Finite and Countable Sets

A set S is said to be *finite* if it consists of a finite number of elements. It is said to be *countable* if there exists a one-to-one function from S into the set of nonnegative integers. Thus, according to our definition, a finite set is also countable but not conversely. A countable set S that is not finite may be represented by listing its elements $x_0, x_1, x_2, \ldots$ (i.e., $S = \{x_0, x_1, x_2, \ldots\}$). A countable union of countable sets is countable, that is, if $A = \{a_0, a_1, \ldots\}$ is a countable set and $S_{a_0}, S_{a_1}, \ldots$ are each countable sets, then $\cup_{k=0}^{\infty} S_{a_k}$ is also a countable set.

Sets of Real Numbers

If a and b are real numbers or $+\infty$, $-\infty$, we denote by $[a, b]$ the set of numbers x satisfying $a \leq x \leq b$ (including the possibility $x = +\infty$ or $x = -\infty$). 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.

If S is a set of real numbers that is bounded above, then there is a smallest real number y such that $x \leq y$ for all $x \in S$. This number is called the *least upper bound* or *supremum* of S and is denoted by $\sup\{x \mid x \in S\}$ or $\max\{x \mid x \in S\}$. (This is somewhat inconsistent with normal mathematical usage, where the use of max in place of sup indicates that the supremum is attained by some element of S.) Similarly, the greatest real number z such that $z \leq x$ for all $x \in S$ is called the *greatest lower bound* or *infimum* of S and is denoted by $\inf\{x \mid x \in S\}$ or $\min\{x \mid x \in S\}$. If S is unbounded above, we write $\sup\{x \mid x \in S\} = +\infty$, and if it is unbounded below, we write $\inf\{x \mid x \in S\} = -\infty$. If S is the empty set, then by convention we write $\inf\{x \mid x \in S\} = +\infty$ and $\sup\{x \mid x \in S\} = -\infty$.

A.2 EUCLIDEAN SPACE

The set of all n-tuples $x = (x_1, \ldots, x_n)$ of real numbers constitutes the *n-dimensional Euclidean space*, denoted by $\Re^n$. The elements of $\Re^n$ are referred to as n-dimensional vectors or simply vectors when confusion cannot arise. The one-dimensional Euclidean space $\Re^1$ consists of all the real numbers and is denoted by $\Re$. Vectors in $\Re^n$ can be added by adding their corresponding components. They can be multiplied by a scalar by multiplication of each component by a scalar. The *inner product* of two vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ is denoted by $x'y$ and is equal to $\sum_{i=1}^n x_i y_i$. The *norm* of a vector $x = (x_1, \ldots, x_n) \in \Re^n$ is denoted by $\|x\|$ and is equal to $(x'x)^{1/2} = (\sum_{i=1}^n x_i^2)^{1/2}$.

A set of vectors $a_1, a_2, \ldots, a_k$ is said to be *linearly dependent* if there exist scalars $\lambda_1, \lambda_2, \ldots, \lambda_k$, not all zero, such that

$$\lambda_1 a_1 + \cdots + \lambda_k a_k = 0.$$

If no such set of scalars exists, the vectors are said to be *linearly independent*.

A.3 MATRICES

An $m \times n$ *matrix* is a rectangular array of numbers, referred to as elements or components, which are arranged in m rows and n columns. If $m = n$ the matrix is said to be *square*. The element in the ith row and jth column of a matrix A is denoted by a subscript ij, such as a_{ij}, in which case we write $A = [a_{ij}]$. The $n \times n$ *identity matrix*, denoted by I, is the matrix with elements $a_{ij} = 0$ for $i \neq j$ and $a_{ii} = 1$, for $i = 1, \ldots, n$. The *sum* of two $m \times n$ matrices A and B is written as $A + B$ and is the matrix whose elements are the sum of the corresponding elements in A and B. The *product of a matrix A and a scalar* λ, written as λA or $A\lambda$, is obtained by multiplying each element of A by λ. The *product* AB of an $m \times n$ matrix A and an $n \times p$ matrix B is the $m \times p$ matrix C with elements $c_{ij} = \sum_{k=1}^n a_{ik} b_{kj}$. If b is an n-dimensional column vector and A is an $m \times n$ matrix, then Ab is an m-dimensional column vector.

