentiable convex functions, there is another characterization of convexity, which is given in the following proposition.

Proposition B.4: (Second Derivative Characterizations) Let C be a convex subset of $\Re^n$ and let $f : \Re^n \mapsto \Re$ be twice continuously differentiable over $\Re^n$.

(a) If $\nabla^2 f(x)$ is positive semidefinite for all $x \in C$, then f is convex over C.

(b) If $\nabla^2 f(x)$ is positive definite for every $x \in C$, then f is strictly convex over C.

(c) If C is open and f is convex over C, then $\nabla^2 f(x)$ is positive semidefinite for all $x \in C$.

(d) If $f(x) = x'Qx$, where Q is a symmetric matrix, then f is convex if and only if Q is positive semidefinite. Furthermore, f is strictly convex if and only if Q is positive definite.

Proof: See Prop. 1.1.10 and Section 1.1.4 of [Ber09]. **Q.E.D.**

The conclusion of Prop. B.4(c) can also be proved if C is assumed to have nonempty interior instead of being open. We now consider a strengthened form of strict convexity for a continuously differentiable function $f : \Re^n \mapsto \Re$. We say that f is *strongly convex* if for some $\sigma > 0$, we have

$$f(y) \geq f(x) + \nabla f(x)'(y - x) + \frac{\sigma}{2}\|x - y\|^2, \qquad \forall\, x, y \in \Re^n. \tag{B.3}$$

It can be shown that an equivalent definition is that

$$\big(\nabla f(x) - \nabla f(y)\big)'(x - y) \geq \sigma\|x - y\|^2, \qquad \forall\, x, y \in \Re^n. \tag{B.4}$$

A proof of this may be found in several sources, including the on-line exercises of Chapter 1 of [Ber09]. By fixing x in the definition (B.3), we see that a strongly convex function majorizes a coercive function, so it is itself coercive. It is also strictly convex, as shown among other properties by the following proposition.

Proposition B.5: (Strong Convexity) Let $f : \Re^n \mapsto \Re$ be a function that is continuously differentiable. Then:

(a) If f strongly convex in the sense that it satisfies Eq. (B.4) for some $\sigma > 0$, then f is strictly convex. If in addition, ∇f satisfies the Lipschitz condition

$$\big\|\nabla f(x) - \nabla f(y)\big\| \leq L\,\|x - y\|, \qquad \forall\, x, y \in \Re^n, \tag{B.5}$$

for some $L > 0$, then we have for all $x, y \in \Re^n$

$$\big(\nabla f(x) - \nabla f(y)\big)'(x - y) \geq \frac{\sigma L}{\sigma + L}\|x - y\|^2 + \frac{1}{\sigma + L}\big\|\nabla f(x) - \nabla f(y)\big\|^2. \tag{B.6}$$

(b) If f is twice continuously differentiable over $\Re^n$, then f satisfies Eq. (B.4) if and only if the matrix $\nabla^2 f(x) - \sigma I$, where I is the identity, is positive semidefinite for every $x \in \Re^n$.

Proof: (a) Fix some $x, y \in \Re^n$ such that $x \neq y$, and define the function $h : [0, 1] \mapsto \Re$ by

$$h(t) = f\big(x + t(y - x)\big).$$

Consider some $t, \bar{t} \in [0,1]$ such that $t < \bar{t}$. Using the chain rule and Eq. (B.4), we have

$$\begin{aligned}\Big(\frac{dh(\bar{t})}{dt} - \frac{dh(t)}{dt}\Big)(\bar{t}-t) \\ &= \Big(\nabla f\big(x+\bar{t}(y-x)\big) - \nabla f\big(x+t(y-x)\big)\Big)'(y-x)(\bar{t}-t) \\ &\geq \sigma(\bar{t}-t)^2\|x-y\|^2 > 0.\end{aligned}$$

Thus, dh/dt is strictly increasing, and for any $t \in (0,1)$

$$\frac{h(t)-h(0)}{t} = \frac{1}{t}\int_0^t \frac{dh(\tau)}{d\tau}\,d\tau < \frac{1}{1-t}\int_t^1 \frac{dh(\tau)}{d\tau}\,d\tau = \frac{h(1)-h(t)}{1-t}.$$

Equivalently, we have $th(1)+(1-t)h(0) > h(t)$, so from the definition of h, we obtain

$$tf(y)+(1-t)f(x) > f\big(ty+(1-t)x\big).$$

Since this inequality was proved for arbitrary $t \in (0,1)$ and $x \neq y$, it follows that f is strictly convex.

We now assume that the Lipschitz condition (B.5) holds, and show Eq. (B.6). From Eqs. (B.4) and (B.5), we have $\sigma \leq L$. If $\sigma = L$, the result follows by combining the relation (iii) of Prop. B.3(c) and the relation

$$\big\|\nabla f(x) - \nabla f(y)\big\| \geq \sigma\|x-y\|, \qquad \forall\ x,y \in \Re^n,$$

which is a consequence of the strong convexity assumption (B.4). For $\sigma < L$ consider the function

$$\phi(x) = f(x) - \frac{\sigma}{2}\|x\|^2.$$

We will show that $\nabla\phi$, which is given by

$$\nabla\phi(x) = \nabla f(x) - \sigma x, \tag{B.7}$$

is Lipschitz continuous with constant $L-\sigma$. To this end, based on the equivalence of statements (i) and (v) of Prop. B.3(c), it is sufficient to show that

$$\big(\nabla\phi(x) - \nabla\phi(y)\big)'(x-y) \leq (L-\sigma)\|x-y\|^2, \qquad \forall\ x,y \in \Re^n,$$

or, using the expression (B.7) for $\nabla\phi$,

$$\big(\nabla f(x) - \nabla f(y) - \sigma(x-y)\big)'(x-y) \leq (L-\sigma)\|x-y\|^2, \qquad \forall\ x,y \in \Re^n.$$

This relation is equivalently written as

$$\big(\nabla f(x) - \nabla f(y)\big)'(x-y) \leq L\|x-y\|^2, \qquad \forall\ x,y \in \Re^n,$$

and is true by the equivalence of statements (i) and (v) of Prop. B.3(c).

Having shown that $\nabla\phi$ is Lipschitz continuous with constant $L-\sigma$, we use the equivalence of statements (i) and (iii) of Prop. B.3(c) to the function ϕ and obtain

$$\big(\nabla\phi(x)-\nabla\phi(y)\big)'(x-y) \geq \frac{1}{L-\sigma}\big\|\nabla\phi(x)-\nabla\phi(y)\big\|^2.$$

Using the expression (B.7) for $\nabla\phi$ in this relation, we have

$$\big(\nabla f(x)-\nabla f(y)-\sigma(x-y)\big)'(x-y) \geq \frac{1}{L-\sigma}\big\|\nabla f(x)-\nabla f(y)-\sigma(x-y)\big\|^2,$$

which after expanding the quadratic and collecting terms, can be verified to be equivalent to the desired relation.

(b) Suppose that f satisfies Eq. (B.4). We fix some $x \in \Re^n$, let d be any vector in $\Re^n$, and let γ be a scalar in $(0,1]$. We use the second order expansion of Prop. A.23(b) twice to obtain

$$f(x+\gamma d) = f(x) + \gamma d'\nabla f(x) + \frac{\gamma^2}{2} d'\nabla^2 f(x+t\gamma d)d,$$

and

$$f(x) = f(x+\gamma d) - \gamma d'\nabla f(x+\gamma d) + \frac{\gamma^2}{2} d'\nabla^2 f(x+s\gamma d)d,$$

for some t and s belonging to $[0,1]$. By adding these two equations and using Eq. (B.4), we obtain

$$\frac{\gamma^2}{2} d'\big(\nabla^2 f(x+s\gamma d)+\nabla^2 f(x+t\gamma d)\big)d = \big(\nabla f(x+\gamma d)-\nabla f(x)\big)'(\gamma d) \geq \sigma\gamma^2\|d\|^2.$$

We divide both sides by γ^2 and then take the limit as $\gamma \to 0$ to conclude that $d'\nabla^2 f(x)d \geq \sigma\|d\|^2$. Since this inequality was proved for every $d \in \Re^n$, it follows that $\nabla^2 f(x) - \sigma I$ is positive semidefinite.

Conversely, assume that $\nabla^2 f(x) - \sigma I$ is positive semidefinite for all $x \in \Re^n$. Fix some $x, y \in \Re^n$ such that $x \neq y$, and consider the function $g : [0,1] \mapsto \Re$ defined by

$$g(t) = \nabla f\big(tx + (1-t)y\big)'(x-y).$$

Using the Mean Value Theorem (Prop. A.22 in Appendix A), we have

$$\big(\nabla f(x)-\nabla f(y)\big)'(x-y) = g(1) - g(0) = \frac{dg(t)}{dt}$$

for some $t \in [0,1]$. Since $\nabla^2 f\big(tx+(1-t)y\big) - \sigma I$ is positive semidefinite, we have

$$\frac{dg(t)}{dt} = (x-y)'\nabla^2 f\big(tx+(1-t)y\big)(x-y) \geq \sigma\|x-y\|^2.$$

By combining the preceding two relations, we obtain Eq. (B.4). **Q.E.D.**

Convex and Affine Hulls

Let X be a subset of $\Re^n$. A *convex combination* of elements of X is a vector of the form $\sum_{i=1}^m \alpha_i x_i$, where $x_1, \ldots, x_m$ belong to X and $\alpha_1, \ldots, \alpha_m$ are scalars such that

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

The *convex hull* of X, denoted $\mathrm{conv}(X)$, is the set of all convex combinations of elements of X. In particular, if X consists of a finite number of vectors $x_1, \ldots, x_m$, its convex hull is

$$\mathrm{conv}\big(\{x_1, \ldots, x_m\}\big) = \left\{ \sum_{i=1}^m \alpha_i x_i \;\middle|\; \alpha_i \geq 0,\, i = 1, \ldots, m,\, \sum_{i=1}^m \alpha_i = 1 \right\}.$$

It is straightforward to verify that $\mathrm{conv}(X)$ is a convex set, and using this, to assert that $\mathrm{conv}(X)$ is the intersection of all convex sets containing X.

We recall that a linear manifold M is a set of the form $x + S = \{z \mid z - x \in S\}$, where S is a subspace, called the subspace parallel to M. If S is a subset of $\Re^n$, the *affine hull* of S, denoted $\mathrm{aff}(S)$, is the intersection of all linear manifolds containing S. Note that $\mathrm{aff}(S)$ is itself a linear manifold and that it contains $\mathrm{conv}(S)$. It can be seen that the affine hull of S and the affine hull of $\mathrm{conv}(S)$ coincide.

Given a nonempty subset X of $\Re^n$, a *nonnegative combination* of elements of X is a vector of the form $\sum_{i=1}^m \alpha_i x_i$, where m is a positive integer, $x_1, \ldots, x_m$ belong to X, and $\alpha_1, \ldots, \alpha_m$ are nonnegative scalars. If the scalars α_i are all positive, $\sum_{i=1}^m \alpha_i x_i$ is said to be a *positive combination*. A set $C \subset \Re^n$ is said to be a *cone* if $ax \in C$ for all $a > 0$ and $x \in C$. The *cone generated by* X, denoted $\mathrm{cone}(X)$, is the set of all nonnegative combinations of elements of X. It is easily seen that $\mathrm{cone}(X)$ is a convex cone containing the origin, although it need not be closed even if X is compact.

The following is a fundamental characterization of convex hulls.

Proposition B.6: (Caratheodory's Theorem) Let X be a nonempty subset of $\Re^n$.

(a) Every nonzero vector from $\mathrm{cone}(X)$ can be represented as a positive combination of linearly independent vectors from X.

(b) Every vector from $\mathrm{conv}(X)$ can be represented as a convex combination of no more than $n + 1$ vectors from X.

Proof: See Prop. 1.2.1 and Section 1.2 of [Ber09]. **Q.E.D.**

Closure and Continuity Properties

We now explore some topological properties of convex sets and functions. Let C be a convex subset of $\Re^n$. We say that x is a *relative interior point* of C, if $x \in C$ and there exists a neighborhood N of x such that $N \cap \text{aff}(C) \subset C$, i.e., if x is an interior point of C relative to aff(C). The *relative interior of* C, denoted ri(C), is the set of all relative interior points of C. For example, if C is a line segment connecting two distinct points in the plane, then ri(C) consists of all points of C except for the end points.

Proposition B.7: Let C be a nonempty convex set.

(a) (*Line Segment Principle*) If $x \in \text{ri}(C)$ and $\bar{x} \in \text{cl}(C)$, then all points on the line segment connecting x and $\bar{x}$, except possibly $\bar{x}$, belong to ri(C).

(b) (*Nonemptiness of Relative Interior*) ri(C) is a nonempty convex set, and has the same affine hull as C. In fact, if m is the dimension of aff(C) and $m > 0$, there exist vectors $x_0, x_1, \ldots, x_m \in \text{ri}(C)$ such that $x_1 - x_0, \ldots, x_m - x_0$ span the subspace parallel to aff(C).

(c) (*Prolongation Lemma*) $x \in \text{ri}(C)$ if and only if every line segment in C having x as one endpoint can be prolonged beyond x without leaving C [i.e., for every $\bar{x} \in C$, there exists a $\gamma > 1$ such that $x + (\gamma - 1)(x - \bar{x}) \in C$].

Proof: See Props. 1.3.1-1.3.3 and Section 1.3 of [Ber09]. **Q.E.D.**

An important property of the closure of a convex set C is that it does not "differ" much from C, in the sense that cl(C) and C have the same relative interior. (This is not true for a nonconvex set; take for example the set of rational numbers.) The next proposition proves this property, together with some additional related facts.

Proposition B.8: (Properties of Closure and Relative Interior)

(a) The closure cl(C) and the relative interior ri(C) of a convex set C are convex. Furthermore $\text{ri}\big(\text{cl}(C)\big) = \text{ri}(C)$.

(b) For a convex set C, we have $\text{cl}(C) = \text{cl}\big(\text{ri}(C)\big)$.

(c) Let C and $\bar{C}$ be nonempty convex sets. Then the following three conditions are equivalent:

(i) C and $\bar{C}$ have the same relative interior.

(ii) C and $\bar{C}$ have the same closure.

(iii) $\mathrm{ri}(C) \subset \bar{C} \subset \mathrm{cl}(C)$.

(d) The vector sum of two closed convex sets at least one of which is compact, is a closed convex set.

(e) The image of a convex and compact set under a linear transformation is a convex and compact set.

(f) The convex hull of a compact set is compact.

(g) If C_1 and C_2 are convex sets then

$$\mathrm{ri}(C_1 \times C_2) = \mathrm{ri}(C_1) \times \mathrm{ri}(C_2).$$

Moreover, if $\mathrm{ri}(C_1)$ and $\mathrm{ri}(C_2)$ have a nonempty intersection, then

$$\mathrm{ri}(C_1 + C_2) = \mathrm{ri}(C_1) + \mathrm{ri}(C_2), \quad \mathrm{ri}(C_1 \cap C_2) = \mathrm{ri}(C_1) \cap \mathrm{ri}(C_2).$$

Proof: See Section 1.3.1 of [Ber09]. **Q.E.D.**

An important property of real-valued convex functions over $\Re^n$ is that they are continuous. Extended real-valued convex functions also have interesting continuity properties; see [Ber09], Sections 1.3.2, 1.3.3, for a fuller account. We have the following proposition.

Proposition B.9: (Continuity of a Convex Function) If $f : \Re^n \mapsto \Re$ is convex, then it is continuous. More generally, if $C \subset \Re^n$ is convex and $f : C \mapsto \Re$ is convex, then f is continuous in the relative interior of C.

Proof: See Section 1.3.2 of [Ber09]. **Q.E.D.**

Another important fact is that in order for all of the level sets of a closed convex function to be compact, it is sufficient that one of its nonempty level sets be compact. This follows from the theory of directions of recession (the specialization to convex functions of the notions of asymptotic sequences and asymptotic directions of Section 3.1.2). This theory is developed in Sections 1.4 and 3.2 of [Ber09], but will not be needed in this book. The following proposition is sufficient for our purposes.

Proposition B.10: (Nonemptiness and Compactness of the Set of Minimizing Points)

(a) The set of minimizing points of a convex function $f : \Re^n \mapsto \Re$ over a closed convex set X is nonempty and compact if and only if all its level sets,

$$L_a = \big\{x \in X \mid f(x) \leq a\big\}, \qquad a \in \Re,$$

are compact.

(b) The set of minimizing points over a closed convex set X of a sum $f_1 + \cdots + f_m$, where $f_1, \ldots, f_m$ are real-valued convex functions on $\Re^n$, is nonempty and compact if either X is compact, or if one of the functions is coercive (for example it is positive definite quadratic).

Proof: See Section 1.4 and Prop. 3.2.3 of [Ber09]. **Q.E.D.**

B.2 HYPERPLANES

A *hyperplane* in $\Re^n$ is a set of the form $\{x \mid a'x = b\}$, where a is nonzero vector in $\Re^n$ and b is a scalar. If $\bar{x}$ is any vector in a hyperplane $H = \{x \mid a'x = b\}$, then we must have $a'\bar{x} = b$, so the hyperplane can be equivalently described as

$$H = \{x \mid a'x = a'\bar{x}\},$$

or

$$H = \bar{x} + \{x \mid a'x = 0\}.$$

Thus, H is an affine set that is parallel to the subspace $\{x \mid a'x = 0\}$. The vector a is orthogonal to this subspace, and consequently, a is called the *normal* vector of H; see Fig. B.1.

The sets

$$\{x \mid a'x \geq b\}, \qquad \{x \mid a'x \leq b\},$$

are called the *closed halfspaces* associated with the hyperplane (also referred to as the *positive and negative halfspaces*, respectively). The sets

$$\{x \mid a'x > b\}, \qquad \{x \mid a'x < b\},$$

are called the *open halfspaces* associated with the hyperplane.

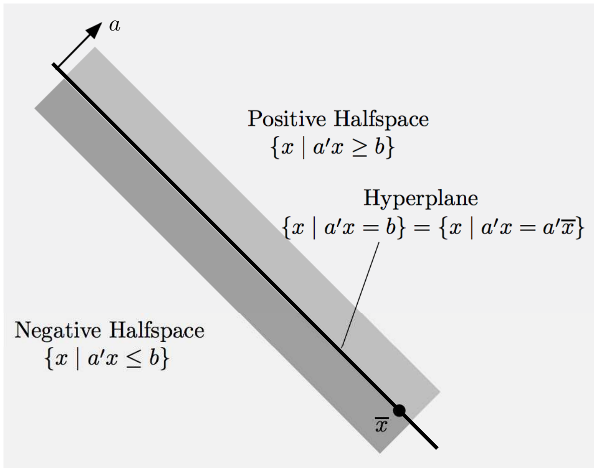

Figure B.1. Illustration of the hyperplane $H = \{x \mid a'x = b\}$. If $\bar{x}$ is any vector in the hyperplane, then the hyperplane can be equivalently described as

$$H = \{x \mid a'x = a'\bar{x}\} = \bar{x} + \{x \mid a'x = 0\}.$$

The hyperplane divides the space into two halfspaces as illustrated.

Proposition B.11: (Supporting Hyperplane Theorem) If $C \subset \Re^n$ is a convex set and $\bar{x}$ is a point that does not belong to the interior of C, there exists a vector $a \neq 0$ such that

$$a'x \geq a'\bar{x}, \qquad \forall\ x \in C.$$

Proof: See Prop. 1.5.1 of [Ber09]. **Q.E.D.**

Proposition B.12: (Separating Hyperplane Theorem) If C_1 and C_2 are two nonempty and disjoint convex subsets of $\Re^n$, there exists a hyperplane that separates them, i.e., a vector $a \neq 0$ such that

$$a'x_1 \leq a'x_2, \qquad \forall\ x_1 \in C_1,\ x_2 \in C_2.$$

Proof: See Prop. 1.5.2 of [Ber09]. **Q.E.D.**

Proposition B.13: (Strict Separation Theorem) If C_1 and C_2 are two nonempty and disjoint convex sets such that C_1 is closed and C_2 is compact, there exists a hyperplane that strictly separates them, i.e., a vector $a \neq 0$ and a scalar b such that

$$a'x_1 < b < a'x_2, \qquad \forall\ x_1 \in C_1,\ x_2 \in C_2.$$

Proof: See Prop. 1.5.3 of [Ber09]. **Q.E.D.**

The preceding proposition may be used to provide a fundamental characterization of closed convex sets, namely that *every closed convex set is the intersection of the halfspaces that contain it*. To see this, let C be the set at issue, and note that C is contained in the intersection of the halfspaces that contain C. To show the reverse inclusion, let $x \notin C$. Applying the Strict Separation Theorem (Prop. B.13) to the sets C and $\{x\}$, we see that there exists a halfspace containing C but not containing x. Hence, if $x \notin C$, then x cannot belong to the intersection of the halfspaces containing C, proving the result.

We finally provide a special type of separation theorem that is particularly useful in convex optimization. The proof is somewhat complicated, and can be found in [Roc70] (Ths. 11.3 and 20.2), in [BNO03] (Props. 2.4.6 and 3.5.1), and in [Ber09] (Props 1.5.6 and 1.5.7).

Proposition B.14: (Proper Separation)

(a) Let C_1 and C_2 be two nonempty convex subsets of $\Re^n$. There exists a hyperplane that separates C_1 and C_2, and does not contain both C_1 and C_2 if and only if

$$\mathrm{ri}(C_1) \cap \mathrm{ri}(C_2) = \emptyset.$$

(b) Let C and P be two nonempty convex subsets of $\Re^n$ such that P is the intersection of a finite number of closed halfspaces. There-exists a hyperplane that separates C and P, and does not contain C if and only if

$$\mathrm{ri}(C) \cap P = \emptyset.$$

Proof: See Props 1.5.6 and 1.5.7 of [Ber09]. **Q.E.D.**

B.3 CONES AND POLYHEDRAL CONVEXITY

We now develop some basic results regarding cones and polyhedral sets, in the context of the objectives of this book. A much broader discussion is found in Ch. 2 of [Ber09]. We introduce three important types of cones.

Given a cone C, the cone given by

$$C^\perp = \{y \mid y'x \leq 0,\ \forall\, x \in C\},$$

is called the *polar cone* of C. Note that the polar cone of a subspace is the orthogonal complement, illustrating that the notion of polarity may be viewed as a generalization of the notion of orthogonality.

A cone C is said to be *finitely generated*, if it has the form

$$C = \left\{ x \;\middle|\; x = \sum_{j=1}^{r} \mu_j a_j,\ \mu_j \geq 0,\ j = 1, \ldots, r \right\},$$

where $a_1, \ldots, a_r$ are some vectors.

A cone C is said to be *polyhedral*, if it has the form

$$C = \{x \mid a_j'x \leq 0,\ j = 1, \ldots, r\},$$

where $a_1, \ldots, a_r$ are some vectors.

It is straightforward to show that the polar cone of any cone, as well as all finitely generated and polyhedral cones are convex, by verifying the definition of convexity of Eq. (B.1). Furthermore, polar and polyhedral cones are closed, since they are intersections of closed halfspaces. Finitely generated cones are also closed as shown in part (b) of the following proposition, which also provides some additional important results.

Proposition B.15:

(a) (*Polar Cone Theorem*) For any nonempty closed convex cone C, we have $(C^\perp)^\perp = C$.

(b) Let $a_1, \ldots, a_r$ be vectors of $\Re^n$. Then the finitely generated cone

$$C = \left\{ x \;\middle|\; x = \sum_{j=1}^{r} \mu_j a_j,\ \mu_j \geq 0,\ j = 1, \ldots, r \right\}$$

is closed and its polar cone is the polyhedral cone given by

$$C^\perp = \{x \mid x'a_j \leq 0,\ j = 1, \ldots, r\}.$$

(c) (*Minkowski-Weyl Theorem*) A cone is polyhedral if and only if it is finitely generated.

(d) (*Farkas' Lemma*) Let x, $e_1, \ldots, e_m$, and $a_1, \ldots, a_r$ be vectors of $\Re^n$. We have $x'y \le 0$ for all vectors $y \in \Re^n$ such that

$$y'e_i = 0, \quad \forall\ i = 1, \ldots, m, \qquad y'a_j \le 0, \quad \forall\ j = 1, \ldots, r,$$

if and only if x can be expressed as

$$x = \sum_{i=1}^{m} \lambda_i e_i + \sum_{j=1}^{r} \mu_j a_j,$$

where λ_i and μ_j are some scalars with $\mu_j \ge 0$ for all j.

Proof: See Props. 2.2.1, 2.3.1, and 2.3.2 of [Ber09]. **Q.E.D.**

Polyhedral Sets

A subset of $\Re^n$ is said to be a *polyhedral set* (or *polyhedron*) if it is nonempty and it is the intersection of a finite number of closed halfspaces, i.e., if it is of the form

$$P = \{x \mid a_j'x \le b_j,\ j = 1, \ldots, r\},$$

where a_j are some vectors and b_j are some scalars.

The following is a fundamental result, showing that a polyhedral set can be represented as the sum of a finitely generated cone and the convex hull of a finite set of points. The proof is based on an interesting construction that can be used to translate results about polyhedral cones to results about polyhedral sets.

Proposition B.16: A set P is polyhedral if and only if there exist a nonempty and finite set of vectors $\{v_1, \ldots, v_m\}$, and a finitely generated cone C such that

$$P = \left\{ x \;\middle|\; x = y + \sum_{j=1}^{m} \mu_j v_j,\ y \in C,\ \sum_{j=1}^{m} \mu_j = 1,\ \mu_j \ge 0,\ j = 1, \ldots, m \right\}.$$

Proof: See Prop. 2.3.3 of [Ber09]. **Q.E.D.**

B.4 EXTREME POINTS AND LINEAR PROGRAMMING

A vector x is said to be an *extreme point* of a convex set C if x belongs to C and there do not exist vectors $y, z \in C$, and a scalar $\alpha \in (0, 1)$ such that

$$y \neq x, \qquad z \neq x, \qquad x = \alpha y + (1 - \alpha)z.$$

An equivalent definition is that x cannot be expressed as a convex combination of some vectors of C, all of which are different from x.

