IMAGE ALGEBRA

G.X. Ritter

CENTER FOR COMPUTER VISION AND VISUALIZATION
Department of Computer and Information Science and Engineering
University of Florida
Gainesville, FL 32611

PREFACE

This document is an initial draft of the developing theory of image algebra. The primary objective of this treatise is to provide the reader with an introduction to the theory and foundations of image algebra. For readers interested in applications and image algebra specification of a wide variety of image processing transforms we recommend the *Handbook of Computer Vision Algorithms in Image Algebra* [46].

Since the discipline of image algebra is in its infancy and a state of flux, this document will undergo various changes before its completion in book format. The book will consist of eight chapters and will be largely self contained. The first chapter will contain all the introductory material; e.g., what is image algebra all about, the history of image algebra, the people involved, organization of the book, etc. Chapters 2 and 3 contain basic background material dealing with point set theory, topology, and abstract algebra. Lack of this background is often a fatal stumbling block to understanding the underlying concepts of image algebra and the mathematics of computer vision in general.

As to this initial draft, we recommend reading the introduction and then proceed to a quick overview of the basic concepts that define image algebra, we suggest to start with Section **3.13** (Chapter **3**) and then proceed directly to Chapter **4**, referring to preceding sections for notation and theorems as the need arises.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

Since the field of image algebra is a recent development it will be instructive to provide some background information. In the broad sense, image algebra is a mathematical theory concerned with the transformation and analysis of images. Although much emphasis is focused on the analysis and transformation of digital images, the main goal is the establishment of a comprehensive and unifying theory of image transformations, image analysis, and image understanding in the discrete as well as the continuous domain [45].

The idea of establishing a unifying theory for the various concepts and operations encountered in image and signal processing is not new. Over thirty years ago, Unger proposed that many algorithms for image processing and image analysis could be implemented in parallel using*cellular array*Cellular array computer computers [61]. These cellular array computers were inspired by the work of von Neumann in the 1950s [63, 64]. Realization of von Neumann's cellular array machines was made possible with the advent of VLSI technology. NASA's massively parallel processor or MPP and the CLIP series of computers developed by Duff and his colleagues represent the classic embodiment of von Neumann's original automaton [2, 18, 16, 17, 20]. A more general class of cellular array computers are pyramids and Thinking Machines Corporation's Connection Machines [59, 60, 25]. In an abstract sense, the various versions of Connection Machines are universal cellular automatons with an additional mechanism added for non-local communication.

Many operations performed by these cellular array machines can be expressed in terms of simple elementary operations. These elementary operations create a mathematical basis for the theoretical formalism capable of expressing a large number of algorithms for image processing and analysis. In fact, a common thread among designers of parallel image processing architectures is the belief that large classes of image transformations can be described by a small set of standard rules that induce these architectures. This belief led to the creation of mathematical formalisms that were used to aid in the design of special-purpose parallel architectures. Matheron and Serra's Texture Analyzer [28], ERIM's (Environmental Research Institute of Michigan) Cytocomputer [32, 57, 31], and Martin Marietta's GAPP [7, 5, 6] are examples of this approach.

The formalism associated with these cellular architectures is that of pixel neighborhood arithmetic and mathematical morphology. Mathematical morphologyMathematical morphologyMorphology is the part of image processing concerned with image filtering and analysis by structuring elements. It grew out of the early work of Minkowski and Hadwiger [39, 40, 22], and entered the modern era through the work of Matheron and Serra of the Ecole des Mines in Fontainebleau, France [37, 52, 53, 54]. Matheron and Serra not only formulated the modern concepts of morphological image transformations, but also designed and built the Texture Analyzer System. Since those early days, morphological operations have been applied from low-level, to intermediate, to high-level vision problems. Among some recent research papers on morphological image processing are Crimmins and Brown [8], Haralick et al. [24, 23], Maragos and Schafer [34, 36, 35], Davidson [14, 13], Dougherty [15], Goutsias [51, 21], and Koskinen and Astola [30].

Serra and Sternberg were the first to unify morphological concepts and methods into a coherent algebraic theory specifically designed for image processing and image analysis. Sternberg was also the first to use the term "image algebra" [56, 58]. In the mid 1980s, Maragos introduced a new theory unifying a large class of linear and nonlinear systems under the theory of mathematical morphology [33]. More recently, Davidson completed the mathematical foundation of mathematical morphology by formulating

its embedding into the lattice algebra known as *Mini-Max algebra* [11, 12]. However, despite these profound accomplishments, morphological methods have some well-known limitations. For example, such fairly common image processing techniques as feature extraction based on convolution, Fourier-like transformations, chain coding, histogram equalization transforms, image rotation, and image registration and rectification are — with the exception of a few simple cases — either extremely difficult or impossible to express in terms of morphological operations. The failure of a morphologically based image algebra to express a fairly straightforward U.S. government-furnished FLIR (forward-looking infrared) algorithm was demonstrated by Miller of Perkin-Elmer [38].

The failure of an image algebra based solely on morphological operations to provide a universal image processing algebra is due to its set-theoretic formulation, which rests on the Minkowski addition and subtraction of sets [22]. These operations ignore the linear domain, transformations between different domains (spaces of different sizes and dimensionality), and transformations between different value sets (algebraic structures), e.g., sets consisting of real, complex, or vector valued numbers. The image algebra discussed in this text includes these concepts and extends the morphological operations [45].

The development of image algebra grew out of a need, by the U.S. Air Force Systems Command, for a common image-processing language. Defense contractors do not use a standardized, mathematically rigorous and efficient structure that is specifically designed for image manipulation. Documentation by contractors of algorithms for image processing and rationale underlying algorithm design is often accomplished via word description or analogies that are extremely cumbersome and often ambiguous. The result of these *ad hoc* approaches has been a proliferation of nonstandard notation and increased research and development cost. In response to this chaotic situation, the Air Force Armament Laboratory (AFATL — now known as Wright Laboratory MNGA) of the Air Force Systems Command, in conjunction with the Defense Advanced Research Project Agency (DARPA — now known as the Advanced Research Project Agency or ARPA), supported the early development of image algebra with the intent that the fully developed structure would subsequently form the basis of a common image-processing language. The goal of AFATL was the development of a complete, unified algebraic structure that provides a common mathematical environment for image-processing algorithm development, optimization, comparison, coding, and performance evaluation. The development of this structure proved highly successful, capable of fulfilling the tasks set forth by the government, and is now commonly known as image algebra.

Because of the goals set by the government, the theory of image algebra provides for a language which, if properly implemented as a standard image processing environment, can greatly reduce research and development costs. Since the foundation of this language is purely mathematical and independent of any future computer architecture or language, the longevity of an image algebra standard is assured. Furthermore, savings due to commonality of language and increased productivity could dwarf any reasonable initial investment for adapting image algebra as a standard environment for image processing.

Although commonality of language and cost savings are two major reasons for considering image algebra as a standard language for image processing, there exists a multitude of other reasons for desiring the broad acceptance of image algebra as a component of all image processing development systems. Premier among these is the predictable influence of an image algebra standard on future image processing technology. In this, it can be compared to the influence on scientific reasoning and the advancement of science due to the replacement of the myriad of different number systems (e.g., Roman, Syrian, Hebrew, Egyptian, Chinese, etc.) by the now common Indo-Arabic notation. Additional benefits provided by the use of image algebra are

- The elemental image algebra operations are small in number, translucent, simple, and provide a method of transforming images that is easily learned and used;
- Image algebra operations and operands provide the capability of expressing all image-to-image transformations;
- Theorems governing image algebra make computer programs based on image algebra notation amenable to both machine dependent and machine independent optimization techniques;
- The algebraic notation provides a deeper understanding of image manipulation operations due to conciseness and brevity of code and is capable of suggesting new techniques;
- The notational adaptability to programming languages allows the substitution of extremely short and concise image algebra expressions for equivalent blocks of code, and therefore increases programmer productivity;
- Image algebra provides a rich mathematical structure that can be exploited to relate image processing problems to other mathematical areas;
- Without image algebra, a programmer will never benefit from the bridge that exists between an image algebra programming language and the multitude of mathematical structures, theorems, and identities that are related to image algebra;
- There is no competing notation that adequately provides all these benefits.

The role of image algebra in computer vision and image processing tasks and theory should not be confused with the government's Ada programming language effort. The goal of the development of the Ada programming language was to provide a single high-order language in which to implement embedded systems. The special architectures being developed nowadays for image processing applications are not often capable of directly executing Ada language programs, often due to support of parallel processing models not accommodated by Ada's tasking mechanism. Hence, most applications designed for such processors are still written in special assembly or microcode languages. Image algebra, on the other hand, provides a level of specification, directly derived from the underlying mathematics on which image processing is based and that is compatible with both sequential and parallel architectures.

Enthusiasm for image algebra must be tempered by the knowledge that image algebra, like any other field of mathematics, will never be a finished product but remain a continuously evolving mathematical theory concerned with the unification of image processing and computer vision tasks. Much of the mathematics associated with image algebra and its implication to computer vision remains largely unchartered territory which awaits discovery. For example, very little work has been done in relating image algebra to computer vision techniques which employ tools from such diverse areas as knowledge representation, graph theory, and surface representation.

