APPENDIX A: Mathematical Background

In Sections A.1-A.3 of this appendix, we provide some basic definitions, notational conventions, and results from linear algebra and real analysis. We assume that the reader is familiar with these subjects, so no proofs are given. For additional related material, we refer to textbooks such as Hoffman and Kunze [HoK71], Lancaster and Tismenetsky [LaT85], and Strang [Str76] (linear algebra), and Ash [Ash72], Ortega and Rheinboldt [OrR70], and Rudin [Rud76] (real analysis).

In Section A.4, we provide a few convergence theorems for deterministic and random sequences, which we will use for various convergence analyses of algorithms in the text. Except for the Supermartingale Convergence Theorem for sequences of random variables (Prop. A.4.5), we provide complete proofs.

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 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 exists. For the empty set, we use the convention

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

If $\sup X$ is equal to a scalar $\overline{x}$ that belongs to the set X, we say that $\overline{x}$ is the *maximum point* of X and we write $\overline{x} = \max X$. Similarly, if $\inf X$ is equal to a scalar $\overline{x}$ that belongs to the set X, we say that $\overline{x}$ is the *minimum point* of X and we write $\overline{x} = \min X$. 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.

Vector Notation

We denote by $\Re^n$ the set of n-dimensional real vectors. For any $x \in \Re^n$, we use x_i (or sometimes x^i) to indicate its ith *coordinate*, also called its ith *component*. 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 transpose of x, which is an n-dimensional row vector. The *inner product* of two vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ 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.

Function Notation and Terminology

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 $\{f(x) \mid x \in U\}$ is called the *image* or *forward image of U under f*, and the set $\{x \in X \mid f(x) \in V\}$ is called the *inverse image of V under f*.

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.1 LINEAR ALGEBRA

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

$$\{x + y \mid x \in X,\, y \in Y\},$$

which is referred to as the *vector sum of X and Y*. 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 $\overline{x}$, we simplify this notation as follows:

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

We also denote by $X - Y$ the set

$$\{x - y \mid x \in X,\, y \in Y\}.$$

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

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

which is viewed as 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* in $\Re^n$ is a translated subspace, i.e., a set X of the form $X = \overline{x} + S = \{\overline{x} + x \mid x \in S\}$, where $\overline{x}$ 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 $X = \overline{x} + \overline{S}$ be two representations of the affine set X. Then, we must have $x = \overline{x} + \overline{s}$ for some $\overline{s} \in \overline{S}$ (since

$x \in X$), so that $X = \overline{x} + \overline{s} + S$. Since we also have $X = \overline{x} + \overline{S}$, it follows that $S = \overline{S} - \overline{s} = \overline{S}$.] A nonempty set X is a subspace if and only if it contains the origin, and every line that passes through any pair of its points that are distinct, i.e., it contains 0 and all points $\alpha x + (1 - \alpha)y$, where $\alpha \in \Re$ and $x, y \in X$ with $x \neq y$. Similarly X is affine if and only if it contains every line that passes through any pair of its points that are distinct. The *span* of a finite collection $\{x_1, \ldots, x_m\}$ of elements of $\Re^n$, denoted by $\text{span}(x_1, \ldots, x_m)$, 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. Every subspace of nonzero dimension has a basis that is orthogonal (i.e., any pair of distinct vectors from the basis is orthogonal). The *dimension of an affine set* $\overline{x} + S$ is the dimension of the corresponding subspace S. An $(n-1)$-dimensional affine subset of $\Re^n$ is called a *hyperplane*, assuming $n \geq 2$. It is a set specified by a single linear equation, i.e., a set of the form $\{x \mid a'x = b\}$, where $a \neq 0$ and $b \in \Re$.

Given any subset X of $\Re^n$, 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 component. The *transpose* of A, denoted by A', is defined by $[A']_{ij} = a_{ji}$. For any two matrices A and B of compatible dimensions, the transpose of the product matrix AB satisfies $(AB)' = B'A'$. The inverse of a square and invertible A is denoted A^{-1}.

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):

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

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

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

Let A be an $m \times n$ matrix. The *range space* of A, denoted by $R(A)$, is the set of all vectors $y \in \Re^m$ such that $y = Ax$ for some $x \in \Re^n$. The *nullspace* of A, denoted by $N(A)$, is the set of all vectors $x \in \Re^n$ such that $Ax = 0$. It is seen that the range space and the nullspace of A are subspaces. The *rank* of A is the dimension of the range space of 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., $R(A) = N(A')^\perp$.