An important fact that forms the basis for the simplex method of linear programming, is that if a linear function f attains a minimum over a polyhedral set C having at least one extreme point, then f attains a minimum at some extreme point of C (as well as possibly at some other nonextreme points). We will prove this fact after considering the more general case where f is concave, and C is closed and convex. We first show a preliminary result.

Proposition B.17: Let C be a nonempty, closed, convex set in $\Re^n$.

(a) If H is a hyperplane that passes through a boundary point of C and contains C in one of its halfspaces, then every extreme point of $C \cap H$ is also an extreme point of C.

(b) C has at least one extreme point if and only if it does not contain a line, i.e., a set L of the form $L = \{x + \alpha d \mid \alpha \in \Re\}$ with $d \neq 0$.

Proof: (a) Let $\bar{x}$ be an element of T which is not an extreme point of C. Then we have $\bar{x} = \alpha y + (1 - \alpha)z$ for some $\alpha \in (0, 1)$, and some $y \in C$ and $z \in C$, with $y \neq x$ and $z \neq x$. Since $\bar{x} \in H$, $\bar{x}$ is a boundary point of C, and the halfspace containing C is of the form $\{x \mid a'x \geq a'\bar{x}\}$, where $a \neq 0$. Then $a'y \geq a'\bar{x}$ and $a'z \geq a'\bar{x}$, which in view of $\bar{x} = \alpha y + (1 - \alpha)z$, implies that $a'y = a'\bar{x}$ and $a'z = a'\bar{x}$. Therefore, $y \in T$ and $z \in T$, showing that $\bar{x}$ cannot be an extreme point of T.

(b) Assume that C has an extreme point x and contains a line $L = \{\bar{x} + \alpha d \mid \alpha \in \Re\}$, where $d \neq 0$. We will arrive at a contradiction. For each integer $n > 0$, the vector

$$x_n = \left(1 - \frac{1}{n}\right) x + \frac{1}{n}(\bar{x} + nd) = x + d + \frac{1}{n}(\bar{x} - x)$$

lies in the line segment connecting x and $\bar{x} + nd$, so it belongs to C. Since C is closed, $x + d = \lim_{n \to \infty} x_n$ must also belong to C. Similarly, we show that $x - d$ must belong to C. Thus $x - d$, x, and $x + d$ all belong to C, contradicting the hypothesis that x is an extreme point.

Conversely, we use induction on the dimension of the space to show that if C does not contain a line, it must have an extreme point. This is true in the real line $\Re^1$, so assume it is true in $\Re^{n-1}$. If a nonempty, closed, convex subset C of $\Re^n$ contains no line, it must have some boundary point $\bar{x}$. Take any hyperplane H passing through $\bar{x}$ and containing C in one of its halfspaces. Then, since H is an $(n-1)$-dimensional manifold, the set $C \cap H$ lies in an $(n-1)$-dimensional space and contains no line, so by the induction hypothesis, it must have an extreme point. By part (a), this extreme point must also be an extreme point of C. **Q.E.D.**

Proposition B.18: Let C be a convex subset of $\Re^n$, and let C^* be the set of minima of a concave function $f : C \mapsto \Re$ over C.

(a) If C^* contains a relative interior point of C, then f must be constant over C, i.e., $C^* = C$.

(b) If C is closed and contains at least one extreme point, and C^* is nonempty, then C^* contains some extreme point of C.

Proof: (a) Let x^* belong to $C^* \cap \mathrm{ri}(C)$, and let x be any vector in C. By the prolongation lemma of Prop. B.7(c), there exists a $\gamma > 1$ such that the vector

$$\hat{x} = x^* + (\gamma - 1)(x^* - x)$$

belongs to C, implying that

$$x^* = \frac{1}{\gamma}\hat{x} + \frac{\gamma - 1}{\gamma}x.$$

By the concavity of the function f, we have

$$f(x^*) \geq \frac{1}{\gamma} f(\hat{x}) + \frac{\gamma - 1}{\gamma} f(x),$$

and since $f(\hat{x}) \geq f(x^*)$ and $f(x) \geq f(x^*)$, we obtain

$$f(x^*) \geq \frac{1}{\gamma} f(\hat{x}) + \frac{\gamma - 1}{\gamma} f(x) \geq f(x^*).$$

Hence $f(x) = f(x^*)$.

(b) Let x^* minimize f over C. If $x^* \in \mathrm{ri}(C)$, by part (a), f must be constant over C, so it attains a minimum at an extreme point of C (since C has at least one extreme point by assumption). If $x^* \notin \mathrm{ri}(C)$, then by Prop. B.14(a), there exists a hyperplane H_1 properly separating x^* and C. Since $x^* \in C$, H_1 must contain x^*, so by the proper separation property, H_1

cannot contain C, and it follows that the intersection $C \cap H_1$ has dimension smaller than the dimension of C.

If $x^* \in \mathrm{ri}(C \cap H_1)$, then f must be constant over $C \cap H_1$, so it attains a minimum at an extreme point of $C \cap H_1$ [since C contains an extreme point, it does not contain a line by Prop. B.17(b), and hence $C \cap H_1$ does not contain a line, which implies that $C \cap H_1$ has an extreme point]. By Prop. B.17(a), this optimal extreme point is also an extreme point of C. If $x^* \notin \mathrm{ri}(C \cap H_1)$, there exists a hyperplane H_2 properly separating x^* and $C \cap H_1$. Again, since $x^* \in C \cap H_1$, H_2 contains x^*, so it cannot contain $C \cap H_1$, and it follows that the intersection $C \cap H_1 \cap H_2$ has dimension smaller than the dimension of $C \cap H_1$.

If $x^* \in \mathrm{ri}(C \cap H_1 \cap H_2)$, then f must be constant over $C \cap H_1 \cap H_2$, etc. Since with each new hyperplane, the dimension of the intersection of C with the generated hyperplanes is reduced, this process will be repeated at most n times, until x^* is a relative interior point of some set $C \cap H_1 \cap \cdots \cap H_k$, at which time an extreme point of $C \cap H_1 \cap \cdots \cap H_k$ will be obtained. Through a reverse argument, repeatedly applying Prop. B.17(a), it follows that this extreme point is an extreme point of C. **Q.E.D.**

As a corollary we have the following:

Proposition B.19: Let C be a closed convex set and let $f : C \mapsto \Re$ be a concave function. Assume that for some invertible $n \times n$ matrix A and some $b \in \Re^n$ we have

$$Ax \geq b, \qquad \forall\, x \in C.$$

Then if f attains a minimum over C, it attains a minimum at some extreme point of C.

Proof: Consider the transformation $x = A^{-1}y$ and the problem of minimizing

$$h(y) = f\big(A^{-1}y\big)$$

over $Y = \{y \mid A^{-1}y \in C\}$. The function h is concave over the closed convex set Y. Furthermore, $y \geq b$ for all $y \in Y$, implying that Y does not contain a line, so that by Prop. B.17(b), Y contains an extreme point. If follows from Prop. B.18(b) that h attains a minimum at some extreme point y^* of Y. Then f attains its minimum over C at $x^* = A^{-1}y^*$, while x^* is an extreme point of C, since it can be verified that invertible transformations of sets map extreme points to extreme points. **Q.E.D.**

Extreme Points of Polyhedral Sets

We now consider a polyhedral set P and we characterize the set of its extreme points (also called *vertices*). By Prop. B.16, P can be represented as

$$P = C + \hat{P},$$

where C is a finitely generated cone C and $\hat{P}$ is the convex hull of some vectors $v_1, \ldots, v_m$:

$$\hat{P} = \left\{ x \;\middle|\; x = \sum_{j=1}^{m} \mu_j v_j,\ \sum_{j=1}^{m} \mu_j = 1,\ \mu_j \geq 0,\ j = 1, \ldots, m \right\}.$$

We note that an extreme point $\bar{x}$ of P cannot be of the form $\bar{x} = c + \hat{x}$, where $c \neq 0$, $c \in C$, and $\hat{x} \in \hat{P}$, since in this case $\bar{x}$ would be the midpoint of the line segment connecting the distinct vectors $\hat{x}$ and $2c + \hat{x}$. Therefore, an extreme point of P must belong to $\hat{P}$, and since $\hat{P} \subset P$, it must also be an extreme point of $\hat{P}$. An extreme point of $\hat{P}$ must be one of the vectors $v_1, \ldots, v_m$, since otherwise this point would be expressible as a convex combination of $v_1, \ldots, v_m$. Thus the set of extreme points of P is either empty or finite. Using Prop. B.17(b), it follows that *the set of extreme points of P is nonempty and finite if and only if P contains no line*.

If P is bounded, then we must have $P = \hat{P}$, and it can be shown that P *is equal to the convex hull of its extreme points* (not just the convex hull of the vectors $v_1, \ldots, v_m$). For a sketch of the proof note that if P is represented as

$$P = \text{conv}\big(\{v_1, \ldots, v_m\}\big) + C,$$

where $v_1, \ldots, v_m$ are some vectors and C is a finitely generated cone (cf. Prop. B.16), then the set of extreme points of P is a subset of $\{v_1, \ldots, v_m\}$. The reason is that an extreme point $\bar{x}$ cannot be of the form $\bar{x} = \tilde{x} + y$, where $\tilde{x} \in \text{conv}\big(\{v_1, \ldots, v_m\}\big)$ and $y \neq 0$, $y \in C$, since in this case $\bar{x}$ would be the midpoint of the line segment connecting the distinct vectors $\tilde{x}$ and $\tilde{x} + 2y$. It thus follows that an extreme point must belong to $\text{conv}\big(\{v_1, \ldots, v_m\}\big)$.

The following proposition gives another and more specific characterization of extreme points of polyhedral sets, and is central in the theory of linear programming.

Proposition B.20: Let P be a polyhedral set in $\Re^n$.

(a) If P has the form

$$P = \big\{ x \mid a_j' x \leq b_j,\ j = 1, \ldots, r \big\},$$

where a_j and b_j are given vectors and scalars, respectively, then a vector $v \in P$ is an extreme point of P if and only if the set

$$A_v = \big\{a_j \mid a_j'v = b_j,\, j \in \{1,\ldots,r\}\big\}$$

contains n linearly independent vectors.

(b) If P has the form

$$P = \{x \mid Ax = b,\, x \geq 0\},$$

where A is a given $m \times n$ matrix and b is a given vector, then a vector $v \in P$ is an extreme point of P if and only if the columns of A corresponding to the nonzero coordinates of v are linearly independent.

(c) *(Fundamental Theorem of Linear Programming)* Assume that P has at least one extreme point. Then if a linear function attains a minimum over P, it attains a minimum at some extreme point of P.

Proof: (a) If the set A_v contains fewer than n linearly independent vectors, then the system of equations

$$a_j'w = 0, \qquad \forall\ a_j \in A_v$$

has a nonzero solution $\bar{w}$. For sufficiently small $\gamma > 0$, we have $v + \gamma\bar{w} \in P$ and $v - \gamma\bar{w} \in P$, thus showing that v is not an extreme point. Thus, if v is an extreme point, A_v must contain n linearly independent vectors.

Conversely, suppose that A_v contains a subset $\bar{A}_v$ consisting of n linearly independent vectors. Suppose that for some $y \in P$, $z \in P$, and $\alpha \in (0,1)$, we have $v = \alpha y + (1-\alpha)z$. Then for all $a_j \in \bar{A}_v$, we have

$$b_j = a_j'v = \alpha a_j'y + (1-\alpha)a_j'z \leq \alpha b_j + (1-\alpha)b_j = b_j.$$

Thus v, y, and z are all solutions of the system of n linearly independent equations

$$a_j'w = b_j, \qquad \forall\ a_j \in \bar{A}_v.$$

Hence $v = y = z$, implying that v is an extreme point.

(b) Let k be the number of zero coordinates of v, and consider the matrix $\bar{A}$, which is the same as A except that the columns corresponding to the zero coordinates of v are set to zero. We write P in the form

$$P = \{x \mid Ax \leq b,\, -Ax \leq -b,\, -x \leq 0\},$$

and apply the result of part (a). We obtain that v is an extreme point if and only if $\bar{A}$ contains $n-k$ linearly independent rows, which is equivalent to the $n-k$ nonzero columns of $\bar{A}$ (corresponding to the nonzero coordinates of v) being linearly independent.

(c) Since P is polyhedral, it has a representation

$$P = \{x \mid Ax \geq b\},$$

for some $m \times n$ matrix A and some $b \in \Re^m$. If A had rank less than n, then its nullspace would contain some nonzero vector $\bar{x}$, so P would contain a line parallel to $\bar{x}$, contradicting the existence of an extreme point [cf. Prop. B.17(b)]. Thus A has rank n and hence it must contain n linearly independent rows that constitute an $n \times n$ invertible submatrix $\hat{A}$. If $\hat{b}$ is the corresponding subvector of b, we see that every $x \in P$ satisfies $\hat{A}x \geq \hat{b}$. The result then follows using Prop. B.19. **Q.E.D.**

B.5 DIFFERENTIABILITY ISSUES

Convex functions have interesting differentiability properties, which we discuss in this section. We first consider real-valued functions. Recall that the directional derivative of a function $f : \Re^n \mapsto \Re$ at a point $x \in \Re^n$ in the direction $y \in \Re^n$ is given by

$$f'(x; y) = \lim_{\alpha \downarrow 0} \frac{f(x + \alpha y) - f(x)}{\alpha},$$

provided that the limit exists, in which case we say that f is directionally differentiable at x in the direction y, and we call $f'(x; y)$ the *directional derivative of f at x in the direction y*. We say that f is *directionally differentiable at x* if it is directionally differentiable at x in all directions. Recall also that f is differentiable at x if it is directionally differentiable at x and $f'(x; y)$ is linear, as a function of y, of the form

$$f'(x; y) = \nabla f(x)'y,$$

where $\nabla f(x)$ is the gradient of f at x. It can be shown that if f is differentiable, then its gradient is continuous over $\Re^n$ (see [Ber15a], Exercise 3.4).

Given a convex function $f : \Re^n \mapsto \Re$, we say that a vector $d \in \Re^n$ is a *subgradient* of f at a point $x \in \Re^n$ if

$$f(z) \geq f(x) + (z - x)'d, \qquad \forall\ z \in \Re^n. \tag{B.8}$$

If instead f is a concave function, we say that d is a subgradient of f at x if $-d$ is a subgradient of the convex function $-f$ at x. The set of all subgradients of a convex (or concave) function f at $x \in \Re^n$ is called the *subdifferential* of f at x, and is denoted by $\partial f(x)$.

The next proposition clarifies the relationship between the directional derivative and the subdifferential, and provides some basic properties of subgradients.

Proposition B.21: Let $f : \Re^n \mapsto \Re$ be a convex function. For every $x \in \Re^n$, the following hold:

(a) A vector d is a subgradient of f at x if and only if

$$f'(x; y) \geq y'd, \qquad \forall\, y \in \Re^n.$$

(b) The subdifferential $\partial f(x)$ is a nonempty, convex, and compact set, and there holds

$$f'(x; y) = \max_{d \in \partial f(x)} y'd, \qquad \forall\, y \in \Re^n.$$

Furthermore, if X is a bounded set, the set $\cup_{x \in X} \partial f(x)$ is bounded.

(c) f is differentiable at x with gradient $\nabla f(x)$, if and only if it has $\nabla f(x)$ as its unique subgradient at x. Moreover, if f is differentiable over $\Re^n$, then $\nabla f(\cdot)$ is a continuous function.

(d) If a sequence $\{x_k\}$ converges to x and $d_k \in \partial f(x_k)$ for all k, the sequence $\{d_k\}$ is bounded and each of its limit points is a subgradient of f at x.

(e) If f is equal to the sum $f_1 + \cdots + f_m$ of convex functions $f_j : \Re^n \mapsto \Re$, $j = 1, \ldots, m$, then $\partial f(x)$ is equal to the vector sum $\partial f_1(x) + \cdots + \partial f_m(x)$.

(f) If f is equal to the composition of a convex function $h : \Re^m \mapsto \Re$ and an $m \times n$ matrix A $[f(x) = h(Ax)]$, then $\partial f(x)$ is equal to $A'\partial h(Ax) = \big\{A'g \mid g \in \partial h(Ax)\big\}$.

(g) A vector $x^* \in X$ minimizes f over a convex set $X \subset \Re^n$ if and only if there exists a subgradient $d \in \partial f(x^*)$ such that

$$d'(z - x^*) \geq 0, \qquad \forall\, z \in X.$$

Proof: See Props. 3.1.1-3.1.4, and Exercise 3.4 of [Ber15a]. **Q.E.D.**

Note that the necessary condition for optimality of part (g) of the preceding proposition generalizes the optimality condition of Section 1.1 for the case where f is differentiable:

$$\nabla f(x^*)'(z - x^*) \geq 0, \qquad \forall\, z \in X.$$

In the special case where $X = \Re^n$, we obtain a basic necessary and sufficient condition for unconstrained optimality of x^*, namely $0 \in \partial f(x^*)$. This optimality condition is also evident from the subgradient inequality (B.8).

Subdifferential of the Maximum of a Convex Function

A case of great interest in optimization involves functions of the form

$$f(x) = \max_{z \in Z} \phi(x, z).$$

The directional derivative and the subdifferential of f can be described in terms of the directional derivative and the subdifferential of ϕ, evaluated at points $\bar{z}$ where the maximum is attained, as shown by the following proposition.

Proposition B.22: (Danskin's Theorem) Let $Z \subset \Re^m$ be a compact set, and let $\phi : \Re^n \times Z \mapsto \Re$ be continuous and such that $\phi(\cdot, z) : \Re^n \mapsto \Re$ is convex for each $z \in Z$.

(a) The function $f : \Re^n \mapsto \Re$ given by

$$f(x) = \max_{z \in Z} \phi(x, z) \tag{B.9}$$

is convex and has directional derivative given by

$$f'(x; y) = \max_{z \in Z(x)} \phi'(x, z; y),$$

where $\phi'(x, z; y)$ is the directional derivative of the function $\phi(\cdot, z)$ at x in the direction y, and $Z(x)$ is the set of maximizing points in Eq. (B.9)

$$Z(x) = \left\{ \overline{z} \;\middle|\; \phi(x, \overline{z}) = \max_{z \in Z} \phi(x, z) \right\}.$$

In particular, if $Z(x)$ consists of a unique point $\overline{z}$ and $\phi(\cdot, \overline{z})$ is differentiable at x, then f is differentiable at x, and $\nabla f(x) = \nabla_x \phi(x, \overline{z})$, where $\nabla_x \phi(x, \overline{z})$ is the vector with coordinates

$$\frac{\partial \phi(x, \overline{z})}{\partial x_i}, \qquad i = 1, \ldots, n.$$

(b) If $\phi(\cdot, z)$ is differentiable for all $z \in Z$ and $\nabla_x \phi(x, \cdot)$ is continuous on Z for each x, then

$$\partial f(x) = \text{conv}\big\{\nabla_x \phi(x, z) \mid z \in Z(x)\big\}, \qquad \forall\ x \in \Re^n. \tag{B.10}$$

In particular, if ϕ is linear in x for all $z \in Z$, i.e.,

$$\phi(x, z) = a_z' x + b_z, \qquad \forall\ z \in Z,$$

then

$$\partial f(x) = \text{conv}\big\{a_z \mid z \in Z(x)\big\}.$$

Proof: See Prop. 4.5.1 of [BNO03] or Exercise 3.5 of [Ber15a] (with solution included). **Q.E.D.**

The preceding proposition derives its origin from a theorem by Danskin [Dan67] that provides a formula for the directional derivative of the maximum of a (not necessarily convex) directionally differentiable function. When adapted to a convex function f, this formula yields the expression (B.10) for $\partial f(x)$.

Subdifferential of the Expected Value of a Convex Function

Another important subdifferential formula relates to the subgradients of an expected value function

$$f(x) = E\big\{F(x, \omega)\big\},$$

where ω is a random variable taking values in a set Ω, and $F(\cdot, \omega) : \Re^n \mapsto \Re$ is a real-valued convex function such that f is real-valued (note that f is easily verified to be convex). If ω takes a finite number of values with probabilities $p(\omega)$, then the formulas

$$f'(x; d) = E\big\{F'(x, \omega; d)\big\}, \qquad \partial f(x) = E\big\{\partial F(x, \omega)\big\}, \tag{B.11}$$

hold because they can be written in terms of finite sums as

$$f'(x; d) = \sum_{\omega \in \Omega} p(\omega) F'(x, \omega; d), \qquad \partial f(x) = \sum_{\omega \in \Omega} p(\omega) \partial F(x, \omega),$$

so Prop. B.21(e) applies. However, the formulas (B.11) hold even in the case where Ω is uncountably infinite, with appropriate mathematical interpretation of the integral of set-valued functions $E\big\{\partial F(x, \omega)\big\}$ as the set of integrals

$$\int_{\omega \in \Omega} g(x, \omega)\, dP(\omega), \tag{B.12}$$

where $g(x, \omega) \in \partial F(x, \omega)$, $\omega \in \Omega$ (measurability issues must be addressed in this context). For a formal proof and analysis, see the author's papers [Ber72], [Ber73], which also provide a necessary and sufficient condition for f to be differentiable, even when $F(\cdot, \omega)$ is not. In this connection, it is important to note that the integration over ω in Eq. (B.12) may smooth out the nondifferentiabilities of $F(\cdot, \omega)$ if ω is a "continuous" random variable. This property can be used in turn in algorithms, including schemes that bring to bear the methodology of differentiable optimization.

Subgradients of Extended Real-Valued Convex Functions

The notion of a subdifferential and a subgradient of a convex extended real-valued function $f : \Re^n \mapsto (-\infty, \infty]$ can be developed along the lines of the present section. In particular, a vector d is a subgradient of f at a vector x such that $f(x) < \infty$ if the subgradient inequality holds, i.e.,

$$f(z) \geq f(x) + (z - x)'d, \qquad \forall\, z \in \Re^n. \tag{B.13}$$

The subdifferential $\partial f(x)$ is the set of all subgradients of the convex function f. By convention, $\partial f(x)$ is considered empty for all x with $f(x) = \infty$.

Note that $\partial f(x)$ is always a closed set, since for any x with $f(x) < \infty$, it is the set of all d that lie in the intersection of the infinite collection of closed halfspaces defined by Eq. (B.13). However, contrary to the case of real-valued functions, $\partial f(x)$ may be empty, or closed but unbounded, even if $f(x) < \infty$. For example, the subdifferential of the extended real-valued convex function

$$f(x) = \begin{cases} -\sqrt{x} & \text{if } 0 \leq x \leq 1, \\ \infty & \text{otherwise,} \end{cases}$$

is given by

$$\partial f(x) = \begin{cases} -\frac{1}{2\sqrt{x}} & \text{if } 0 < x < 1, \\ [-1/2, \infty) & \text{if } x = 1, \\ \varnothing & \text{if } x \leq 0 \text{ or } 1 < x. \end{cases}$$

Thus, $\partial f(x)$ can be empty and can be unbounded at points x that belong to the effective domain of f (as in the cases $x = 0$ and $x = 1$, respectively, of the above example). However, it can be shown that $\partial f(x)$ is nonempty and compact at points x that are *interior* points of the effective domain of f, as also illustrated by the above example. Also $\partial f(x)$ is nonempty at points x that are *relative interior* points of the effective domain of f. These facts are shown in [Ber09], Prop. 5.4.1.

There are generalized versions of some of the preceding results within the context of extended real-valued convex functions, but with appropriate adjustments and additional assumptions to deal with cases where $\partial f(x)$ may be empty or noncompact. For example the sum differentiation formula

$$\partial(f_1 + \cdots + f_m)(x) = \partial f_1(x) + \cdots + \partial f_m(x)$$

[cf. Prop. B.21(e)] may fail even for x in the effective domain of $f_1+\cdots+f_m$; a condition such as that the relative interiors of the effective domains of the extended real-valued convex functions $f_1,\ldots,f_m$ have a point in common is necessary for the formula to hold for all $x\in\Re^n$ (see the books [Roc70] and [Ber09]). There is a similar result for the subdifferential of the composition $f(x)=h(Ax)$ [cf. Prop. B.21(f)], for the case where h is extended real-valued convex and A is a matrix: we have

$$\partial f(x)=A'\partial h(Ax),\qquad \forall\ x\in\Re^n,$$

if the range of A contains a point in the relative interior of $\mathrm{dom}(h)$.

Danskin's Theorem for Extended Real-Valued Convex Functions

Let us finally note an extension of Danskin's Theorem [Prop. B.22(b)], which provides a more general formula for the subdifferential $\partial f(x)$ of the function

$$f(x)=\sup_{z\in Z}\phi(x,z), \tag{B.14}$$

where Z is a compact set. This version of the theorem does not require that $\phi(\cdot,z)$ is differentiable. Instead it assumes that $\phi(\cdot,z)$ is an extended real-valued closed proper convex function for each $z\in Z$, that $\mathrm{int}\big(\mathrm{dom}(f)\big)$ [the interior of the set $\mathrm{dom}(f)=\big\{x\mid f(x)<\infty\big\}$] is nonempty, and that ϕ is continuous on the set $\mathrm{int}\big(\mathrm{dom}(f)\big)\times Z$. Then for all $x\in\mathrm{int}\big(\mathrm{dom}(f)\big)$, we have

$$\partial f(x)=\mathrm{conv}\big\{\partial\phi(x,z)\mid z\in Z(x)\big\}, \tag{B.15}$$

where $\partial\phi(x,z)$ is the subdifferential of $\phi(\cdot,z)$ at x for any $z\in Z$, and $Z(x)$ is the set of maximizing points in Eq. (B.14); for a formal statement and proof of this result, see Prop. A.22 of the author's Ph.D. thesis, which may be found on-line [Ber71].

