APPENDIX A:
Mathematical Background

In this appendix, we list 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). We also provide in Section A.4, with some proofs, a few sequence convergence theorems that we will use.

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 $\mathbb{R}$. The set $\mathbb{R}$ 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 $\infty$. 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

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

If $\sup X$ is equal to a scalar $\pi$ that belongs to the set $X$, we say that $\pi$ is the *maximum point* of $X$ and we write $\pi = \max X$. Similarly, if $\inf X$ is equal to a scalar $\pi$ that belongs to the set $X$, we say that $\pi$ is the *minimum point* of $X$ and we write $\pi = \min X$. Thus, when we write $\max X$ (or $\min X$, 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 $\mathbb{R}^n$ the set of $n$-dimensional real vectors. For any $x \in \mathbb{R}^n$, we use $x_i$ to indicate its $i$th coordinate, also called its $i$th component. Vectors in $\mathbb{R}^n$ will be viewed as column vectors, unless the contrary is explicitly stated. For any $x \in \mathbb{R}^n$, $x'$ denotes the transpose of $x$, which is an $n$-dimensional row vector. The *inner product* of two vectors $x, y \in \mathbb{R}^n$ is defined by $x'y = \sum_{i=1}^n x_i y_i$. Two vectors $x, y \in \mathbb{R}^n$ satisfying $x'y = 0$ are called orthogonal.

If $x$ is a vector in $\mathbb{R}^n$, 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 \( f : \mathbb{R}^n \to \mathbb{R} \) is said to be affine if it has the form \( f(x) = a'x + b \) for some \( a \in \mathbb{R}^n \) and \( b \in \mathbb{R} \). Similarly, a function \( f : \mathbb{R}^n \to \mathbb{R}^m \) is said to be affine if it has the form \( f(x) = Ax + b \) for some \( m \times n \) matrix \( A \) and some \( b \in \mathbb{R}^m \). 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 set and \( \lambda \) is a scalar, we denote by \( \lambda X \) the set \( \{ \lambda x \mid x \in X \} \). If \( X_1 \) and \( X_2 \) are two subsets of \( \mathbb{R}^n \), 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 \( x \), we simplify this notation as follows:
\[
\mathcal{T} + X = \{ \mathcal{T} + x \mid x \in X \}.
\]
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 \( X_i \subset \mathbb{R}^{n_i}, i = 1, \ldots, m \), 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 \( \mathbb{R}^{n_1+\cdots+n_m} \).

### Subspaces and Linear Independence

A nonempty subset \( S \) of \( \mathbb{R}^n \) is called a subspace if \( ax + by \in S \) for every \( x, y \in S \) and every \( a, b \in \mathbb{R} \). An affine set in \( \mathbb{R}^n \) is a translated subspace, i.e., a set \( X \) of the form \( X = \mathcal{T} + S = \{ \mathcal{T} + x \mid x \in S \} \), where \( \mathcal{T} \) is a vector in \( \mathbb{R}^n \) and \( S \) is a subspace of \( \mathbb{R}^n \), 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 \( X = \mathcal{T} + \mathcal{S} \) be two representations of the affine set \( X \). Then, we must have \( x = \mathcal{T} + \mathcal{S} \) for some \( \mathcal{T} \in \mathcal{S} \) (since \( x \in X \)), so that \( X = \mathcal{T} + \mathcal{S} + S \). Since we also have \( X = \mathcal{T} + \mathcal{S} \), it follows that \( S = \mathcal{S} - \mathcal{T} = \mathcal{S} \).] 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 \( \alpha \in \mathbb{R} \) and \( x, y \in X \) with \( x \neq y \). Similarly \( X \) is affine if and only
Mathematical Background

Appendix A

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 \( \mathbb{R}^n \), 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 \( \alpha_k \) is a scalar.

The vectors \( x_1, \ldots, x_m \in \mathbb{R}^n \) 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 \( \mathbb{R}^n \) 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 \( \mathbf{x} + S \) is the dimension of the corresponding subspace \( S \). An \((n - 1)\)-dimensional affine set is called a hyperplane. 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 \( b \in \mathbb{R} \).