Several image algebra programming languages have been developed. These include image algebra Fortran (IAF) [68], an image algebra Ada (IAA) translator [65], image algebra Connection Machine *Lisp [67, 19], an image algebra language (IAL) implementation on transputers [9, 10], and an image algebra C++ class library (`iac++`) [66, 62]. Unfortunately, there is often a tendency among engineers to confuse or equate these languages with image algebra. An image algebra programming language is *not* image algebra, which is a mathematical theory. An image algebra-based programming language typically implements a particular subalgebra of the full image algebra. In addition, simplistic implementations can result in poor computational performance. Restrictions and limitations in implementation are usually due to a combination of factors, the most pertinent being development costs and hardware and software environment constraints. They are not limitations of image algebra, and they should not be confused with the capability of image algebra as a mathematical tool for image manipulation.

Image algebra is a *heterogeneous* or *many-valued algebra*Heterogeneous algebra in the sense of Birkhoff and Lipson [3, 45], with multiple sets of operands and operators. Manipulation of images for purposes of image enhancement, analysis, and understanding involves operations not only on images, but also on different types of values and quantities associated with these images. Thus, the basic operands of image algebra are images and the values and quantities associated with these images. Roughly speaking, an image consists of two things, a collection of *points* and a set of *values* associated with these points. Images are therefore endowed with two types of information, namely the spatial relationship of the points, and also some type of numeric or other descriptive information associated with these points. Consequently, the field of image algebra bridges two broad mathematical areas, the theory of point sets and the algebra of value sets, and investigates their interrelationship. In the sections that follow we discuss point and value sets as well as images, templates, and neighborhoods that characterize some of their interrelationships.

CHAPTER 2
ELEMENTS OF POINT SET TOPOLOGY

In order to read anything about our subject, the reader will have to learn the language that is used in it. We shall try to keep the number of technical terms as small as possible, but there is a certain minimum vocabulary that is essential. Much of the standard language is taken from point set topology and the theory of algebraic systems, subjects with which we are not concerned for their own sake. Both subjects are, indeed, independent branches of mathematics. Point set topology and set theory have their own basic undefined concepts, subject to various axioms; one of these undefined concepts is the notion of a *set* itself. It is not our intention to formally define the required axioms that govern the use of sets but deal with sets on an intuitive basis.

2.1 Sets

Intuitively, we think of a set as something made up by all the objects that satisfy some given condition, such as the set of integers, the set of pages in this book, or the set of objects named in a list. The objects making up the set are called the elements, or members, of the set and may themselves be sets, as in the case of all subsets of a given set.

We adopt the convention of denoting sets by capital letters and the elements of sets by small letters. The following is a brief summary of some of the things we shall simply assume about sets.

2.1.1 A set *X* is comprised of *elements*, and if x is one of the elements, we shall denote this fact by "$x \in X$." The notation "$x \notin X$" shall denote the fact that x is an object which is not an element of *X*.

2.1.2 There is exactly one set with no elements. It is the *empty set*, and is denoted by the symbol $\varnothing$.

Throughout this book, the notation of symbolic logic will be used to shorten statements. If *p* and *q* are propositions, then the statement "$p \Rightarrow q$" means that *p* implies *q* or, equivalently, if *p* is true, then *q* is true. The statement "$p \iff q$" is read: "*p* if and only if *q*," and means that *p and q are logically equivalent*; i.e. "$p \Rightarrow q$ and $q \Rightarrow p$."

An expression $p(x)$ that becomes a proposition whenever values from a specified domain of discourse are substituted for x is called a propositional function or, equivalently, a condition on x; and *p* is called a property, or predicate. The assertion "x has property *p*" means that "$p(x)$" is true. Thus, if $p(x)$ is the propositional function "x is an integer," then *p* denotes the property "is an integer," and "$p(2)$" is true, whereas "$p(1/2)$" is false.

The quantifier "there exists" is denoted by $\exists$, and the quantifier "for each" is denoted by $\forall$. The assertion "$\forall x\ \exists y\ s.t. \forall z :\ p(x, y, z)$" reads "for each x there exists a y such that for each z, $p(x, y, z)$ is true."

A set may be described either by giving a characterizing property or by listing the elements. The standard way to describe a set by listing elements is to enclose the designations of the elements, separated by commas, in braces, e.g. {1,2,3,4,5}. In terms of a characterizing property this set could be written

as $\{x : p(x)\}$, which reads "the set of all x such that $p(x)$," where *p* denotes the property "is a positive integer less than 6."

If X and Y are sets, then "$X = Y$" will mean that X and Y have the same elements; that is, $\forall x : (x \in X) \iff (x \in Y)$.

$X \subset Y$, read "*X* is a subset of *Y*," signifies that each element of *X* is an element of *Y*, that is, $\forall x : (x \in X) \Rightarrow (x \in Y)$. Equality is not excluded — we call *X* a *proper subset* of *Y* whenever $X \subset Y$ and $X \neq Y$. The set whose elements are all the subsets of a given set *X* is called the *power set* of *X* and is denoted by 2^X. The following statements are evident:

2.1.3 $X \subset X$ for every set *X*.

2.1.4 If $X \subset Y$ and $Y \subset Z$, then $X \subset Z$ (i.e. $\subset$ is transitive).

2.1.5 $X = Y$ if and only if both $X \subset Y$ and $Y \subset X$.

2.1.6 $\varnothing \subset X$ for every set *X*.

2.1.7 $\varnothing \in 2^X$ and $X \in 2^X$.

2.1.8 $Y \subset X \iff Y \in 2^X$, and $x \in X \iff \{x\} \in 2^X$.

Of these, **2.1.5** is very important: the equality of two sets is usually proven by showing that each of the two inclusions is valid.

Throughout this text, various familiar sets of numbers will occur naturally. For convenience, we shall now reserve:

$\mathbb{Z}$ to denote the set of integers,

$\mathbb{Q}$ the set of rational numbers,

$\mathbb{R}$ the set of real numbers,

and

$\mathbb{C}$ the set of complex numbers.

The notation $\mathbb{Z}^+$, $\mathbb{Z}^-$, $\mathbb{R}^+$, and $\mathbb{R}^-$ will refer to the set of all positive integers, the set of all negative integers, the set of all positive real numbers, and the set of all negative real numbers, respectively. Observe that the number 0 is not an element of any of these four sets. The set whose elements are the number 0 and the positive integers will be denoted by $\mathbb{N}$. Thus,

$$\mathbb{N} = \{0, 1, 2, 3, \ldots\}.$$

Three finite subsets of $\mathbb{Z}$ that will occur throughout much of this text are

$$\mathbb{Z}_n = \{0, 1, \ldots, n-1\},\ \mathbb{Z}_n^+ = \{1, 2, \ldots, n\},\ \text{and}\ \mathbb{Z}_{\pm n} = \{-n, \ldots, -1, 0, 1, \ldots, n\},$$

where $n \in \mathbb{Z}^+$.

2.2 The Algebra of Sets

Intuitively, an *algebra* is simply a collection of nonempty sets together with a finite number of operations (rules) for transforming one or more elements of the sets into another element of one of the sets. The formal definition of an algebra will be given in Chapter **3**. In this section we define the elementary set-theoretic operations and present a list of standard formulae which are convenient in symbolic work.

When defining operations on and between sets it is customary to view the sets under consideration as subsets of some larger set *U*, called a *universal set* or the *universe of discourse*. For instance, in plane geometry, the universal set consists of all the points in the plane.

2.2.1 **Example**: Consider the equation

$$(x+1)(2x-3)\left(x^2+1\right)=0$$

whose solution set is the set $X=\left\{-1,\frac{3}{2},i,-i\right\}$, where $i=\sqrt{-1}$, provided that $\mathbb{C}$ is the universal set. However, if $\mathbb{R}$ is the universal set, then $X=\left\{-1,\frac{3}{2}\right\}$.

2.2.2 **Definition**. Let *X* and *Y* be given sets. The *union* of *X* and *Y*, written $X\cup Y$, is defined as the set whose elements are either in *X* or in *Y* (or in both *X* and *Y*). Thus,

$$X\cup Y=\{z:\ z\in X \text{ or } z\in Y\}.$$

The *intersection* of *X* and *Y*, written $X\cap Y$, is defined as the set of all elements that belong to both *X* and *Y*. Thus, $X\cap Y=\{z:\ z\in X \text{ and } z\in Y\}$. For example, $\mathbb{N}\cup\mathbb{Z}^-=\mathbb{Z}$ and $\mathbb{N}\cap\{-2,-1,0,1,2\}=\{0,1,2\}$.

Two sets *X* and *Y* are called *disjoint* if they have no elements in common, that is, if $X\cap Y=\varnothing$. Obviously, $\mathbb{Z}^+$ and $\mathbb{Z}^-$ are disjoint.

If $X\subset U$, then the *complement* of *X* (with respect to *U*) is denoted by X' and is defined as $X'=\{x:\ x\in U,\ x\notin X\}$. The *difference* of two sets $X,Y\subset U$ is denoted by $X\backslash Y$ and defined as $X\backslash Y=\{x\in X:\ x\notin Y\}$. Note that $X'=U\backslash X$.

For future reference we list below (**2.2.1**) some of the more important laws governing operations with sets. Here *X*, *Y*, and *Z* are subsets of some given universal set *U*.