The *transpose* of an $m \times n$ matrix A is the $n \times m$ matrix A' with elements $a'_{ij} = a_{ji}$. The elements of a given row (or column) of A constitute a vector called a row vector (or column vector, respectively) of A. A square matrix A is *symmetric* if $A' = A$. An $n \times n$ matrix A is called *nonsingular* or *invertible* if there is an $n \times n$ matrix called the *inverse* of A and denoted by A^{-1}, such that $A^{-1}A = I = AA^{-1}$, where I is the $n \times n$ identity matrix. An $n \times n$ matrix is nonsingular if and only if its n row vectors are linearly independent or, equivalently, if its n column vectors are linearly independent. Thus, an $n \times n$ matrix A is nonsingular if and only if the relation $Av = 0$, where $v \in \Re^n$, implies that $v = 0$.

Rank of a Matrix

The *rank* of a matrix A is equal to the maximum number of linearly independent row vectors of A. It is also equal to the maximum number of linearly independent column vectors. Thus, the rank of an $m \times n$ matrix is at most equal to the minimum of the dimensions m and n. An $m \times n$ matrix is said to be of *full rank* if its rank is maximal, that is, if its rank is equal to the minimum of m and n. A square matrix is of full rank if and only if it is nonsingular.

Eigenvalues

Given a square $n \times n$ matrix A, the determinant of the matrix $\gamma I - A$, where I is the $n \times n$ identity matrix and γ is a scalar, is an nth degree polynomial. The n roots of this polynomial are called the *eigenvalues* of A. Thus, γ is an eigenvalue of A if and only if the matrix $\gamma I - A$ is singular, or equivalently, if and only if there exists a nonzero vector v such that $Av = \gamma v$. Such a vector v is called an *eigenvector* corresponding to γ. The eigenvalues and eigenvectors of A can be complex even if A is real. A matrix A is singular if and only if it has an eigenvalue that is equal to zero. If A is nonsingular, then the eigenvalues of A^{-1} are the reciprocals of the eigenvalues of A. The eigenvalues of A and A' coincide.

If $\gamma_1, \ldots, \gamma_n$ are the eigenvalues of A, then the eigenvalues of $cI + A$, where c is a scalar and I is the identity matrix, are $c + \gamma_1, \ldots, c + \gamma_n$. The eigenvalues of A^k, where k is any positive integer, are equal to $\gamma_1^k, \ldots, \gamma_n^k$. From this it follows that $\lim_{k \to 0} A^k = 0$ if and only if all the eigenvalues of A lie strictly within the unit circle of the complex plane. Furthermore, if the latter condition holds, the iteration

$$x_{k+1} = Ax_k + b,$$

where b is a given vector, converges to

$$\overline{x} = (I - A)^{-1} b,$$

which is the unique solution of the equation $x = Ax + b$.

If all the eigenvalues of A are distinct, then their number is exactly n, and there exists a set of corresponding linearly independent eigenvectors. In this case, if $\gamma_1, \ldots, \gamma_n$ are the eigenvalues and $v_1, \ldots, v_n$ are such eigenvectors, every vector $x \in \Re^n$ can be decomposed as

$$x = \sum_{i=1}^{n} \xi_i v_i,$$

where ξ_i are some unique (possibly complex) numbers. Furthermore, we have for all positive integers k,

$$A^k x = \sum_{i=1}^{n} \gamma_i^k \xi_i v_i.$$

If A is a transition probability matrix, that is, all the elements of A are nonnegative and the sum of the elements of each of its rows is equal to 1, then all the eigenvalues of A lie within the unit circle of the complex plane. Furthermore, 1 is an eigenvalue of A and the unit vector $(1, 1, \ldots, 1)$ is a corresponding eigenvector.

Positive Definite and Semidefinite Symmetric Matrices

A square symmetric $n \times n$ matrix A is said to be *positive semidefinite* if $x'Ax \geq 0$ for all $x \in \Re^n$. It is said to be *positive definite* if $x'Ax > 0$ for all nonzero $x \in \Re^n$. The matrix A is said to be *negative semidefinite (definite)* if $-A$ is *positive semidefinite (definite)*. In this book, the notions of positive definiteness and semidefiniteness will be used only in connection with symmetric matrices.