Matrices

For any matrix \( A \), we use \( A_{ij} \), \( [A]_{ij} \), or \( a_{ij} \) to denote its \( ij \)th 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 \( \mathbb{R}^n \) 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 \( \mathbb{R}^m \), 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 \( y \in \mathbb{R}^m \) such that \( y = Ax \) for some \( x \in \mathbb{R}^n \). The
nullspace of $A$, denoted by $N(A)$, is the set of all vectors $x \in \mathbb{R}^n$ such that $Ax = 0$. It is seen that the range space and the null space 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 $\phi$ 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 $\phi$ are called the eigenvalues of $A$. A vector $x$ (with possibly complex coordinates) such that $Ax = \lambda x$, where $\lambda$ is an eigenvalue of $A$, is called an eigenvector of $A$ associated with $\lambda$.

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 $x \in \mathbb{R}^n$, we have $Ax \neq 0$. 
(iv) For every \( y \in \mathbb{R}^n \), there is a unique \( x \in \mathbb{R}^n \) 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 \( (AB)^{-1} = B^{-1}A^{-1} \).

\textbf{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**

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
\[ \lambda x'x \leq x'Ax \leq \bar{\lambda} x'x, \quad \forall x \in \mathbb{R}^n, \]
where \( \lambda \) 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 \( x \in \mathbb{R}^n, x \neq 0 \). It is called **positive semidefinite** if \( x'Ax \geq 0 \) for all \( x \in \mathbb{R}^n \).

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 $\mathbb{R}^n$ is a function that assigns a scalar $\|x\|$ to every $x \in \mathbb{R}^n$ and that has the following properties:

(a) $\|x\| \geq 0$ for all $x \in \mathbb{R}^n$.

(b) $\|\alpha x\| = |\alpha| \cdot \|x\|$ for every scalar $\alpha$ and every $x \in \mathbb{R}^n$.

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

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

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

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

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. 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 $\alpha$. 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 $\ell_\infty$-norm), defined by

$$\|x\|_\infty = \max_{i=1, \ldots, n} |x_i|,$$
The $\ell_1$-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 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 $\epsilon$). 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 integer $K$ (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, **bounded 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
\[
y_m = \sup \{ x_k \mid k \geq m \}, \quad z_m = \inf \{ x_k \mid k \geq m \}.
\]

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

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

(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 $\mathbb{R}^n$ is said to converge to some $x \in \mathbb{R}^n$ if the $i$th component of $x_k$ converges to the $i$th 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 $\{x_k\} \subset \mathbb{R}^n$ 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\}_k will also be used).

A vector \(x \in \mathbb{R}^n\) 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 \(\mathbb{R}^n\) has at least one limit point.

\[ o(\cdot) \text{ Notation} \]

For a function \(h : \mathbb{R}^n \to \mathbb{R}^m\) we write \(h(x) = o(\|x\|^{p})\), 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 \(\mathbb{R}^n\) if there exists a sequence \(\{x_k\} \subset X\) that converges to \(x\). The closure of \(X\), denoted \(\text{cl}(X)\), is the set of all closure points of \(X\).

**Definition A.2.2:** A subset \(X\) of \(\mathbb{R}^n\) 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 \(x^* \in \mathbb{R}^n\) and \(\epsilon > 0\), the sets \(\{x \mid \|x - x^*\| < \epsilon\}\) and \(\{x \mid \|x - x^*\| \leq \epsilon\}\) are called an open sphere and a closed sphere centered at \(x^*\). Sometimes the terms open ball and closed ball are used, respectively. A consequence of the definitions, is that a subset \(X\) of \(\mathbb{R}^n\) 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 $\mathbb{R}^n$ 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 $\mathbb{R}^n$ is closed.
(g) A set $X$ 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 sets such that $X_{k+1} \subset X_k$ for all $k$, then the intersection $\bigcap_{k=0}^{\infty} X_k$ is nonempty and compact.

The topological properties of sets in $\mathbb{R}^n$, 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 $\mathbb{R}^n$, there exists a scalar $c$ such that
$$
\| x \| \leq c \| x \|', \quad \forall \ x \in \mathbb{R}^n.
$$
(b) If a subset of $\mathbb{R}^n$ 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 \( f : X \mapsto \mathbb{R}^m \) be a function, where \( X \) is a subset of \( \mathbb{R}^n \), and let \( x \) be a vector in \( X \). If there exists a vector \( y \in \mathbb{R}^m \) 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 \( y \in \mathbb{R}^m \) 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 subset of \( \mathbb{R}^n \).