Because of associativity, we can designate $X\cup(Y\cup Z)$ simply by $X\cup Y\cup Z$. Similarly, a union (or intersection) of four sets, say $(W\cup X)\cup(Z\cup Y)$, can be written as $W\cup X\cup Y\cup Z$ because, by associativity, the distribution of parentheses is irrelevant, and by commutativity, the order of terms plays no role. By induction, the same remarks apply to the union (or intersection) of any finite number of sets. The union of *n* sets, $X_1,\ldots,X_n$, is written $\bigcup_{i=1}^{n}X_i$, and the intersection is $\bigcap_{i=1}^{n}X_i$.

The relation between $\cap$, $\cup$, and $\subset$ is given by:

2.2.3 The statements

(i) $X\subset Y$, (ii) $X=X\cap Y$, (iii) $Y=X\cup Y$, (iv) $Y'\subset X'$, and (v) $X\cap Y'=\varnothing$

are all equivalent.

Identity Laws	
$X \cup U = U$	$X \cap U = X$
$X \cup \varnothing = X$	$X \cap \varnothing = \varnothing$
Idempotent Laws	
$X \cup X = X$	$X \cap X = X$
Complement Laws	
$(X')' = X$	$\varnothing' = U,\ U' = \varnothing$
$X \cup X' = U$	$X \cap X' = \varnothing$
Associative Laws	
$(X \cup Y) \cup Z = X \cup (Y \cup Z)$	$(X \cap Y) \cap Z = X \cap (Y \cap Z)$
Commutative Laws	
$X \cup Y = Y \cup X$	$X \cap Y = Y \cap X$
Distributive Laws	
$X \cup (Y \cap Z) = (X \cup Y) \cap (X \cup Z)$	$X \cap (Y \cup Z) = (X \cap Y) \cup (X \cap Z)$
DeMorgan's Laws	
$(X \cup Y)' = X' \cap Y'$	$(X \cap Y)' = X' \cup Y'$

Figure 2.2.1 Laws of Operations with Sets

2.3 Cartesian Products

Let $X = \{x, y\}$ and $Y = \{w, y, z\}$. The set of distinct ordered pairs

$$Z = \{(x, w), (x, y), (x, z), (y, w), (y, y), (y, z)\}$$

in which the first component of each ordered pair is an element of *X* while the second component is an element of *Y* is called the *Cartesian product* of *X* and *Y*. Ordered pairs are subject to the one condition: $(x, y) = (z, w) \iff x = z \ and \ y = w$; in particular, $(x, y) = (y, x) \iff x = y$. Since the Cartesian product is one of the most important constructions of set theory — enabling us to express many concepts in terms of sets — we define it formally.

2.3.1 **Definition**. Let *X, Y* be two sets. The *Cartesian product* of *X* and *Y,* denoted by $X \times Y$, is the set of all ordered pairs $\{(x, y) : \ x \in X, y \in Y\}$.

2.3.2 **Example**: View the elements of $X = \{1, 2, 3\}$ as the coordinates of points on the *x*-axis and the elements of $Y = \{1, 2, 3, 4\}$ as the coordinates of points on the *y*-axis. Then the elements of $X \times Y$ are the rectangular coordinates of the twelve points shown (Figure **2.3.1**).

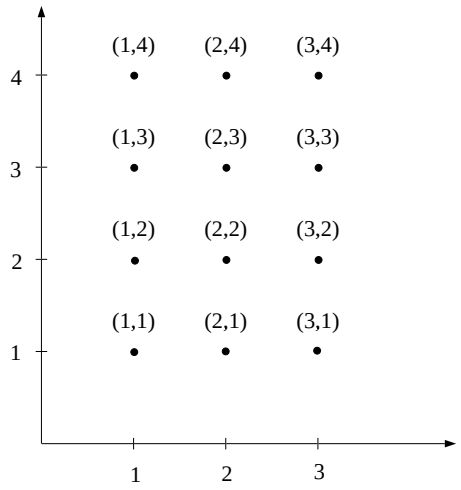


Figure 2.3.1 The Cartesian product $X \times Y$ viewed as a subset of the plane.

If in the above example we would have let $X = Y = \mathbb{N}$ instead, then the elements of the set $X \times Y = \mathbb{N} \times \mathbb{N}$ are all the points in the first quadrant having integral coordinates. Similarly, when $X = Y = \mathbb{R}$, the set $X \times Y = \mathbb{R} \times \mathbb{R}$ consists of all pairs (x, y) of real numbers x and y, and represents the usual (x, y)-coordinate plane.

The Cartesian product of three sets *X, Y,* and *Z* is defined by $X \times Y \times Z = (X \times Y) \times Z$, and that of n sets by induction: $X_1 \times \cdots \times X_n = (X_1 \times \cdots \times X_{n-1}) \times X_n$. The Cartesian product of *n* sets is also denoted by $\prod_{i=1}^{n} X_i$; an element of $\prod_{i=1}^{n} X_i$ is written $(x_1, \ldots, x_n)$ and x_i is called *the ith coordinate*. If $X = X_i$ for $i = 1, \ldots, n$, then we define $X^n = \prod_{i=1}^{n} X_i$. In particular, the (x, y)-coordinate plane, also known as *two dimensional Euclidean space*, is denoted by $\mathbb{R}^2$, while $\mathbb{Z}^2$ denotes the discrete subset of $\mathbb{R}^2$ consisting of all points having integral coordinates.

The notion of Cartesian product extends the set of elementary set-theoretic operations. In comparison to $\cup$ and $\cap$, however, the Cartesian product is neither commutative nor associative: in general, $X \times Y \neq Y \times X$ and $(X \times Y) \times Z \neq X \times (Y \times Z)$. Also, $X \times Y = \varnothing \iff X = \varnothing$ or $Y = \varnothing$ (or both). The relation between the Cartesian product and the operations of union and intersection can be summarized as follows:

2.3.3 $X = (Y \cup Z) = (X \times Y) \cup (X \times Z)$.
$X \times (Y \cap Z) = (X \times Y) \cap (X \times Z)$.

2.4 Families of Sets

If for each element λ of some nonempty set Λ there corresponds a set X_λ, then the collection $\{X_\lambda : \lambda \in \Lambda\}$ is called a *family* of sets, and Λ is called an *indexing set* for the family. We also write $\{X_\lambda\}_{\lambda \in \Lambda}$ for $\{X_\lambda : \lambda \in \mathbf{\Lambda}\}$. If the indexing set $\Lambda = \mathbb{N}$, then the indexed family $\{X_i\}_{i \in \mathbb{N}}$ is called a *sequence* (of sets) and may also be denoted by $\{X_i\}_{i=1}^{\infty}$.

The notion of union and intersection can be generalized to any arbitrary indexed family of subsets of some universal set *U*.

2.4.1 **Definition.** Let $\{X_\lambda\}_{\lambda\in\Lambda}$ be a family of subsets of a universal set *U*. The *union* of this family is denoted by $\bigcup_{\lambda\in\Lambda} X_\lambda$ and is the set

$$\{x \in U : x \in X_\lambda \text{ for at least one } \lambda \in \Lambda\} .$$

The *intersection* is denoted by $\bigcap_{\lambda\in\Lambda} X_\lambda$ and is the set

$$\{x \in U : x \in X_\lambda \text{ for every } \lambda \in \Lambda\} .$$

For a sequence $\{X_i\}_{i=1}^{\infty}$ of sets we also use the notation

$$\bigcup_{i=1}^{\infty} X_i = X_1 \cup X_2 \cup \ldots, \text{ and } \bigcap_{i=1}^{\infty} X_i = X_1 \cap X_2 \cap \ldots$$

to denote union and intersection, respectively.

2.4.2 **Example:** Let $\Lambda = \{\lambda : \lambda \in \mathbb{R} : 0 \le \lambda \le 1\}$. For each $\lambda \in \Lambda$, let $X_\lambda = \{r : r \in \mathbb{R},\ 0 \le r \le \lambda\}$. Then

$$\bigcup_{\lambda\in\Lambda} X_\lambda = \{r : 0 \le r \le 1\} = \Lambda, \text{ and } \bigcap_{\lambda\in\Lambda} X_\lambda = \{0\} .$$

Example **2.4.2** can be generalized to the following useful fact: let *X* be any set and, for each $x \in X$, let X_x be a subset of *X* such that $x \in X_x \subset X$. Then $X = \bigcup_{x\in X} X_x$.

It follows from the definition that the union and intersection of a family of sets does not depend on how the family is indexed. That is, union and intersection are unrestrictive commutative and associative. The complement laws, distributive laws, and DeMorgan's laws also hold for these generalized operations. In particular, we have

2.4.3 If $\{X_\lambda\}_{\lambda\in\Lambda}$ is any family of subsets of some universal set *U* and $Y \subset U$, then

(1) $Y \cup \left(\bigcap_{\lambda\in\Lambda} X_\lambda\right) = \bigcap_{\lambda\in\Lambda} (Y \cup X_\lambda)$

(2) $Y \cap \left(\bigcup_{\lambda\in\Lambda} X_\lambda\right) = \bigcup_{\lambda\in\Lambda} (Y \cap X_\lambda)$

(3) $\left(\bigcup_{\lambda\in\Lambda} X_\lambda\right)' = \bigcap_{\lambda\in\Lambda} X_\lambda'$ and $\left(\bigcap_{\lambda\in\Lambda} X_\lambda\right)' = \bigcup_{\lambda\in\Lambda} X_\lambda'$

(4) $Y \times \left(\bigcup_{\lambda\in\Lambda} X_\lambda\right) = \bigcup_{\lambda\in\mathbf{\Lambda}} \left(Y \times X_\lambda\right)$

(5) $Y \times \left(\bigcap_{\lambda\in\Lambda} X_\lambda\right) = \bigcap_{\lambda\in\Lambda} \left(Y \times X_\lambda\right)$

(6) $\bigcap_{\lambda\in\Lambda} 2^{X_\lambda} = 2^{\left(\bigcap_{\lambda\in\Lambda} X_\lambda\right)}$ and $\bigcup_{\lambda\in\Lambda} 2^{X_\lambda} \subset 2^{\left(\bigcup_{\lambda\in\Lambda} X_\lambda\right)}$.