Square Matrices

By a square matrix we mean any $n \times n$ matrix, where $n \geq 1$. The determinant of a square matrix A is denoted by $\det(A)$.

Definition A.1.1: A square matrix A is called *singular* if its determinant is zero. Otherwise it is called *nonsingular* or *invertible*.

Definition A.1.2: 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 nonzero 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.

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.1.1:

(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 is a unique $x \in \Re^n$ such that $Ax = y$.

(v) There is 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.

(viii) All eigenvalues of A are nonzero.

(b) Assuming that A is nonsingular, the matrix B of statement (v) (called the *inverse* of A and denoted by A^{-1}) is unique.

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

Proposition A.1.2: Let A be an $n \times n$ matrix.

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

(b) For any scalar 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.

(c) 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.

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

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

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.

Symmetric and Positive Definite Matrices

A square matrix A is said to be *symmetric* if $A = A'$. Symmetric matrices have several special properties, particularly regarding their eigenvalues and eigenvectors.

Proposition A.1.3: 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) There holds

$$\underline{\lambda}\, x'x \leq x'Ax \leq \bar{\lambda}\, x'x, \qquad \forall\ x \in \Re^n,$$

where $\underline{\lambda}$ and $\bar{\lambda}$ are the smallest and largest eigenvalues of A, respectively.

Definition A.1.3: 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 *positive semidefinite* if $x'Ax \geq 0$ for all $x \in \Re^n$.

Throughout this book, the notion of positive definiteness applies exclusively to symmetric matrices. Thus *whenever we say that a matrix is positive (semi)definite, we implicitly assume that the matrix is symmetric*, although we usually add the term "symmetric" for clarity.

Proposition A.1.4:

(a) A square matrix is symmetric and positive definite if and only if it is invertible and its inverse is symmetric and positive definite.

(b) The sum of two symmetric positive semidefinite matrices is positive semidefinite. If one of the two matrices is positive definite, the sum is positive definite.

(c) If A is a symmetric positive semidefinite $n \times n$ matrix and T is an $m \times n$ matrix, then the matrix TAT' is positive semidefinite. If A is positive definite and T is invertible, then TAT' is positive definite.

(d) If A is a symmetric positive definite $n \times n$ matrix, there exists a unique symmetric positive definite matrix that yields A when multiplied with itself. This matrix is called the *square root of A*. It is denoted by $A^{1/2}$, and its inverse is denoted by $A^{-1/2}$.

A.2 TOPOLOGICAL PROPERTIES

Definition A.2.1: A *norm* $\|\cdot\|$ on $\Re^n$ is a function 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) $\|\alpha x\| = |\alpha| \cdot \|x\|$ for every scalar α and every $x \in \Re^n$.

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

(d) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \Re^n$ (this is referred to as the *triangle inequality*).

The *Euclidean norm* of a vector $x = (x_1, \ldots, x_n)$ is defined by

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

Except for specialized contexts, we use this norm. In particular, *in the absence of a clear indication to the contrary,* $\|\cdot\|$ *will denote the Euclidean norm*. The *Schwarz inequality* states that for any two vectors x and y, we have

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

with equality holding if and only if $x = \alpha y$ for some scalar α. The *Pythagorean Theorem* states that for any two vectors x and y that are orthogonal, we have

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

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

$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|,$$

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 sometimes 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 scalar sequence $\{x_k \mid k = 1, 2, \ldots\}$ (or $\{x_k\}$ for short) 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 $x_k \geq b$ for all $k \geq K$, 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 $x_k \leq b$ for all $k \geq K$, 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 said to be *bounded above* (respectively, *below*) if there exists some scalar b such that $x_k \leq b$ (respectively, $x_k \geq b$) 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 $x_{k+1} \leq x_k$ (respectively, $x_{k+1} \geq x_k$) 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.2.1: 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.2.1). 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.2.2: 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. A sequence $\{x_k\} \subset \Re^n$ is said to be a *Cauchy sequence* if $\|x_m - x_n\| \to 0$ as $m, n \to \infty$, i.e., given any $\epsilon > 0$, there exists N such that $\|x_m - x_n\| \leq \epsilon$ for all $m, n \geq N$. A sequence is Cauchy if and only if it converges to some vector. The sequence $\{x_k\}$ is called bounded if each of its corresponding component sequences is bounded. 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 $\{x_k\}$. Thus 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).

A vector $x \in \Re^n$ is said to be a *limit point* of a sequence $\{x_k\}$ if there exists a subsequence of $\{x_k\}$ that converges to x. The following is a classical result that will be used often.