Note that the nonemptiness of $\mathrm{int}\big(\mathrm{dom}(f)\big)$ is an essential assumption for the formula (B.15) to hold. In particular, the formula may not hold if instead we just assume that the relative interior of $\mathrm{dom}(f)$ is nonempty. For an example, consider the two spheres in $\Re^2$

$$S_1=\big\{(x_1,x_2)\mid (x_1-1)^2+x_2^2\le 1\big\},\qquad S_2=\big\{(x_1,x_2)\mid (x_1+1)^2+x_2^2\le 1\big\},$$

let f_1 and f_2 be the indicator functions of S_1 and S_2, respectively,

$$f_1(x)=\begin{cases}0 & \text{if } x\in S_1,\\ \infty & \text{if } x\notin S_1,\end{cases},\qquad f_2(x)=\begin{cases}0 & \text{if } x\in S_2,\\ \infty & \text{if } x\notin S_2,\end{cases}$$

and let

$$f(x)=\max\big\{f_1(x),f_2(x)\big\}=\begin{cases}0 & \text{if } x=0,\\ \infty & \text{if } x\neq 0.\end{cases}$$

Then it can be seen that the formula (B.15) does not hold at $x=0$.

APPENDIX C:
Line Search Methods

In this appendix we describe algorithms for one-dimensional minimization. These are iterative algorithms, used to implement (approximately) the line minimization stepsize rules.

We briefly present three practical methods. The first two use polynomial interpolation, one requiring derivatives, the second only function values. The third, the Golden Section method, also requires just function values. By contrast with the interpolation methods, it does not depend on the existence of derivatives of the minimized function and may be applied even to discontinuous functions. Its validity depends, however, on a certain unimodality assumption.

In our presentation of the interpolation methods, we consider minimization of the function

$$g(\alpha) = f(x + \alpha d),$$

where f is continuously differentiable. By the chain rule, we have

$$g'(\alpha) = \frac{dg(\alpha)}{d\alpha} = \nabla f(x + \alpha d)'d.$$

We assume that

$$g'(0) = \nabla f(x)'d < 0,$$

i.e., that d is a descent direction at x. We give no convergence or rate of convergence results, but under some fairly natural assumptions, it can be shown that the interpolation methods converge superlinearly.

C.1 CUBIC INTERPOLATION

The cubic interpolation method successively determines at each iteration an appropriate interval $[a, b]$ within which a local minimum of g is guaranteed to exist. It then fits a cubic polynomial to the values $g(a)$, $g(b)$, $g'(a)$, $g'(b)$. The minimizing point $\bar{\alpha}$ of this cubic polynomial lies within $[a, b]$ and replaces one of the two points a or b for the next iteration.

Cubic Interpolation

Step 1: (Determination of the Initial Interval) Let $s > 0$ be some scalar. (Note: If d "approximates well" the Newton direction, then we take $s = 1$.) Evaluate $g(\alpha)$ and $g'(\alpha)$ at the points $\alpha = 0$, s, $2s$, $4s$, $8s, \ldots$, until two successive points a and b are found such that either $g'(b) \geq 0$ or $g(b) \geq g(a)$. Then, it can be seen that a local minimum of g exists within the interval $(a, b]$. [Note: If $g(s)$ is "much larger" than $g(0)$, it is advisable to replace s by βs, where $\beta \in (0, 1)$, for example $\beta = \frac{1}{2}$ or $\beta = \frac{1}{5}$, and repeat this step.] One can show that this step can be carried out if $\lim_{\alpha \to \infty} g(\alpha) > g(0)$.

Step 2: (Updating of the Current Interval) Given the current interval $[a, b]$, a cubic polynomial is fitted to the four values $g(a)$, $g'(a)$, $g(b)$, $g'(b)$. The cubic can be shown to have a unique minimum $\bar{\alpha}$ in the interval $(a, b]$ given by

$$\bar{\alpha} = b - \frac{g'(b) + w - z}{g'(b) - g'(a) + 2w}(b - a),$$

where

$$z = \frac{3\big(g(b) - g(a)\big)}{b - a} + g'(a) + g'(b),$$

$$w = \sqrt{z^2 - g'(a)g'(b)}.$$

If $g'(\bar{\alpha}) \geq 0$ or $g(\bar{\alpha}) \geq g(a)$ replace b by $\bar{\alpha}$. If $g'(\bar{\alpha}) < 0$ and $g(\bar{\alpha}) < g(a)$ replace a by $\bar{\alpha}$. (Note: In practice the computation is terminated once the length of the current interval becomes smaller than a prespecified tolerance or else we obtain $\bar{\alpha} = b$.)

C.2 QUADRATIC INTERPOLATION

This method uses three points a, b, and c such that $a < b < c$, and $g(a) > g(b)$ and $g(b) < g(c)$. Such a set of points is referred to as a *three-point pattern*. It can be seen that a local minimum of g must lie between the extreme points a and c of a three-point pattern a, b, c. At each iteration, the method fits a quadratic polynomial to the three values $g(a)$, $g(b)$, and $g(c)$, and replaces one of the points a, b, and c by the minimizing point of this quadratic polynomial (see Fig. C.1).

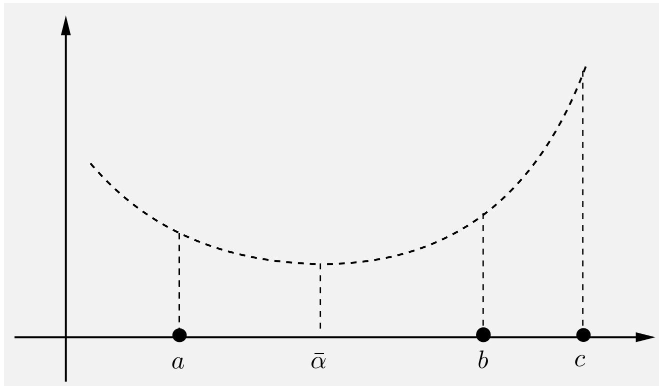

Figure C.1. A three-point pattern and the associated quadratic polynomial. If $\bar{\alpha}$ minimizes the quadratic, a new three point pattern is obtained using $\bar{\alpha}$ and two of the three points a, b, and c ($\bar{\alpha}$, and a, b in the example of the figure).

Quadratic Interpolation

Step 1: (Determination of Initial Three-Point Pattern) We search along the line as in the cubic interpolation method until we find three successive points a, b, and c with $a < b < c$ such that $g(a) > g(b)$ and $g(b) < g(c)$. As for the cubic interpolation method, we assume that this stage can be carried out, and we can show that this is guaranteed if $\lim_{\alpha\to\infty} g(\alpha) > g(0)$.

Step 2: (Updating the Current Three-Point Pattern) Given the current three-point pattern a, b, c, we fit a quadratic polynomial to the values $g(a)$, $g(b)$, and $g(c)$, and we determine its unique minimum $\bar{\alpha}$. It can be shown that $\bar{\alpha} \in (a, c)$ and that

$$\bar{\alpha} = \frac{1}{2}\frac{g(a)(c^2 - b^2) + g(b)(a^2 - c^2) + g(c)(b^2 - a^2)}{g(a)(c - b) + g(b)(a - c) + g(c)(b - a)}.$$

Then, we form a new three-point pattern as follows. If $\bar{\alpha} > b$, we replace a or c by $\bar{\alpha}$ depending on whether $g(\bar{\alpha} < g(b)$ or $g(\bar{\alpha}) > g(b)$, respectively. If $\bar{\alpha} < b$, we replace c or a by $\bar{\alpha}$ depending on whether $g(\bar{\alpha}) < g(b)$ or $g(\bar{\alpha}) > g(b)$, respectively. [Note: If $g(\bar{\alpha}) = g(b)$ then a special local search near $\bar{\alpha}$ should be conducted to replace $\bar{\alpha}$ by a point $\bar{\alpha}'$ with $g(\bar{\alpha}') \neq g(b)$. The computation is terminated when the length of the three-point pattern is smaller than a certain tolerance.]

An alternative possibility for quadratic interpolation is to determine the minimum $\bar{a}$ of the quadratic polynomial that has the same value as g at the points 0 and a, and the same first derivative as g at 0. It can be

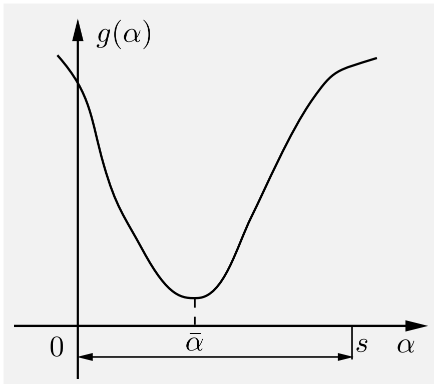

Figure C.2. A strictly unimodal function g over an interval $[0, s]$ is defined as a function that has a unique global minimum α^* in $[0, s]$ and if α_1, α_2 are two points in $[0, s]$ such that $\alpha_1 < \alpha_2 < \alpha^*$ or $\alpha^* < \alpha_1 < \alpha_2$, then

$$g(\alpha_1) > g(\alpha_2) > g(\alpha^*)$$

or

$$g(\alpha^*) < g(\alpha_1) < g(\alpha_2),$$

respectively. An example of a strictly unimodal function, is a function which is strictly convex over $[0, s]$.

verified that this minimum is given by

$$\bar{a} = \frac{g'(0)a^2}{2\big(g'(0)a + g(0) - g(a)\big)}.$$

C.3 THE GOLDEN SECTION METHOD

Here, we assume that $g(\alpha)$ is *strictly unimodal* in the interval $[0, s]$, as defined in Fig. C.2. The Golden Section method minimizes g over $[0, s]$ by determining at the kth iteration an interval $[\alpha_k, \bar{\alpha}_k]$ containing α^*. These intervals are obtained using the number

$$\tau = \frac{3 - \sqrt{5}}{2},$$

which satisfies $\tau = (1-\tau)^2$ and is related to the Fibonacci number sequence. The significance of this number will be seen shortly.

Initially, we take

$$[\alpha_0, \bar{\alpha}_0] = [0, s].$$

Given $[\alpha_k, \bar{\alpha}_k]$, we determine $[\alpha_{k+1}, \bar{\alpha}_{k+1}]$ so that $\alpha^* \in [\alpha_{k+1}, \bar{\alpha}_{k+1}]$ as follows. We calculate

$$b_k = \alpha_k + \tau(\bar{\alpha}_k - \alpha_k)$$
$$\bar{b}_k = \bar{\alpha}_k - \tau(\bar{\alpha}_k - \alpha_k)$$

and $g(b_k)$, $g(\bar{b}_k)$. Then:

(1) If $g(b_k) < g(\bar{b}_k)$ we set

$$\alpha_{k+1} = \alpha_k, \quad \bar{\alpha}_{k+1} = b_k \qquad \text{if} \qquad g(\alpha_k) \le g(b_k)$$

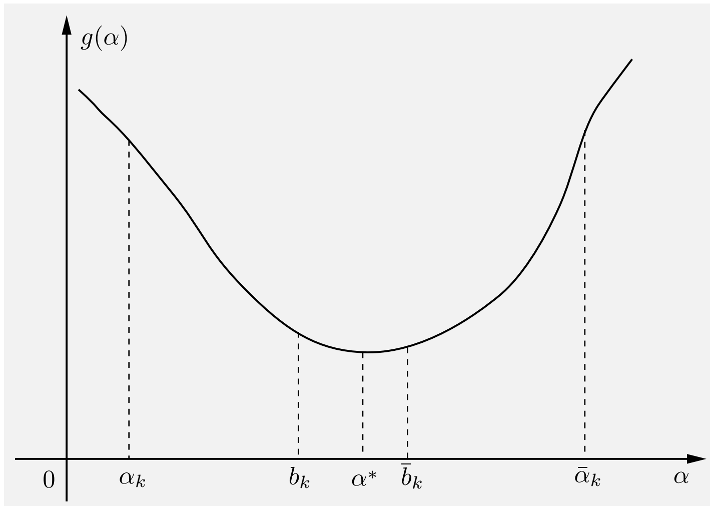

Figure C.3. Golden Section search. Given the interval $[\alpha_k, \bar{\alpha}_k]$ containing the minimum α^*, we calculate

$$b_k = \alpha_k + \tau(\bar{\alpha}_k - \alpha_k)$$

and

$$\bar{b}_k = \bar{\alpha}_k - \tau(\bar{\alpha}_k - \alpha_k).$$

The new interval $[\alpha_{k+1}, \bar{\alpha}_{k+1}]$ has either b_k or $\bar{b}_k$ as one of its endpoints.

$$\alpha_{k+1} = \alpha_k, \quad \bar{\alpha}_{k+1} = \bar{b}_k \qquad \text{if} \qquad g(\alpha_k) > g(b_k).$$

(2) If $g(b_k) > g(\bar{b}_k)$ we set

$$\alpha_{k+1} = \bar{b}_k, \quad \bar{\alpha}_{k+1} = \bar{\alpha}_k \qquad \text{if} \qquad g(\bar{b}_k) \geq g(\bar{\alpha}_k)$$

$$\alpha_{k+1} = b_k, \quad \bar{\alpha}_{k+1} = \bar{a}_k \qquad \text{if} \qquad g(\bar{b}_k) < g(\alpha_k).$$

(3) If $g(b_k) = g(\bar{b}_k)$ we set

$$\alpha_{k+1} = b_k, \qquad \bar{\alpha}_{k+1} = \bar{b}_k.$$

Based on the definition of a strictly unimodal function it can be shown (see Fig. C.3) that the intervals $[\alpha_k, \bar{\alpha}_k]$ contain α^* and their lengths converge to zero. In practice, the computation is terminated once $(\bar{\alpha}_k - \alpha_k)$ becomes smaller than a prespecified tolerance.

An important fact, which rests on the choice of the particular number τ is that

$$[\alpha_{k+1}, \bar{\alpha}_{k+1}] = [\alpha_k, \bar{b}_k] \qquad \Longrightarrow \qquad \bar{b}_{k+1} = b_k,$$

$$[\alpha_{k+1}, \bar{\alpha}_{k+1}] = [b_k, \bar{\alpha}_k] \qquad \Longrightarrow \qquad b_{k+1} = \bar{b}_k.$$

In other words, a trial point b_k or $\bar{b}_k$ that is not used as the end point of the next interval continues to be a trial point for the next iteration. The reader can verify this, using the property

$$\tau = (1 - \tau)^2.$$

Thus, in either of the above situations, the values $\bar{b}_{k+1}$, $g(\bar{b}_{k+1})$ or b_{k+1}, $g(b_{k+1})$ are available and need not be recomputed at the next iteration, requiring a single function evaluation instead of two.

APPENDIX D: *Implementation of Newton's Method*

In this appendix we describe a globally convergent version of Newton's method based on the modified Cholesky factorization approach discussed in Section 1.4. A computer code implementing the method can be freely obtained from the author's web page or through the book's web page.

D.1 CHOLESKY FACTORIZATION

We will give an algorithm for factoring a positive definite symmetric matrix A as

$$A = LL',$$

where L is lower triangular. This is the *Cholesky factorization*. Let a_{ij} be the elements of A and let A_i be the ith leading principal submatrix of A, i.e., the submatrix

$$A_i = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1i} \\ a_{21} & a_{22} & \cdots & a_{2i} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ii} \end{bmatrix}.$$

It is seen that this submatrix is positive definite, since for any $y \in \Re_i$, $y \neq 0$, we have by the positive definiteness of A

$$y' A_i y = [\, y' \quad 0\,]\, A \begin{bmatrix} y \\ 0 \end{bmatrix} > 0.$$

The factorization of A is obtained by successive factorization of $A_1, A_2, \ldots$. Indeed we have $A_1 = L_1 L_1'$, where $L_1 = [\sqrt{a_{11}}]$. Suppose we have the Cholesky factorization of A_{i-1},

$$A_{i-1} = L_{i-1} L_{i-1}'. \tag{D.1}$$

Let us write

$$A_i = \begin{bmatrix} A_{i-1} & \beta_i \\ \beta_i' & a_{ii} \end{bmatrix}, \tag{D.2}$$

where β_i is the column vector

$$\beta_i = \begin{bmatrix} a_{1i} \\ \vdots \\ a_{i-1,i} \end{bmatrix}. \tag{D.3}$$

Based on Eqs. (D.1)-(D.3), it can be verified that

$$A_i = L_i L_i',$$

where

$$L_i = \begin{bmatrix} L_{i-1} & 0 \\ l_i' & \lambda_{ii} \end{bmatrix}, \tag{D.4}$$

and

$$l_i = L_{i-1}^{-1} \beta_i, \qquad \lambda_{ii} = \sqrt{a_{ii} - l_i' l_i}. \tag{D.5}$$

The scalar λ_{ii} is well defined because it can be shown that $a_{ii} - l_i' l_i > 0$. This is seen by defining $b = A_{i-1}^{-1} \beta_i$, and by using the positive definiteness of A_i to write

$$\begin{aligned} 0 < [\, b' \quad -1 \,]\, A_i \begin{bmatrix} b \\ -1 \end{bmatrix} &= b' A_{i-1} b - 2b' \beta_i + a_{ii} \\ &= b' \beta_i - 2b' \beta_i + a_{ii} = a_{ii} - b' \beta_i \\ &= a_{ii} - \beta_i' A_{i-1}^{-1} \beta_i = a_{ii} - \beta_i' (L_{i-1} L_{i-1}')^{-1} \beta_i \\ &= a_{ii} - (L_{i-1}^{-1} \beta_i)' (L_{i-1}^{-1} \beta_i) = a_{ii} - l_i' l_i. \end{aligned}$$

The preceding construction can also be used to show that the Cholesky factorization is unique among factorizations involving lower triangular matrices with positive elements along the diagonal. Indeed, A_1 has a unique such factorization, and if A_{i-1} has a unique factorization $A_{i-1} = L_{i-1} L_{i-1}'$, then L_i is uniquely determined from the requirement $A_i = L_i L_i'$ with the diagonal elements of L_i positive, and Eqs. (D.4) and (D.5).

Cholesky Factorization by Columns

In the preceding algorithm, we calculate L by rows, i.e., we first calculate the first row of L, then the second row, etc. An alternative and equivalent method is to calculate L by columns, i.e., first calculate the first column of L, then the second column, etc. To see how this can be done, we note that the first column of A is equal to the first column of L multiplied with l_{11}, i.e.,

$$a_{i1} = l_{11} l_{i1}, \qquad i = 1, \ldots, n,$$

from which we obtain

$$l_{11} = \sqrt{a_{11}},$$

$$l_{i1} = \frac{a_{i1}}{l_{11}}, \qquad i = 2, \ldots, n.$$

Similarly, given columns $1, 2, \ldots, j-1$ of L, we equate the elements of the jth column of A with the corresponding elements of LL' and we obtain the elements of the jth column of L as follows:

$$l_{jj} = \sqrt{a_{jj} - \sum_{m=1}^{j-1} l_{jm}^2},$$

$$l_{ij} = \frac{a_{ij} - \sum_{m=1}^{j-1} l_{jm} l_{im}}{l_{jj}}, \qquad i = j+1, \ldots, n.$$

D.2 APPLICATION TO A MODIFIED NEWTON METHOD

Consider now adding to A a diagonal correction E and simultaneously factoring the matrix

$$F = A + E,$$

where E is such that F is positive definite. The elements of E are introduced sequentially during the factorization process as some diagonal elements of the triangular factor are discovered, which are either negative or are close to zero, indicating that A is either not positive definite or is nearly singular. As discussed in Section 1.4, this is a principal method by which Newton's method is modified to enhance its global convergence properties. The precise mechanization is as follows:

We first fix positive scalars μ_1 and μ_2, where $\mu_1 < \mu_2$. We calculate the first column of the triangular factor L of F by

$$l_{11} = \begin{cases} \sqrt{a_{11}} & \text{if } \mu_1 < a_{11}, \\ \sqrt{\mu_2} & \text{otherwise,} \end{cases}$$

$$l_{i1} = \frac{a_{i1}}{l_{11}}, \qquad i = 2, \ldots, n.$$

Similarly, given columns $1, 2, \ldots, j-1$ of L, we obtain the elements of the jth column from the equations

$$l_{jj} = \begin{cases} \sqrt{a_{jj} - \sum_{m=1}^{j-1} l_{jm}^2} & \text{if } \mu_1 < a_{11} - \sum_{m=1}^{j-1} l_{jm}^2, \\ \sqrt{\mu_2} & \text{otherwise,} \end{cases}$$

$$l_{ij} = \frac{a_{ij} - \sum_{m=1}^{j-1} l_{jm} l_{im}}{l_{jj}}, \qquad i = j+1, \ldots, n.$$

In words, if the diagonal element of LL' comes out less than μ_1, we bring it up to μ_2.

Note that the jth diagonal element of the correction matrix E is equal to zero if $\mu_1 < a_{jj} - \sum_{m=1}^{j-1} l_{jm}^2$ and is equal to

$$\mu_2 - \left(a_{jj} - \sum_{m=1}^{j-1} l_{jm}^2 \right)$$

otherwise.

The preceding scheme can be used to modify Newton's method, where at the kth iteration, we add a diagonal correction Δ^k to the Hessian $\nabla^2 f(x^k)$ and simultaneously obtain the Cholesky factorization $L^k L^{k\prime}$ of $\nabla^2 f(x^k) + \Delta^k$ as described above. A modified Newton direction d^k is then obtained by first solving the triangular system

$$L^k y = -\nabla f(x^k),$$

and then solving the triangular system

$$L^{k\prime} d^k = y.$$

Solving the first system is called *forward elimination* and is accomplished in $O(n^2)$ arithmetic operations using the equations

$$y_1 = -\frac{\partial f(x^k)/\partial x_1}{l_{11}},$$

$$y_i = -\frac{\partial f(x^k)/\partial x_i + \sum_{m=1}^{i-1} l_{im} y^m}{l_{ii}}, \qquad i = 2, \ldots, n,$$

where l_{im} is the imth element of L^k. Solving the second system is called *back substitution* and is accomplished again in $O(n^2)$ arithmetic operations using the equations

$$d^n = \frac{y^n}{l_{nn}},$$

$$d_i = \frac{y_i - \sum_{m=i+1}^{n} l_{mi} d^m}{l_{ii}}, \qquad i = 1, \ldots, n-1.$$

The next point x^{k+1} is obtained from

$$x^{k+1} = x^k + \alpha^k d^k,$$

where α^k is chosen by the Armijo rule with unity initial step whenever the Hessian is not modified ($\Delta^k = 0$) and by means of a line minimization otherwise.

Assuming fixed values of μ_1 and μ_2, the following may be verified for the modified Newton's method just described:

(a) The algorithm is globally convergent in the sense that every limit point of $\{x^k\}$ is a stationary point of f. This can be shown using Prop. 1.2.1 in Section 1.2.

(b) For each local minimum x^* with positive definite Hessian, there exist scalars $\mu > 0$ and $\epsilon > 0$ such that if $\mu_1 \leq \mu$ and $\|x^0 - x^*\| \leq \epsilon$, then $x^k \to x^*$, $\Delta^k = 0$, and $\alpha^k = 1$ for all k. In other words if μ_1 is not chosen too large, the Hessian will never be modified near x^*, the method will be reduced to the pure form of Newton's method, and the convergence to x^* will be superlinear. The theoretical requirement that μ_1 be sufficiently small can be eliminated by making μ_1 dependent on the norm of the gradient (e.g. $\mu_1 = c\|\nabla f(x^k)\|$, where c is some positive scalar).

Practical Choice of Parameters and Stepsize Selection

We now address some practical issues. As discussed earlier, one should try to choose μ_1 small in order to avoid detrimental modification of the Hessian. Some trial and error with one's particular problem may be required here. As a practical matter, we recommend choosing initially $\mu_1 = 0$ and increasing μ_1 only if difficulties arise due to roundoff error or extremely large norm of calculated direction. (Choosing $\mu_1 = 0$, runs counter to our convergence theory because the generated directions are not guaranteed to be gradient related, but the practical consequences of this are typically insignificant.)

The parameter μ_2 should generally be chosen considerably larger than μ_1. It can be seen that choosing μ_2 very small can make the modified Hessian matrix $L^k L^{k\prime}$ nearly singular. On the other hand, choosing μ_2 very large has the effect of making nearly zero the coordinates of d^k that correspond to nonzero diagonal elements of the correction matrix Δ^k. Generally, some trial and error is necessary to determine a proper value of μ_2. A good guideline is to try a relatively small value of μ_2 and to increase μ_2 if the stepsize generated by the line minimization algorithm is substantially smaller than unity. The idea here is that small values of μ_2 tend to produce directions d^k with large value of norm and hence small values of stepsize. Thus a small value of stepsize indicates that μ_2 is chosen smaller than appropriate, and suggests that an increase of μ_2 is desirable. It is also possible to construct along these lines an adaptive scheme that changes the values of μ_1 and μ_2 in the course of the algorithm.

The following scheme to set and adjust μ_1 and μ_2 has worked well for the author. At each iteration k, we determine the maximal absolute diagonal element of the Hessian, i.e.,

$$w^k = \max\left\{\left|\frac{\partial^2 f(x^k)}{(x_1)^2}\right|, \ldots, \left|\frac{\partial^2 f(x^k)}{(x_n)^2}\right|\right\},$$

and we set μ_1 and μ_2 to

$$\mu_1 = r_1 w^k, \qquad \mu_2 = r_2 w^k.$$

The scalar r_1 is set at some "small" (or zero) value. The scalar r_2 is changed each time the Hessian is modified; it is multiplied by 5 if the stepsize obtained by the minimization rule is less than 0.2, and it is divided by 5 each time the stepsize is larger than 0.9.

Finally, regarding stepsize selection, any of a large number of possible line minimization algorithms can be used for those iterations where the Hessian is modified (in other iterations the Armijo rule with unity initial stepsize is used). One possibility is to use quadratic interpolation based on function values; see Section C.2 in Appendix C.

It is worth noting that if the cost function is quadratic, then it can be shown that a unity stepsize results in cost reduction for any values of μ_1 and μ_2. In other words if f is quadratic (not necessarily positive definite), we have

$$f\big(x^k - (F^k)^{-1}\nabla f(x^k)\big) \leq f(x^k),$$

where $F^k = \nabla^2 f(x^k) + \Delta^k$ and Δ^k is any positive definite matrix such that F^k is positive definite. As a result, a stepsize near unity is appropriate for initiating the line minimization algorithm. This fact can be used to guide the implementation of the line minimization routine.