2.5 Functions

The notion of a *function* (or *map*) is basic in all mathematics. Intuitively, a function *f* from a set *X* into a set *Y*, written $f : X \to Y$, is a rule which assigns to each $x \in X$ some element y of *Y*, where the assignment of x to y by the rule *f* is denoted by $f(x) = y$ or $f : x \mapsto y$. However, we shall define the notion of a function formally in terms of the primitive concept "set" by identifying functions with their graphs.

2.5.1 **Definition.** Let *X* and *Y* be two sets. A *function f* from *X* to *Y*, denoted by $f : X \to Y$ is a subset $f \subset X \times Y$ with the property: for each $x \in X$, there is one, and only one, $y \in Y$ such that $(x, y) \in f$. The set of all functions from *X* into *Y* will be denoted by Y^X. Thus $Y^X = \{f : f \text{ is a function from } X \text{ to } Y\}$.

We write $f(x) = y$ for $(x, y) \in f$ and say that y is the *value f* assumes at x, or that y is the *evaluation of f at* x. For instance, defining $X = Y = \mathbb{N}$, then the set $f \subset X \times Y$ defined by

(i) $f = \{(x, y) : y = 2x + 1,\ x \in X\}$ or

(ii) $f = \{(x, 2x + 1) :\ x \in X\}$

is a function $f : X \to Y$. Observe that this function is completely specified by the rule

(iii) $f(x) = 2x + 1$ or, equivalently, by either $y = 2x + 1$ or $x \mapsto 2x + 1$.

Throughout much of this book we shall specify functions by assignment rules (as in (iii)) and call the set $\{(x, y) : y = f(x) \text{ and } x \in X\}$, where $f : X \to Y$, the *graph* of the assignment rule $f(x) = y$ or, simply, the graph of *f*. Note that under this definition the statement "the graph of *f* " is synonymous with the statement "the function *f*." We will also periodically refer to certain special properties and types of functions. In particular, it will be important to distinguish between the following types of functions:

2.5.2 A function $f : X \to Y$ is said to be *one-to-one* (or 1–1) if distinct elements in *X* have distinct evaluations; i.e. $x \neq z \Rightarrow f(x) \neq f(z)$ or, equivalently, $f(x) = f(z) \Rightarrow x = z$.

2.5.3 A function $f : X \to Y$ is said to be *onto* (or *f* is a function from *X onto Y*) if every $y \in Y$ corresponds to an evaluation $f(x)$ for some $x \in X$; i.e. if $y \in Y \Rightarrow \exists x \in X \text{ such that } f(x) = y$.

2.5.4 A function $f : X \to Y$ assigning all $x \in X$ to the same single element $y \in Y$ is called a *constant* function.

2.5.5 The function $f : X \to X$ with the property $f(x) = x \ \ \forall x \in X$ is called the *identity* function on *X* and will be denoted by 1_X. If $A \subset X$, the function $i : A \to X$ given by $i(a) = a$ is called the *inclusion* of *A* into *X*.

2.5.6 Given sets $X_1, \ldots, X_n$, the function $p_j : \prod_{i=1}^{n} X_i \to X_j$, where $1 \leq j \leq n$, defined by $p_j(x_1, \ldots, x_j, \ldots, x_n) = x_j$ is called the *projection* onto the *j*th coordinate.

2.5.7 Suppose $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$, and X are nonempty sets, and $\forall i = 1, \ldots, n$ there exist functions $f_i : X_i \to Y_i$ and $g_i : X \to X_i$. Then the families $\{f_i : i = 1, \ldots, n\}$ and $\{g_i : i = 1, \ldots, n\}$ induce new functions $f : \prod_{i=1}^{n} X_i \to \prod_{i=1}^{n} Y_i$ and $g : X \to \prod_{i=1}^{n} X_i$ that are defined by $f(x_1, \ldots, x_n) = (f_1(x_1), \ldots, f_n(x_n))$ and $g(x) = (g_1(x), \ldots, g_n(x))$, respectively. The functions f_i and g_i are called the *ith-coordinate functions* of *f* and *g*, respectively.

2.5.8 Given a function $f : X \to Y$ and a subset $A \subset X$, then the function *f* considered *only* on *A* (i.e. the function $\{ (x, f(x)) : x \in A \}$) is called the *restriction* of *f* to *A* and is denoted by $f|_A$. Thus, $f|_A = f \cap (A \times Y)$.

2.5.9 In the reverse direction, if $A \subset X$ and $f : A \to Y$, then any function $F : X \to Y$ with the property $F|_A = f$, is called *an extension of f over X relative to Y.*

2.5.10 Let $X = A \cup B$. Given functions $f : A \to Y$ and $g : B \to Y$, then the function $F : X \to Y$ defined by

$$F(x) = \begin{cases} f(x) & if\ x \in A \\ g(x) & if\ x \in X \backslash A \end{cases}$$

is called *the extension of f to g over X relative to Y.* We will use the symbol $f|^g$ to denote this extension.

The difference between **2.5.9** and **2.5.10** is in the definition of *an* extension and *the* extension of the function *f*. Observe also that in 2.5.10, $F|_A = (f|^g)\ \big|_A = f$.

The following examples should help in clarifying the important concepts of "one-to-one" and "onto" functions. The function $f : \mathbb{N} \to \mathbb{N}$ defined by $f(x) = 2x+1$ is not onto since, for example, $f(x) \neq 2 \in \mathbb{N}$ for any $x \in \mathbb{N}$. However, *f* is one-to-one since $2x + 1 = 2z + 1 \Rightarrow x = z$. On the other hand, the function $g : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$ defined by $g(x) = x^2$ is onto since for every $y \in \mathbb{R}^+ \cup \{0\}$ $\exists x \in \mathbb{R}$ (namely $x = \pm\sqrt{y}$) such that $g(x) = y$. But g is not one-to-one since $g(-2) = g(2)$ and $-2 \neq 2$.

Given two functions $f : X \to Y$ and $g : Y \to Z$, then the *composition* $g \circ f : X \to Z$ is defined by $(g \circ f)(x) = g(f(x))$ $\forall x \in X$. The following theorem indicates a simple method for establishing that a given function *f* (respectively g) is one-to-one (respectively onto).

2.5.11 **Theorem.** *Let $f : X \to Y$ and $g : Y \to X$ satisfy $g \circ f = 1_X$. Then f is one-to-one and g is onto.*

Proof: Since $f(x) = f(z) \Rightarrow x = (g \circ f)(x) = g(f(x)) = g(f(z)) = (g \circ f)(z) = z$, we have that *f* is one-to-one. The function g is onto since for any $x \in X$ $\exists y \in Y$, namely $y = f(x)$, such that $x = (g \circ f)(x) = g(f(x)) = g(y)$.

Q.E.D.

As a simple illustration of Theorem **2.5.11** we show that for any function $h : X \to Z$, the function $f : X \to Y$, where $Y = X \times Z$ and $f(x) = (x, h(x))$ is one-to-one. Let $p_1 : X \times Z \to X$ be the

projection onto the first coordinate. Then $p_1 \circ f : X \to X$ is 1_X. Hence, by Theorem **2.5.11**, *f* is one-to-one (and p_1 is onto).

The fact that the composite of one-to-one and onto functions is again a one-to-one and onto function follows from the next theorem.

2.5.12 Theorem. *Suppose* $f : X \to Y$ *and* $g : Y \to Z$.

(i) *If f and g are onto, then* $g \circ f : X \to Z$ *is onto.*

(ii) *If f and g are one-to-one, then* $g \circ f : X \to Z$ *is one-to-one.*

Proof: (i) Let $z \in Z$. Since g is onto, $\exists y \in Y$ such that $g(y) = z$. Since *f* is onto, $\exists x \in X$ such that $f(x) = y$. But then $(g \circ f)(x) = g(f(x)) = z$. Therefore, $g \circ f$ is onto.

(ii) Suppose $(g \circ f)(x) = (g \circ f)(x')$, i.e. $g(f(x)) = g(f(x'))$. Then $f(x) = f(x')$ since g is one-to-one. But then $x = x'$ since *f* is one-to-one. Accordingly, $g \circ f$ is one-to-one.

Q.E.D.

2.6 Induced Set Functions

If $f : X \to Y$, then the set *X* is called the *domain* of *f* and is denoted by $domain(f)$. The *range* of *f*, denoted by $range(f)$, is defined as $range(f) = \{f(x) : x \in X\}$. It follows that $domain(f|_A) = A$, $range(f|_A) = \{f(x) : x \in A\}$, and that $f : X \to range(f)$ is onto.