A positive definite symmetric matrix is invertible and its inverse is also positive definite symmetric. Also, an invertible positive semidefinite symmetric matrix is positive definite. Analogous results hold for negative definite and semidefinite symmetric matrices. If A and B are $n \times n$ positive semidefinite (definite) symmetric matrices, then the matrix $\lambda A + \mu B$ is also positive semidefinite (definite) symmetric for all $\lambda \geq 0$ and $\mu \geq 0$. If A is an $n \times n$ positive semidefinite symmetric matrix and C is an $m \times n$ matrix, then the matrix CAC' is positive semidefinite symmetric. If A is positive definite symmetric, and C has rank m (equivalently, $m \leq n$ and C has full rank), then CAC' is positive definite symmetric.

An $n \times n$ positive definite symmetric matrix A can be written as CC' where C is a square invertible matrix. If A is positive semidefinite symmetric and its rank is m, then it can be written as CC', where C is an $n \times m$ matrix of full rank.

A symmetric $n \times n$ matrix A has real eigenvalues and a set of n real linearly independent eigenvectors, which are orthogonal (the inner product of any pair is 0). If A is positive semidefinite (definite) symmetric, its eigenvalues are nonnegative (respectively, positive).

Partitioned Matrices

It is often convenient to partition a matrix into submatrices. For example, the matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix}$$

may be partitioned into

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

where

$$A_{11} = \begin{pmatrix} a_{11} & a_{12} \end{pmatrix}, \qquad A_{12} = \begin{pmatrix} a_{13} & a_{14} \end{pmatrix},$$

$$A_{21} = \begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}, \qquad A_{22} = \begin{pmatrix} a_{23} & a_{24} \\ a_{33} & a_{34} \end{pmatrix}.$$

We separate the components of a partitioned matrix by a space, as in $(B\ C)$, or by a comma, as in (B, C). The transpose of the partitioned matrix A is

$$A' = \begin{pmatrix} A'_{11} & A'_{21} \\ A'_{12} & A'_{22} \end{pmatrix}.$$

Partitioned matrices may be multiplied just as nonpartitioned matrices, provided the dimensions involved in the partitions are compatible. Thus if

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \qquad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

then

$$AB = \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{pmatrix},$$

provided the dimensions of the submatrices are such that the preceding products $A_{ij}B_{jk}$, $i, j, k = 1, 2$ can be formed.

Matrix Inversion Formulas

Let A and B be square invertible matrices, and let C be a matrix of appropriate dimension. Then, if all the following inverses exist, we have

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

The equation can be verified by multiplying the right-hand side by

$$A + CBC'$$

and showing that the product is the identity matrix.

Consider a partitioned matrix M of the form

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

Then we have

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

where

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

provided all the inverses exist. The proof is obtained by multiplying M with the expression given for M^{-1} and verifying that the product yields the identity matrix.

A.4 ANALYSIS

Convergence of Sequences

A sequence of vectors $x_0, x_1, \ldots, x_k, \ldots$ in $\Re^n$, denoted by $\{x_k\}$, is said to converge to a *limit* x if $\|x_k - x\| \to 0$ as $k \to \infty$ (i.e., if, given any $\epsilon > 0$, there is an integer N such that for all $k \geq N$ we have $\|x_k - x\| < \epsilon$). If $\{x_k\}$ converges to x, we write $x_k \to x$ or $\lim_{k\to\infty} x_k = x$. We have $Ax_k + By_k \to Ax + By$ if $x_k \to x$, $y_k \to y$, and A, B are matrices of appropriate dimension.

A vector x is said to be a *limit point* of a sequence $\{x_k\}$ if there is a subsequence of $\{x_k\}$ that converges to x, that is, if there is an infinite subset $\mathcal{K}$ of the nonnegative integers such that for any $\epsilon > 0$, there is an integer N such that for all $k \in \mathcal{K}$ with $k \geq N$ we have $\|x_k - x\| < \epsilon$.

A sequence of real numbers $\{r_k\}$, which is monotonically nondecreasing (nonincreasing), that is, satisfies $r_k \leq r_{k+1}$ for all k, must either converge to a real number or be unbounded above (below). In the latter case we write $\lim_{k\to\infty} r_k = \infty$ $(-\infty)$. Given any bounded sequence of real numbers $\{r_k\}$, we may consider the sequence $\{s_k\}$, where $s_k = \sup\{r_i \mid i \geq k\}$. Since this sequence is monotonically nonincreasing and bounded, it must have a limit. This limit is called the *limit superior* of $\{r_k\}$ and is denoted by $\limsup_{k\to\infty} r_k$. The *limit inferior* of $\{r_k\}$ is similarly defined and is denoted by $\liminf_{k\to\infty} r_k$. If $\{r_k\}$ is unbounded above, we write $\limsup_{k\to\infty} r_k = \infty$, and if it is unbounded below, we write $\liminf_{k\to\infty} r_k = -\infty$. We also use this notation if $r_k \in [-\infty, \infty]$ for all k.