References

[AFB06] Ahn, S., Fessler, J., Blatt, D., and Hero, A. O., 2006. “Convergent Incremental Optimization Transfer Algorithms: Application to Tomography,” IEEE Transactions on Medical Imaging, Vol. 25, pp. 283-296.

[AHR93] Anstreicher, K. M., den Hertog, D., Roos, C., and Terlaky, T., 1993. “A Long Step Barrier Method for Convex Quadratic Programming,” Algorithmica, Vol. 10, pp. 365-382.

[AHR97] Auslender, A., Cominetti, R., and Haddou, M., 1997. “Asymptotic Analysis for Penalty and Barrier Methods in Convex and Linear Programming,” Math. Operations Res., Vol. 22, pp. 43-62.

[AHU58] Arrow, K. J., Hurwicz, L., and Uzawa, H., (Eds.), 1958. Studies in Linear and Nonlinear Programming, Stanford Univ. Press, Stanford, CA.

[AHU61] Arrow, K. J., Hurwicz, L., and Uzawa, H., 1961. “Constraint Qualifications in Maximization Problems,” Naval Research Logistics Quarterly, Vol. 8, pp. 175-191.

[AMO91] Ahuja, R. K., Magnanti, T. L., and Orlin, J. B., 1991. “Some Recent Advances in Network Flows,” SIAM Review, Vol. 33, pp. 175-219.

[AaL97] Aarts, E., and Lenstra, J. K., 1997. Local Search in Combinatorial Optimization, Wiley, N. Y.

[Aba67] Abadie, J., 1967. “On the Kuhn-Tucker Theorem,” in Nonlinear Programming, Abadie, J., (Ed.), North Holland, Amsterdam.

[Ali92] Alizadeh, F., 1992. “Optimization over the Positive-Definite Cone: Interior Point Methods and Combinatorial Applications,” in Pardalos, P., (Ed.), Advances in Optimization and Parallel Computing, North Holland, Amsterdam.

[Ali95] Alizadeh, F., 1995. “Interior-Point Methods in Semidefinite Programming with Applications in Combinatorial Applications,” SIAM J. on Optimization, Vol. 5, pp. 13-51.

[AnH13] Andersen, M. S., and Hansen, P. C., 2013. “Generalized Row-Action Methods for Tomographic Imaging,” Numerical Algorithms, Vol. 67, pp. 1-24.

[AnV94] Anstreicher, K. M., and Vial, J.-P., 1994. “On the Convergence of an Infeasible Primal-Dual Interior-Point Method for Convex Programming,” Optimization Methods and Software, Vol. 3, pp. 273-283.

[Arm66] Armijo, L., 1966. “Minimization of Functions Having Continuous Partial Derivatives,” Pacific J. Math., Vol. 16, pp. 1-3.

[Ash72] Ash, R. B., 1972. Real Analysis and Probability, Academic Press, N. Y.

[AtV95] Atkinson, D. S., and Vaidya, P. M., 1995. “A Cutting Plane Algorithm for Convex Programming that Uses Analytic Centers,” Math. Programming, Vol. 69, pp. 1-44.

[AuC90] Auslender, A., and Cominetti, R., 1990. "First and Second Order Sensitivity Conditions," Optimization, Vol. 21, pp. 1-13.

[AuE76] Aubin, J. P., and Ekeland, I., 1976. "Estimates of the Duality Gap in Nonconvex Optimization," Math. Operations Res., Vol. 1, pp. 225-245.

[AuT03] Auslender, A., and Teboulle, M., 2003. Asymptotic Cones and Functions in Optimization and Variational Inequalities, Springer, N. Y.

[Aus76] Auslender, A., 1976. Optimization: Methodes Numeriques, Mason, Paris.

[Aus92] Auslender, A., 1992. "Asymptotic Properties of the Fenchel Dual Functional and Applications to Decomposition Properties," Vol. 73, pp. 427-449.

[Aus96] Auslender, A., 1996. "Non Coercive Optimization Problems," Math. of Operations Research, Vol. 21, pp. 769-782.

[Aus97] Auslender, A., 1997. "How to Deal with the Unbounded in Optimization: Theory and Algorithms," Math. Programing, Vol. 79, pp. 3-18.

[Avr76] Avriel, M., 1976. Nonlinear Programming: Analysis and Methods, Prentice-Hall, Englewood Cliffs, N. J.

[BCS99] Bonnans, J. F., Cominetti, R., and Shapiro, A., 1999. "Second Order Optimality Conditions Based on Parabolic Second Order Tangent Sets," SIAM J. on Optimization, Vol. 9, pp. 466-492.

[BGG84] Bertsekas, D. P., Gafni, E. M., and Gallager, R. G., 1984. "Second Derivative Algorithms for Minimum Delay Distributed Routing in Networks," IEEE Trans. on Communications, Vol. 32, pp. 911-919.

[BGI95] Burachik, R., Grana Drummond, L. M., Iusem, A. N., and Svaiter, B. F., 1995. "Full Convergence of the Steepest Descent Method with Inexact Line Searches," Optimization, Vol. 32, pp. 137-146.

[BGL06] Bonnans, J. F., Gilbert, J. C., Lemaréchal, C., and Sagastizábal, S. C., 2006. Numerical Optimization: Theoretical and Practical Aspects, Springer, N. Y.

[BGS72] Bazaraa, M. S., Goode, J. J., and Shetty, C. M., 1972. "Constraint Qualifications Revisited," Management Science, Vol. 18, pp. 567-573.

[BGT81] Bland, R. G., Goldfarb, D., and Todd, M. J., 1981. "The Ellipsoid Method: A Survey," Operations Research, Vol. 29, pp. 1039-91.

[BHG08] Blatt, D., Hero, A. O., Gauchman, H., 2008. "A Convergent Incremental Gradient Method with a Constant Step Size," SIAM J. Optimization, Vol. 18, pp. 29-51.

[BHT87] Bertsekas, D. P., Hossein, P., and Tseng, P., 1987. "Relaxation Methods for Network Flow Problems with Convex Arc Costs," SIAM J. on Control and Optimization, Vol. 25, pp. 1219-1243.

[BJS90] Bazaraa, M. S., Jarvis, J. J., and Sherali, H. D., 1990. Linear Programming and Network Flows, 2nd edition, Wiley, N. Y.

[BLY15] Bragin, M. A., Luh, P. B., Yan, J. H., Yu, N., and Stern, G. A., 2015. "Convergence of the Surrogate Lagrangian Relaxation Method," J. of Optimization Theory and Applications, Vol. 164, pp. 173-201.

[BMM95a] Ball, M. O., Magnanti, T. L., Monma, C. L., and Nemhauser, G. L., 1995. Network Models, Handbooks in OR and MS, Vol. 7, North-Holland, Amsterdam.

[BMM95b] Ball, M. O., Magnanti, T. L., Monma, C. L., and Nemhauser, G. L., 1995. Network Routing, Handbooks in OR and MS, Vol. 8, North-Holland, Amsterdam.

[BMR00] Birgin, E. G., Martinez, J. M., and Raydan, M., 2000. "Nonmonotone Spectral

Projected Gradient Methods on Convex Sets," SIAM J. on Optimization, Vol. 10, pp. 1196-1211.

[BMS99] Boltyanski, V., Martini, H., and Soltan, V., 1999. Geometric Methods and Optimization Problems, Kluwer, Boston.

[BMT90] Burke, J. V., Moré, J. J., and Toraldo, G., 1990. "Convergence Properties of Trust Region Methods for Linear and Convex Constraints," Math. Programming, Vol. 47, pp. 305-336.

[BNO03] Bertsekas, D. P., Nedić, A., and Ozdaglar, A. E., 2003. Convex Analysis and Optimization, Athena Scientific, Belmont, MA.

[BOT06] Bertsekas, D. P., Ozdaglar, A. E., and Tseng, P., 2006 "Enhanced Fritz John Optimality Conditions for Convex Programming," SIAM J. on Optimization, Vol. 16, pp. 766-797.

[BPC11] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J., 2011. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers, Now Publishers Inc, Boston, MA.

[BPT92] Bonnans, J. F., Panier, E. R., Tits, A. L., and Zhou, J. L., 1992. "Avoiding the Maratos Effect by Means of a Nonmonotone Line Search II. Inequality Constrained Problems – Feasible Iterates," SIAM J. Numer. Anal., Vol. 29, pp. 1187-1202.

[BPT97a] Bertsekas, D. P., Polymenakos, L. C., and Tseng, P., 1997. "An ϵ-Relaxation Method for Separable Convex Cost Network Flow Problems," SIAM J. on Optimization, Vol. 7, pp. 853-870.

[BPT97b] Bertsekas, D. P., Polymenakos, L. C., and Tseng, P., 1997. "Epsilon-Relaxation and Auction Methods for Separable Convex Cost Network Flow Problems," in Network Optimization, Pardalos, P. M., Hearn, D. W., and Hager, W. W., (Eds.), Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, N. Y., pp. 103-126.

[BSL14] Bergmann, R., Steidl, G., Laus, F., and Weinmann, A., 2014. "Second Order Differences of Cyclic Data and Applications in Variational Denoising," arXiv preprint arXiv:1405.5349.

[BSS93] Bazaraa, M. S., Sherali, H. D., and Shetty, C. M., 1993. Nonlinear Programming Theory and Algorithms, 2nd edition, Wiley, N. Y.

[BST14] Bolte, J., Sabach, S., and Teboulle, M., 2014. "Proximal Alternating Linearized Minimization for Nonconvex and Nonsmooth Problems," Math. Programming, Vol. 146, pp. 1-36.

[BTW82] Boggs, P. T., Tolle, J. W., and Wang, P., 1982. "On the Local Convergence of Quasi-Newton Methods for Constrained Optimization," SIAM J. on Control and Optimization, Vol. 20, pp. 161-171.

[BaB88] Barzilai, J., and Borwein, J. M., 1988. "Two-Point Step Size Gradient Methods," IMA J. Numerical Analysis, Vol. 8, pp. 141-148.

[BaC11] Bauschke, H. H., and Combettes, P. L., 2011. Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer, NY.

[BaD93] Barzilai, J., and Dempster, M. A. H., 1993. "Measuring Rates of Convergence of Numerical Algorithms," J. Opt. Theory and Appl., Vol. 78, pp. 109-125.

[BaL89] Bayer, D. A., and Lagarias, J. C., 1989. "The Nonlinear Geometry of Linear Programming. I. Affine and Projective Scaling Trajectories. II. Legendre Transform Coordinates and Central Trajectories. III. Projective Legendre Transform Coordinates and Hilbert Geometry," Trans. Amer. Math. Soc., Vol. 314, pp. 499-581.

[BaT85] Balas, E., and Toth, P., 1985. "Branch and Bound Methods," in The Traveling

Salesman Problem, Lawler, E., Lenstra, J. K., Rinnoy Kan, A. H. G., and Shmoys, D. B., (Eds.), Wiley, N. Y., pp. 361-401.

[BaW75] Balinski, M., and Wolfe, P., (Eds.), 1975. Nondifferentiable Optimization, Math. Programming Study 3, North-Holland, Amsterdam.

[Bac14] Bacak, M., 2014. "Computing Medians and Means in Hadamard Spaces," arXiv preprint arXiv:1210.2145v3.

[Bac16] Bacak, M., 2016. "A variational Approach to Stochastic Minimization of Convex Functionals," arXiv preprint arXiv:1605.03289.

[BeE88] Bertsekas, D. P., and Eckstein, J., 1988. "Dual Coordinate Step Methods for Linear Network Flow Problems," Math. Programming, Vol. 42, pp. 203-243.

[BeG82] Bertsekas, D. P., and Gafni, E., 1982. "Projection Methods for Variational Inequalities with Application to the Traffic Assignment Problem," Math. Programming Studies, Vol. 17, pp. 139-159.

[BeG83] Bertsekas, D. P., and Gafni, E., 1983. "Projected Newton Methods and Optimization of Multicommodity Flows," IEEE Trans. Automat. Control, Vol. AC-28, pp. 1090-1096.

[BeG92] Bertsekas, D. P., and Gallager, R. G., 1992. Data Networks, 2nd edition, Prentice-Hall, Englewood Cliffs, N. J.

[BeM71] Bertsekas, D. P, and Mitter, S. K., 1971. "Steepest Descent for Optimization Problems with Nondifferentiable Cost Functionals," Proc. 5th Annual Princeton Confer. Inform. Sci. Systems, Princeton, N. J., pp. 347-351.

[BeM73] Bertsekas, D. P., and Mitter, S. K., 1973. "A Descent Numerical Method for Optimization Problems with Nondifferentiable Cost Functionals," SIAM J. on Control, Vol. 11, pp. 637-652.

[BeN01] Ben-Tal, A., and Nemirovski, A., 2001. Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications, SIAM, Philadelphia.

[BeO02] Bertsekas, D. P., and Ozdaglar, A. E., 2002. "Pseudonormality and a Lagrange Multiplier Theory for Constrained Optimization," J. Opt. Th. and Appl., Vol. 114, pp. 287-343.

[BeS15] Beck, A., and Shtern, S., 2015. "Linearly Convergent Away-Step Conditional Gradient for Non-Strongly Convex Functions," arXiv preprint arXiv:1504.05002.

[BeT88] Bertsekas, D. P., and Tseng, P., 1988. "Relaxation Methods for Minimum Cost Ordinary and Generalized Network Flow Problems," Operations Research, Vol. 36, pp. 93-114.

[BeT89] Bertsekas, D. P., and Tsitsiklis, J. N., 1989. Parallel and Distributed Computation: Numerical Methods, Prentice-Hall, Englewood Cliffs, N. J; republished by Athena Scientific, Belmont, MA, 1997.

[BeT91] Bertsekas, D. P., and Tsitsiklis, J. N., 1991. "Some Aspects of Parallel and Distributed Iterative Algorithms - A Survey," Automatica, Vol. 27, pp. 3-21.

[BeT94] Bertsekas, D. P., and Tseng, P., 1994. "Partial Proximal Minimization Algorithms for Convex Programming," SIAM J. on Optimization, Vol. 4, pp. 551-572.

[BeT96] Bertsekas, D. P., and Tsitsiklis, J. N., 1996. Neuro-Dynamic Programming, Athena Scientific, Belmont, MA.

[BeT97] Bertsimas, D., and Tsitsiklis, J. N., 1997. Introduction to Linear Optimization, Athena Scientific, Belmont, MA.

[BeT00] Bertsekas, D. P., and Tsitsiklis, J. N., 2000. "Gradient Convergence of Gradient Methods with Errors," SIAM J. on Optimization, Vol. 36, pp. 627-642.

[BeT07] Bertsekas, D. P., and Tseng, P., 2007. "Set Intersection Theorems and Existence of Optimal Solutions," Mathematical Programming, Vol. 110, pp. 287-314.

[BeT08] Bertsekas, D. P., and Tsitsiklis, J. N., 2008. Introduction to Probability, 2nd Edition, Athena Scientific, Belmont, MA.

[BeT13] Beck, A., and Tetruashvili, L., 2013. "On the Convergence of Block Coordinate Descent Type Methods," SIAM J. on Optimization, Vol. 23, pp. 2037-2060.

[BeY09] Bertsekas, D. P., and Yu, H., 2009. "Projected Equation Methods for Approximate Solution of Large Linear Systems," J. of Computational and Applied Mathematics, Vol. 227, pp. 27-50.

[BeY10] Bertsekas, D. P., and Yu, H., 2010. "Asynchronous Distributed Policy Iteration in Dynamic Programming," Proc. of Allerton Conf. on Communication, Control and Computing, Allerton Park, Ill, pp. 1368-1374.

[BeY11] Bertsekas, D. P., and Yu, H., 2011. "A Unifying Polyhedral Approximation Framework for Convex Optimization," SIAM J. on Optimization, Vol. 21, pp. 333-360.

[BeZ82] Ben-Tal, A., and Zowe, J., 1982. "A Unified Theory of First and Second-Order Conditions for Extremum Problems in Topological Vector Spaces," Math. Programming Studies, Vol. 19, pp. 39-76.

[BeZ97] Ben-Tal, A., and Zibulevsky, M., 1997. "Penalty/Barrier Multiplier Methods for Convex Programming Problems," SIAM J. on Optimization, Vol. 7, pp. 347-366.

[Ben62] Benders, J. F., 1962. "Partitioning Procedures for Solving Mixed Variables Programming Problems," Numer. Math., Vol. 4, pp. 238-252.

[Ber71] Bertsekas, D. P., 1971. "Control of Uncertain Systems With a Set-Membership Description of the Uncertainty," Ph.D. Thesis, Dept. of EECS, MIT; may be downloaded from http://web.mit.edu/dimitrib/www/publ.html.

[Ber72] Bertsekas, D. P., 1972. "Stochastic Optimization Problems with Nondifferentiable Cost Functionals with an Application in Stochastic Programming," Proc. 1972 IEEE Conf. Decision and Control, pp. 555-559.

[Ber73] Bertsekas, D. P., 1973. "Stochastic Optimization Problems with Nondifferentiable Cost Functionals," J. of Optimization Theory and Applications, Vol. 12, pp. 218-231.

[Ber74] Bertsekas, D. P., 1974. "Partial Conjugate Gradient Methods for a Class of Optimal Control Problems," IEEE Trans. Automat. Control, Vol. 19, pp. 209-217.

[Ber75a] Bertsekas, D. P., 1975. "Necessary and Sufficient Conditions for a Penalty Method to be Exact," Math. Programming, Vol. 9, pp. 87-99.

[Ber75b] Bertsekas, D. P., 1975. "Combined Primal-Dual and Penalty Methods for Constrained Optimization, SIAM J. on Control, Vol. 13, pp. 521-544.

[Ber75c] Bertsekas, D. P., 1975. "Nondifferentiable Optimization Via Approximation," Math. Programming Study 3, Balinski, M., and Wolfe, P., (Eds.), North-Holland, Amsterdam, pp. 1-25.

[Ber75d] Bertsekas, D. P., 1975. "On the Method of Multipliers for Convex Programming," IEEE Transactions on Aut. Control, Vol. 20, pp. 385-388.

[Ber76a] Bertsekas, D. P., 1976. "On Penalty and Multiplier Methods for Constrained Optimization," SIAM J. on Control and Optimization, Vol. 14, pp. 216-235.

[Ber76b] Bertsekas, D. P., 1976. "Multiplier Methods: A Survey," Automatica, Vol. 12, pp. 133-145.

[Ber76c] Bertsekas, D. P., 1976. "On the Goldstein-Levitin-Poljak Gradient Projection Method," IEEE Trans. Automat. Control, Vol. 21, pp. 174-184.

[Ber77] Bertsekas, D. P., 1977. "Approximation Procedures Based on the Method of Multipliers," J. Opt. Th. and Appl., Vol. 23, pp. 487-510.

[Ber78] Bertsekas, D. P., 1978. "Local Convex Conjugacy and Fenchel Duality," Preprints of Triennial World Congress of IFAC, Helsinki, Vol. 2, pp. 1079-1084.

[Ber79a] Bertsekas, D. P., 1979. "Convexification Procedures and Decomposition Algorithms for Large-Scale Nonconvex Optimization Problems," J. Opt. Th. and Appl., Vol. 29, pp. 169-197.

[Ber79b] Bertsekas, D. P., 1979. "A Distributed Algorithm for the Assignment Problem," Lab. for Information and Decision Systems Working Paper, M.I.T.

[Ber80a] Bertsekas, D. P., 1980. "A Class of Optimal Routing Algorithms for Communication Networks," Proc. of the Fifth International Conference on Computer Communication, Atlanta, Ga., pp. 71-76.

[Ber80b] Bertsekas, D. P., 1980. "Variable Metric Methods for Constrained Optimization Based on Differentiable Exact Penalty Functions," Proc. Allerton Conference on Communication, Control, and Computation, Allerton Park, Ill., pp. 584-593.

[Ber81] Bertsekas, D. P., 1981. "A New Algorithm for the Assignment Problem," Math. Programming, Vol. 21, pp. 152-171.

[Ber82a] Bertsekas, D. P., 1982. Constrained Optimization and Lagrange Multiplier Methods, Academic Press, N. Y.; republished by Athena Scientific, Belmont, MA, 1997.

[Ber82b] Bertsekas, D. P., 1982. "Projected Newton Methods for Optimization Problems with Simple Constraints," SIAM J. on Control and Optimization, Vol. 20, pp. 221-246.

[Ber82c] Bertsekas, D. P., 1982. "Enlarging the Region of Convergence of Newton's Method for Constrained Optimization," J. Opt. Th. and Appl., Vol. 36, pp. 221-252.

[Ber82d] Bertsekas, D. P., 1982. "Distributed Dynamic Programming," IEEE Trans. Aut. Control, Vol. AC-27, pp. 610-616.

[Ber82e] Bertsekas, D. P., 1982. "Notes on Nonlinear Programming and Discrete-Time Optimal Control," Lab. for Information and Decision Systems Report LIDS-P-919, MIT.

[Ber83] Bertsekas, D. P., 1983. "Distributed Asynchronous Computation of Fixed Points," Math. Programming, Vol. 27, pp. 107-120.

[Ber85] Bertsekas, D. P., 1985. "A Unified Framework for Minimum Cost Network Flow Problems," Math. Programming, Vol. 32, pp. 125-145.

[Ber86] Bertsekas, D. P., 1986. "Distributed Relaxation Methods for Linear Network Flow Problems," Proceedings of 25th IEEE Conference on Decision and Control, pp. 2101-2106.

[Ber91] Bertsekas, D. P., 1991. Linear Network Optimization: Algorithms and Codes, M.I.T. Press, Cambridge, MA.

[Ber92] Bertsekas, D. P., 1992. "Auction Algorithms for Network Problems: A Tutorial Introduction," Computational Optimization and Applications, Vol. 1, pp. 7-66.

[Ber96a] D. P. Bertsekas, 1996. "Thevenin Decomposition and Network Optimization," J. Opt. Theory and Appl., Vol. 89, pp. 1-15.

[Ber96b] Bertsekas, D. P., 1996. "Incremental Least Squares Methods and the Extended Kalman Filter," SIAM J. on Optimization, Vol. 6, pp. 807-822.

[Ber97] Bertsekas, D. P., 1997. "A New Class of Incremental Gradient Methods for Least Squares Problems," SIAM J. on Optimization, Vol. 7, pp. 913-926.

[Ber98] Bertsekas, D. P., 1998. Network Optimization: Continuous and Discrete Models, Athena Scientific, Belmont, MA.

[Ber99] Bertsekas, D. P., 1999. "A Note on Error Bounds for Convex and Nonconvex Problems," Computational Optimization and Applications, Vol. 12, pp. 41-51.

[Ber05a] Bertsekas, D. P., 2005. Dynamic Programming and Optimal Control, 3rd Edition, Vol. I, Athena Scientific, Belmont, MA.

[Ber05b] Bertsekas, D. P., 2005. "Dynamic Programming and Suboptimal Control: A Survey from ADP to MPC," Fundamental Issues in Control, Special Issue for the CDC-ECC-05, European J. of Control, Vol. 11, Nos. 4-5.

[Ber05c] Bertsekas, D. P., 2005. "Lagrange Multipliers with Optimal Sensitivity Properties in Constrained Optimization," Lab. for Information and Decision Systems Report 2632, MIT; in Proc. of the 2004 Erice Workshop on Large Scale Nonlinear Optimization, Erice, Italy, Kluwer.

[Ber09] Bertsekas, D. P., 2009. Convex Optimization Theory, Athena Scientific, Belmont, MA.

[Ber10a] Bertsekas, D. P., 2010. "Extended Monotropic Programming and Duality," Lab. for Information and Decision Systems Report LIDS-P-2692, MIT, March 2006, corrected in Feb. 2010.

[Ber10b] Bertsekas, D. P., 2010. "Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey," Lab. for Information and Decision Systems Report LIDS-P-2848, MIT.

[Ber11a] Bertsekas, D. P., 2011. "Incremental Proximal Methods for Large Scale Convex Optimization," Math. Programming, Vol. 129, pp. 163-195.

[Ber11b] Bertsekas, D. P., 2011. "Centralized and Distributed Newton Methods for Network Optimization and Extensions," Lab. for Information and Decision Systems Report LIDS-P-2866, MIT.

[Ber12] Bertsekas, D. P., 2012. Dynamic Programming and Optimal Control: Approximate Dynamic Programming, 4th Edition, Vol. II, Athena Scientific, Belmont, MA.

[Ber13] Bertsekas, D. P., 2013. Abstract Dynamic Programming, Athena Scientific, Belmont, MA.

[Ber15a] Bertsekas, D. P., 2015. Convex Optimization Algorithms, Athena Scientific, Belmont, MA.

[Ber15b] Bertsekas, D. P., 2015. "Incremental Aggregated Proximal and Augmented Lagrangian Algorithms," Lab. for Information and Decision Systems Report LIDS-P-3176, MIT, September 2015.

[BiL97] Birge, J. R., and Louveaux, 1997. Introduction to Stochastic Programming, Springer-Verlag, New York, N. Y.

[Bia15] Bianchi, P., 2015. "Ergodic Convergence of a Stochastic Proximal Point Algorithm," arXiv preprint arXiv:1504.05400.

[Bis95] Bishop, C. M, 1995. Neural Networks for Pattern Recognition, Oxford University Press, N. Y.

[BoL00] Borwein, J. M., and Lewis, A. S., 2000. Convex Analysis and Nonlinear Optimization, Springer-Verlag, N. Y.

[BoL05] Bottou, L., and LeCun, Y., 2005. "On-Line Learning for Very Large Datasets," Applied Stochastic Models in Business and Industry, Vol. 21, pp. 137-151.

[BoS00] Bonnans, J. F., and Shapiro, A., 2000. Perturbation Analysis of Optimization Problems, Springer-Verlag, N. Y.

[BoT80] Boggs, P. T., and Tolle, J. W., 1980. "Augmented Lagrangians which are Quadratic in the Multiplier," J. Opt. Th. and Appl., Vol. 31, pp. 17-26.