Proposition A.2.3: (Bolzano-Weierstrass Theorem) A bounded sequence in $\Re^n$ has at least one limit point.

$o(\cdot)$ Notation

For a function $h : \Re^n \mapsto \Re^m$ we write $h(x) = o\big(\|x\|^p\big)$, where p is a positive integer, 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 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.2.2: 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.

Given $x^* \in \Re^n$ and $\epsilon > 0$, the sets $\big\{x \mid \|x - x^*\| < \epsilon\big\}$ and $\big\{x \mid \|x - x^*\| \leq \epsilon\big\}$ are called an *open sphere* and a *closed sphere* centered at x^*, respectively. Sometimes the terms *open ball* and *closed ball* are used. A consequence of the definitions, is that a subset X of $\Re^n$ is open if and only if for every $x \in X$ there is an open sphere that is centered at x and is contained in X. A *neighborhood* of a vector x is an open set containing x.

Definition A.2.3: We say that x is an *interior point* of a subset X of $\Re^n$ if there exists a neighborhood of x that is contained in X. The set of all interior points of X is called the *interior* of X, and is denoted

by $\text{int}(X)$. A vector $x \in \text{cl}(X)$ 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.

Proposition A.2.4:

(a) The union of a finite collection of closed sets is closed.

(b) The intersection of any collection of closed sets is closed.

(c) The union of any collection of open sets is open.

(d) The intersection of a finite collection of 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 set $X \subset \Re^n$ is compact if and only if every sequence of elements of X has a subsequence that converges to an element of X.

(h) If $\{X_k\}$ is a sequence of nonempty and compact subsets of $\Re^n$ such that $X_{k+1} \subset X_k$ for all k, then the intersection $\cap_{k=0}^{\infty} X_k$ is nonempty and compact.

The topological properties of sets in $\Re^n$, such as being open, closed, or compact, do not depend on the norm being used. This is a consequence of the following proposition.

Proposition A.2.5: (Norm Equivalence Property)

(a) For any two norms $\|\cdot\|$ and $\|\cdot\|'$ on $\Re^n$, there exists a scalar c such that

$$\|x\| \leq c\|x\|', \qquad \forall\ x \in \Re^n.$$

(b) If a subset of $\Re^n$ is open (respectively, closed, bounded, or compact) with respect to some norm, it is open (respectively, closed, bounded, or compact) with respect to all other norms.

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 $\{f(x_k)\}$ 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, $\lim_{z \downarrow x} f(z) = y$].

Definition A.2.4: Let X be a nonempty 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 function $f : X \mapsto \Re^m$ is called *Lipschitz continuous* over X if there exists a scalar L such that

$$\big\|f(x) - f(y)\big\| \leq L\|x - y\|, \qquad \forall\, x, y \in X.$$

(d) 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.

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* (without qualification). We use similar terminology for right-continuous, left-continuous, Lipschitz continuous, upper semicontinuous, and lower semicontinuous functions.

Proposition A.2.6:

(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, $\{x \in \Re^n \mid f(x) \in Y\}$, 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, $\{f(x) \mid x \in X\}$, is compact.

If $f : \Re^n \mapsto \Re$ is a continuous function and $X \subset \Re^n$ is compact, by Prop. A.2.6(c), the sets

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

are nonempty and compact for all $\gamma \in \Re$ with $\gamma > f^*$, where

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

Since the set of minima of f is the intersection of the nonempty and compact sets V_{γ_k} for any sequence $\{\gamma_k\}$ with $\gamma_k \downarrow f^*$ and $\gamma_k > f^*$ for all k, it follows from Prop. A.2.4(h) that the set of minima is nonempty. This proves the following classical theorem of Weierstrass.

Proposition A.2.7: (Weierstrass' Theorem for Continuous Functions) A continuous function $f : \Re^n \mapsto \Re$ attains a minimum over any compact subset of $\Re^n$.

A.3 DERIVATIVES

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

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

For any $d \in \Re^n$, we define the one-sided *directional derivative* of f at a vector x in the direction d by

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

provided that the limit exists.