Open, Closed, and Compact Sets

A subset S of $\Re^n$ is said to be *open* if for every vector $x \in S$ one can find an $\epsilon > 0$ such that $\{z \mid \|z - x\| < \epsilon\} \subset S$. A set S is *closed* if and only if every convergent sequence $\{x_k\}$ with elements in S converges to a point that also belongs to S. A set S is said to be *compact* if and only if it is both closed and bounded (i.e., it is closed and for some $M > 0$ we have $\|x\| \leq M$ for all $x \in S$). A set S is compact if and only if every sequence $\{x_k\}$ with elements in S has at least one limit point that belongs to S. Another important fact is that if $S_0, S_1, \ldots, S_k, \ldots$ is a sequence of nonempty compact sets in $\Re^n$ such that $S_k \supset S_{k+1}$ for all k, then the intersection $\cap_{k=0}^{\infty} S_k$ is a nonempty and compact set.

Continuous Functions

A function f mapping a set S_1 into a set S_2 is denoted by $f : S_1 \to S_2$. A function $f : \Re^n \to \Re^m$ is said to be *continuous* if for all x, $f(x_k) \to f(x)$ whenever $x_k \to x$. Equivalently, f is continuous if, given $x \in \Re^n$ and $\epsilon > 0$,

there is a $\delta > 0$ such that whenever $\|y-x\| < \delta$, we have $\|f(y)-f(x)\| < \epsilon$. The function

$$(a_1 f_1 + a_2 f_2)(\cdot) = a_1 f_1(\cdot) + a_2 f_2(\cdot)$$

is continuous for any two scalars a_1, a_2 and any two continuous functions $f_1, f_2 : \Re^n \to \Re^m$. If S_1, S_2, S_3 are any sets and $f_1 : S_1 \to S_2$, $f_2 : S_2 \to S_3$ are functions, the function $f_2 \cdot f_1 : S_1 \to S_3$ defined by $(f_2 \cdot f_1)(x) = f_2\big(f_1(x)\big)$ is called the *composition* of f_1 and f_2. If $f_1 : \Re^n \to \Re^m$ and $f_2 : \Re^m \to \Re^p$ are continuous, then $f_2 \cdot f_1$ is also continuous.

Derivatives

Let $f : \Re^n \mapsto \Re$ be some function. For a fixed $x \in \Re^n$, the first partial derivative of f at the point x with respect to the ith coordinate is defined by

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

where e_i is the ith unit vector, and we assume that the above limit exists. If the partial derivatives with respect to all coordinates exist, f is called differentiable at x and its *gradient* at x is defined to be the column vector

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

The function f is called differentiable if it is differentiable at every $x \in \Re^n$. If $\nabla f(x)$ exists for every x and is a continuous function of x, f is said to be *continuously differentiable*. Such a function admits, for every fixed x, the first order expansion

$$f(x+y) = f(x) + y' \nabla f(x) + o(\|y\|),$$

where $o(\|y\|)$ is a function of y with the property $\lim_{\|y\| \to 0} o(\|y\|)/\|y\| = 0$.

A vector-valued function $f : \Re^n \mapsto \Re^m$ is called differentiable (respectively, continuously differentiable) if each component f_i of f is differentiable (respectively, continuously differentiable). The *gradient matrix* of f, denoted by $\nabla f(x)$, is the $n \times m$ matrix whose ith column is the gradient $\nabla f_i(x)$ of f_i. Thus,

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

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

If the gradient $\nabla f(x)$ is itself a differentiable function, then f is said to be twice differentiable. We denote by $\nabla^2 f(x)$ the Hessian matrix of f at x, that is, the matrix

$$\nabla^2 f(x) = \left[\frac{\partial^2 f(x)}{\partial x^i \partial x^j}\right]$$

the elements of which are the second partial derivatives of f at x.

Let $f : \Re^k \mapsto \Re^m$ and $g : \Re^m \mapsto \Re^n$ be continuously differentiable functions, and let $h(x) = g\big(f(x)\big)$. The *chain rule* for differentiation states that

$$\nabla h(x) = \nabla f(x) \nabla g\big(f(x)\big), \qquad \text{for all } x \in \Re^k.$$

For example, if A and B are given matrices, then if $h(x) = Ax$, we have $\nabla h(x) = A'$ and if $h(x) = ABx$, we have $\nabla h(x) = B'A'$.