[BoV04] Boyd, S., and Vandenbergue, L., 2004. Convex Optimization, Cambridge Univ. Press, Cambridge, U.K.

[Bog90] Bogart, K. P., 1990. Introductory Combinatorics, Harcourt Brace Jovanovich, Inc., New York, N. Y.

[Bon89a] Bonnans, J. F., 1989. "A Variant of a Projected Variable Metric Method for Bound Constrained Optimization Problems," Report, INRIA, France.

[Bon89b] Bonnans, J. F., 1989. "Asymptotic Admissibility of the Unit Stepsize in Exact Penalty Methods," SIAM J. on Control and Optimization, Vol. 27, pp. 631-641.

[Bon92] Bonnans, J. F., 1992. "Directional Derivatives of Optimal Solutions in Smooth Nonlinear Programming," J. Opt. Theory and Appl., Vol. 73, pp. 27-45.

[Bon94] Bonnans, J. F., 1994. "Local Analysis of Newton Type Methods for Variational Inequalities and Nonlinear Programming," J. Applied Math. Optimization, Vol. 29, pp. 161-186.

[Bor08] Borkar, V. S., 2008. Stochastic Approximation: A Dynamical Systems Viewpoint, Cambridge Univ. Press.

[Brä93] Brännlund, U., 1993. "On Relaxation Methods for Nonsmooth Convex Optimization," Doctoral Thesis, Royal Institute of Technology, Stockholm, Sweden.

[Bro70] Broyden, C. G., 1970. "The Convergence of a Class of Double Rank Minimization Algorithms," J. Inst. Math. Appl., Vol. 6, pp. 76-90.

[BuM88] Burke, J. V., and Moré, J. J., 1988. "On the Identification of Active Constraints," SIAM J. Numer. Anal., Vol. 25, pp. 1197-1211.

[BuQ98] Burke, J. V., and Qian, M., 1998. "A Variable Metric Proximal Point Algorithm for Monotone Operators," SIAM J. on Control and Optimization, Vol. 37, pp. 353-375.

[CCP70] Canon, M. D., Cullum, C. D., and Polak, E., 1970. Theory of Optimal Control and Mathematical Programming, McGraw-Hill, N. Y.

[CCP98] Cook, W., Cunningham, W., Pulleyblank, W., and Schrijver, A., 1998. Combinatorial Optimization, Wiley, N. Y.

[CFM75] Camerini, P. M., Fratta, L., and Maffioli, F., 1975. "On Improving Relaxation Methods by Modified Gradient Techniques," Math. Programming Studies, Vol. 3, pp. 26-34.

[CGT91] Conn, A. R., Gould, N. I. M., and Toint, P. L., 1991. "A Globally Convergent Augmented Lagrangian Algorithm for Optimization with General Constraints and Simple Bounds," SIAM J. Numer. Anal., Vol. 28, pp. 545-572.

[CGT92] Conn, A. R., Gould, N. I. M., and Toint, P. L., 1992. "LANCELOT: A FORTRAN Package for Large-Scale Nonlinear Optimization," Springer-Verlag, N. Y.

[CGT00] Conn, A. R., Gould, N. I. M., and Toint, P. L., 2000. Trust Region Methods, SIAM, Philadelphia, PA.

[CHY16] Chen, C., He, B., Ye, Y., and Yuan, X., 2016. "The Direct Extension of ADMM for Multi-Block Convex Minimization Problems is not Necessarily Convergent," Math. Programming, Series A, Vol. 155, pp. 57-79.

[CPS92] Cottle, R., Pang, J. S., and Stone, R. E., 1992. The Linear Complementarity Problem, Academic Press, Boston.

[CPS11] Choi, S. C. T., Paige, C. C., and Saunders, M. A., 2011. "MINRES-QLP: A Krylov Subspace Method for Indefinite or Singular Symmetric Systems," SIAM Journal on Scientific Computing, Vol. 33, pp. 1810-1836.

[CaC68] Canon, M. D., and Cullum, C. D., 1968. "A Tight Upper Bound on the Rate of Convergence of the Frank-Wolfe Algorithm," SIAM J. on Control, Vol. 6, pp. 509-516.

[CaF97] Caprara, A., and Fischetti, M., 1997. "Branch and Cut Algorithms," in Annotated Bibliographies in Combinatorial Optimization, Dell'Amico, M., Maffioli, F., and Martello, S., (Eds.), Wiley, Chisester, Chapter 4.

[CaG74] Cantor, D. G., Gerla, M., 1974. "Optimal Routing in Packet Switched Computer Networks," IEEE Trans. on Computing, Vol. C-23, pp. 1062-1068.

[CaM87] Calamai, P. H., and Moré, J. J., 1987. "Projected Gradient Methods for Linearly Constrained Problems," Math. Programming, Vol. 39, pp. 98-116.

[Cam94] Cameron, P. J., 1994. Combinatorics: Topics, Techniques, Algorithms, Cambridge Univ. Press.

[Car61] Carroll, C. W., 1961. "The Created Response Surface Technique for Optimizing Nonlinear Restrained Systems," Operations Research, Vol. 9, pp. 169-184.

[Cau47] Cauchy, M. A., 1847. "Analyse Mathématique–Méthode Générale Pour La Résolution des Systémes d' Eq́uations Simultanées," Comptes Vendus Acad. Sc., Paris.

[CeH87] Censor, Y., and Herman, G. T., 1987. "On Some Optimization Techniques in Image Reconstruction from Projections," Applied Numer. Math., Vol. 3, pp. 365-391.

[CeZ92] Censor, Y., and Zenios, S. A., 1992. "The Proximal Minimization Algorithm with D-Functions," J. Opt. Theory and Appl., Vol. 73, pp. 451-464.

[CeZ97] Censor, Y., and Zenios, S. A., 1997. Parallel Optimization: Theory, Algorithms, and Applications, Oxford University Press, N. Y.

[ChG59] Cheney, E. W., and Goldstein, A. A., 1959. "Newton's Method for Convex Programming and Tchebycheff Approximation," Numer. Math., Vol. I, pp. 253-268.

[ChT93] Chen, G., and Teboulle, M., 1993. "Convergence Analysis of a Proximal-Like Minimization Algorithm Using Bregman Functions," SIAM J. on Optimization, Vol. 3, pp. 538-543.

[ChT94] Chen, G., and Teboulle, M., 1994. "A Proximal-Based Decomposition Method for Convex Minimization Problems," Math. Programming, Vol. 64, pp. 81-101.

[Cla83] Clarke, F. H., 1983. Nonsmooth Analysis and Optimization, Wiley-Interscience, N. Y.

[CoC82a] Coleman, T. F., and Conn, A. R., 1982. "Nonlinear Programming Via an Exact Penalty Function: Asymptotic Analysis," Math. Progamming, Vol. 24, pp. 123-136.

[CoC82b] Coleman, T. F., and Conn, A. R., 1982. "Nonlinear Programming Via an Exact Penalty Function: Global Analysis," Math. Programming, Vol. 24, pp. 137-161.

[CoL94] Correa, R., and Lemarechal, C., 1994. "Convergence of Some Algorithms for Convex Minimization," Math. Programming, Vol. 62, pp. 261-276.

[CoT13] Couellan, N. P., and Trafalis, T. B., 2013. "On-line SVM Learning via an

Incremental Primal-Dual Technique," Optimization Methods and Software, Vol. 28, pp. 256-275.

[Coh80] Cohen, G., 1980. "Auxiliary Problem Principle and Decomposition of Optimization Problems," J. Opt. Theory and Appl., Vol. 32, pp. 277-305.

[Cro58] Croes, G. A., 1958. "A Method for Solving Traveling Salesman Problems," Operations Research, Vol. 6, pp. 791-812.

[Cry71] Cryer, C. W., 1971. "The Solution of a Quadratic Programming Problem Using Systematic Overrelaxation," SIAM J. on Control, Vol. 9, pp. 385-392.

[Cul71] Cullum, J., 1971. "An Explicit Procedure for Discretizing Continuous Optimal Control Problems," J. Opt. Theory and Appl., Vol. 8, pp. 15-34.

[Cyb89] Cybenko, 1989. "Approximation by Superpositions of a Sigmoidal Function," Math. of Control, Signals, and Systems, Vol. 2, pp. 303-314.

[DCD14] Defazio, A. J., Caetano, T. S., and Domke, J., 2014. "Finito: A Faster, Permutable Incremental Gradient Method for Big Data Problems," Proceedings of the 31st ICML, Beijing.

[DCR15] Duchi, J. C., Chaturapruek, S., and Re, R., 2015. "Asynchronous Stochastic Convex Optimization," arXiv preprint arXiv:1508.00882.

[DES82] Dembo, R. S., Eisenstadt, S. C., and Steihaug, T., 1982. "Inexact Newton Methods," SIAM J. Numer. Anal., Vol. 19, pp. 400-408.

[DFJ54] Dantzig, G. B., Fulkerson, D. R., and Johnson, S. M., 1954. "Solution of a Large-Scale Traveling-Salesman Problem," Operations Research, Vol. 2, pp. 393-410.

[DHS06] Dai, Y. H., Hager, W. W., Schittkowski, K., and Zhang, H., 2006. "The Cyclic Barzilai-Borwein Method for Unconstrained Optimization," IMA J. of Numerical Analysis, Vol. 26, pp. 604-627.

[DKK83] Decker, D. W., Keller, H. B., and Kelley, C. T., 1983. "Convergence Rates for Newton's Method at Singular Points," SIAM J. Numer. Anal., Vol. 20, pp. 296-314.

[DaW60] Dantzig, G. B., and Wolfe, P., 1960. "Decomposition Principle for Linear Programs," Operations Research, Vol. 8, pp. 101-111.

[DaY14a] Davis, D., and Yin, W., 2014. "Convergence Rate Analysis of Several Splitting Schemes," arXiv preprint arXiv:1406.4834.

[DaY14b] Davis, D., and Yin, W., 2014. "Convergence Rates of Relaxed Peaceman-Rachford and ADMM Under Regularity Assumptions, arXiv preprint arXiv:1407.5210.

[Dan67] Danskin, J. M., 1967. The Theory of Max-Min and its Application to Weapons Allocation Problems, Springer, NY.

[Dan71] Daniel, J. W., 1971. The Aproximate Minimization of Functionals, Prentice-Hall, Englewood Cliffs, N. J.

[Dav59] Davidon, W. C., 1959. "Variable Metric Method for Minimization," Argonne National Lab., Report ANL-5990 (Rev.), Argonne, Ill. Reprinted with a new preface in SIAM J. on Optimization, Vol. 1, 1991, pp. 1-17.

[Dav76] Davidon, W. C., 1976. "New Least Squares Algorithms," J. Opt. Theory and Appl., Vol. 18, pp. 187-197.

[DeK80] Decker, D. W., and Kelley, C. T., 1980. "Newton's Method at Singular Points, Parts I and II," SIAM J. Numer. Anal., Vol. 17, pp. 66-70, 465-471.

[DeM77] Dennis, J. E., and Moré, J. J., 1977. "Quasi-Newton Methods: Motivation and Theory," SIAM Review, Vol. 19, pp. 46-89.

[DeR70] Demjanov, V. F., and Rubinov, A. M., 1970. Approximate Methods in Optimization Problems, American Elsevier, N. Y.

[DeS83] Dennis, J. E., and Schnabel, R. E., 1983. Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Prentice-Hall, Englewood Cliffs, N. J.

[DeT83] Dembo, R. S., and Tulowitzki, U., 1983. "On the Minimization of Quadratic Functions Subject to Box Constraints," Working Paper Series B No. 71, School of Organization and Management, Yale Univ., New Haven, Conn.

[DeT91] Dennis, J. E., and Torczon, V., 1991. "Direct Search Methods on Parallel Machines," SIAM J. on Optimization, Vol. 1, pp. 448-474.

[DeT93] De Angelis, P. L., and Toraldo, G., 1993. "On the Identification Property of a Projected Gradient Method," SIAM J. Numer. Anal., Vol. 30, pp. 1483-1497.

[DeV85] Demjanov, V. F., and Vasilév, L. V., 1985. Nondifferentiable Optimization, Optimization Software, N. Y.

[Del12] Delfour, M. C., 2012. Introduction to Optimization and Semidifferential Calculus, SIAM, Phila.

[Deu12] Deuflhard, P., 2012. "A Short History of Newton's Method," Documenta Mathematica, Optimization Stories, pp. 25-30.

[DiG79] DiPillo, G., and Grippo, L., 1979. "A New Class of Augmented Lagrangians in Nonlinear Programming," SIAM J. on Control and Optimization, Vol. 17, pp. 618-628.

[DiG89] DiPillo, G., and Grippo, L., 1989. "Exact Penalty Functions in Constrained Optimization," SIAM J. on Control and Optimization, Vol. 27, pp. 1333-1360.

[Dix72a] Dixon, L. C. W., 1972. "Quasi-Newton Algorithms Generate Identical Points," Math. Programming, Vol. 2, pp. 383-387.

[Dix72b] Dixon, L. C. W., 1972. "Quasi-Newton Algorithms Generate Identical Points. II. The Proofs of Four New Theorems," Math. Programming, Vol. 3, pp. 345-358.

[DoJ86] Dontchev, A. L., and Jongen, H. Th., 1986. "On the Regularity of the Kuhn-Tucker Curve," SIAM J. on Control and Optimization, Vol. 24, pp. 169-176.

[DoR09] Dontchev, A. L., and Rockafellar, R. T., 2009. Implicit Functions and Solution Mappings," 2nd edition, Springer, N. Y.

[DuB89] Dunn, J. C., and Bertsekas, D. P., 1989. "Efficient Dynamic Programming Implementations of Newton's Method for Unconstrained Optimal Control Problems," J. Opt. Theory and Appl., Vol. 63, pp. 23-38.

[DuM65] Dubovitskii, M. D., and Milyutin, A. A., 1965. "Extremum Problems in the Presence of Restriction," USSR Comp. Math. and Math. Phys., Vol. 5, pp. 1-80.

[DuS83] Dunn, J. C., and Sachs, E., 1983. "The Effect of Perturbations on the Convergence Rates of Optimization Algorithms," Appl. Math. Optim., Vol. 10, pp. 143-157.

[DuZ89] Du, D.-Z., and Zhang, X.-S., 1989. "Global Convergence of Rosen's Gradient Projection Method," Math. Programming, Vol. 44, pp. 357-366.

[Dun79] Dunn, J. C., 1979. "Rates of Convergence for Conditional Gradient Algorithms Near Singular and Nonsingular Extremals," SIAM J. on Control and Optimization, Vol. 17, pp. 187-211.

[Dun80a] Dunn, J. C., 1980. "Convergence Rates for Conditional Gradient Sequences Generated by Implicit Step Length Rules," SIAM J. on Control and Optimization, Vol. 18, pp. 473-487.

[Dun80b] Dunn, J. C., 1980. "Newton's Method and the Goldstein Step Length Rule for

Constrained Minimization Problems," SIAM J. on Control and Optimization, Vol. 18, pp. 659-674.

[Dun81] Dunn, J. C., 1981. "Global and Asymptotic Convergence Rate Estimates for a Class of Projected Gradient Processes," SIAM J. on Control and Optimization, Vol. 19, pp. 368-400.

[Dun87] Dunn, J. C., 1987. "On the Convergence of Projected Gradient Processes to Singular Critical Points," J. Opt. Theory and Appl., Vol. 55, pp. 203-216.

[Dun88a] Dunn, J. C., 1988. "Gradient Projection Methods for Systems Optimization Problems," Control and Dynamic Systems, Vol. 29, pp. 135-195.

[Dun88b] Dunn, J. C., 1988. "A Projected Newton Method for Minimization Problems with Nonlinear Inequality Constraints," Numer. Math., Vol. 53, pp. 377-409.

[Dun91a] Dunn, J. C., 1991. "Scaled Gradient Projection Methods for Optimal Control Problems and Other Structured Nonlinear Programs," in New Trends in Systems Theory, Conte, G., et al, (Eds.), Birkhöuser, Boston, MA.

[Dun91b] Dunn, J. C., 1991. "A Subspace Decomposition Principle for Scaled Gradient Projection Methods: Global Theory," SIAM J. on Control and Optimization, Vol. 29, pp. 219-246.

[Dun93a] Dunn, J. C., 1993. "A Subspace Decomposition Principle for Scaled Gradient Projection Methods: Local Theory," SIAM J. on Control and Optimization, Vol. 31, pp. 219-246.

[Dun93b] Dunn, J. C., 1993. "Second-Order Multiplier Update Calculations for Optimal Control Problems and Related Large Scale Nonlinear Programs," SIAM J. on Optimization, Vol. 3, pp. 489-502.

[Dun93c] Dunn, J. C., 1993. Private Communication.

[Dun94] Dunn, J. C., 1994. "Gradient-Related Constrained Minimization Algorithms in Function Spaces: Convergence Properties and Computational Implications," in Large Scale Optimization: State of the Art, Hager, W. W., Hearn, D. W., and Pardalos, P. M., (Eds.), Kluwer, Boston.

[Eas58] Eastman, W. L., 1958. Linear Programming with Pattern Constraints, Ph.D. Thesis, Harvard University, Cambridge, MA.

[EcB90] Eckstein, J., and Bertsekas, D. P., 1990. "An Alternating Direction Method for Linear Programming," Report LIDS-P-1967, Lab. for Info. and Dec. Systems, M.I.T.

[EcB92] Eckstein, J., and Bertsekas, D. P., 1992. "On the Douglas-Rachford Splitting Method and the Proximal Point Algorithm for Maximal Monotone Operators," Math. Programming, Vol. 55, pp. 293-318.

[Eck94a] Eckstein, J., 1994. "Nonlinear Proximal Point Algorithms Using Bregman Functions, with Applications to Convex Programming," Math. Operations Res., Vol. 18, pp. 202-226.

[Eck94b] Eckstein, J., 1994. "Parallel Alternating Direction Multiplier Decomposition of Convex Programs," J. Opt. Theory and Appl., Vol. 80, pp. 39-62.

[EkT76] Ekeland, I., and Teman, R., 1976. Convex Analysis and Variational Problems, North-Holland Publ., Amsterdam.

[ElM75] Elzinga, J., and Moore, T. G., 1975. "A Central Cutting Plane Algorithm for the Convex Programming Problem," Math. Programming, Vol. 8, pp. 134-145.

[Eve63] Everett, H., 1963. "Generalized Lagrange Multiplier Method for Solving Problems of Optimal Allocation of Resources," Operations Research, Vol. 11, pp. 399-417.

[Evt85] Evtushenko, Y. G., 1985. Numerical Optimization Techniques, Optimization Software, N. Y.

[FAJ14] Feyzmahdavian, H. R., Aytekin, A., and Johansson, M., 2014. "A Delayed Proximal Gradient Method with Linear Convergence Rate," in Prop. of 2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-6.

[FGK73] Fratta, L., Gerla, M., and Kleinrock, L., 1973. "The Flow Deviation Method: An Approach to Store-and-Forward Communication Network Design," Networks, Vol. 3, pp. 97-133.

[FGW02] Forsgren, A., Gill, P. E., and Wright, M. H., 2002. "Interior Methods for Nonlinear Optimization," SIAM Review, Vol. 44, pp. 525-597.

[FaF63] Fadeev, D. K., and Fadeeva, V. N., 1963. Computational Methods of Linear Algebra, Freeman, San Francisco, CA.

[FaP03] Facchinei, F., and Pang, J.-S., 2003. Finite-Dimensional Variational Inequalities and Complementarity Problems, Springer Verlag, N. Y.

[Fab73] Fabian, V., 1973. "Asymptotically Efficient Stochastic Approximation: The RM Case," Ann. Statist., Vol. 1, pp. 486-495.

[FeM91] Ferris, M. C., and Mangasarian, O. L., 1991. "Parallel Constraint Distribution," SIAM J. on Optimization, Vol. 1, pp. 487-500.

[Fen49] Fenchel, W., 1949. "On Conjugate Convex Functions," Canad. J. Math., Vol. 1, pp. 73-77.

[Fen51] Fenchel, W., 1951. "Convex Cones, Sets, and Functions," Mimeographed Notes, Princeton Univ.

[Fey16] Feyzmahdavian, H. R., 2016. Performance Analysis of Positive Systems and Optimization Algorithms with Time-Delays, Doctoral Thesis, KTH, Sweden.

[FiM68] Fiacco, A. V., and McCormick, G. P., 1968. Nonlinear Programming: Sequential Unconstrained Minimization Techniques, Wiley, N. Y.

[Fia83] Fiacco, A. V., 1983. Introduction to Sensitivity and Stability Analysis in Nonlinear Programming, Academic Press, N. Y.

[FlH95] Florian, M. S., and Hearn, D., 1995. "Network Equilibrium Models and Algorithms," Handbooks in OR and MS, Ball, M. O., Magnanti, T. L., Monma, C. L., and Nemhauser, G. L., (Eds.), Vol. 8, North-Holland, Amsterdam, pp. 485-550.

[FlP63] Fletcher, R., and Powell, M. J. D., 1963. "A Rapidly Convergent Descent Algorithm for Minimization," Comput. J., Vol. 6, pp. 163-168.

[FlP95] Floudas, C., and Pardalos, P. M., (Eds.), 1995. State of the Art in Global Optimization: Computational Methods and Applications, Kluwer, Boston.

[Fla92] Flam, S. D., 1992. "On Finite Convergence and Constraint Identification of Subgradient Projection Methods," Math. Programming, Vol. 57, pp 427-437.

[Fle70a] Fletcher, R., 1970. "A New Approach to Variable Metric Algorithms," Computer J., Vol. 13, pp. 317-322.

[Fle70b] Fletcher, R., 1970. "A Class of Methods for Nonlinear Programming with Termination and Convergence Properties," in Integer and Nonlinear Programming, Abadie, J., (Ed.), pp. 157-173, North-Holland Publ., Amsterdam.

[Fle00] Fletcher, R., 2000. Practical Methods of Optimization, 2nd edition, Wiley, NY.

[Flo95] Floudas, C. A., 1995. Nonlinear and Mixed-Integer Optimization: Fundamentals and Applications, Oxford University Press, N. Y.

[FoG83] Fortin, M., and Glowinski, R., (Eds.), 1983. Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems, North-Holland, Amsterdam.

[FoS11] Fong, D. C. L., and Saunders, M., 2011. "LSMR: An Iterative Algorithm for Sparse Least-Squares Problems," SIAM Journal on Scientific Computing, Vol. 33, pp. 2950-2971.

[FoS12] Fong, D. C. L., and Saunders, M., 2012. "CG Versus MINRES: An Empirical Comparison," SQU Journal for Science, Vol. 17, pp. 44-62.

[FrG16] Freund, R., and Grigas, P., 2016. "New Analysis and Results for the Frank-Wolfe Method," Math. Programming, Series A, Vol. 155, pp. 199-230.

[FrW56] Frank, M., and Wolfe, P., 1956. "An Algorithm for Quadratic Programming," Naval Research Logistics Quarterly, Vol. 3, pp. 95-110.

[Fra12] Frauendorfer, K., 2012. Stochastic Two-Stage Programming, Springer, N. Y.

[Fre91] Freund, R. M., 1991. "Theoretical Efficiency of a Shifted Barrier Function Algorithm in Linear Programming," Linear Algebra and Appl., Vol. 152, pp. 19-41.

[Fri56] Frisch, M. R., 1956. "La Resolution des Problemes de Programme Lineaire par la Methode du Potential Logarithmique," Cahiers du Seminaire D'Econometrie, Vol. 4, pp. 7-20.

[Fuk92] Fukushima, M., 1992. "Application of the Alternating Direction Method of Multipliers to Separable Convex Programming," Comp. Opt. and Appl., Vol. 1, pp. 93-111.

[GFJ15] Ghadimi, E., Feyzmahdavian, H. R., and Johansson, M., 2015. "Global Convergence of the Heavy-Ball Method for Convex Optimization," in European Control Conference (ECC), pp. 310-315.

[GHV92] Goffin, J. L., Haurie, A., and Vial, J. P., 1992. "Decomposition and Nondifferentiable Optimization with the Projective Algorithm," Management Science, Vol. 38, pp. 284-302.

[GKT51] Gale, D., Kuhn, H. W., and Tucker, A. W., 1951. "Linear Programming and the Theory of Games," in Activity Analysis of Production and Allocation, Koopmans, T. C., (Ed.), Wiley, N. Y.

[GKX10] Gupta, M. D., Kumar, S., and Xiao, J. 2010. "L1 Projections with Box Constraints," arXiv preprint arXiv:1010.0141.

[GLL91] Grippo, L., Lampariello, F., and Lucidi, S., 1991. "A Class of Nonmonotone Stabilization Methods in Unconstrained Minimization," Numer. Math., Vol. 59, pp. 779-805.

[GLY94] Goffin, J. L., Luo, Z.-Q., and Ye, Y., 1994. "On the Complexity of a Column Generation Algorithm for Convex or Quasiconvex Feasibility Problems," in Large Scale Optimization: State of the Art, Hager, W. W., Hearn, D. W., and Pardalos, P. M., (Eds.), Kluwer, Boston.

[GLY96] Goffin, J. L., Luo, Z.-Q., and Ye, Y., 1996. "Complexity Analysis of an Interior Cutting Plane Method for Convex Feasibility Problems," SIAM J. on Optimization, Vol. 6, pp. 638-652.

[GMW81] Gill, P. E., Murray, W., and Wright, M. H., 1981. Practical Optimization, Academic Press, N. Y.

[GMW91] Gill, P. E., Murray, W., and Wright, M. H., 1991. Numerical Linear Algebra and Optimization, Vol. I, Addison-Wesley, Redwood City, CA.

[GOP15a] Gurbuzbalaban, M., Ozdaglar, A., and Parrilo, P., 2015. "On the Convergence Rate of Incremental Aggregated Gradient Algorithms," arXiv preprint arXiv:1506.02081.

[GOP15b] Gurbuzbalaban, M., Ozdaglar, A., and Parrilo, P., 2015. "Convergence Rate of Incremental Gradient and Newton Methods," arXiv preprint arXiv:1510.08562.