If the directional derivative of f at a vector x exists in all directions and $f'(x;d)$ is a linear function of d, we say that f is *differentiable* at x. It can be seen that f is differentiable at x if and only if the gradient $\nabla f(x)$ exists and satisfies $\nabla f(x)'d = f'(x;d)$ for all $d \in \Re^n$, or equivalently

$$f(x+\alpha d) = f(x) + \alpha \nabla f(x)'d + o(|\alpha|), \qquad \forall\ \alpha \in \Re.$$

The function f is called *differentiable over a subset S of* $\Re^n$ if it is differentiable at every $x \in S$. 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 S and $\nabla f(\cdot)$ is continuous at all $x \in S$, f is said to be *continuously differentiable over S*. It can then be shown that for any $x \in S$ and norm $\|\cdot\|$,

$$f(x+d) = f(x) + \nabla f(x)'d + o(\|d\|), \qquad \forall\ d \in \Re^n.$$

The function f is called *continuously differentiable* (without qualification) if it is differentiable and $\nabla f(\cdot)$ is continuous at all $x \in \Re^n$. In our development, whenever we assume that f is differentiable, we also assume that it is continuously differentiable. Part of the reason is that a convex differentiable function is automatically continuously differentiable over $\Re^n$ (see Section 3.1).

If each one of the partial derivatives of a function $f : \Re^n \mapsto \Re$ is a continuously differentiable function of x over an open set S, we say that f is *twice continuously differentiable* over S. We then denote by

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$$

the ith partial derivative of $\partial f/\partial x_j$ at a vector $x \in \Re^n$. The *Hessian* of f at x, denoted by $\nabla^2 f(x)$, is the matrix whose components are the above second derivatives. The matrix $\nabla^2 f(x)$ is symmetric. In our development, whenever we assume that f is twice differentiable, we also assume that it is twice continuously differentiable.

We now state some theorems relating to differentiable functions.

Proposition A.3.1: (Mean Value Theorem) Let $f : \Re^n \mapsto \Re$ be continuously differentiable over an open sphere S, and let x be a vector in S. Then for all y such that $x+y \in S$, there exists an $\alpha \in [0,1]$ such that

$$f(x+y) = f(x) + \nabla f(x+\alpha y)'y.$$

Proposition A.3.2: (Second Order Expansions) Let $f : \Re^n \mapsto \Re$ be twice continuously differentiable over an open sphere S, and let x be a vector in S. Then for all y such that $x + y \in S$:

(a) There exists an $\alpha \in [0, 1]$ such that

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

(b) We have

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

A.4 CONVERGENCE THEOREMS

We will now discuss a few convergence theorems relating to iterative algorithms. Given a mapping $T : \Re^n \mapsto \Re^n$, the iteration

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

aims at finding a fixed point of T, i.e., a vector x^* such that $x^* = T(x^*)$. A common criterion for existence of a fixed point is that T is a *contraction mapping* (or contraction for short) with respect to some norm, i.e., for some $\beta < 1$, and some norm $\|\cdot\|$ (not necessarily the Euclidean norm), we have

$$\big\|T(x) - T(y)\big\| \leq \beta\|x - y\|, \qquad \forall\ x, y \in \Re^n.$$

When T is a contraction, it has a unique fixed point and the iteration $x_{k+1} = T(x_k)$ converges to the fixed point. This is shown in the following classical theorem.

Proposition A.4.1: (Contraction Mapping Theorem) Let $T : \Re^n \mapsto \Re^n$ be a contraction mapping. Then T has a unique fixed point x^*, and the sequence generated by the iteration $x_{k+1} = T(x_k)$ converges to x^*, starting from any $x_0 \in \Re^n$.

Proof: We first note that T can have at most one fixed point (if $\tilde{x}$ and $\hat{x}$ are two fixed points, we have

$$\|\tilde{x} - \hat{x}\| = \big\|T(\tilde{x}) - T(\hat{x})\big\| \leq \beta\|\tilde{x} - \hat{x}\|,$$

which implies that $\tilde{x} = \hat{x}$). Using the contraction property, we have for all $k, m > 0$

$$\|x_{k+m} - x_k\| \leq \beta^k \|x_m - x_0\| \leq \beta^k \sum_{\ell=1}^{m} \|x_\ell - x_{\ell-1}\| \leq \beta^k \sum_{\ell=0}^{m-1} \beta^\ell \|x_1 - x_0\|,$$

and finally,

$$\|x_{k+m} - x_k\| \leq \frac{\beta^k(1-\beta^m)}{1-\beta}\|x_1 - x_0\|.$$

Thus $\{x_k\}$ is a Cauchy sequence, and hence converges to some x^*. Taking the limit in the equation $x_{k+1} = T(x_k)$ and using the continuity of T (implied by the contraction property), we see that x^* must be a fixed point of T. **Q.E.D.**

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.