A.5 CONVEX SETS AND FUNCTIONS

A subset C of $\Re^n$ is said to be *convex* if for every $x, y \in C$ and every scalar α with $0 \le \alpha \le 1$, we have $\alpha x + (1-\alpha)y \in C$. In words, C is convex if the line segment connecting any two points in C belongs to C. A function $f : C \to \Re$, defined over a convex subset C of $\Re^n$, is said to be *convex* if for every $x, y \in C$ and every scalar α with $0 \le \alpha \le 1$ we have

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

The function f is said to be *concave* if $(-f)$ is convex, or equivalently if for every $x, y \in C$ and every scalar α with $0 \le \alpha \le 1$ we have

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

If $f : C \to \Re$ is convex, then the sets $\Gamma_\lambda = \big\{x \mid x \in C,\ f(x) \le \lambda\big\}$ are convex for every scalar λ. An important property is that a real-valued convex function defined over $\Re^n$ is continuous.

If $f_1, f_2, \ldots, f_m$ are convex functions defined over a convex subset C of $\Re^n$ and $\alpha_1, \alpha_2, \ldots, \alpha_m$ are nonnegative scalars, then the function $\alpha_1 f_1 + \cdots + \alpha_m f_m$ is also convex over C. If $f : \Re^m \to \Re$ is convex, A is an $m \times n$ matrix, and b is a vector in $\Re^m$, the function $g : \Re^n \to \Re$ defined by $g(x) = f(Ax + b)$ is also convex. If $f : \Re^n \to \Re$ is convex, then the function $g(x) = E_w\big\{f(x+w)\big\}$, where w is a random vector in $\Re^n$, is a convex function provided the expected value is finite for every $x \in \Re^n$.

For functions $f : \Re^n \to \Re$ that are differentiable, there are alternative characterizations of convexity. Thus, f is convex if and only if

$$f(y) \ge f(x) + \nabla f(x)'(y-x), \qquad \text{for all } x, y \in \Re^n.$$

If f is twice continuously differentiable, then f is convex if and only if $\nabla^2 f(x)$ is a positive semidefinite symmetric matrix for every $x \in \Re^n$.

For accounts of convexity and its applications in optimization, see Bertsekas [BNO03] and Rockafellar [Roc70].

APPENDIX C:
On Probability Theory

This appendix lists selectively some of the basic probabilistic notions that we will be using. Its main purpose is to familiarize the reader with some of our terminology. It is not meant to be exhaustive, and the reader should consult textbooks such as Ash [Ash70], Feller [Fel68], Papoulis [Pap65], Ross [Ros85], Stirzaker [Sti94], and Bertsekas and Tsitsiklis [BeT02] for detailed accounts. For fairly accessible treatments of measure theoretic probability, see Adams and Guillemin [AdG86], and Ash [Ash72].

C.1 PROBABILITY SPACES

A *probability space* consists of

(a) A set Ω.

(b) A collection $\mathcal{F}$ of subsets of Ω, called *events*, which includes Ω and has the following properties:

(1) If A is an event, then the complement $\overline{A} = \{\omega \in \Omega \mid \omega \notin A\}$ is also an event. (The complement of Ω is the empty set and is considered to be an event.)

(2) If $A_1, A_2, \ldots, A_k, \ldots$ are events, then $\cup_{k=1}^{\infty} A_k$ is also an event.

(3) If $A_1, A_2, \ldots, A_k, \ldots$ are events, then $\cap_{k=1}^{\infty} A_k$ is also an event.

(c) A function $P(\cdot)$ assigning to each event A a real number $P(A)$, called the *probability of the event* A, and satisfying:

(1) $P(A) \geq 0$ for every event A.

(2) $P(\Omega) = 1$.

(3) $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ for every pair of disjoint events A_1, A_2.

(4) $P(\cup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k)$ for every sequence of mutually disjoint events $A_1, A_2, \ldots, A_k, \ldots$

The function P is referred to as a *probability measure*.