Since for each $A \in 2^X$, $range(f|_A) \in 2^Y$, *f* induces a function $\check{f} : 2^X \to 2^Y$ defined by

$$\check{f}(A) = \{\, f(x) : x \in A \,\}.$$

In addition, *f* also induces a function $\check{f}^{-1} : 2^Y \to 2^X$ defined by

$$\check{f}^{-1}(B) = \{\, x : x \in X \text{ and } f(x) \in B \,\}.$$

The set $\check{f}^{-1}(B)$ is called the *inverse image* of *B*.

It is common practice to let *f*(A) and $f^{-1}(B)$ denote the evaluation $\check{f}(A)$ and $\check{f}^{-1}(B)$, respectively. Since in most cases the context of discussion avoids possible misinterpretation, this economizes notational overhead. Although we shall follow this convention to some extent, we point out that in machine implementation and algorithm description the maps f, f^{-1}, $\check{f}$, and $\check{f}^{-1}$ must be distinguished as they do represent different processes.

2.6.1 Example: Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^2$, $A = \{1, -\sqrt{2}, \sqrt{2}, 3\}$, and $B = \{1, 4, 9\}$. Then $f(A) = \{1, 2, 9\}$ and $f^{-1}(B) = \{1, -1, 2, -2, 3, -3\}$.

The induced set-valued functions $\check{f}$ and $\check{f}^{-1}$ are called *set functions* since they are functions of sets into sets. Induced set functions possess various properties. In particular we state:

2.6.2 **Theorem.** *Let $f : X \to Y$, $A \subset B \subset X$, and $\{A_\lambda\}_{\lambda\in\Lambda}$ be any family of subsets of X. Then:*

(i) $f(A) \subset f(B)$

(ii) $f\left(\bigcup_{\lambda\in\Lambda} A_\lambda\right) = \bigcup_{\lambda\in\Lambda} f(A_\lambda)$

(iii) $f\left(\bigcap_{\lambda\in\Lambda} A_\lambda\right) \subset \bigcap_{\lambda\in\Lambda} f(A_\lambda)$

The following example shows that the inclusion (iii) cannot, in general, be replaced by equality.

2.6.3 **Example:** Let $A = \{\,(x,y) : 1 \le x \le 2,\ 1 \le y \le 2\,\}$, $B = \{\,(x,y) : 1 \le x \le 2,\ 3 \le y \le 4\,\}$, and $p_1 : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ the projection onto the first coordinate (i.e. the *x*-axis). Since $A \cap B = \varnothing$, $\check{p}_1(A \cap B) = \varnothing \neq \{x : 1 \le x \le 2\} = \check{p}_1(A) \cap \check{p}_1(B)$.

Of the two induced set functions, $\check{f}^{-1}$ is the more important one as it is much more *well-behaved* in the sense that it preserves the elementary set operations.

2.6.4 **Theorem.** *Let $f : X \to Y$, $A \subset B \subset Y$, and $\{B_\lambda\}_{\lambda\in\Lambda}$ any family of subsets of Y. Then:*

(i) $f^{-1}(A) \subset f^{-1}(B)$

(ii) $f^{-1}(B') = \big(f^{-1}(B)\big)'$

(iii) $f^{-1}\left(\bigcup_{\lambda\in\Lambda} B_\lambda\right) = \bigcup_{\lambda\in\Lambda} f^{-1}(B_\lambda)$

(iv) $f^{-1}\left(\bigcap_{\lambda\in\Lambda} B_\lambda\right) = \bigcap_{\lambda\in\Lambda} f^{-1}(B_\lambda)$

Proof: We only prove (iv); the proofs of (i) through (iii) are just as simple. Observe that

$$x \in f^{-1}\left(\bigcap_{\lambda\in\Lambda} B_\lambda\right) \iff f(x) \in \bigcap_{\lambda\in\Lambda} B_\lambda \iff f(x) \in B_\lambda\ \forall\lambda \in \Lambda$$
$$\iff x \in f^{-1}(B_\lambda)\ \ \forall\lambda\in\Lambda \iff x \in \bigcap_{\lambda\in\Lambda} f^{-1}(B_\lambda)$$

$$x \in f^{-1}\left(\bigcap_{\lambda\in\Lambda} B_\lambda\right) \iff f(x) \in \bigcap_{\lambda\in\Lambda} B_\lambda \iff f(x) \in B_\lambda\ \forall\lambda \in \Lambda$$
$$\iff x \in f^{-1}(B_\lambda)\ \ \forall\lambda\in\Lambda \iff x \in \bigcap_{\lambda\in\Lambda} f^{-1}(B_\lambda)$$

Q.E.D.

The following important relationship between the two set functions is also simple to verify:

2.6.5 Theorem. *Let $f : X \to Y$, $A \subset X$, and $B \subset Y$. Then:*
(i) $A \subset f^{-1} \circ f(A)$ *and*
(ii) $f \circ f^{-1}(B) \subset B$.

The inclusions (i) and (ii) cannot, in general, be replaced by equality.

2.6.6 Example: Let *A* and p_1 be as in Example **2.6.3**. Then

$$A \neq p_1^{-1}(p_1(A)) = p_1^{-1}(\{\, x : 1 \leq x \leq 2 \,\}) = \{(x, y) : 1 \leq x \leq 2,\ -\infty < y < \infty\}\,.$$

If $f : X \to Y$ is onto, then equality holds in **2.6.5** (ii). In particular, *f* is onto $\iff \check{f}\big(\check{f}^{-1}(B)\big) = B\ \forall B \subset Y$. Similarly, *f* is one-to-one $\iff \forall y \in \check{f}(X)$ the set $\check{f}^{-1}(\{y\})$ consists of a single element. Thus, with each one-to-one function $f : X \to Y$ there is associated a function $f^{-1} : \check{f}(X) \to X$, called the *inverse* of *f*, which is defined by $f^{-1}(y) = x$, where $\{x\} = \check{f}^{-1}(\{y\})$; that is, $f^{-1}(y) \in \check{f}^{-1}(\{y\})\ \forall y \in \check{f}(X)$. For example, if $f : \mathbb{R}^+ \to \mathbb{R}$ is given by $f(x) = \sqrt{x}$, then $f^{-1}(2) \in \check{f}^{-1}(\{2\}) = \{4\}$ and, hence, $f^{-1}(2) = 4$. Similarly, $f^{-1}\big(\sqrt{2}\big) = 2$ and, in general, $f^{-1}(x) = x^2\ \forall x \in \check{f}(\mathbb{R}^+)$. If $f : X \to Y$ is both one-to-one and onto, then the inverse of *f* is a function $f^{-1} : Y \to X$ since $f(X) = Y$. This differs from $\check{f}^{-1} : 2^Y \to 2^X$ even though, as pointed out earlier, it is standard practice to use the same functional notation.

There are several useful observations concerning one-to-one and onto functions. Suppose that $f : X \to Y$ is one-to-one and onto. Then $f^{-1} : Y \to X$ is one-to-one and onto, and $(f^{-1})^{-1} = f$. In addition, equality holds in **2.6.5** (i) and (ii). If we also have a one-to-one and onto function $g : Y \to Z$, then both $g \circ f$ and $(g \circ f)^{-1}$ are one-to-one and onto functions. This follows from Theorem **2.5.12** and the fact that $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

2.7 Finite, Countable, and Uncountable Sets

Two sets are said to be *equivalent* if there exists a one-to-one and onto function $f : X \to Y$. Hence the idea that the two sets *X* and *Y* are equivalent means that they are identical except for the names of the elements. That is, we can view *Y* as being obtained from *X* by renaming an element x in *X* with the name of a certain unique element y in *Y*, namely $y = f(x)$. If two sets *X* and *Y* are finite, then they are equivalent if and only if they contain the same number of elements. Indeed, the idea of a finite set *X* is the same as saying that *X* is equivalent to a subset $\{1, 2, \ldots, n\}$ of the natural numbers for some $n \in \mathbb{Z}^+$. To make this and related ideas more precise, we list the following definitions involving a set *X* and $\mathbb{Z}^+$.

2.7.1 Definition. A set *X* is
finite if and only if either $X = \varnothing$ or $\exists n \in \mathbb{Z}^+$ such that *X* is equivalent to $\{1, 2, \ldots, n\}$,
infinite if it is not finite,
denumerable if and only if it is equivalent to $\mathbb{Z}^+$,
countable if and only if it is finite or denumerable, and
uncountable if it is not countable.

Some properties of countable sets are listed in the next example.

2.7.2 Example:

(i) $X \subset \mathbb{Z}^+ \Rightarrow X$ is countable.

(ii) X is countable $\Longleftrightarrow \exists$ a set $Y \subset \mathbb{Z}^+$ such that *X* is equivalent to *Y*.

(iii) If *X* is equivalent to *Y* and *Y* is countable, then *X* is countable.

(iv) If $X \subset Y$ and *Y* is countable, then *X* is countable.

(v) If $f : X \to Y$ is onto and *X* is countable, then *Y* is countable.

(vi) If $f : X \to Y$ is one-to-one and *Y* is countable, then *X* is countable.

(vii) If Λ is countable and $\{X_\lambda : \lambda \in \Lambda\}$ is a collection of countable sets, then $\bigcup_{\lambda\in\Lambda} X_\lambda$ is countable; i.e. the countable union of countable sets is countable.