The next theorem applies to a mapping that is nonexpansive with respect to the Euclidean norm. It shows that a fixed point of such a mapping can be found by an interpolated iteration, provided at least one fixed point exists. The idea underlying the theorem is quite intuitive: if x^* is a fixed point of T, the distance $\big\|T(x_k) - x^*\big\|$ cannot be larger than the distance $\|x_k - x^*\|$ (by nonexpansiveness of T):

$$\big\|T(x_k) - x^*\big\| = \big\|T(x_k) - T(x^*)\big\| \leq \|x_k - x^*\|.$$

Hence, if $x_k \neq T(x_k)$, any point obtained by strict interpolation between x_k and $T(x_k)$ must be strictly closer to x^* than x_k (by Euclidean geometry). Note, however, that for this argument to work, we need to know that T has at least one fixed point. If T is a contraction, this is automatically guaranteed, but if T is just nonexpansive, there may not exist a fixed point [as an example, just let $T(x) = 1 + x$].

Proposition A.4.2: (Krasnosel'skii-Mann Theorem for Nonexpansive Iterations [Kra55], [Man53]) Consider a mapping $T : \Re^n \mapsto \Re^n$ that is nonexpansive with respect to the Euclidean norm $\|\cdot\|$, i.e.,

$$\big\|T(x) - T(y)\big\| \leq \|x - y\|, \qquad \forall\ x, y \in \Re^n,$$

and has at least one fixed point. Then the iteration

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

where $\alpha_k \in [0,1]$ for all k and

$$\sum_{k=0}^{\infty} \alpha_k(1-\alpha_k) = \infty,$$

converges to a fixed point of T, starting from any $x_0 \in \Re^n$.

Proof: We will use the identity

$$\big\|\alpha x + (1-\alpha)y\big\|^2 = \alpha\|x\|^2 + (1-\alpha)\|y\|^2 - \alpha(1-\alpha)\|x-y\|^2, \tag{A.2}$$

which holds for all $x, y \in \Re^n$, and $\alpha \in [0,1]$, as can be verified by a straightforward calculation. For any fixed point x^* of T, we have

$$\begin{aligned}\|x_{k+1}-x^*\|^2 &= \big\|(1-\alpha_k)(x_k-x^*) + \alpha_k\big(T(x_k)-T(x^*)\big)\big\|^2 \\ &= (1-\alpha_k)\|x_k-x^*\|^2 + \alpha_k\big\|T(x_k)-T(x^*)\big\|^2 \\ &\quad - \alpha_k(1-\alpha_k)\big\|T(x_k)-x_k\big\|^2 \\ &\le \|x_k-x^*\|^2 - \alpha_k(1-\alpha_k)\big\|T(x_k)-x_k\big\|^2,\end{aligned} \tag{A.3}$$

where for the first equality we use iteration (A.1) and the fact $x^* = T(x^*)$, for the second equality we apply the identity (A.2), and for the inequality we use the nonexpansiveness of T. By adding Eq. (A.3) for all k, we obtain

$$\sum_{k=0}^{\infty} \alpha_k(1-\alpha_k)\big\|T(x_k)-x_k\big\|^2 \le \|x_0-x^*\|^2.$$

In view of the hypothesis $\sum_{k=0}^{\infty}\alpha_k(1-\alpha_k) = \infty$, it follows that

$$\lim_{k\to\infty,\, k\in\mathcal{K}} \big\|T(x_k)-x_k\big\| = 0, \tag{A.4}$$

for some subsequence $\{x_k\}_{\mathcal{K}}$. Since from Eq. (A.3), $\{x_k\}_{\mathcal{K}}$ is bounded, it has at least one limit point, call it $\overline{x}$, so $\{x_k\}_{\overline{\mathcal{K}}} \to \overline{x}$ for an infinite index set $\overline{\mathcal{K}} \subset \mathcal{K}$. Since T is nonexpansive it is continuous, so $\big\{T(x_k)\big\}_{\overline{\mathcal{K}}} \to T(\overline{x})$, and in view of Eq. (A.4), it follows that $\overline{x}$ is a fixed point of T. Letting $x^* = \overline{x}$ in Eq. (A.3), we see that $\big\{\|x_k-\overline{x}\|\big\}$ is nonincreasing and hence converges, necessarily to 0, so the entire sequence $\{x_k\}$ converges to the fixed point $\overline{x}$. **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.4.3: Let $\{\alpha_k\}$ be a nonnegative sequence satisfying