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

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

(c) A real-valued function \( f : X \mapsto \mathbb{R} \) 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 \( f : X \mapsto \mathbb{R}^m \) is continuous at every vector in a subset of its domain \( X \), we say that \( f \) is **continuous over that subset**. If \( f : X \mapsto \mathbb{R}^m \) 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, upper semicontinuous, and lower semicontinuous functions.

**Proposition A.2.6:**

(a) Any vector norm on \( \mathbb{R}^n \) is a continuous function.

(b) Let \( f : \mathbb{R}^m \mapsto \mathbb{R}^p \) and \( g : \mathbb{R}^n \mapsto \mathbb{R}^m \) be continuous functions. The composition \( f \circ g : \mathbb{R}^n \mapsto \mathbb{R}^p \), defined by \( (f \circ g)(x) = f(g(x)) \), is a continuous function.

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

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

If \( f : \mathbb{R}^n \mapsto \mathbb{R} \) is a continuous function and \( X \subset \mathbb{R}^n \) is compact, by
Prop. A.2.6(c), the sets
\[ V_{\gamma} = \{ x \in X \mid f(x) \leq \gamma \} \]
are nonempty and compact for all \( \gamma \in \mathbb{R} \) 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 \( V_{\gamma_k} \) 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 \( f : \mathbb{R}^n \to \mathbb{R} \) attains a minimum over any compact subset of \( \mathbb{R}^n \).

### A.3 DERIVATIVES

Let \( f : \mathbb{R}^n \to \mathbb{R} \) be some function, fix some \( x \in \mathbb{R}^n \), and consider the expression
\[
\lim_{\alpha \to 0} \frac{f(x + \alpha e_i) - f(x)}{\alpha},
\]
where \( e_i \) is the \( i \)th unit vector (all components are 0 except for the \( i \)th component which is 1). If the above limit exists, it is called the \( i \)th partial derivative of \( f \) at the vector \( x \) and it is denoted by \( \partial f / \partial x_i \) \( (x) \) or \( \partial f / \partial x_i \) \( (x) \). 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 \( d \in \mathbb{R}^n \), 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 \( d \in \mathbb{R}^n \), or equivalently

\[
f(x + \alpha d) = f(x) + \alpha \nabla f(x)'d + o(|\alpha|), \quad \forall \alpha \in \mathbb{R}.
\]

The function \( f \) is called differentiable over a subset \( S \) of \( \mathbb{R}^n \) if it is differentiable at every \( x \in S \). The function \( f \) is called differentiable (without qualification) if it is differentiable at all \( x \in \mathbb{R}^n \).

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 \( \| \cdot \| \),

\[
f(x + d) = f(x) + \nabla f(x)'d + o(\|d\|), \quad \forall d \in \mathbb{R}^n.
\]

If each one of the partial derivatives of a function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) 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 \( i \)th partial derivative of \( \partial f/\partial x_j \) at a vector \( x \in \mathbb{R}^n \). 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.

We now state some theorems relating to differentiable functions.

**Proposition A.3.1: (Mean Value Theorem)** Let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) 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 $f : \mathbb{R}^n \to \mathbb{R}$ 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

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

(b) We have

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

Strong Convexity

If $f : \mathbb{R}^n \to \mathbb{R}$ is a function that is continuous over a closed convex set $C \subset \text{dom}(f)$, and $\sigma > 0$, we say that $f$ is strongly convex over $C$ with coefficient $\sigma$ if for all $x, y \in C$ and all $\alpha \in [0, 1]$, we have

$$f(\alpha x + (1 - \alpha) y) + \frac{\sigma}{2} \alpha (1 - \alpha) \|x - y\|^2 \leq \alpha f(x) + (1 - \alpha) f(y).$$

If $f$ is strongly convex over $C$ with coefficient $\sigma$, then $f$ is strictly convex over $C$. Furthermore, there exists a unique $x^* \in C$ that minimizes $f$ over $C$, and we have

$$f(x) \geq f(x^*) + \frac{\sigma}{2} \|x - x^*\|^2, \quad \forall \ x \in C.$$ 

If $\text{int}(C)$, the interior of $C$, is nonempty, and $f$ is continuously differentiable over $\text{int}(C)$, the following are equivalent:

(i) $f$ is strongly convex with coefficient $\sigma$ over $C$.