Assuming Example **2.7.2** as a fact, it is not difficult to show that the set $\mathbb{Z}^2$ and the set $\mathbb{Q}$ of rational numbers are countable sets. First, note that the function $f : \mathbb{Z}^+ \to \mathbb{Z}$ defined by $f(2n) = n$ and $f(2n-1) = -n+1$, where $n = 1, 2, 3, \ldots$ is one-to-one and onto. Thus, $\mathbb{Z}$ is equivalent to $\mathbb{Z}^+$ and hence countable. Now for each $i \in \mathbb{Z}$ define $X_i = \{(i, n) : n \in \mathbb{Z}\}$. Then, since $\mathbb{Z}$ is countable, it follows from **2.7.2** (vii) that $\mathbb{Z}^2 = \bigcup_{i\in\mathbb{Z}} X_i$ is countable.

To show that $\mathbb{Q}$ is countable, define $f : \mathbb{Z}^2 \to \mathbb{Q}$ by

$$f(i,j) = \begin{cases} i/j & if\ j \neq 0 \\ 0 & if\ j = 0\,. \end{cases}$$

Then *f* is obviously onto. Hence by **2.7.2** (v), $\mathbb{Q}$ is countable.

We wish to attach a label *card*(*X*) to each set *X*, called the *cardinality* of *X*, which will provide us with a measure of the "size" of *X*. In particular, the label should distinguish in some way if one or two given sets has more members than the other. Assigning $card(X) = n$ to a set *X* equivalent to $\{1, 2, \ldots, n\}$ will satisfy this requirement for finite sets (using the convention $card(X) = 0$ if and only if $X = \varnothing$), for if *Y* is equivalent to *X*, then it follows from Theorem **2.5.12** that *Y* is also equivalent to $\{1, 2, \ldots, n\}$. Therefore, $card(Y) = n$ and hence, $card(X) = card(Y)$. As can be seen from the finite case, counting is not needed for the purpose of determining whether or not two sets have the same cardinality. We need only to pair off each member of one set with a member of the other set and see if any elements are left over. Thus the notion of two sets having the "same size" can be formalized as follows: two sets *X* and *Y* have the same *cardinality* if and only if they are equivalent.

2.7.3 Examples:

(i) Let $X = \{2n : n \in \mathbb{Z}^+\} \subset \mathbb{Z}^+$ and $f : \mathbb{Z}^+ \to X$ be defined by $f(n) = 2n$. It is easy to verify that *f* is one-to-one and onto. Thus *X* and $\mathbb{Z}^+$ have the same cardinality even though *X* is a proper subset of $\mathbb{Z}^+$. This is not possible for finite sets. No finite set can be equivalent to one of its proper subsets.

(ii) Any open interval $(a, b) = \{x \in \mathbb{R} : a < x < b\} \subset \mathbb{R}$ is equivalent to the open interval $(-1, 1)$, since the function $f : (-1, 1) \to (a, b)$ defined by $f(x) = \frac{1}{2}a(1-x) + \frac{1}{2}b(1+x)$ is one-to-one and onto. Furthermore, by defining $g : (-1, 1) \to \mathbb{R}$ by $g(x) = \frac{x}{1-|x|}$ it can be seen that (–1,1)

is equivalent to $\mathbb{R}$. Thus, $card((-1,1)) = card(\mathbb{R})$. It therefore follows that any open interval (a,b) has "just as many points" as $\mathbb{R}$ itself.

In order to compare the size of two sets, we make the following definition:

2.7.4 Definition. For two sets, *X* and *Y*, we write $card(X) \leq card(Y)$ if there exists a one-to-one function from *X* to *Y*.

Note that we use the symbol "≤" rather than the word "smaller." Obviously, if $X \subset Y$, then $card(X) \leq card(Y)$. However, the existence of a one-to-one function from *X* to *Y* does not exclude the possibility that there exists also a one-to-one and onto function from *X* to *Y*, as Example **2.7.3** (i) and (ii) show. Example **2.7.2** (iv) shows that, roughly speaking, countable infinite sets represent the *smallest* infinity: No uncountable set can be a subset of a countable set. In fact, it can be shown that $card(\mathbb{Z}^+) = card(\mathbb{Q})$ and $card(\mathbb{Z}^+) \neq card(\mathbb{R})$ [27]. Since $\mathbb{Z}^+ \subset \mathbb{R}$, $card(\mathbb{Z}^+) \leq card(\mathbb{R})$. Thus, $\mathbb{R}$ is in a sense much larger than $\mathbb{Z}^+$ or $\mathbb{Q}$, which are of the same infinite size.

2.8 Algebra of Real-Valued Functions

If $f \in \mathbb{R}^X$ (i.e., $f : X \to \mathbb{R}$), then *f* is called a *real valued* function on *X*. Many of the common arithmetic operations of $\mathbb{R}$ are inherited by $\mathbb{R}^X$. We provide a quick review of these inherited operations as they are of fundamental importance in image algebra. Specifically, let $f, g \in \mathbb{R}^X$, $\mathrm{k} \in \mathbb{R}$, and |k| denote the absolute of k. Then we define:

2.8.1
$$
\begin{array}{ll}
(i) & (f+g) : X \to \mathbb{R} \text{ by } (f+g)(x) = f(x) + g(x) \\
(ii) & (\mathrm{k} \cdot f) : X \to \mathbb{R} \text{ by } (\mathrm{k} \cdot f)(x) = \mathrm{k} \cdot (f(x)) \\
(iii) & (|f|) : X \to \mathbb{R} \text{ by } (|f|)(x) = |f(x)| \\
(iv) & (f \cdot g) : X \to \mathbb{R} \text{ by } (f \cdot g)(x) = f(x) \cdot g(x)
\end{array}
$$

Observe that $(f \cdot g) : X \to \mathbb{R}$ is not the composition $f \circ g$ defined previously. For example, if $f : \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = x^2$ and $g : \mathbb{R} \to \mathbb{R}$ by $g(x) = 2x$, then $(f \cdot g)(x) = 2x^3$, while $(f \circ g)(x) = 4x^2$. Note also that $f \cdot g = g \cdot f$, but $f \circ g \neq g \circ f$.

It is also convenient to identify the real number $\mathrm{k} \in \mathbb{R}$ with the constant function $k : X \to \mathbb{R}$ defined by $k(x) = \mathrm{k}\ \forall x \in X$. In particular, if $f : X \to \mathbb{R}$, then $\forall \mathrm{k} \in \mathbb{R}$ we define scalar addition of *f* by k as the function $(f + \mathrm{k}) : X \to \mathbb{R}$, where $(f + \mathrm{k})(x) = f(x) + \mathrm{k}$. Thus, $f + \mathrm{k} = f + k$.

The set $\mathbb{R}^X$, together with the above operations, possesses various important properties, some of which are included in the next theorem.

2.8.2 Theorem. *The set $\mathbb{R}^X$ together with the operations defined in* **2.8.1** *satisfies the following properties:*

1. *The operation of addition of functions f and g satisfies:*

(i) $\quad (f+g)+h = f+(g+h)$

(ii) $f + g = g + f$

(iii) $\exists 0 \in \mathbb{R}^X$, *i.e. a zero function* $0 : X \to \mathbb{R}$, *such that* $f + 0 = f$

(iv) $\forall f \in \mathbb{R}^X \ \exists - f \in \mathbb{R}^X$, *i.e., a function* $-f : X \to \mathbb{R}$, *such that* $f + (-f) = 0$, *the zero function.*

2. *The operation of scalar multiplication* $\mathrm{k} \cdot f$ *of a function f by the real number* k *satisfies:*

(i) $\mathrm{k} \cdot (\mathrm{k}' \cdot f) = (\mathrm{k}\mathrm{k}') \cdot f$

(ii) $1 \cdot f = f$

(iii) $0 \cdot f = 0$

3. *The operations of addition and scalar multiplication satisfy:*

(i) $\mathrm{k} \cdot (f + g) = (\mathrm{k} \cdot f) + (\mathrm{k} \cdot g)$

(ii) $(\mathrm{k} + \mathrm{k}') \cdot f = (\mathrm{k} \cdot f) + (\mathrm{k}' \cdot f)$

Proof: We only prove part **3** of the theorem. The remaining parts are just as simple to prove and are left as an exercise.

$$(i) \quad [\mathrm{k} \cdot (f + g)](x) = \mathrm{k} \cdot [(f + g)(x)] = \mathrm{k} \cdot [f(x) + g(x)]$$
$$= \mathrm{k}(f(x)) + \mathrm{k}(g(x)) = (\mathrm{k} \cdot f)(x) + (\mathrm{k} \cdot g)(x)$$

for all $x \in X$; hence, $\mathrm{k} \cdot (f + g) = (\mathrm{k} \cdot f) + (\mathrm{k} \cdot g)$. Observe that we use the fact that k, $f(x)$ and $g(x)$ are real numbers and satisfy the distributive law.

$$(ii) \quad \big((\mathrm{k} + \mathrm{k}') \cdot f\big)(x) = \big(\mathrm{k} + \mathrm{k}'\big) f(x)$$
$$= \mathrm{k}(f(x)) + \mathrm{k}'(f(x)) = (\mathrm{k} \cdot f)(x) + \big(\mathrm{k}' \cdot f\big)(x)$$

for all $x \in X$; so $(\mathrm{k} + \mathrm{k}') \cdot f = (\mathrm{k} \cdot f) + (\mathrm{k}' \cdot f)$.

Q.E.D.

It follows from properties **1** through **3** of Theorem **2.8.2** that $\mathbb{R}^X$ together with operations defined in **2.8.1** forms a *real linear vector space* as defined in Chapter **3**. If *X* is a finite set consisting of *n* elements, then $\mathbb{R}^X$ may be viewed as the well-known vector space $\mathbb{R}^n = \prod_{i=1}^{n} \mathbb{R}$.