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

where $\beta_k \geq 0$, $\gamma_k > 0$ 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 $\overline{k}$ be such that $\alpha_k \geq \epsilon$ and $\beta_k/\gamma_k \leq \epsilon/2$ for all $k \geq \overline{k}$, we would have for all $k \geq \overline{k}$

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

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

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

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

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

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

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

As an 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.4.3, 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 now give two theorems relating to *supermartingale convergence* analysis (the term refers to a collection of convergence theorems for sequences of nonnegative scalars or random variables, which satisfy certain inequalities implying that the sequences are "almost" nonincreasing). The first theorem relates to deterministic sequences, while the second theorem relates to sequences of random variables. We prove the first theorem, and we refer to the literature on stochastic processes and iterative methods for the proof of the second.

Proposition A.4.4: Let $\{Y_k\}$, $\{Z_k\}$, $\{W_k\}$, and $\{V_k\}$ be four scalar sequences such that

$$Y_{k+1} \leq (1+V_k)Y_k - Z_k + W_k, \qquad k = 0, 1, \ldots, \tag{A.5}$$

$\{Z_k\}$, $\{W_k\}$, 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

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

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

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

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

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

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.5) for the index k set to $0, \ldots, k$, and adding, we have

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

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

$$\prod_{\ell=0}^{\infty}(1 + V_\ell) < \infty. \tag{A.6}$$

Define

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

Multiplying Eq. (A.5) with $\prod_{\ell=0}^{k}(1 + V_\ell)^{-1}$, we obtain

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

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

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

and $\prod_{\ell=0}^{k-1}(1 + V_\ell)$ converges to a finite value by the nonnegativity of $\{V_k\}$ and Eq. (A.6), 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.**

The next theorem has a long history. The particular version we give here is due to Robbins and Sigmund [RoS71]. Their proof assumes the special case of the theorem where $V_k \equiv 0$ (see Neveu [Nev75], p. 33, for a proof of this special case), and then uses the line of proof of the preceding proposition. Note, however, that contrary to the preceding proposition, the following theorem requires nonnegativity of the sequence $\{Y_k\}$.

Proposition A.4.5: (Supermartingale Convergence Theorem) Let $\{Y_k\}$, $\{Z_k\}$, $\{W_k\}$, and $\{V_k\}$ be four nonnegative sequences of random variables, and let $\mathcal{F}_k$, $k = 0, 1, \ldots$, be sets of random variables such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k. Assume that:

(1) For each k, Y_k, Z_k, W_k, and V_k are functions of the random variables in $\mathcal{F}_k$.

(2) We have

$$E\big\{Y_{k+1} \mid \mathcal{F}_k\big\} \leq (1 + V_k)Y_k - Z_k + W_k, \qquad k = 0, 1, \ldots.$$

(3) There holds, with probability 1,

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

Then $\{Y_k\}$ converges to a nonnegative random variable Y, and we have $\sum_{k=0}^{\infty} Z_k < \infty$, with probability 1.

Fejér Monotonicity

The 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, as well as the Krasnosel'skii-Mann Theorem [Prop. A.4.2; see Eq. (A.3)].

The following theorem is appropriate for our purposes. There are several related but somewhat different theorems in the literature, and for complementary discussions, we refer to [BaB96], [Com01], [BaC11], [CoV13].

Proposition A.4.6: (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 $\overline{x}$ that belongs to X^*, then the entire sequence $\{x_k\}$ converges to $\overline{x}$.

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

$$\phi(x; x^*) = 0 \qquad \text{if and only if} \qquad x \in X^*. \tag{A.7}$$

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.4.4 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.4.4, define for all k,

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

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

$$\overline{Y}_{k+1} \leq \overline{Y}_k + \overline{\delta}_k, \qquad \forall\ k = 0, 1, \ldots, \tag{A.8}$$

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

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

so that by adding Eq. (A.8), we obtain for all $k > \overline{k}$,

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

Since ϵ is arbitrarily small, it follows that $\overline{Y}_k \to 0$. We now note that as in Eq. (A.6),

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

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

(c) From Prop. A.4.4, 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 $\overline{x}$, and by the lower semicontinuity of $\phi(\cdot; x^*)$, we must have

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

which in view of the nonnegativity of ϕ, implies that $\phi(\overline{x}; x^*) = 0$. Using the hypothesis (A.7), it follows that $\overline{x} \in X^*$, so by part (b), the entire sequence $\{x_k\}$ converges to $\overline{x}$. **Q.E.D.**