[GOP15c] Gurbuzbalaban, M., Ozdaglar, A., and Parrilo, P., 2015. "Why Random Reshuffling Beats Stochastic Gradient Descent," arXiv preprint arXiv:1510.08560.

[GaB82] Gafni, E. M., and Bertsekas, D. P., 1982. "Convergence of a Gradient Projection Method," Report LIDS-P-1201, Lab. for Info. and Dec. Systems, M.I.T.

[GaB84] Gafni, E. M., and Bertsekas, D. P., 1984. "Two-Metric Projection Methods for Constrained Optimization," SIAM J. on Control and Optimization, Vol. 22, pp. 936-964.

[GaD88] Gawande, M., and Dunn, J. C., 1988. "Variable Metric Gradient Projection Processes in Convex Feasible Sets Defined by Nonlinear Inequalities," Appl. Math. Optim., Vol. 17, pp. 103-119.

[GaJ88] Gauvin, J., and Janin, R., 1988. "Directional Behavior of Optimal Solutions in Nonlinear Mathematical Programming," Math. of Operations Res., Vol. 13, pp. 629-649.

[GaM76] Gabay, D., and Mercier, B., 1976. "A Dual Algorithm for the Solution of Nonlinear Variational Problems via Finite-Element Approximations," Comp. Math. Appl., Vol. 2, pp. 17-40.

[GaM92] Gaudioso, M., and Monaco, M. F, 1992. "Variants to the Cutting Plane Approach for Convex Nondifferentiable Optimization," Optimization, Vol. 25, pp. 65-75.

[Gab79] Gabay, D., 1979. Methodes Numeriques pour l'Optimization Non Lineaire, These de Doctorat d'Etat et Sciences Mathematiques, Univ. Pierre at Marie Curie (Paris VI).

[Gab82] Gabay, D., 1982. "Reduced Quasi-Newton Methods with Feasibility Improvement for Nonlinearly Constrained Optimization," Math. Programming Studies, Vol. 16, pp. 18-44.

[Gab83] Gabay, D., 1983. "Applications of the Method of Multipliers to Variational Inequalities," in M. Fortin and R. Glowinski, eds., Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems, North-Holland, Amsterdam.

[Gai94] Gaivoronski, A. A., 1994. "Convergence Analysis of Parallel Backpropagation Algorithm for Neural Networks," Optimization Methods and Software, Vol. 4, pp. 117-134.

[Gal77] Gallager, R. G., 1977. "A Minimum Delay Routing Algorithm Using Distributed Computation," IEEE Trans. on Communications, Vol. 25, pp. 73-85.

[Gau77] Gauvin, J., 1977. "A Necessary and Sufficient Condition to Have Bounded Multipliers in Convex Programming," Math. Programming., Vol. 12, pp. 136-138.

[Geo70] Geoffrion, A. M., 1970. "Elements of Large-Scale Mathematical Programming, I, II," Management Science, Vol. 16, pp. 652-675, 676-691.

[Geo74] Geoffrion, A. M., 1974. "Lagrangian Relaxation for Integer Programming," Math. Programming Studies, Vol. 2, pp. 82-114.

[Geo77] Geoffrion, A. M., 1977. "Objective Function Approximations in Mathematical Programming," Math. Programming, Vol. 13, pp. 23-27.

[GiK95] Gilmore, P., and Kelley, C. T., 1995. "An Implicit Filtering Algorithm for Optimization of Functions with Many Local Minima," SIAM J. on Optimization, Vol. 5, pp. 269-285.

[GiM74] Gill, P. E., and Murray, W., (Eds.), 1974. Numerical Methods for Constrained Optimization, Academic Press, N. Y.

[GlL97] Glover, F., and Laguna, M., 1997. Tabu Search, Kluwer, Boston.

[GlM75] Glowinski, R. and Marrocco, A., 1975. "Sur l' Approximation par Elements Finis d' Ordre un et la Resolution par Penalisation-Dualite d'une Classe de Problemes de Dirichlet Non Lineaires" Revue Francaise d'Automatique Informatique Recherche Operationnelle, Analyse Numerique, R-2, pp. 41-76.

[GlP79] Glad, T., and Polak, E., 1979. "A Multiplier Method with Automatic Limitation of Penalty Growth," Math. Programming, Vol. 17, pp. 140-155.

[Gla79] Glad, T., 1979. "Properties of Updating Methods for the Multipliers in Augmented Lagrangians," J. Opt. Th. and Appl., Vol. 28, pp. 135-156.

[GoP79] Gonzaga, C., and Polak, E., 1979. "On Constraint Dropping Schemes and Optimality Functions for a Class of Outer Approximations Algorithms," SIAM J. on Control and Optimization, Vol. 17, pp.477-493.

[GoT71] Gould, F. J., and Tolle, J., 1971. "A Necessary and Sufficient Condition for Constrained Optimization," SIAM J. Applied Math., Vol. 20, pp. 164-172.

[GoT72] Gould, F. J., and Tolle, J., 1972. "Geometry of Optimality Conditions and Constraint Qualifications," Math. Programming, Vol. 2, pp. 1-18.

[GoV90] Goffin, J. L., and Vial, J. P., 1990. "Cutting Planes and Column Generation Techniques with the Projective Algorithm," J. Opt. Th. and Appl., Vol. 65, pp. 409-429.

[GoV99] Goffin, J. L., and Vial, J. P., 1999. "Convex Nondifferentiable Optimization: A Survey Focussed on the Analytic Center Cutting Plane Method," Logilab Technical Report, Department of Management Studies, University of Geneva, Switzerland; also GERAD Tech. Report G-99-17, McGill Univ., Montreal, Canada.

[Gof77] Goffin, J. L., 1977. "On Convergence Rates of Subgradient Optimization Methods," Math. Programming, Vol. 13, pp. 329-347.

[Gol62] Goldstein, A. A., 1962. "Cauchy's Method of Minimization," Numer. Math., Vol. 4, pp. 146-150.

[Gol64] Goldstein, A. A., 1964. "Convex Programming in Hibert Space," Bull. Amer. Math. Soc., Vol. 70, pp. 709-710.

[Gol67] Goldstein, A. A., 1967. Constructive Real Analysis, Harper and Row, N. Y.

[Gol70] Goldfarb, D., 1970. "A Family of Variable-Metric Methods Derived by Variational Means," Math. Comp., Vol. 24, pp. 23-26.

[Gol85] Golshtein, E. G., 1985. "A Decomposition Method for Linear and Convex Programming Problems," Matecon, Vol. 21, pp. 1077-1091.

[Gol89] Goldberg, D. E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning, Addison Wesley, Reading, MA.

[Gom58] Gomory, R. E., 1958. "Outline of an Algorithm for Integer Solutions to Linear Programs," Bulletin of the American Mathematical Society, Vol. 64, pp. 275-278.

[Gon91] Gonzaga, C. C., 1991. "Large Step Path-Following Methods for Linear Programming, Part I: Barrier Function Method," SIAM J. on Optimization, Vol. 1, pp. 268-279.

[Gon92] Gonzaga, C. C., 1992. "Path Following Methods for Linear Programming," SIAM Review, Vol. 34, pp. 167-227.

[Gon00] Gonzaga, C. C., 2000. "Two Facts on the Convergence of the Cauchy Algorithm," J. of Optimization Theory and Applications, Vol. 107, pp. 591-600.

[GrS00] Grippo, L., and Sciandrone, M., 2000. "On the Convergence of the Block Nonlinear Gauss-Seidel Method Under Convex Constraints," Operations Research Letters, Vol. 26, pp. 127-136.

[GrW08] Griewank, A., and Walther, A., 2008. Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, SIAM.

[Gri94] Grippo, L., 1994. "A Class of Unconstrained Minimization Methods for Neural Network Training," Optimization Methods and Software, Vol. 4, pp. 135-150.

[GuM86] Guelat, J., and Marcotte, P., 1986. "Some Comments on Wolfe's 'Away Step'," Math. Programming, Vol. 35, pp. 110-119.

[Gui69] Guignard, M., 1969. "Generalized Kuhn-Tucker Conditions for Mathematical Programming Problems in a Banach Space," SIAM J. on Control, Vol. 7, pp. 232-241.

[Gul92] Guler, O., 1992. "New Proximal Point Algorithms for Convex Minimization," SIAM J. on Optimization, Vol. 2, pp. 649-664.

[Gul94] Guler, O., 1994. "Limiting Behavior of Weighted Central Paths in Linear Programming," Math. Programming, Vol. 65, pp. 347-363.

[HDY12] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior A., et al. 2012. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," Signal Processing Magazine, IEEE, Vol. 29, pp. 82-97.

[HKR95] den Hertog, D., Kaliski, J., Roos, C., and Terlaky, T., 1995. "A Path-Following Cutting Plane Method for Convex Programming," Annals of Operations Research, Vol. 58, pp. 69-98.

[HLV87] Hearn, D. W., Lawphongpanich, S., and Ventura, J. A., 1987. "Restricted Simplicial Decomposition: Computation and Extensions," Math. Programming Studies, Vol. 31, pp. 119-136.

[HPT00] Horst, R., Pardalos, P. M., and Thoai, N. V., 2000. Introduction to Global Optimization, 2nd Edition, Kluwer, Boston.

[HDR13] Hong, M., Wang, X., Razaviyayn, M., and Luo, Z. Q., 2013. "Iteration Complexity Analysis of Block Coordinate Descent Methods," arXiv preprint arXiv:1310.6957.

[HaB70] Haarhoff, P. C., and Buys, J. D, 1970. "A New Method for the Optimization of a Nonlinear Function Subject to Nonlinear Constraints," Computer J., Vol. 13, pp. 178-184.

[HaH93] Hager, W. W., and Hearn, D. W., 1993. "Application of the Dual Active Set Algorithm to Quadratic Network Optimization," Computational Optimization and Applications, Vol. 1, pp. 349-373.

[HaM79] Han, S. P., and Mangasarian, O. L., 1979. "Exact Penalty Functions in Nonlinear Programming," Math. Programming, Vol. 17, pp. 251-269.

[Hag99] Hager, W. W., 1999. "Stabilized Sequential Quadratic Programming," Computational Optimization and Applications, Vol. 12.

[Han77] Han, S. P., 1977. "A Globally Convergent Method for Nonlinear Programming," J. Opt. Th. and Appl., Vol. 22, pp. 297-309.

[Hay11] Haykin, S., 2011. Neural Networks and Learning Machines, (3rd Ed.), Pearson Education, Upper Saddle River, N. J.

[HeK70] Held, M., and Karp, R. M., 1970. "The Traveling Salesman Problem and Minimum Spanning Trees," Operations Research, Vol. 18, pp. 1138-1162.

[HeK71] Held, M., and Karp, R. M., 1971. "The Traveling Salesman Problem and Minimum Spanning Trees: Part II," Math. Programming, Vol. 1, pp. 6-25.

[HeL89] Hearn, D. W., and Lawphongpanich, S., 1989. "Lagrangian Dual Ascent by Generalized Linear Programming," Operations Res. Letters, Vol. 8, pp. 189-196.

[HeS52] Hestenes, M. R., and Stiefel, E. L., 1952. "Methods of Conjugate Gradients for Solving Linear Systems," J. Res. Nat. Bur. Standards Sect. B, Vol. 49, pp. 409-436.

[Hei96] Heinkenschloss, M., 1996. "Projected Sequential Quadratic Programming Methods," SIAM J. on Optimization, Vol. 6, pp. 373-417.

[Her94] den Hertog, D., 1994. Interior Point Approach to Linear, Quadratic, and Convex Programming, Kluwer, Dordrecht, The Netherlands.

[Hes69] Hestenes, M. R., 1969. "Multiplier and Gradient Methods," J. Opt. Th. and Appl., Vol. 4, pp. 303-320.

[Hes75] Hestenes, M. R., 1975. Optimization Theory: The Finite Dimensional Case, Wiley, N. Y.

[Hes80] Hestenes, M. R., 1980. Conjugate Direction Methods in Optimization, Springer-Verlag, Berlin and N. Y.

[HiL93] Hiriart-Urruty, J.-B., and Lemarechal, C., 1993. Convex Analysis and Minimization Algorithms, Vols. I and II, Springer-Verlag, Berlin and N. Y.

[Hil57] Hildreth, C., 1957. "A Quadratic Programming Procedure," Naval Res. Logist. Quart., Vol. 4, pp. 79-85. See also "Erratum," Naval Res. Logist. Quart., Vol. 4, p. 361.

[HoJ61] Hooke, R., and Jeeves, T. A., 1961. "Direct Search Solution of Numerical and Statistical Problems," J. Assoc. Comp. Mach., Vol. 8, pp. 212-221.

[HoK71] Hoffman, K., and Kunze, R., 1971. Linear Algebra, Prentice-Hall, Englewood Cliffs, N. J.

[HoL13] Hong, M., and Luo, Z. Q., 2013. "On the Linear Convergence of the Alternating Direction Method of Multipliers," arXiv preprint arXiv:1208.3922.

[Hoh77] Hohenbalken, B. von, 1977. "Simplicial Decomposition in Nonlinear Programming," Math. Programming, Vol. 13, pp. 49-68.

[Hol74] Holloway, C. A., 1974. "An Extension of the Frank and Wolfe Method of Feasible Directions," Math. Programming, Vol. 6, pp. 14-27.

[HuD84] Hughes, G. C., and Dunn, J. C., 1984. "Newton-Goldstein Convergence Rates for Convex Constrained Minimization Problems with Singular Solutions," Appl. Math. Optim., Vol. 12, pp. 203-230.

[HuL88] Huang, C., and Litzenberger, R. H., 1988. Foundations of Financial Economics, Prentice-Hall, Englewood Cliffs, N. J.

[IJT15a] Iusem, A., Jofre, A., and Thompson, P., 2015. "Approximate Projection Methods for Monotone Stochastic Variational Inequalities," Math. Programming, to appear.

[IJT15b] Iusem, A., Jofre, A., and Thompson, P., 2015. "Incremental Constraint Projection Methods for Monotone Stochastic Variational Inequalities," Math. Operations Res., to appear.

[IST94] Iusem, A. N., Svaiter, B., and Teboulle, M., 1994. "Entropy-Like Proximal Methods in Convex Programming," Math. Operations Res., Vol. 19, pp. 790-814.

[IbF96] Ibaraki, S., and Fukushima, M., 1996. "Partial Proximal Method of Multipliers for Convex Programming Problems," J. of Operations Research Society of Japan, Vol. 39, pp. 213-229.

[IbK88] Ibaraki, T., and Katoh, N., 1988. Resource Allocation Problems: Algorithmic Approaches, M.I.T. Press, Cambridge, MA.

[Iof94] Ioffe, A., 1994. "On Sensitivity Analysis of Nonlinear Programs in Banach Spaces: The Approach via Composite Unconstrained Minimization," SIAM J. on Optimization, Vol. 4, pp. 1-43.

[IuT95] Iusem, A. N., and Teboulle, M., 1995. "Convergence Rate Analysis of Nonquadratic Proximal Methods for Convex and Linear Programming," Math. Operations Res., Vol. 20.

[Ius99] Iusem, A. N., 1999. "Augmented Lagrangian Methods and Proximal Point Methods for Convex Minimization," Investigacion Operativa, Vol. 8, pp. 11-49 .

[JRT95] Junger, M., Reinelt, G., and Rinaldi, G., 1995. "Practical Problem Solving with Cutting Plane Algorithms in Combinatorial Oprimization," in Combinatorial Optimization, Cook, W. J., Lovasz, L., and Seymour, P., (Eds.), DIMACS Series in Discrete Mathematics and Computer Science, AMS, pp. 11-152.

[JaS95] Jarre, F., and Saunders, M. A., 1995. "A Practical Interior-Point Method for Convex Programming," SIAM J. Optimization, Vol. 5, pp. 149-171.

[Jag13] Jaggi, M., 2013. "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization," Proc. of ICML 2013.

[JeW92] Jeyakumar, V., and Wolkowicz, H., 1992. "Generalizations of Slater's Constraint Qualification for Infinite Convex Programs," Math. Programming, Vol. 57, pp. 85-101.

[Joh48] John, F., 1948. "Extremum Problems with Inequalities as Subsidiary Conditions," in Studies and Essays: Courant Anniversary Volume, K. O. Friedrichs, Neugebauer, O. E., and Stoker, J. J., (Eds.), Wiley-Interscience, N. Y., pp. 187-204.

[KAK89] Korst, J., Aarts, E. H., and Korst, A., 1989. Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing, Wiley, N. Y.

[KLT03] Kolda, T. G., Lewis, R. M., and Torczon, V., 2003. "Optimization by Direct Search: New Perspectives on Some Classical and Modern Methods," SIAM Review, Vol. 45, pp. 385-482.

[KMN91] Kojima, M., Meggido, N., Noma, T., and Yoshise, A., 1991. A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems, Springer-Verlag, Berlin.

[KMY89] Kojima, M., Mizuno, S., and Yoshise, A., 1989. "A Primal-Dual Interior Point Algorithm for Linear Programming," in Progress in Mathematical Programming, Interior Point and Related Methods, Meggido, N., (Ed.), Springer-Verlag, N. Y., pp. 29-47.

[KaW94] Kall, P., and Wallace, S. W., 1994. Stochastic Programming, Wiley, Chichester, UK.

[Kan39] Kantorovich, L. V., 1939. "The Method of Successive Approximations for Functional Equations," Acta Math., Vol. 71, pp. 63-97.

[Kan45] Kantorovich, L. V., 1945. "On an Effective Method of Solution of Extremal Problems for a Quadratic Functional," Dokl. Akad. Nauk SSSR, Vol. 48, pp. 483-487.

[Kan49] Kantorovich, L. V., 1949. "On Newton's Method," Trudy Mat. Inst. Steklov, Vol. 28, pp. 104-144. Translated in Selected Articles in Numerical Analysis by C. D. Benster, 104.1-144.2.

[Kar39] Karush, W., 1939. "Minima of Functions of Several Variables with Inequalities as Side Conditions," M.S. Thesis, Department of Math., University of Chicago.

[Kar84] Karmarkar, N., 1984. "A New Polynomial-Time Algorithm for Linear Programming," Combinatorica, Vol. 4, pp. 373-395.

[KeG88] Keerthi, S. S., and Gilbert, E. G., 1988. "Optimal, Infinite Horizon Feedback Laws for a General Class of Constrained Discete Time Systems: Stability and Moving-Horizon Approximations," J. Optimization Theory Appl., Vo. 57, pp. 265-293.

[Kel60] Kelley, J. E., 1960. "The Cutting-Plane Method for Solving Convex Programs," J. Soc. Indust. Appl. Math., Vol. 8, pp. 703-712.

[Kel99] Kelley, C. T., 1999. Iterative Methods for Optimization, SIAM, Philadelphia.

[Kha79] Khachiyan, L. G., 1979. "A Polynomial Algorithm for Linear Programming," Soviet Math. Doklady, Vol. 20, pp. 191-194.

[KiN91] Kim, S., and Nazareth, J. L., 1991. "The Decomposition Principle and Algorithms for Linear Programming," Linear Algebra and its Applications, Vol. 152, pp. 119-133.

[KiR92] King, A., and Rockafellar, R. T., 1992. "Sensitivity Analysis for Nonsmooth Generalized Equations," Math. Programming, Vol. 55, pp. 193-212.

[KlH98] Klatte, D., and Henrion, R., 1998. "Regularity and Stability in Nonlinear Semi-Infinite Optimization," in Semi-Infinite Programming, Reemtsen, R., and Ruckman, J. J., (Eds.), Kluwer, Boston, pp. 69-102.

[KoB72] Kort, B. W., and Bertsekas, D. P., 1972. "A New Penalty Function Method for Constrained Minimization," Proc. 1972 IEEE Confer. Decision Control, New Orleans, LA, pp. 162-166.

[KoB76] Kort, B. W., and Bertsekas, D. P., 1976. "Combined Primal-Dual and Penalty Methods for Convex Programming," SIAM J. on Control and Optimization, Vol. 14, pp. 268-294.

[KoM98] Kontogiorgis, S., and Meyer, R. R., 1998. "A Variable-Penalty Alternating Directions Method for Convex Optimization," Math. Programming, Vol. 83, pp. 29-53.

[KoN93] Kortanek, K. O., and No, H., 1993. "A Central Cutting Plane Algorithm for Convex Semi-Infinite Programming Problems," SIAM J. on Optimization, Vol. 3, pp. 901-918.

[KoZ93] Kortanek, K. O., and Zhu, J., 1993. "A Polynomial Barrier Algorithm for Linearly Constrained Convex Programming Problems," Math. Operations Res., Vol. 18, pp. 116-127.

[KoZ95] Kortanek, K. O., and Zhu, J., 1995. "On Controlling the Parameter in the Logarithmic Barrier Term for Convex Programming Problems," J. Opt. Th. and Appl., Vol. 84, pp. 117-143.

[Koh74] Kohonen, T., 1974. "An Adaptive Associative Memory Principle," IEEE Trans. on Computers, Vol. C-23, pp. 444-445.

[Kor75] Kort, B. W., 1975. "Combined Primal-Dual and Penalty Function Algorithms for Nonlinear Programming," Ph.D. Thesis, Dept. of Engineering-Economic Systems, Stanford Univ., Stanford, Ca.

[Kor76] Korpelevich, G. M., 1976. "The Extragradient Method for Finding Saddle Points and Other Problems," Matecon, Vol. 12, pp. 747-756.

[KuC78] Kushner, H. J., and Clark, D. S., 1978. Stochastic Approximation Methods for Constrained and Unconstrained Systems, Springer-Verlag, N. Y.

[KuT51] Kuhn, H. W., and Tucker, A. W., 1951. "Nonlinear Programming," in Proc. of the Second Berkeley Symposium on Math. Statistics and Probability, Neyman, J., (Ed.), Univ. of California Press, Berkeley, CA, pp. 481-492.

[KuY97] Kushner, H. J., and Yin, G., 1997. Stochastic Approximation Methods, Springer-Verlag, N. Y.

[Kuh76] Kuhn, H. W., 1976. "Nonlinear Programming: A Historical View," in Nonlinear Programming, Cottle, R. W., and Lemke, C. E., (Eds.), SIAM-AMS Proc., Vol. IX, American Math. Soc., Providence, RI, pp. 1-26.

[LJS12] Lacoste-Julien, S., Jaggi, M., Schmidt, M., and Pletscher, P., 2012. "Block-Coordinate Frank-Wolfe Optimization for Structural SVMs," arXiv preprint arXiv:1207-.4747.

[LMS92] Lustig, I. J., Marsten, R. E., and Shanno, D. F., 1992. "On Implementing Mehrotra's Predictor-Corrector Interior-Point Method for Linear Programming," SIAM J. on Optimization, Vol. 2, pp. 435-449.

[LPW92] Ljung, L., Pflug, G., and Walk, H., 1992. Stochastic Approximation and Optimization of Random Systems, Birkhauser, Boston.

[LRW98] Lagarias, J. C., Reeds, J. A., Wright, M. H., and Wright, P. E., 1998. "Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions," SIAM J. on Optimization, Vol. 9, pp. 112-147.

[LST01] Lucidi, S., Sciandrone, M., and Tseng, P., 2001. "Objective-Derivative-Free Methods for Constrained Optimization," Math. Programming, Vol. 92, pp. 37-59.

[LVB98] Lobo, M. S., Vandenberghe, L., Boyd, S., and Lebret, H., 1998. "Applications of Second-Order Cone Programming," Linear Algebra and Applications, Vol. 284, pp. 193-228.

[LaD60] Land, A. H., and Doig, A. G., 1960. "An Automatic Method for Solving Discrete Programming Problems," Econometrica, Vol. 28, pp. 497-520.

[LaJ13] Lacoste-Julien, S., and Jaggi, M., 2013. "An Affine Invariant Linear Convergence Analysis for Frank-Wolfe Algorithms," arXiv preprint arXiv:1312.7864.

[LaT85] Lancaster, P., and Tismenetsky, M., 1985. The Theory of Matrices, Academic Press, N. Y.

[LaW78] Lasdon, L. S., and Waren, A. D., 1978. "Generalized Reduced Gradient Software for Linearly and Nonlinearly Constrained Problems," in Design and Implementation of Optimization Software, Greenberg, H. J., (Ed.), Sijthoff and Noordhoff, Holland, pp. 335-362.

[Lan15] Landi, G., 2015. "A Modified Newton Projection Method for ℓ_1-Regularized Least Squares Image Deblurring," J. of Mathematical Imaging and Vision, Vol. 51, pp. 195-208.

[Las70] Lasdon, L. S., 1970. Optimization Theory for Large Systems, Macmillian, N. Y.; republished by Dover Pubs, N. Y., 2002.

[Law76] Lawler, E., 1976. Combinatorial Optimization: Networks and Matroids, Holt, Reinhart, and Winston, N. Y.

[LeL10] Leventhal, D., and Lewis, A. S., 2010. "Randomized Methods for Linear Constraints: Convergence Rates and Conditioning," Math. of Operations Research, Vol. 35, pp. 641-654.

[LeO16] Lewis, A. S., and Overton, M. L., 2013. "Nonsmooth Optimization via Quasi-Newton Methods," Math. Programming, Vol. 141, pp. 135-163.

[LeP65] Levitin, E. S., and Poljak, B. T., 1965. “Constrained Minimization Methods,” Ž. Vyčisl. Mat. i Mat. Fiz., Vol. 6, pp. 787-823.

[LeS93] Lemaréchal, C., and Sagastizábal, C., 1993. “An Approach to Variable Metric Bundle Methods,” in Systems Modelling and Optimization, Proc. of the 16th IFIP-TC7 Conference, Compiègne, Henry, J., and Yvon, J.-P., (Eds.), Lecture Notes in Control and Information Sciences 197, pp. 144-162.

[Lem74] Lemarechal, C., 1974. “An Algorithm for Minimizing Convex Functions,” in Information Processing ’74, Rosenfeld, J. L., (Ed.), pp. 552-556, North-Holland, Amsterdam.

[LiM79] Lions, P. L., and Mercier, B., 1979. “Splitting Algorithms for the Sum of Two Nonlinear Operators,” SIAM J. on Numerical Analysis, Vol. 16, pp. 964-979.