Convention for Finite and Countable Probability Spaces

The case of a probability space where the set Ω is a countable (possibly finite) set is encountered frequently in this book. When we specify that Ω is finite or countable, we implicitly assume that the associated collection of events is the collection of *all* subsets of Ω (including Ω and the empty set). Then, if Ω is a finite set, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, the probability space is specified by the probabilities $p_1, p_2, \ldots, p_n$, where p_i denotes the probability of the event consisting of just ω_i. Similarly, if $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k, \ldots\}$, the probability space is specified by the corresponding probabilities $p_1, p_2, \ldots, p_k, \ldots$ In either case we refer to $(p_1, p_2, \ldots, p_n)$ or $(p_1, p_2, \ldots, p_k, \ldots)$ as a *probability distribution over* Ω.

C.2 RANDOM VARIABLES

A *random variable* on a probability space $(\Omega, \mathcal{F}, P)$ is a function $x : \Omega \to \Re$ such that for every scalar λ the set

$$\{\omega \in \Omega \mid x(\omega) \leq \lambda\}$$

is an event (i.e., belongs to the collection $\mathcal{F}$). An n-dimensional *random vector* $x = (x_1, x_2, \ldots, x_n)$ is an n-tuple of random variables $x_1, x_2, \ldots, x_n$, each defined on the same probability space.

We define the *distribution function* $F : \Re \to \Re$ [or *cumulative distribution function* (CDF for short)] of a random variable x by

$$F(z) = P\Big(\{\omega \in \Omega \mid x(\omega) \leq z\}\Big);$$

that is, $F(z)$ is the probability that the random variable takes a value less than or equal to z. We define the distribution function $F : \Re^n \to \Re$ of a random vector $x = (x_1, x_2, \ldots, x_n)$ by

$$F(z_1, z_2, \ldots, z_n) = P\Big(\{\omega \in \Omega \mid x_1(\omega) \leq z_1, x_2(\omega) \leq z_2, \ldots, x_n(\omega) \leq z_n\}\Big).$$

Given the distribution function of a random vector $x = (x_1, \ldots, x_n)$, the (marginal) distribution function of each random variable x_i is obtained from

$$F_i(z_i) = \lim_{z_j \to \infty, j \neq i} F(z_1, z_2, \ldots, z_n).$$

The random variables $x_1, \ldots, x_n$ are said to be *independent* if

$$F(z_1, z_2, \ldots, z_n) = F_1(z_1)F_2(z_2)\cdots F_n(z_n),$$

for all scalars $z_1, \ldots, z_n$.

The *expected value of a random variable* x with distribution function F is defined by

$$E\{x\} = \int_{-\infty}^{\infty} z dF(z)$$

provided the integral is well-defined. The *expected value of a random vector* $x = (x_1, \ldots, x_n)$ is the vector

$$E\{x\} = \big(E\{x_1\}, E\{x_2\}, \ldots, E\{x_n\}\big).$$

The *covariance matrix* of a random vector $x = (x_1, \ldots, x_n)$ with expected value $E\{x\} = (\overline{x}_1, \ldots, \overline{x}_n)$ is defined to be the $n \times n$ positive semidefinite symmetric matrix

$$\begin{pmatrix} E\big\{(x_1 - \overline{x}_1)^2\big\} & \cdots & E\big\{(x_1 - \overline{x}_1)(x_n - \overline{x}_n)\big\} \\ \vdots & \vdots & \vdots \\ E\big\{(x_n - \overline{x}_n)(x_1 - \overline{x}_1)\big\} & \cdots & E\big\{(x_n - \overline{x}_n)^2\big\} \end{pmatrix},$$

provided the expected values are well-defined.

Two random vectors x and y are said to be *uncorrelated* if

$$E\Big\{\big(x - E\{x\}\big)\big(y - E\{y\}\big)'\Big\} = 0,$$

where $\big(x - E\{x\}\big)$ is viewed as a column vector and $\big(y - E\{y\}\big)'$ is viewed as a row vector.

The random vector $x = (x_1, \ldots, x_n)$ is said to be characterized by a *probability density function* $f : \Re^n \to \Re$ if

$$F(z_1, z_2, \ldots, z_n) = \int_{-\infty}^{z_1} \int_{-\infty}^{z_2} \cdots \int_{-\infty}^{z_n} f(y_1, \ldots, y_n) dy_1 \ldots dy_n,$$

for every $z_1, \ldots, z_n$.

C.3 CONDITIONAL PROBABILITY

We restrict ourselves to the case where the underlying probability space Ω is a countable (possibly finite) set and the set of events is the set of all subsets of Ω.