(ii) We have

$$(\nabla f(x) - \nabla f(y))(x - y) \geq \sigma \|x - y\|^2, \quad \forall \ x, y \in \text{int}(C).$$

Furthermore, if $f$ is twice continuously differentiable over $\text{int}(C)$, the above two properties are equivalent to:

(iii) The matrix $\nabla^2 f(x) - \sigma I$ is positive semidefinite for every $x \in \text{int}(C)$, where $I$ is the identity matrix.

A proof may be found in many sources, including the on-line exercises of Chapter 1 of [Ber09].
A.4 CONVERGENCE THEOREMS

In our analysis, we will use a few convergence theorems relating to sequences, in the context of iterative algorithms. Given a mapping \( G : \mathbb{R}^n \to \mathbb{R}^n \), the iteration
\[
x_{k+1} = G(x_k),
\]
aims at finding a fixed point of \( G \), i.e., a vector \( x^* \) such that \( x^* = G(x^*) \).

A common criterion for existence of a fixed point is that \( G \) is a contraction mapping (or contraction for short) with respect to some norm, i.e., for some \( \rho < 1 \), and some norm \( \| \cdot \| \) (not necessarily the Euclidean norm), we have
\[
\| G(x) - G(y) \| \leq \rho \| x - y \|, \quad \forall x, y \in \mathbb{R}^n.
\]

Note that a contraction mapping is necessarily continuous. When \( G \) is a contraction, it has a unique fixed point and the iteration \( x_{k+1} = G(x_k) \) converges to the fixed point. This is shown in the following classical theorem.

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

**Proof:** We first note that \( G \) can have at most one fixed point (if \( \tilde{x} \) and \( \hat{x} \) are two fixed points, we have
\[
\| \tilde{x} - \hat{x} \| = \| G(\tilde{x}) - G(\hat{x}) \| \leq \rho \| \tilde{x} - \hat{x} \|,
\]
which implies that \( \tilde{x} = \hat{x} \)). Using the contraction property, we have for all \( k, m > 0 \)
\[
\| x_{k+m} - x_k \| \leq \rho^k \| x_m - x_0 \| \leq \rho^k \sum_{\ell=1}^{m} \| x_{\ell} - x_{\ell-1} \| \leq \rho^k \sum_{\ell=0}^{m-1} \rho^\ell \| x_1 - x_0 \|,
\]
and finally,
\[
\| x_{k+m} - x_k \| \leq \rho^k (1 - \rho^m) \| x_1 - x_0 \|.
\]

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

In the case of a linear mapping
\[
G(x) = Ax + b,
\]
where $A$ is an $n \times n$ matrix and $b \in \mathbb{R}^n$, it can be shown that $G$ is a contraction mapping (with respect to some norm) if and only if all the eigenvalues of $A$ lie strictly within the unit circle.

For nonstationary iterations of the form $x_{k+1} = G_k(x_k)$, where the function $G_k$ depends on $k$, the following proposition is often useful.

**Proposition A.4.2:** Let $\{\alpha_k\}$ be a nonnegative sequence satisfying

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

where $\{\beta_k\}$ and $\{\gamma_k\}$ are nonnegative sequences such that

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

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 $k \geq \overline{k}$

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

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

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

This contradicts the nonnegativity of $\{\alpha_k\}$ and the assumption $\sum_{k=0}^{\infty} \gamma_k = \infty$.

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

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

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

As an example, consider a sequence of contraction mappings $G_k : \mathbb{R}^n \mapsto \mathbb{R}^n$ with

$$
\|G_k(x) - G_k(y)\| \leq (1 - \gamma_k)\|x - y\|, \quad \forall \ x, y \in \mathbb{R}^n,
$$
where for all $k$,
\[ \gamma_k \in (0, 1], \quad \gamma_k \to 0, \quad \sum_{k=0}^{\infty} \gamma_k = \infty. \]
Assume also that all the functions $G_k$ have the same fixed point $x^*$. Then we have
\[ \| x_{k+1} - x^* \| = \| G_k(x_k) - G_k(x^*) \| \leq (1 - \gamma_k) \| x_k - x^* \|, \]
and from Prop. A.4.2, it follows that the sequence $\{x_k\}$ generated by the iteration $x_{k+1} = G_k(x_k)$ converges to $x^*$ starting from any $x_0 \in \mathbb{R}^n$.