[LiP87] Lin, Y. Y., and Pang, J.-S., 1987. “Iterative Methods for Large Convex Quadratic Programs: A Survey,” SIAM J. on Control and Optimization, Vol. 18, pp. 383-411.

[Lin07] Lin, C. J., 2007. “Projected Gradient Methods for Nonnegative Matrix Factorization,” Neural Computation, Vol. 19, pp. 2756-2779.

[LuB96] Lucena, A., and Beasley, J. E., 1996. “Branch and Cut Algorithms,” in Advances in Linear and Integer Programming, Beasley, J. E., (Ed.), Oxford University Press, N. Y., Chapter 5.

[LuT91] Luo, Z. Q., and Tseng, P., 1991. “On the Convergence of a Matrix-Splitting Algorithm for the Symmetric Monotone Linear Complementarity Problem,” SIAM J. on Control and Optimization, Vol. 29, pp. 1037-1060.

[LuT92a] Luo, Z. Q., and Tseng, P., 1992. “On the Convergence of the Coordinate Descent Method for Convex Differentiable Minimization,” J. Opt. Th. and Appl., Vol. 72, pp. 7-35.

[LuT92b] Luo, Z. Q., and Tseng, P., 1992. “On the Linear Convergence of Descent Methods for Convex Essentially Smooth Minimization,” SIAM J. on Control and Optimization, Vol. 30, pp. 408-425.

[LuT93a] Luo, Z. Q., and Tseng, P., 1993. “Error Bounds and Convergence Analysis of Feasible Descent Methods: A General Approach,” Annals of Operations Res., Vol. 46, pp. 157-178.

[LuT93b] Luo, Z. Q., and Tseng, P., 1993. “Error Bound and Reduced Gradient Projection Algorithms for Convex Minimization over a Polyhedral Set,” SIAM J. on Optimization, Vol. 3, pp. 43-59.

[LuT93c] Luo, Z. Q., and Tseng, P., 1993. “On the Convergence Rate of Dual Ascent Methods for Linearly Constrained Convex Minimization,” Math. of Operations Res., Vol. 18, pp. 846-867.

[LuT94a] Luo, Z. Q., and Tseng, P., 1994. “Analysis of an Approximate Gradient Projection Method with Applications to the Backpropagation Algorithm,” Optimization Methods and Software, Vol. 4, pp. 85-101.

[LuT94b] Luo, Z. Q., and Tseng, P., 1994. “On the Rate of Convergence of a Distributed Asynchronous Routing Algorithm,” IEEE Transactions on Automatic Control, Vol. 39, pp. 1123-1129.

[LuX15] Lu, Z., and Xiao, L., 2015. “On the Complexity Analysis of Randomized Block-Coordinate Descent Methods,” Math. Programming, Vol. 152, pp. 615-642.

[LuY16] Luenberger, D. G., and Ye, Y., 1916. Introduction to Linear and Nonlinear Programming, 4th edition, Springer, N. Y.

[Lue69] Luenberger, D. G., 1969. Optimization by Vector Space Methods, Wiley, N. Y.

[Lue84] Luenberger, D. G., 1984. Introduction to Linear and Nonlinear Programming, 2nd edition, Addison-Wesley, Reading, MA.

[Lue98] Luenberger, D. G., 1998. Investment Science, Oxford University Press, N. Y.

[Luo91] Luo, Z. Q., 1991. “On the Convergence of the LMS Algorithm with Adaptive Learning Rate for Linear Feedforward Networks,” Neural Computation, Vol. 3, pp. 226-245.

[Luq84] Luque, F.J., 1984. “Asymptotic Convergence Analysis of the Proximal Point Algorithm,” SIAM J. on Control and Optimization, Vol. 22, pp. 277-293.

[MMS91] McShane, K. A., Monma, C. L., and Shanno, D., 1991. “An Implementation of a Primal-Dual Interior Point Method for Linear Programming,” ORSA J. on Computing, Vol. 1, pp. 70-83.

[MMZ95] McKenna, M. P., Mesirov, J. P., and Zenios, S. A., 1995. “Data Parallel Quadratic Programming on Box-Constrained Problems,” SIAM J. on Optimization, Vol. 5, pp. 570-589.

[MOT95] Mahey, P., Oualibouch, S., and Tao, P. D., 1995. “Proximal Decomposition on the Graph of a Maximal Monotone Operator,” SIAM J. on Optimization, Vol. 5, pp. 454-466.

[MRR00] Mayne, D. Q., Rawlings, J. B., Rao, C. V., and Scokaert, P. O. M., 2000. “Constrained Model Predictive Control: Stability and Optimality,” Automatica, Vol. 36, pp. 789-814.

[MSQ98] Mifflin, R., Sun, D., and Qi, L., 1998. “Quasi-Newton Bundle-Type Methods for Nondifferentiable Convex Optimization,” SIAM J. on Optimization, Vol. 8, pp. 583-603.

[MTW93] Monteiro, R. D. C., Tsuchiya, T., and Wang, Y., 1993. “A Simplified Global Convergence Proof of the Affine Scaling Algorithm,” Annals of Operations Res., Vol. 47, pp. 443-482.

[MYF03] Moriyama, H., Yamashita, N., and Fukushima, M., 2003. “The Incremental Gauss-Newton Algorithm with Adaptive Stepsize Rule,” Computational Optimization and Applications, Vol. 26, pp. 107-141.

[MaD87] Mangasarian, O. L., and De Leone, R., 1987. “Parallel Successive Overelaxation Methods for Symmetric Linear Complementarity Problems and Linear Programs,” J. of Optimization Th. and Appl., Vol. 54, pp. 437-446.

[MaD88] Mangasarian, O. L., and De Leone, R., 1988. “Parallel Gradient Projection Successive Overrelaxation for Symmetric Linear Complementarity Problems,” Annals of Operations Res., Vol. 14, pp. 41-59.

[MaF67] Mangasarian, O. L., and Fromovitz, S., 1967. “The Fritz John Necessary Optimality Conditions in the Presence of Equality and Inequality Constraints,” J. Math. Anal. and Appl., Vol. 17, pp. 37-47.

[MaP82] Mayne, D. Q., and Polak, E., 1982. “A Superlinearly Convergent Algorithm for Constrained Optimization Problems,” Math. Programming Studies, Vol. 16, pp. 45-61.

[MaS94] Mangasarian, O. L., and Solodov, M. V., 1994. “Serial and Parallel Backpropagation Convergence Via Nonmonotone Perturbed Minimization,” Optimization Methods and Software, Vol. 4, pp. 103-116.

[Mai13] Mairal, J., 2013. “Optimization with First-Order Surrogate Functions,” arXiv preprint arXiv:1305.3120.

[Mai14] Mairal, J., 2014. “Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning,” arXiv preprint arXiv:1402.4419.

[Man69] Mangasarian, O. L., 1969. Nonlinear Programming, Prentice-Hall, Englewood Cliffs, N. J.; also SIAM, Classics in Applied Mathematics 10, Phila., PA., 1994.

[Mar70] Martinet, B., 1970. "Regularisation d'Inequations Variationnelles par Approximations Successives," Rev. Francaise Inf. Rech. Oper., Vol. 4, pp. 154-158.

[Mar72] Martinet, B., 1972. "Determination Approchee d'un Point Fixe d'une Application Pseudo-Contractante," C. R. Acad. Sci. Paris, 274A, pp. 163-165.

[Mar78] Maratos, N., 1978. "Exact Penalty Function Algorithms for Finite Dimensional and Control Optimization Problems," Ph.D. Thesis, Imperial College Sci. Tech, Univ. of London.

[McK98] McKinnon, K. I. M., 1998. "Convergence of the Nelder-Mead Simplex Method to a Non-Stationary Point," SIAM J. on Optimization, Vol. 9, pp. 148-158.

[McL80] McLinden, L., 1980. "The Complementarity Problem for Maximal Monotone Multifunctions," in Variational Inequalities and Complementarity Problems, Cottle, R., Giannessi, F., and Lions, J.-L., (Eds.), Wiley, N. Y., pp. 251-270.

[McS73] McShane, E. J., 1973. "The Lagrange Multiplier Rule," American Mathematical Monthly, Vol. 80, pp. 922-925.

[Meg88] Megiddo, N., 1988. "Pathways to the Optimal Set in Linear Programming," in Progress in Mathematical Programming, Megiddo, N., (Ed.), Springer-Verlag, N. Y., pp. 131-158.

[Meh92] Mehrotra, S., 1992. "On the Implementation of a Primal-Dual Interior Point Method," SIAM J. on Optimization, Vol. 2, pp. 575-601.

[Mey79] Meyer, R. R., 1979. "Two-Segment Separable Programming," Management Science, Vol. 25, pp. 385-395.

[Mey07] Meyn, S., 2007. Control Techniques for Complex Networks, Cambridge Univ. Press, N. Y.

[Mif96] Mifflin, R., 1996. "A Quasi-Second-Order Proximal Bundle Algorithm," Math. Programming, Vol. 73, pp. 51-72.

[Mig94] Migdalas, A., 1994. "A Regularization of the Frank-Wolfe Method and Unification of Certain Nonlinear Programming Methods," Math. Programming, Vol. 65, pp. 331-345.

[Min60] Minty, G. J., 1960. "Monotone Networks," Proc. Roy. Soc. London, A, Vol. 257, pp. 194-212.

[Min86] Minoux, M., 1986. Mathematical Programming: Theory and Algorithms, Wiley, N. Y.

[Mit66] Mitter, S. K., 1966. "Successive Approximation Methods for the Solution of Optimal Control Problems," Automatica, Vol. 3, pp. 135-149.

[MoA89a] Monteiro, R. D. C., and Adler, I., 1989. "Interior Path Following Primal-Dual Algorithms, Part I: Linear Programming," Math. Programming, Vol. 44, pp. 27-41.

[MoA89b] Monteiro, R. D. C., and Adler, I., 1989. "Interior Path Following Primal-Dual Algorithms, Part II: Convex Quadratic Programming," Math. Programming, Vol. 44, pp. 43-66.

[MoL99] Morari, M., and Lee, J. H., 1999. "Model Predictive Control: Past, Present, and Future," Computers and Chemical Engineering, Vol. 23, pp. 667-682.

[MoS83] Moré, J. J., and Sorensen, D. C., 1983. "Computing a Trust Region Step," SIAM J. on Scientific and Statistical Computing, Vol. 4, pp. 553-572.

[MoT89] Moré, J. J., and Toraldo, G., 1989. "Algorithms for Bound Constrained Quadratic Programming Problems," Numer. Math., Vol. 55, pp. 377-400.

[MoW93] Moré, J. J., and Wright, S. J., 1993. Optimization Software Guide, SIAM, Frontiers in Applied Mathematics 14, Phila., PA.

[Mor88] Mordukhovich, B. S., 1988. Approximation Methods in Problems of Optimization and Control, Nauka, Moscow.

[Mor06] Mordukhovich, B. S., 2006. Variational Analysis and Generalized Differentiation I: Basic Theory, Springer, N. Y.

[MuP75] Mukai, H., and Polak, E., 1975. "A Quadratically Convergent Primal-Dual Algorithm with Global Convergence Properties for Solving Optimization Problems with Equality Constraints," Math. Programming, Vol. 9, pp. 336-349.

[MuR95] Mulvey, J. M., and Ruszcynski, A., 1995. "A New Scenario Decomposition Method for Large Scale Stochastic Optimization," Operations Research, Vol. 43, pp. 477-490.

[MuS87] Murtagh, B. A., and Saunders, M. A., 1987. "MINOS 5.1 User's Guide," Technical Report SOL-83-20R, Stanford Univ.

[Mur92] Murty, K. G., 1992. Network Programming, Prentice-Hall, Englewood Cliffs, N. J.

[NBB01] Nedić, A., Bertsekas, D. P., and Borkar, V. S., 2001. "Distributed Asynchronous Incremental Subgradient Methods," in Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, Butnariu, D., Censor, Y., and Reich, S., (Eds.), Elsevier Science, Amsterdam, Netherlands.

[NSL15] Nutini, J., Schmidt, M., Laradji, I. H., Friedlander, M., and Koepke, H., 2015. "Coordinate Descent Converges Faster with the Gauss-Southwell Rule Than Random Selection," arXiv preprint arXiv:1506.00552.

[NaQ96] Nazareth, J. L., and Qi, L., 1996. "Globalization of Newton's Method for Solving Nonlinear Equations," Numerical Linear Algebra with Applications, Vol. 3, pp. 239-249.

[NaS89] Nash, S. G., and Sofer, 1989. "Block Truncated-Newton Methods for Parallel Optimization," Math. Programming, Vol. 45, pp. 529-546.

[NaT02] Nazareth, L., and Tseng, P., 2002. "Gilding the Lily: A Variant of the Nelder-Mead Algorithm Based on Golden-Section Search," Computational Optimization and Applications, Vol. 22, pp. 133-144.

[Nag93] Nagurney, A., 1993. Network Economics: A Variational Inequality Approach, Kluwer, Dordrecht, The Netherlands.

[Nas85] Nash, S. G., 1985. "Preconditioning of Truncated-Newton Methods," SIAM J. on Scientific and Statistical Computing, Vol. 6, pp. 599-616.

[Naz94] Nazareth, J. L., 1994. The Newton-Cauchy Framework: A Unified Approach to Unconstrained Nonlinear Minimization, Lecture Notes in Computer Science No. 769, Springer-Verlag, Berlin and New York.

[Naz96] Nazareth, J. L., 1996. "Lagrangian Globalization: Solving Nonlinear Equations via Constrained Optimization," in Mathematics of Numerical Analysis, Renegar, J., Shub, M., and Smale, S., (Eds.), Lectures in Applied Mathematics, Vol. 32, The American Mathematical Society, Providence, RI, pp. 533-542.

[NeB00] Nedić, A., and Bertsekas, D. P., 2000. "Convergence Rate of Incremental Subgradient Algorithms," Stochastic Optimization: Algorithms and Applications," S. Uryasev and P. M. Pardalos, Eds., Kluwer, pp. 263-304.

[NeB01] Nedić, A., and Bertsekas, D. P., 2001. "Incremental Subgradient Methods for Nondifferentiable Optimization," SIAM J. on Optimization, Vol. 12, pp. 109-138.

[NeB10] Nedić, A., and Bertsekas, D. P., 2010. "The Effect of Deterministic Noise in Subgradient Methods," Math. Programming, Ser. A, Vol. 125, pp. 75-99.

[NeN94] Nesterov, Y., and Nemirovskii, A., 1994. Interior Point Polynomial Algorithms in Convex Programming, SIAM, Studies in Applied Mathematics 13, Phila., PA.

[NeW88] Nemhauser, G. L., and Wolsey, L. A., 1988. Integer and Combinatorial Optimization, Wiley, N. Y.

[NeY83] Nemirovsky, A., and Yudin, D. B., 1983. Problem Complexity and Method Efficiency, Wiley, N. Y.

[Ned11] Nedić, A., 2011. "Random Algorithms for Convex Minimization Problems," Math. Programming, Vol. 129, pp. 225?253.

[Nes83] Nesterov, Y., 1983. "A Method for Unconstrained Convex Minimization Problem with the Rate of Convergence $O(1/k^2)$," Doklady AN SSSR, Vol. 269, pp. 543-547; translated as Soviet Math. Dokl.

[Nes95] Nesterov, Y., 1995. "Complexity Estimates of Some Cutting Plane Methods Based on Analytic Barrier," Math. Programming, Vol. 69, pp. 149-176.

[Nes04] Nesterov, Y., 2004. Introductory Lectures on Convex Optimization, Kluwer Academic Publisher, Dordrecht, The Netherlands.

[Nes12] Nesterov, Y., 2012. "Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems," SIAM J. on Optimization, Vol. 22, pp. 341-362.

[Neu28] Neumann, J. von, 1928. "Zur Theorie der Gesellschaftsspiele," Math. Ann., Vol. 100, pp. 295-320.

[NgS79] Nguyen, V. H., and Strodiot, J. J., 1979. "On the Convergence Rate of a Penalty Function Method of Exponential Type," J. Opt. Th. and Appl., Vol. 27, pp. 495-508.

[NoW06] Nocedal, J., and Wright, S. J., 2006. Numerical Optimization, 2nd Edition, Springer, NY.

[Noc80] Nocedal, J., 1980. "Updating Quasi-Newton Matrices with Limited Storage," Math. of Computation, Vol. 35, pp. 773-782.

[OLR85] O'hEigeartaigh, M., Lenstra, S. K., and Rinnoy Kan, A. H. G., (Eds.), 1985. Combinatorial Optimization: Annotated Bibliographies, Wiley, N. Y.

[OrL74] Oren, S. S., and Luenberger, D. G., 1974. "Self-Scaling Variable Metric Algorithm, Part I," Management Science, Vol. 20, pp. 845-862.

[OrR70] Ortega, J. M., and Rheinboldt, W. C., 1970. Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, N. Y.

[Ore73] Oren, S. S., 1973. "Self-Scaling Variable Metric Algorithm, Part II," Management Science, Vol. 20, pp. 863-874.

[OzB03] Ozdaglar, A. E., and Bertsekas, D. P., 2003. "Routing and Wavelength Assignment in Optical Networks," IEEE Trans. on Networking, pp. 259-272.

[OzB04] Ozdaglar, A. E., and Bertsekas, D. P., 2004. "The Relation Between Pseudonormality and Quasiregularity in Constrained Optimization," Optimization Methods and Software, Vol. 19, pp. 493–506.

[PCC93] Pulleyblank, W., Cook, W., Cunningham, W., and Schrijver, A., 1993. An Introduction to Combinatorial Optimization, Wiley, N. Y.

[PaM89] Pantoja, J. F. A. D., and Mayne, D. Q., 1989. "Sequential Quadratic Programming Algorithm for Discrete Optimal Control Problems with Control Inequality Constraints," Intern. J. on Control, Vol. 53, pp. 823-836.

[PaR87] Pardalos, P. M., and Rosen, J. B., 1987. Constrained Global Optimization: Algorithms and Applications, Springer-Verlag, N. Y.

[PaS82] Papadimitriou, C. H., and Steiglitz, K., 1982. Combinatorial Optimization: Algorithms and Complexity, Prentice-Hall, Englewood Cliffs, N. J.

[PaT91] Panier, E. R., and Tits, A. L., 1991. "Avoiding the Maratos Effect by Means of a Nonmonotone Line Search. I.," SIAM J. on Numer. Anal., Vol. 28, pp. 1183-1195.

[Pan84] Pang, J.-S., 1984. "On the Convergence of Dual Ascent Methods for Large-Scale Linearly Constrained Optimization Problems," Unpublished Manuscript, School of Management, Univ. of Texas, Dallas, Texas.

[Pap81] Papavassilopoulos, G., 1981. "Algorithms for a Class of Nondifferentiable Problems," J. Opt. Th. and Appl., Vol. 34, pp. 41-82.

[Pap82] Pappas, T. N., 1982. "Solution of Nonlinear Equations by Davidon's Least Squares Method," M.S. Thesis, Dept. of Electrical Engineering and Computer Science, M.I.T., Cambridge, MA.

[Pat93] Patriksson, M., 1993. "Partial Linearization Methods for Nonlinear Programming," J. Opt. Th. and Appl., Vol. 78, pp. 227-246.

[Pat98] Patriksson, M., 1998. Nonlinear Programming and Variational Inequalities: A Unified Approach, Kluwer, Dordtrecht, The Netherlands.

[Per78] Perry, A., 1978. "A Modified Conjugate Gradient Algorithm," Operations Research, Vol. 26, pp. 1073-1078.

[Pfl96] Pflug, G. C., 1996. Optimization of Stochastic Models, Kluwer, Boston.

[PiP73] Pironneau, O., and Polak, E., 1973. "Rate of Convergence of a Class of Methods of Feasible Directions," SIAM J. Numer. Anal., Vol. 10, pp. 161-173.

[PiZ92] Pinar, M. C., and Zenios, S. A., 1992. "Parallel Decomposition of Multicommodity Network Flows Using a Linear-Quadratic Penalty Algorithm," ORSA J. on Computing, Vol. 4, pp. 235-249.

[PiZ94] Pinar, M. C., and Zenios, S. A., 1994. "On Smoothing Exact Penalty Functions for Convex Constrained Problems," SIAM J. on Optimization, Vol. 4, pp. 486-511.

[PoH91] Polak, E., and He, L., 1991. "Finite-Termination Schemes for Solving Semi-infinite Satisfycing Problems," J. Opt. Theory and Appl., Vol. 70, pp. 429-442.

[PoR69] Polak, E., and Ribiere, G., 1969. "Note sur la Convergence de Methodes de Directions Conjugees," Rev. Fr. Inform. Rech. Oper., Vol. 16-R1, pp. 35-43.

[PoT73a] Poljak, B. T., and Tsypkin, Y. Z., 1973. "Pseudogradient Adaptation and Training Algorithms," Automation and Remote Control, pp. 45-68.

[PoT73b] Poljak, B. T., and Tretjakov, N. V., 1973. "The Method of Penalty Estimates for Conditional Extremum Problems," Z. VyČisl. Mat. i Mat. Fiz., Vol. 13, pp. 34-46.

[PoT80a] Polak, E., and Tits, A. L., 1980. "A Globally Convergent, Implementable Multiplier Method with Automatic Penalty Limitation," Applied Math. and Optimization, Vol. 6, pp. 335-360.

[PoT80b] Poljak, B. T., and Tsypkin, Y. Z., 1980. "Adaptive Estimation Algorithms (Convergence, Optimality, Stability)," Automation and Remote Control, Vol. 40, pp. 378-389.

[PoT81] Poljak, B. T., and Tsypkin, Y. Z., 1981. "Optimal Pseudogradient Adaptation Algorithms," Automation and Remote Control, Vol. 41, pp. 1101-1110.

[PoT97] Polyak, R., and Teboulle, M., 1997. "Nonlinear Rescaling and Proximal-Like Methods in Convex Optimization," Math. Programming, Vol. 76, pp. 265-284.

[Pol64] Poljak, B. T., 1964. "Some Methods of Speeding up the Convergence of Iteration Methods," Z. VyČisl. Mat. i Mat. Fiz., Vol. 4, pp. 1-17.

[Pol69a] Poljak, B. T., 1969. "The Conjugate Gradient Method in Extremal Problems," Z. Vyčisl. Mat. i Mat. Fiz., Vol. 9, pp. 94-112.

[Pol69b] Poljak, B. T., 1969. "Minimization of Unsmooth Functionals," Z. Vyčisl. Mat. i Mat. Fiz., Vol. 9, pp. 14-29.

[Pol70] Poljak, B. T., 1970. "Iterative Methods Using Lagrange Multipliers for Solving Extremal Problems with Constraints of the Equation Type," Z. VyČisl. Mat. i Mat. Fiz., Vol. 10, pp. 1098-1106.

[Pol71] Polak, E., 1971. Computational Methods in Optimization: A Unified Approach, Academic Press, N. Y.

[Pol73] Polak, E., 1973. "A Historical Survey of Computational Methods in Optimal Control," SIAM Review, Vol. 15, pp. 553-584.

[Pol79] Poljak, B. T., 1979. "On Bertsekas' Method for Minimization of Composite Functions," Internat. Symp. Systems Opt. Analysis, Benoussan, A., and Lions, J. L., (Eds.), pp. 179-186, Springer-Verlag, Berlin and N. Y.

[Pol87] Poljak, B. T., 1987. Introduction to Optimization, Optimization Software Inc., N. Y.

[Pol92] Polyak, R., 1992. "Modified Barrier Functions (Theory and Methods)," Math. Programming, Vol. 54, pp. 177-222.

[Pol97] Polak, E., 1997. Optimization: Algorithms and Consistent Approximations, Springer-Verlag, N. Y.

[Pot94] Potra, F. A., 1994. "A Quadratically Convergent Predictor-Corrector Method for Solving Linear Programs from Infeasible Starting Points," Math. Programming, Vol. 67, pp. 383–406.

[Pow64] Powell, M. J. D., 1964. "An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives," The Computer Journal, Vol. VII, pp. 155-162.

[Pow69] Powell, M. J. D., 1969. "A Method for Nonlinear Constraints in Minimizing Problems," in Optimization, Fletcher, R., (Ed.), Academic Press, N. Y, pp. 283-298.

[Pow73] Powell, M. J. D., 1973. "On Search Directions for Minimization Algorithms," Math. Programming, Vol. 4, pp. 193-201.

[Pre95] Prekopa, A., 1995. Stochastic Programming, Kluwer, Boston.

[PsD75] Pschenichny, B. N., and Danilin, Y. M., 1975. "Numerical Methods in Extremal Problems," MIR, Moscow, (Engl. trans., 1978).

[Psc70] Pschenichny, B. N., 1970. "Algorithms for the General Problem of Mathematical Proramming," Kibernetika (Kiev), Vol. 6, pp. 120-125.

[Pyt98] Pytlak, R., 1998. "An Efficient Algorithm for Large-Scale Nonlinear Programming Problems with Simple Bounds on the Variables," SIAM J. on Optimization, Vol. 8, pp. 532-560.

[RGV14] Richard, E., Gaiffas, S., and Vayatis, N., 2014. "Link Prediction in Graphs with Autoregressive Features," J. of Machine Learning Research, Vol. 15, pp. 565-593.

[RHL13] Razaviyayn, M., Hong, M., and Luo, Z. Q., 2013. "A Unified Convergence Analysis of Block Successive Minimization Methods for Nonsmooth Optimization," SIAM J. on Optimization, Vol. 23, pp. 1126-1153.

[RSP16] Reddi, S. J., Sra, S., Poczos, B., and Smola, A., 2016. "Fast Incremental Method for Nonconvex Optimization," arXiv preprint arXiv:1603.06159.

[Ray93] Raydan, M., 1993. "On the Barzilai and Borwein Choice of Steplength for the Gradient Method," IMA J. Num. Anal., Vol. 13, pp. 321-326.

[Ray97] Raydan, M., 1997. "The Barzilai and Borwein Gradient Method for the Large Scale Unconstrained Minimization Problem," SIAM J. on Optimization, Vol. 7, pp. 26-33.