2.8.3 **Example:** Let $X = \{1, 2, \ldots, n\}$ and let $\nu : \mathbb{R}^X \to \mathbb{R}^n$ be defined by $\nu(f) = (f(1), f(2), \ldots, f(n))$. If $\nu(g) = \nu(f)$, then $(g(1), g(2), \ldots, g(n)) = (f(1), f(2), \ldots, f(n))$ and, hence, $g(i) = f(i)$ $\forall i \in X$. Therefore, $g = f$ and ν is one-to-one. Furthermore, if $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, and $f : X \to \mathbb{R}$ is the function defined by $f(i) = x_i$, then $\nu(f) = (x_1, x_2, \ldots, x_n)$ and, therefore, ν is also onto. It follows that any function $f : X \to \mathbb{R}$ can be uniquely identified with the ordered *n*-tuple $(f(1), f(2), \ldots, f(n))$. In addition, if *f* and *g* correspond to the *n*-tuples

$$f = (x_1, x_2, \ldots, x_n) \ \text{ and } \ g = (y_1, y_2, \ldots, y_n),$$

then

$$f + g = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$$

and for any $k \in \mathbb{R}$,

$$k \cdot f = (kx_1, kx_2, \ldots, kx_n)\,.$$

Example **2.8.3** shows that $\mathbb{R}^X$ and $\mathbb{R}^n$ are, from an algebraic point of view, essentially the same vector space. This fundamental fact provides a key component in the study of the image algebra and its relationship to other algebraic structures. The precise definition of algebraic equivalence of two vector spaces is given in the next chapter.

2.9 Distance Functions

A type of real valued function of particular importance in image processing and pattern recognition is the *distance function.* Distance functions induce geometric structures on sets through the notion of nearness of one element to another. The general definition of a distance or metric on a set *X* is as follows.

2.9.1 Definition. Let *X* be a nonempty set. A *distance function* or *metric* on *X* is a function $d : X \times X \to \mathbb{R}$ that satisfies the following three conditions:
(1) $d(x, y) \geq 0 \quad \forall x, y \in X$ and $d(x, y) = 0 \iff x = y\,.$
(2) $d(x, y) = d(y, x) \quad \forall x, y \in X\,.$
(3) $d(x, z) \leq d(x, y) + d(y, z) \quad \forall x, y, z \in X\,.$

When speaking of sets on which metrics are defined, we refer to the elements of the set as *points* and to $d(x, y)$ as the *distance* between the points x and y. Property (1) of the function *d* characterizes it as strictly nonnegative, and (2) as a symmetric function of x and y. Property (3) is known as the *triangle inequality*. Excellent examples of distance functions are three metrics commonly used in image processing. These are the *Euclidean* distance, the *city block* or *diamond* distance, and the *chessboard* distance on $\mathbb{R}^n$. For arbitrary points $\mathbf{x} = (x_1, \ldots, x_n)$, and $\mathbf{y} = (y_1, \ldots, y_n)$ of $\mathbb{R}^n$ we define the *Euclidean* distance between $\mathbf{x}$ and $\mathbf{y}$ by

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{k=1}^{n} (x_k - y_k)^2\right]^{\frac{1}{2}},$$

the *city block* distance by

$$d_1(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{n} |x_k - y_k|\,,$$

and the *chessboard* distance by

$$d_2(\mathbf{x}, \mathbf{y}) = max\{|x_k - y_k| : 1 \leq k \leq n\}\,.$$

Given a set *X*, a distance *d* on *X*, and a number $r > 0$, then for any point $x \in X$, we can define the set $N_{d,r}(x)$ of points $y \in X$ that are within distance *r* of x : $N_{d,r}(x) = \{y \in X : d(x, y) < r\}$. In order to simplify notation we let $N_r(x)$ denote the set $N_{d,r}(x)$ if it is clear from the discussion as to which metric *d* is being used. We may think of $N_{d,r}(x)$ as a "sphere" with center *x* and radius *r*. Geometrically, however, these sets need not look like spheres. Figure **2.9.1** provides examples of spheres in $\mathbb{R}^2$ of radius $r = 1$ about the point **0**=(0,0) determined by the Euclidean, city block, and chessboard distances.

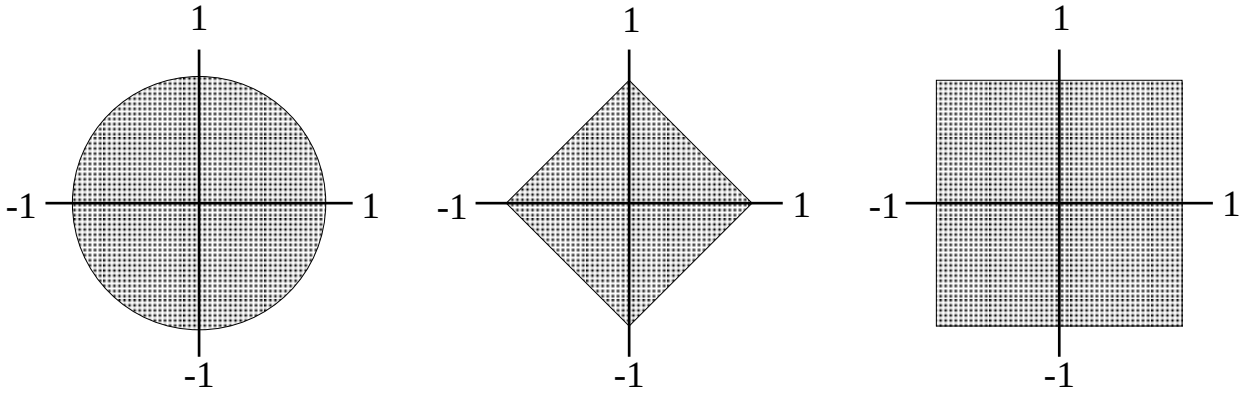


Figure 2.9.1 The spheres $(a)\ N_{d,r}(\mathbf{0}),\ \ (b)\ N_{d_1,r}(\mathbf{0}),\ and\ \ (c)\ N_{d_2,r}(\mathbf{0})$.

2.10 Point Sets in $\mathbb{R}^n$

A large portion of the material introduced in the previous sections dealt with *abstract* sets, that is, sets of arbitrary objects whose nature is immaterial. In this section we briefly review properties of special sets in $\mathbb{R}^n$. A nonempty subset of $\mathbb{R}^n$ will be referred to as a *point set*. Elements of a point set *X*, called the *points* of *X*, will be denoted by small bold letters. In particular, the origin of $\mathbb{R}^n$, which is the *n*-tuple $(0, 0, \ldots, 0)$ consisting only of zeros, will be denoted by **0**.

2.10.1 Definition. Let $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Then the *length* or *norm* of $\mathbf{x}$ is defined as

$$||\mathbf{x}|| = \sqrt{x_1^2 + \cdots + x_n^2}\ .$$

It follows from this definition that the norm of the difference of two points is the same as the Euclidean distance between the points; i.e., if $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, then

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{k=1}^{n} (x_k - y_k)^2\right]^{\frac{1}{2}} = ||\mathbf{x} - \mathbf{y}||\ .$$

It is now easy to establish the essential properties of the norm.

2.10.2 Theorem. *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be points in* $\mathbb{R}^n$. *Then we have*

(i) $\quad ||\mathbf{x}|| \geq 0, \quad and \ ||\mathbf{x}|| = 0 \ if, \ and \ only \ if, \ \mathbf{x} = \mathbf{0}.$

(ii) $\quad ||\mathbf{x} - \mathbf{y}|| = ||\mathbf{y} - \mathbf{x}||.$

(iii) $\quad \left| \sum_{k=1}^{n} x_k y_k \right| \leq ||\mathbf{x}|| \, ||\mathbf{y}||.$

(iv) $\quad ||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$

Proof: Statements (i) and (ii) are immediate from the definition. Statement (iii) is known as the *Cauchy-Schwartz inequality* which can be rewritten as

$$\left(\sum_{k=1}^{n} x_k y_k \right)^2 \leq \left(\sum_{k=1}^{n} x_k^2 \right) \left(\sum_{k=1}^{n} y_k^2 \right).$$

Since a sum of squares can never be negative, we have

$$\sum_{k=1}^{n} (x_k r + y_k)^2 \geq 0$$

for every real number *r*. This inequality can be written in the form

$$Ar^2 + 2Br + C \geq 0,$$

where

$$A = \sum_{k=1}^{n} x_k^2, \quad B = \sum_{k=1}^{n} x_k y_k, \quad C = \sum_{k=1}^{n} y_k^2.$$

If $A > 0$, then let $r = -B/A$ in order to obtain $B^2 - AC \leq 0$, which is the desired inequality. If $A = 0$, the proof becomes trivial.

Statement (iv), known as the *triangle inequality,* follows directly from the Cauchy-Schwartz inequality:

$$\begin{aligned} ||\mathbf{x} + \mathbf{y}||^2 &= \sum_{k=1}^{n} (x_k + y_k)^2 = \sum_{k=1}^{n} \left(x_k^2 + 2x_k y_k + y_k^2 \right) \\ &= ||\mathbf{x}||^2 + ||\mathbf{y}||^2 + 2 \sum_{k=1}^{n} x_k y_k \\ &\leq ||\mathbf{x}||^2 + ||\mathbf{y}||^2 + 2||\mathbf{x}|| \, ||\mathbf{y}|| = (||\mathbf{x}|| + ||\mathbf{y}||)^2. \end{aligned}$$

Thus

$$||\mathbf{x} + \mathbf{y}||^2 \leq (||\mathbf{x}|| + ||\mathbf{y}||)^2.$$

By taking the square root we obtain (iv).