**Supermartingale Convergence**

We now give two theorems relating to supermartingale convergence analysis (the term refers to a collection of theorems that assert convergence of sequences of nonnegative scalars or random variables, which satisfy certain inequalities that imply that the sequences are “almost” nonincreasing). The first theorem relates to deterministic sequences, and is a special case of the second theorem, which 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.4.3:** Let $\{Y_k\}$, $\{Z_k\}$, $\{W_k\}$, and $\{V_k\}$ be four nonnegative scalar sequences such that
\[ Y_{k+1} \leq (1 + V_k)Y_k - Z_k + W_k, \quad k = 0, 1, \ldots, \quad (A.1) \]
and
\[ \sum_{k=0}^{\infty} W_k < \infty, \quad \sum_{k=0}^{\infty} V_k < \infty. \]
Then $\{Y_k\}$ converges to a nonnegative scalar and $\sum_{k=0}^{\infty} Z_k < \infty$.

**Proof:** We will first give the proof assuming that $V_k \equiv 0$, and then generalize. In this case, we have $Y_{k+1} \leq Y_k + W_k$ for all $k$, which implies that
\[ Y_{k+1} \leq Y_0 + \sum_{\ell=0}^{k} W_\ell. \]
Thus the nonnegative sequence $\{Y_k\}$ is bounded above, so it must have a limit point. Similarly, we have
\[ Y_{k+m} \leq Y_k + \sum_{\ell=k}^{k+m} W_\ell, \quad \forall \; k, m \geq 0, \quad (A.2) \]
so if \( \{Y_k\} \) had two limit points \( \bar{y} > y \), we would have \( Y_k \) arbitrarily close to \( \bar{y} \) and \( y \), respectively, for arbitrarily large \( k \), which is a contradiction of Eq. (A.2) and the hypothesis \( \sum_{k=0}^{\infty} W_k < \infty \). Thus \( Y_k \) has a unique limit point. Moreover, by adding Eq. (A.1), we have

\[
\sum_{\ell=0}^{k} Z_\ell \leq Y_{k+1} - Y_0 + \sum_{\ell=0}^{k} W_\ell, \quad \forall \; k = 0, 1, \ldots ,
\]

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 \( \{V_k\} \). Define

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

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

\[
\underline{Y}_{k+1} \leq \underline{Y}_k - \underline{Z}_k + \underline{W}_k.
\]

Since \( \underline{W}_k \leq W_k \), the hypothesis \( \sum_{k=0}^{\infty} W_k < \infty \) implies that \( \sum_{k=0}^{\infty} \underline{W}_k < \infty \), so from the special case of the result already shown, we have that \( \{\underline{Y}_k\} \) converges to a nonnegative scalar and \( \sum_{k=0}^{\infty} \underline{Z}_k < \infty \).

Next we note that

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

since we generally have \( (1 + a) \leq e^a \) and \( \log(1 + a) \leq a \) for any \( a \geq 0 \), so that

\[
\sum_{k=0}^{\infty} V_k < \infty \quad \Rightarrow \quad \prod_{\ell=0}^{\infty} (1 + V_\ell) < \infty.
\]

Since \( \overline{Y}_k = \underline{Y}_k \prod_{\ell=0}^{k-1} (1 + V_\ell) \) and \( \{\overline{Y}_k\} \) converges to a nonnegative scalar, it follows that \( \overline{Y}_k \) also converges to a nonnegative scalar. Finally, we have

\[
Z_k = Z_k \prod_{\ell=0}^{k} (1 + V_\ell) \leq \underline{Z}_k \prod_{\ell=0}^{\infty} (1 + V_\ell),
\]

while \( \sum_{k=0}^{\infty} Z_k < \infty \) and \( \prod_{\ell=0}^{\infty} (1 + V_\ell) < \infty \), so that \( \sum_{k=0}^{\infty} Z_k < \infty \).

Q.E.D.

The next theorem has a long history. The particular version we give here is due to Robbins and Sigmund [RoS71], who prove it assuming the
special case where $V_k \equiv 0$ (see Neveu [Nev75], p. 33, for a proof of this special case).

**Proposition A.4.4: (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
   \[ E\{Y_{k+1} \mid \mathcal{F}_k\} \leq (1 + V_k)Y_k - Z_k + W_k, \quad k = 0, 1, \ldots \]
3. There holds, with probability 1,
   \[ \sum_{k=0}^{\infty} W_k < \infty, \quad \sum_{k=0}^{\infty} V_k < \infty. \]

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