[ReR98] Reemtsen, R., and Ruckman, J. J., (Eds.), 1998. Semi-Infinite Programming, Kluwer, Boston.

[Ren01] Renegar, J., 2001. A Mathematical View of Interior-Point Methods in Convex Optimization, SIAM, Phila.

[RiT14] Richtarik, P., and Takac, M., 2014. "Iteration Complexity of Randomized Block-Coordinate Descent Methods for Minimizing a Composite Function," Math. Programming, Vol. 144, pp. 1-38.

[RoW91] Rockafellar, R. T., and Wets, R. J.-B., 1991. "Scenarios and Policy Aggregation in Optimization under Uncertainty," Math. of Operations Res., Vol. 16, pp. 119-147.

[RoW98] Rockafellar, R. T., and Wets, R. J.-B., 1998. Variational Analysis, Springer-Verlag, Berlin.

[Rob74] Robinson, S. M., 1974. "Perturbed Kuhn-Tucker Points and Rates of Convergence for a Class of Nonlinear Programming Algorithms," Math. Programming, Vol. 7, pp. 1-16.

[Rob87] Robinson, S. M., 1987. "Local Structure of Feasible Sets in Nonlinear Programming, Part III. Stability and Sensitivity," Math. Programming Studies, Vol. 30, pp. 45-66.

[Roc67] Rockafellar, R. T., 1967. "Convex Programming and Systems of Elementary Monotonic Relations," J. of Math. Analysis and Applications, Vol. 19, pp. 543-564.

[Roc70] Rockafellar, R. T., 1970. Convex Analysis, Princeton Univ. Press, Princeton, N. J.

[Roc73a] Rockafellar, R. T., 1973. "A Dual Approach to Solving Nonlinear Programming Problems by Unconstrained Optimization," Math. Programming, pp. 354-373.

[Roc73b] Rockafellar, R. T., 1973. "The Multiplier Method of Hestenes and Powell Applied to Convex Programming," J. Opt. Th. and Appl., Vol. 12, pp. 555-562.

[Roc74] Rockafellar, R. T., 1974. "Augmented Lagrange Multiplier Functions and Duality in Nonconvex Programming," SIAM J. on Control, Vol. 12, pp. 268-285.

[Roc76a] Rockafellar, R. T., 1976. "Monotone Operators and the Proximal Point Algorithm," SIAM J. on Control and Optimization, Vol. 14, pp. 877-898.

[Roc76b] Rockafellar, R. T., 1976. "Augmented Lagrangians and Applications of the Proximal Point Algorithm in Convex Programming," Math. Operations Res., Vol. 1, pp. 97-116.

[Roc76c] Rockafellar, R. T., 1976. "Solving a Nonlinear Programming Problem by Way of a Dual Problem," Symp. Matematica, Vol. 27, pp. 135-160.

[Roc81] Rockafellar, R. T., 1981. "Monotropic Programming: Descent Algorithms and Duality," in Nonlinear Programming 4, by Mangasarian, O. L., Meyer, R. R., and Robinson, S. M., (Eds.), Academic Press, N. Y., pp. 327-366.

[Roc84] Rockafellar, R. T., 1984. Network Flows and Monotropic Optimization, Wiley, N. Y.; republished by Athena Scientific, Belmont, MA, 1998.

[Roc90] Rockafellar, R. T., 1990. "Computational Schemes for Solving Large-Scale Problems in Extended Linear-Quadratic Programming," Math. Programming, Vol. 48, pp. 447-474.

[Roc93] Rockafellar, R. T., 1993. "Lagrange Multipliers and Optimality," SIAM Review, Vol. 35, pp. 183-238.

[Ros60a] Rosenbrock, H. H., 1960. "An Automatic Method for Finding the Greatest or Least Value of a Function," Computer J., Vol. 3, pp. 175-184.

[Ros60b] Rosen, J. B., 1960. "The Gradient Projection Method for Nonlinear Programming, Part I, Linear Constraints," SIAM J. Applied Math., Vol. 8, pp. 514-553.

[RuK04] Rubinstein, R. Y., and Kroese, D. P., 2004. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Springer, N. Y.

[Rud76] Rudin, W., 1976. Real Analysis, McGraw-Hill, N. Y.

[Rus86] Ruszczynski, A., 1986. "A Regularized Decomposition Method for Minimizing a Sum of Polyhedral Functions," Math. Programming, Vol. 35, pp. 309-333.

[Rus89] Ruszczynski, A., 1989. "An Augmented Lagrangian Decomposition Method for Block Diagonal Linear Programming Problems," Operations Res. Letters, Vol. 8, pp. 287-294.

[Rus95] Ruszczynski, A., 1995. "On Convergence of an Augmented Lagrangian Decomposition Method for Sparse Convex Optimization," Math. of Operations Res., Vol. 20, pp. 634-656.

[Rus97] Ruszczynski, A., 1997. "Decomposition Methods in Stochastic Programming," Math. Programming, Vol. 79, pp. 333-353.

[Rus06] Ruszczynski, A., 2006. Nonlinear Optimization, Princeton Univ. Press, Princeton, N. J.

[SBC93] Saarinen, S., Bramley, R., and Cybenko, G., 1993. "Ill-Conditioning in Neural Network Training Problems," SIAM J. Sci. Comput., Vol. 14, pp. 693-714.

[SBK64] Shah, B., Buehler, R., and Kempthorne, O., 1964. "Some Algorithms for Minimizing a Function of Several Variables," J. Soc. Indust. Appl. Math., Vol. 12, pp. 74-92.

[SFR09] Schmidt, M., Fung, G., and Rosales, R., 2009. "Optimization Methods for ℓ_1-Regularization," Univ. of British Columbia, Technical Report TR-2009-19.

[SHH62] Spendley, W. G., Hext, G. R., and Himsworth, F. R., 1962. "Sequential Application of Simplex Designs in Optimisation and Evolutionary Operation," Technometrics, Vol. 4, pp. 441-461.

[SHM16] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser J., et al. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search," Nature, Vol. 529, pp. 484-489.

[SKS12] Schmidt, M., Kim, D., and Sra, S., 2012. "Projected Newton-Type Methods in Machine Learning," in Optimization for Machine Learning, by Sra, S., Nowozin, S., and Wright, S. J., (eds.), MIT Press, Cambridge, MA, pp. 305-329.

[SLB13] Schmidt, M., Le Roux, N., and Bach, F., 2013. "Minimizing Finite Sums with the Stochastic Average Gradient," arXiv preprint arXiv:1309.2388.

[SaS86] Saad, Y., and Schultz, M. H., 1986. "GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems," SIAM J. Sci. Statist. Comput., Vol. 7, pp. 856-869.

[Sah96] Sahinidis, N. V., 1996. "BARON: A General Purpose Global Optimization Software Package," Journal of Global Optimization, Vol. 8, pp. 201-205.

[Sah04] Sahinidis, N. V., 2004. "Optimization Under Uncertainty: State-of-the-Art and Opportunities," Computers and Chemical Engineering, Vol. 28, pp. 971-983.

[Sak66] Sakrison, D. T., 1966. "Stochastic Approximation: A Recursive Method for Solving Regression Problems," in Advances in Communication Theory and Applications, 2, A. V. Balakrishnan, ed., Academic Press, NY, pp. 51-106.

[ScF14] Schmidt, M., and Friedlander, M. P., 2014. "Coordinate Descent Converges Faster with the Gauss-Southwell Rule than Random Selection," Advances in Neural Information Processing Systems 27 (NIPS 2014).

[Sch82] Schnabel, R. B., 1982. "Determining Feasibility of a Set of Nonlinear Inequality Constraints," Math. Programming Studies, Vol. 16, pp. 137-148.

[Sch86] Schrijver, A., 1986. Theory of Linear and Integer Programming, Wiley, N. Y.

[Sch93] Schrijver, A., 1993. Combinatorial Optimization: Polyhedra and Efficiency, Springer, N. Y.

[Sch10] Schmidt, M., 2010. "Graphical Model Structure Learning with L1-Regularization," PhD Thesis, Univ. of British Columbia.

[Sch12] Schittkowski, K., 2012. Nonlinear Programming Codes: Information, Tests, Performance, Springer Science and Business Media.

[SeS86] Sen, S., and Sherali, H. D., 1986. "A Class of Convergent Primal-Dual Subgradient Algorithms for Decomposable Convex Programs," Math. Programming, Vol. 35, pp. 279-297.

[ShA99] Sherali, H. D., and Adams, W. P., 1999. A Reformulation-Linearization Technique for Solving Discrete and Continuous Nonconvex Problems, Kluwer, Boston.

[ShD14] Shapiro, A., and Dentcheva D., 2014. Lectures on Stochastic Programming: Modeling and Theory, SIAM, Phila.

[Sha70] Shanno, D. F., 1970. "Conditioning of Quasi-Newton Methods for Function Minimization," Math. Comput., Vol. 27, pp. 647-656.

[Sha78] Shanno, D. F., 1978. "Conjugate Gradient Methods with Inexact Line Searches," Math. of Operations Res., Vol. 3, pp. 244-256.

[Sha79] Shapiro, J. E., 1979. Mathematical Programming Structures and Algorithms, Wiley, N. Y.

[Sha88] Shapiro, A., 1988. "Sensitivity Analysis of Nonlinear Programs and Differentiability Properties of Metric Projections," SIAM J. on Control and Optimization, Vol. 26, pp. 628-645.

[Sho85] Shor, N. Z., 1985. Minimization Methods for Nondifferentiable Functions, Springer-Verlag, Berlin.

[Sho98] Shor, N. Z., 1998. Nondifferentiable Optimization and Polynomial Problems, Kluwer, Dordrecht, the Netherlands.

[Sla50] Slater, M., 1950. "Lagrange Multipliers Revisited: A Contribution to Non-Linear Programming," Cowles Commission Discussion Paper, Math. 403.

[Sol98] Solodov, M. V., 1998. "Incremental Gradient Algorithms with Stepsizes Bounded Away from Zero," Computational Optimization and Applications, Vol. 11, pp. 23-35.

[Son86] Sonnevend, G., 1986. "An "Analytical Centre" for Polyhedrons and New Classes of Global Algorithms for Linear (Smooth, Convex) Programming," Lecture Notes in Control and Information Sciences, Vol. 84, pp. 866-878.

[Spa03] Spall, J. C., 2003. Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control, J. Wiley, Hoboken, N. J.

[Spa12] Spall, J. C., 2012. "Cyclic Seesaw Process for Optimization and Identification," J. of Optimization Theory and Applications, Vol. 154, pp. 187-208.

[Spi85] Spingarn, J. E., 1985. "Applications of the Method of Partial Inverses to Convex Programming: Decomposition," Math. Programming, Vol. 32, pp. 199-223.

[StW75] Stephanopoulos, G., and Westerberg, A. W., 1975. "The Use of Hestenes' Method of Multipliers to Resolve Dual Gaps in Engineering System Optimization," J. Opt. Th. and Applications, Vol. 15, pp. 285-309.

[Str76] Strang, G., 1976. Linear Algebra and Its Applications, Academic Press, N. Y.

[TBA86] Tsitsiklis, J. N., Bertsekas, D. P., and Athans, M., 1986. "Distributed Asynchronous Deterministic and Stochastic Gradient Optimization Algorithms," IEEE Trans. on Aut. Control, Vol. AC-31, pp. 803-812.

[TBT90] Tseng, P., Bertsekas, D. P., and Tsitsiklis, J. N., 1990. "Partially Asynchronous Algorithms for Network Flow and Other Problems," SIAM J. on Control and Optimization, Vol. 28, pp. 678-710.

[TTA15] Toulis, P., Tran, D., and Airoldi, E. M., 2015. "Stability and Optimality in Stochastic Gradient Descent," arXiv preprint arXiv:1505.02417.

[TZY95] Tapia, R. A., Zhang, Y., and Ye, Y., 1995. "On the Convergence of the Iteration Sequence in Primal-Dual Interior-Point Methods," Math. Programming, Vol. 68, pp. 141-154.

[TaM85] Tanikawa, A., and Mukai, H., 1985. "A New Technique for Nonconvex Primal-Dual Decomposition," IEEE Trans. on Aut. Control, Vol. AC-30, pp. 133-143.

[TaP13] Talischi, C., and Paulino, G. H., 2013. "A Consistent Operator Splitting Algorithm and a Two-Metric Variant: Application to Topology Optimization," arXiv preprint arXiv:1307.5100.

[Tap77] Tapia, R. A., 1977. "Diagonalized Multiplier Methods and Quasi-Newton Methods for Constrained Minimization," J. Opt. Th. and Applications, Vol. 22, pp. 135-194.

[Teb92] Teboulle, M., 1992. "Entropic Proximal Mappings with Applications to Nonlinear Programming," Math. Operations Res., Vol. 17, pp. 1-21.

[ToT90] Toint, P. L., and Tuyttens, D., 1990. "On Large Scale Nonlinear Network Optimization," Math. Programming, Vol. 48, pp. 125-159.

[ToV67] Topkis, D. M., and Veinott, A. F., 1967. "On the Convergence of Some Feasible Directions Algorithms for Nonlinear Programming," SIAM J. on Control, Vol. 5, pp. 268-279.

[Tor91] Torczon, V., 1991. "On the Convergence of the Multidimensional Search Algorithm," SIAM J. on Optimization, Vol. 1, pp. 123-145.

[TrW80] Traub, J. F., and Wozniakowski, H., 1980. A General Theory of Optimal Algorithms, Academic Press, N. Y.

[TsB86] Tsitsiklis, J. N., and Bertsekas, D. P., 1986. "Distributed Asynchronous Optimal Routing in Data Networks," IEEE Trans. on Automatic Control, Vol. 31, pp. 325-331.

[TsB87] Tseng, P., and Bertsekas, D. P., 1987. "Relaxation Methods for Problems with Strictly Convex Separable Costs and Linear Constraints," Math. Programming, Vol. 38, pp. 303-321.

[TsB90] Tseng, P., and Bertsekas, D. P., 1990. "Relaxation Methods for Monotropic Programs," Math. Programming, Vol. 46, pp. 127-151.

[TsB91] Tseng, P., and Bertsekas, D. P., 1991. "Relaxation Methods for Problems with Strictly Convex Costs and Linear Constraints," Math. Operations Res., Vol. 16, pp. 462-481.

[TsB93] Tseng, P., and Bertsekas, D. P., 1993. "On the Convergence of the Exponential Multiplier Method for Convex Programming," Math. Programming, Vol. 60, pp. 1-19.

[TsB00] Tseng, P., and Bertsekas, D. P., 2000. "An Epsilon-Relaxation Method for Separable Convex Cost Generalized Network Flow Problems," Math. Programming, Vol. 88, pp. 85-104.

[Tse89] Tseng, P., 1989. "A Simple Complexity Proof for a Polynomial-Time Linear Programming Algorithm," Operations Res. Letters, Vol. 8, pp. 155-159.

[Tse90] Tseng, P., 1990. "Dual Ascent Methods for Problems with Strictly Convex Costs and Linear Constraints: A Unified Approach," SIAM J. on Control and Optimization, Vol. 28, pp. 214-242.

[Tse91a] Tseng, P., 1991. "On the Rate of Convergence of a Partially Asynchronous Gradient Projection Algorithm," SIAM J. on Optimization, Vol. 4, pp. 603-619.

[Tse91b] Tseng, P., 1991. "Relaxation Method for Large Scale Linear Programming using Decomposition," Math. of Operations Res., Vol. 17, pp. 859-880.

[Tse91c] Tseng, P., 1991. "Decomposition Algorithm for Convex Differentiable Minimization," J. Opt. Theory and Appl., Vol. 70, pp. 109-135.

[Tse92] Tseng, P., 1992. "Complexity Analysis of a Linear Complementarity Algorithm Based on a Lyapunov Function," Math. Programming, Vol. 53, pp. 297-306.

[Tse93] Tseng, P., 1993. "Dual Coordinate Ascent Methods for Non-Strictly Convex Minimization," Math. Programming, Vol. 59, pp. 231-247.

[Tse95a] Tseng, P., 1995. "Fortified-Descent Simplicial Search Method," Report, Dept. of Math., University of Washington, Seattle, Wash.; also in SIAM J. on Optimization, Vol. 10, 2000, pp. 269-288.

[Tse95b] Tseng, P., 1995. "Simplified Analysis of an $O(nL)$-Iteration Infeasible Predictor-Corrector Path Following Method for Monotone LCP,"in Recent Trends in Optimization Theory and Applications, Agarwal, R. P., (Ed.), World Scientific, pp. 423-434.

[Tse98] Tseng, P., 1998. "Incremental Gradient(-Projection) Method with Momentum Term and Adaptive Stepsize Rule," SIAM J. on Optimization, Vol. 8, pp. 506-531.

[Tse00] Tseng, P., 2000. "A Modified Forward-Backward Splitting Method for Maximal Monotone Mappings," SIAM J. on Control and Optimization, Vol. 38, pp. 431-446.

[Tse01a] Tseng, P., 2001. "Convergence of Block Coordinate Descent Methods for Non-differentiable Minimization," J. Optim. Theory Appl., Vol. 109, pp. 475-494.

[Tse01b] Tseng, P., 2001. "An Epsilon Out-of-Kilter Method for Monotropic Programming," Math. of Operations Research, Vol. 26, pp. 221-233.

[Tse04] Tseng, P., 2004. "An Analysis of the EM Algorithm and Entropy-Like Proximal Point Methods," Math. Operations Research, Vol. 29, pp. 27-44.

[Tse08] Tseng, P., 2008. "On Accelerated Proximal Gradient Methods for Convex-Concave Optimization," Report, Math. Dept., Univ. of Washington.

[VaB95] Vandenberghe, L., and Boyd, S., 1995. “A Primal-Dual Potential Reduction Method for Problems Involving Matrix Inequalities,” Math. Programming, Vol. 69, pp. 205-236.

[VaW89] Varaiya, P., and Wets, R. J-B., 1989. “Stochastic Dynamic Optimization Approaches and Computation,” Mathematical Programming: State of the Art, M. Iri and K. Tanabe (eds.), Kluwer, Boston, pp. 309-332.

[VeH93] Ventura, J. A., and Hearn, D. W., 1993. “Restricted Simplicial Decomposition for Convex Constrained Problems,” Math. Programming, Vol. 59, pp. 71-85.

[Ven67] Venter, J. H., 1967. “An Extension of the Robbins-Monro Procedure,” Ann. Math. Statist., Vol. 38, pp. 181-190.

[WDS13] Weinmann, A., Demaret, L., and Storath, M., 2013. “Total Variation Regularization for Manifold-Valued Data,” arXiv preprint arXiv:1312.7710.

[WHM13] Wang, X., Hong, M., Ma, S., Luo, Z. Q., 2013. “Solving Multiple-Block Separable Convex Minimization Problems Using Two-Block Alternating Direction Method of Multipliers,” arXiv preprint arXiv:1308.5294.

[WQB98] Wei, Z., Qi, L., and Birge, J. R., 1998. “New Method for Nonsmooth Convex Optimization,” J. of Inequalities and Applications, Vol. 2, pp. 157-179.

[WSK14] Wytock, M., Sra, S., and Kolter, J. K., 2014. “Fast Newton Methods for the Group Fused Lasso,” Proc. of 2014 Conf. on Uncertainty in Artificial Intelligence.

[WaB13a] Wang, M., and Bertsekas, D. P., 2013. “Incremental Constraint Projection-Proximal Methods for Nonsmooth Convex Optimization,” Lab. for Information and Decision Systems Report LIDS-P-2907, MIT; SIAM Journal on Optimization, Vol. 26, 2016, pp. 681-717.

[WaB13b] Wang, M., and Bertsekas, D. P., 2013. “Convergence of Iterative Simulation-Based Methods for Singular Linear Systems,” Stochastic Systems, Vol. 3, pp. 38-95.

[WaB15] Wang, M., and Bertsekas, D. P., 2015. “Incremental Constraint Projection Methods for Variational Inequalities.” Math. Programming,” Vol. 150.2, pp. 321-363.

[WaB16] Wang, M., and Bertsekas, D. P., 2016. “Stochastic First-Order Methods with Random Constraint Projection.” SIAM J. on Optimization, Vol. 26, pp. 681-717.

[WeO13] Wei, E., and Ozdaglar, A., 2013. “On the $O(1/k)$ Convergence of Asynchronous Distributed Alternating Direction Method of Multipliers,” arXiv preprint arXiv:1307.8254.

[Web29] Weber, A., 1929. Theory of Location of Industries, (Engl. Transl. by C. J. Friedrich), Univ. of Chicago Press, Chicago, Ill.

[WiH60] Widrow, B., and Hoff, M. E., 1960. “Adaptive Switching Circuits,” Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, Part 4, pp. 96-104.

[Wil63] Wilson, R. B., 1963. “A Simplicial Algorithm for Concave Programming,” Ph.D. Thesis, Grad. Sch. Business Admin., Harvard Univ., Cambridge, MA.

[Wol98] Wolsey, L. A., 1998. Integer Programming, Wiley, N. Y.

[Wri92] Wright, S. J., 1992. “An Interior Point Algorithm for Linearly Constrained Optimization,” SIAM J. on Optimization, Vol. 2, pp. 450-473.

[Wri93a] Wright, S. J., 1993. “Identifiable Surfaces in Constrained Optimization,” SIAM J. on Control and Optimization, Vol. 31, pp. 1063-1079.

[Wri93b] Wright, S. J., 1993. “Interior Point Methods for Optimal Control of Discrete Time Systems,” J. Opt. Theory and Appl., Vol. 77, pp. 161-187.

[Wri93c] Wright, S. J., 1993. "A Path-Following Infeasible-Interior-Point Algorithm for Linear Complementarity Problems," Optimization Methods and Software, Vol. 2, pp. 79-106.

[Wri94] Wright, S. J., 1994. "An Infeasible-Interior-Point Algorithm for Linear Complementarity Problems," Math. Programming, Vol. 67, pp. 29-52.

[Wri96] Wright, S. J., 1996. "A Path-Following Interior-Point Algorithm for Linear and Quadratic Problems," Annals of Operations Res., Vol. 62, pp. 103-130.

[Wri97a] Wright, S. J., 1997. Primal-Dual Interior Point Methods, SIAM, Phila., PA.

[Wri97b] Wright, S. J., 1997. "Applying New Optimization Algorithms to Model Predictive Control," Chemical Process Control-V, CACHE, AIChE Symposium Series No. 316, Vol. 93, pp. 147-155.

[Wri98] Wright, S. J., 1998. "Superlinear Convergence of a Stabilized SQP Method to a Degenerate Solution," Computational Optimization and Applications, Vol. 11, pp. 253-275.

[WuB01] Wu, C., and Bertsekas, D. P., 2001. "Distributed Power Control Algorithms for Wireless Networks," IEEE Trans. on Vehicular Technology, Vol. 50, pp. 504-514.

[YSQ14] You, K., Song, S., and Qiu, L., 2014. "Randomized Incremental Least Squares for Distributed Estimation Over Sensor Networks," Preprints of the 19th World Congress The International Federation of Automatic Control Cape Town, South Africa.

[YVG10] Yu, J., Vishwanathan, S. V. N., Gunter, S., and Schraudolph, N. N., 2010. "A Quasi-Newton Approach to Nonsmooth Convex Optimization Problems in Machine Learning," J. of Machine Learning Research, Vol. 11, pp. 1145-1200.

[Ye92] Ye, Y., 1992. "A Potential Reduction Algorithm Allowing Column Generation," SIAM J. on Optimization, Vol. 2, pp. 7-20.

[Ye97] Ye, Y., 1997. Interior Point Algorithms: Theory and Analysis, Wiley Interscience, N. Y.

[Ypm95] Ypma, T. J., 1995. "Historical Development of the Newton-Raphson Method," SIAM Review, Vol. 37, pp. 531-551.

[You15] Yousefpour, R., 2015. "Combination of Steepest Descent and BFGS Methods for Nonconvex Nonsmooth Optimization," Numerical Algorithms, pp. 1-34.

[ZLW99] Zhao, X., Luh, P. B., and Wang, J., 1999. "Surrogate Gradient Algorithm for Lagrangian Relaxation," J. Opt. Theory and Appl., Vol. 100, pp. 699-712.

[ZTP93] Zhang, Y., Tapia, R. A., and Potra, F., 1993. "On the Superlinear Convergence of Interior-Point Algorithms for a General Class of Problems," SIAM J. on Optimization, Vol. 3 pp. 413-422.

[Zal02] Zalinescu, C., 2002. Convex Analysis in General Vector Spaces, World Scientific, Singapore.

[Zan67a] Zangwill, W. I., 1967. "Minimizing a Function Without Calculating Derivatives," The Computer Journal, Vol. X, pp. 293-296.

[Zan67b] Zangwill, W. I., 1967. "Nonlinear Programming via Penalty Functions," Management Science, Vol. 13, pp. 344-358.

[Zan69] Zangwill, W. I., 1969. Nonlinear Programming, Prentice-Hall, Englewood Cliffs, N. J.

[ZhT92] Zhang, Y., and Tapia, R. A., 1992. "Superlinear and Quadratic Convergence of Primal-Dual Interior-Point Algorithms for Linear Programming Revisited," J. Opt. Theory and Appl., Vol. 73, pp. 229-242.

[ZhT93] Zhang, Y., and Tapia, R. A., 1993. "A Superlinearly Convergent Polynomial Primal-Dual Interior-Point Algorithm for Linear Programming," SIAM J. on Optimization, Vol. 3, pp. 118-133.

[Zho93] Zhou, L., 1993. "A Simple Proof of the Shapley-Folkman Theorem," Economic Theory, Vol. 3, pp. 371-372.

[Zhu95] Zhu, C., 1995. "On the Primal-Dual Steepest Descent Algorithm for Extended Linear-Quadratic Programming," SIAM J. on Optimization, Vol. 5, pp. 114-128.

[Zou60] Zoutendijk, G., 1960. Methods of Feasible Directions, Elsevier Publ. Co., Amsterdam.

[Zou76] Zoutendijk, G., 1976. Mathematical Programming Methods, North Holland, Amsterdam.