Given two events A and B, we define the *conditional probability of B given A* by

$$P(B \mid A) = \begin{cases} \frac{P(A \cap B)}{P(A)} & \text{if } P(A) > 0, \\ 0 & \text{if } P(A) = 0. \end{cases}$$

We also use the notation $P\{B \mid A\}$ in place of $P(B \mid A)$. If $B_1, B_2, \ldots$ are a countable (possibly finite) collection of mutually exclusive and exhaustive events (i.e., the sets B_i are disjoint and their union is Ω) and A is an event, then we have

$$P(A) = \sum_i P(A \cap B_i).$$

From the two preceding relations, we obtain the *total probability theorem:*

$$P(A) = \sum_i P(B_i)P(A \mid B_i).$$

We thus obtain for every k,

$$P(B_k \mid A) = \frac{P(A \cap B_k)}{P(A)} = \frac{P(B_k)P(A \mid B_k)}{\sum_i P(B_i)P(A \mid B_i)},$$

assuming that $P(A) > 0$. This relation is referred to as *Bayes' rule.*

Consider now two random vectors x and y taking values in $\Re^n$ and $\Re^m$, respectively [i.e., $x(\omega) \in \Re^n$, $y(\omega) \in \Re^m$ for all $\omega \in \Omega$]. Given two subsets X and Y of $\Re^n$ and $\Re^m$, respectively, we denote

$$P(X \mid Y) = P\Big(\big\{\omega \mid x(\omega) \in X\big\} \mid \big\{\omega \mid y(\omega) \in Y\big\}\Big).$$

For a fixed vector $v \in \Re^n$, we define the *conditional distribution function* of x given v by

$$F(z \mid v) = P\Big(\big\{\omega \mid x(\omega) \leq z\big\} \mid \big\{\omega \mid y(\omega) = v\big\}\Big),$$

and the *conditional expectation* of x given v by

$$E\{x \mid v\} = \int_{\Re^n} z dF(z \mid v),$$

assuming that the integral is well-defined. Note that $E\{x \mid v\}$ is a function mapping v into $\Re^n$.

Finally, let us provide Bayes' rule for random vectors. If $\omega_1, \omega_2, \ldots$ are the elements of Ω, denote

$$z_i = x(\omega_i), \qquad v_i = y(\omega_i), \qquad i = 1, 2, \ldots$$

Also, for any vectors $z \in \Re^n$, $v \in \Re^m$, let us denote

$$P(z) = P\Big(\{\omega \mid x(\omega) = z\}\Big), \qquad P(v) = P\Big(\{\omega \mid y(\omega) = v\}\Big).$$

We have $P(z) = 0$ if $z \neq z_i$, $i = 1, 2, \ldots$, and $P(v) = 0$ if $v \neq v_i$, $i = 1, 2, \ldots$ Denote also

$$P(z \mid v) = P\Big(\{\omega \mid x(\omega) = z\} \mid \{\omega \mid y(\omega) = v\}\Big),$$

$$P(v \mid z) = P\Big(\{\omega \mid y(\omega) = v\} \mid \{\omega \mid x(\omega) = z\}\Big).$$

Then, for all $k = 1, 2, \ldots$, Bayes' rule yields

$$P(z_k \mid v) = \begin{cases} \frac{P(z_k)P(v|z_k)}{\sum_i P(z_i)P(v|z_i)} & \text{if } P(v) > 0, \\ 0 & \text{if } P(v) = 0. \end{cases}$$

APPENDIX D:
On Finite-State Markov Chains

This appendix provides some of the basic probabilistic notions related to stationary Markov chains with a finite number of states. For detailed presentations, see Ash [Ash70], Bertsekas and Tsitsiklis [BeT02], Chung [Chu60], Gallager [Gal99], Kemeny and Snell [KeS60], and Ross [Ros85].

D.1 STATIONARY MARKOV CHAINS

A square $n \times n$ matrix $[p_{ij}]$ is said to be a *stochastic* matrix if all its elements are nonnegative, that is, $p_{ij} \geq 0$, $i, j = 1, \ldots, n$, and the sum of the elements of each of its rows is equal to 1, that is, $\sum_{j=1}^n p_{ij} = 1$ for all $i = 1, \ldots, n$.

Suppose we are given a stochastic $n \times n$ matrix P together with a finite set of states $S = \{1, \ldots, n\}$. The pair (S, P) will be referred to as a *stationary finite-state Markov chain*. We associate with (S, P) a process whereby an initial state $x_0 \in S$ is chosen in accordance with some initial probability distribution

$$r_0 = (r_0^1, r_0^2, \ldots, r_0^n).$$

Subsequently, a transition is made from state x_0 to a new state $x_1 \in S$ in accordance with a probability distribution specified by P as follows. The

probability that the new state will be j is equal to p_{ij} whenever the initial state is i, i.e.,

$$P(x_1 = j \mid x_0 = i) = p_{ij}, \qquad i, j = 1, \ldots, n.$$

Similarly, subsequent transitions produce states $x_2, x_3, \ldots$ in accordance with

$$P(x_{k+1} = j \mid x_k = i) = p_{ij}, \qquad i, j = 1, \ldots, n. \tag{D.1}$$

The probability that after the kth transition the state x_k will be j, given that the initial state x_0 is i, is denoted by

$$p_{ij}^k = P(x_k = j \mid x_0 = i), \qquad i, j = 1, \ldots, n. \tag{D.2}$$

A straightforward calculation shows that these probabilities are equal to the elements of the matrix P^k (P raised to the kth power), in the sense that p_{ij}^k is the element in the ith row and jth column of P^k:

$$P^k = [p_{ij}^k]. \tag{D.3}$$

Given the initial probability distribution p_0 of the state x_0 (viewed as a row vector in $\Re^n$), the probability distribution of the state x_k after k transitions

$$r_k = (r_k^1, r_k^2, \ldots, r_k^n)$$

(viewed again as a row vector) is given by

$$r_k = r_0 P^k, \qquad k = 1, 2, \ldots \tag{D.4}$$

This relation follows from Eqs. (D.2) and (D.3) once we write

$$r_k^j = \sum_{i=1}^n P(x_k = j \mid x_0 = i) r_0^i = \sum_{i=1}^n p_{ij}^k r_0^i.$$

D.2 CLASSIFICATION OF STATES

Given a stationary finite-state Markov chain (S, P), we say that two states i and j *communicate* if there exist two positive integers k_1 and k_2 such that $p_{ij}^{k_1} > 0$ and $p_{ji}^{k_2} > 0$. In words, states i and j communicate if one can be reached from the other with positive probability.

Let $\tilde{S} \subset S$ be a subset of states such that:

1. All states in $\tilde{S}$ communicate.

2. If $i \in \tilde{S}$ and $j \notin \tilde{S}$, then $p_{ij}^k = 0$ for all k.

Then we say that $\tilde{S}$ forms a *recurrent class* of states.

If S forms by itself a recurrent class (i.e., all states communicate with each other), then we say that the Markov chain is *irreducible*. It is possible that there exist several recurrent classes. It can also be proved that at least one recurrent class must exist. A state that belongs to some recurrent class is called *recurrent*; otherwise it is called *transient*. We have

$$\lim_{k \to \infty} p_{ii}^k = 0 \quad \text{if and only if} \quad i \text{ is transient.}$$

In other words, if the process starts at a transient state, the probability of returning to the same state after k transitions diminishes to zero as k tends to infinity.

The definitions imply that if the process starts within a recurrent class, it stays within that class. If it starts at a transient state, it eventually (with probability one) enters a recurrent class after a number of transitions, and subsequently remains there.

D.3 LIMITING PROBABILITIES

An important property of any stochastic matrix P is that the matrix P^* defined by

$$P^* = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k \tag{D.5}$$

exists [in the sense that the sequences of the elements of $(1/N) \sum_{k=0}^{N-1} P^k$ converge to the corresponding elements of P^*]. A proof of this is given in Prop. A.1 of Appendix A in Vol. II. The elements p_{ij}^* of P^* satisfy

$$p_{ij}^* \geq 0, \qquad \sum_{j=1}^n p_{ij}^* = 1, \qquad i, j = 1, \ldots, n.$$

Thus, P^* is a stochastic matrix.

Note that the (i, j)th element of the matrix P^k is the probability that the state will be j after k transitions starting from state i. With this in mind, it can be seen from the definition (D.5) that p_{ij}^* can be interpreted as the long term fraction of time that the state is j given that the initial state is i. This suggests that for any two states i and i' in the same recurrent class we have $p_{ij}^* = p_{i'j}^*$, and this can indeed be proved. In particular, if a Markov chain is irreducible, the matrix P^* has identical rows. Also, if j is a transient state, we have

$$p_{ij}^* = 0, \qquad \text{for all } i = 1, \ldots, n,$$

so the columns of the matrix P^* corresponding to transient states consist of zeroes.