Q.E.D.

With the notion of distance, we can now proceed to define what is meant by a neighborhood of a point in $\mathbb{R}^n$.

2.10.3 Definition. Let $\mathbf{x}_0 \in \mathbb{R}^n$ and $r \in \mathbb{R}^+$. The set

$$N_r(\mathbf{x}_0) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{x}_0\| < r\}$$

is called an *open sphere of radius r and center* $\mathbf{x}_0$. The set

$$\overline{N}_r(\mathbf{x}_0) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{x}_0\| \leq r\}$$

is called the *closed* sphere of radius *r* and center $\mathbf{x}_0$. Any open sphere with center $\mathbf{x}_0$ is called a *neighborhood* of $\mathbf{x}_0$ and is denoted by $N(\mathbf{x}_0)$. An open sphere with its center $\mathbf{x}_0$ removed is called a *deleted neighborhood* of $\mathbf{x}_0$ and is denoted by $N'(\mathbf{x}_0)$.

Observe that on the line $\mathbb{R} = \mathbb{R}^1$, a neighborhood of a point $\mathbf{x}_0$ is a symmetric open interval centered at $\mathbf{x}_0$, while in the plane $\mathbb{R}^2$ it is a disc centered at $\mathbf{x}_0$ with its boundary, which is the set of all points $\mathbf{x}$ satisfying the equation $\|\mathbf{x} - \mathbf{x}_0\| = r$, removed.

With respect to a set $X \subset \mathbb{R}^n$, each point $\mathbf{x}_0 \in \mathbb{R}^n$ has one of three relations, and for each we use a familiar word in a precise way:

1. $\mathbf{x}_0$ is an *interior* point of *X* if there exists a neighborhood $N(\mathbf{x}_0)$ such that $N(\mathbf{x}_0) \subset X$,
2. $\mathbf{x}_0$ is an *exterior* point of *X* if there exists a neighborhood $N(\mathbf{x}_0)$ such that $N(\mathbf{x}_0) \cap X = \varnothing$, and
3. $\mathbf{x}_0$ is a *boundary* point of *X* if $\mathbf{x}_0$ is neither an exterior nor interior point of *X*.

The set of all interior points of *X* is called the *interior* of *X* and is denoted by *intX*. The set of all boundary points of *X* is called the *boundary* of *X* and is denoted by ∂X. Note that a boundary point may or may not belong to *X*: If we let $X = N_r(\mathbf{x}_0)$ and $Y = \overline{N}_r(\mathbf{x}_0)$, then $\partial X = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{x}_0\| = r\}$ and, hence, $\partial X \cap X = \varnothing$ while $\partial Y = \partial X \subset Y$.

Beginning students of multidimensional calculus often confuse the two distinct notions of limit point and boundary point. Limit points of sets in $\mathbb{R}^n$ are defined as follows:

2.10.4 Definition. A point **y** is a *limit point* of *X* if for every neighborhood $N(\mathbf{y})$ of **y**, $N'(\mathbf{y}) \cap X \neq \varnothing$.

It follows from the definition that every interior point of *X* is a limit point of *X*. Thus, limit points need not be boundary points. The next example shows that the converse is also true.

2.10.5 Example: Let $X = \{\mathbf{x} \in \mathbb{R}^2 : 0 < \|\mathbf{x}\| \leq 1\} \cup \{(0,2)\}$. The boundary of *X* consists of three separate pieces: The circumference, where $||\mathbf{x}|| = 1$, and the two points $(0,0)$ and $(0,2)$. The interior of *X* is the set of points **x** with $0 < ||\mathbf{x}|| < 1$, and the set of all limit points of *X* is the set $\{\mathbf{x} \in \mathbb{R}^2 : 0 < \|\mathbf{x}\| \leq 1\} \cup \{(0,0)\} = N_1(\mathbf{0})$. In particular, $(0,2)$ is a boundary point which is not a limit point. A boundary point of *X* which is not a limit point is also called an *isolated point* of *X*. The reason for this definition is that for isolated points one can always find neighborhoods which intersect no other points of *X*.

A subset *X* of $\mathbb{R}^n$ is called an *open set* in $\mathbb{R}^n$ if $X = intX$ and *closed* if every limit point of *X* is a point of *X*. Thus, open sets are the unions of neighborhoods, and closed sets are sets that contain all their

limit points. We wish to stress the difference between having a limit point and containing one. The set $X = \{1/n : n \in \mathbb{Z}^+\} \subset \mathbb{R}$ has a limit point (namely, $x = 0$) but no point of *X* is a limit point of *X*. In every day usage, "open" and "closed" are antonyms; this is not true for the technical meaning. The set *X* just described is neither open nor closed in $\mathbb{R}$. The reader should also be cautioned that an open interval (a, b) in $\mathbb{R}^1$ is no longer an open set when considered as a subset of the plane. In fact, no nonempty subset of $\mathbb{R}^1$ can be open in $\mathbb{R}^2$, because such a set can contain no two-dimensional neighborhoods.

It is important to note that we defined the concept of "open" for sets *in* $\mathbb{R}^n$. This concept can be extended to "open in *X*." Suppose $Y \subset X \subset \mathbb{R}^n$. Then *Y* is said to be *open in X* if there exists an open set *W* in $\mathbb{R}^n$ such that $Y = W \cap X$. Similarly, *Y* is *closed in X* if there is a closed set *W* in $\mathbb{R}^n$ such that $Y = W \cap X$. Thus the set $(0, 1] = \{r \in \mathbb{R} : 0 < r \leq 1\}$ is open in the closed set $[-1, 1] = \{r : -1 \leq r \leq 1\}$ since $(0, 1] = (0, 2) \cap [-1, 1]$ and $(0, 2)$ is open in $\mathbb{R}^1$.

The following list summarizes some important facts about openness and closedness.

2.10.6 Every neighborhood $N(\mathbf{x})$ is an open set and every closed neighborhood $\overline{N}(\mathbf{x})$ is a closed set.

2.10.7 The union of any collection of open sets is open and the intersection of any finite collection of open sets is open. The intersection of an infinite collection of open sets need not be open: $\{0\} = \bigcap_{k=1}^{\infty} \left(-\frac{1}{k}, \frac{1}{k}\right)$ is closed.

2.10.8 The intersection of any collection of closed sets is closed and the union of any finite collection of closed sets is closed. The union of an infinite number of closed sets need not be closed: $(-1, 1) = \bigcup_{k=1}^{\infty} \left[-\frac{1}{k} - 1,\ 1 - \frac{1}{k}\right]$ is open even though $\left[\frac{1}{k} - 1,\ 1 - \frac{1}{k}\right] = \overline{N}_{1/k}(0)$ is closed for each *k*.

2.10.9 A set is open if and only if its complement is closed.

2.10.10 A point **x** belongs to the boundary of *X* if and only if every neighborhood of **x** contains a point that belongs to *X* and a point that does not belong to *X*.

2.10.11 $X \cup \partial X$ is a closed set.

It follows from **2.10.4** that the set $\overline{X} = X \cup \{\mathbf{x} : \mathbf{x} \text{ is a limit point of } X\}$, called the *closure of X*, is a closed set. In particular, *X* is closed if and only if $X = \overline{X}$. The relation between $\overline{X}$ and the closed set in **2.10.11** is given by the equality $\overline{X} = X \cup \partial X$. Also, from the definition of boundary, a point **x** belongs to the boundary of *X* if and only if every neighborhood of **x** contains a point that belongs to *X* and a point that does not belong to *X*. In fact, this observation forms the usual definition of boundary points of regions in digital images. This is in contrast to limit points: Neighborhoods of limit points need not contain points which do not belong to *X*. Another distinguishing feature concerning limit points and boundary points is provided by the next theorem and its corollary.

2.10.12 **Theorem.** *If* **p** *is a limit point of X, then every neighborhood of X contains infinitely many points of X.*

Proof: Suppose there is a neighborhood $N(\mathbf{p})$ which contains only a finite number of points of *X*. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ denote the points of $N(\mathbf{p}) \cap X$, which are distinct from **p** and let

$$r = \underset{1 \le i \le n}{minimum}\{||\mathbf{p} - \mathbf{x}_i|| : \mathbf{x}_i \in N'(\mathbf{p}) \cap X\}.$$

The minimum of a finite set of positive numbers is clearly positive, so that $r > 0$. But then the neighborhood $N_r(\mathbf{p})$ contains no point **x** of *X* such that $\mathbf{x} \neq \mathbf{p}$. Hence, according to Definition **2.10.4**, **p** is not a limit point of *X*. This contradiction establishes the theorem.

Q.E.D.

2.10.13 Corollary. *A finite subset of* $\mathbb{R}^n$ *has no limit points.*

According to the corollary, finite sets are always closed sets. In ordinary usage, the words "finite" and "bounded" are sometimes synonymous. In mathematics they are used to describe quite different aspects of a set. A set *X* is said to be *bounded* if there exists a number *r* such that $||\mathbf{x}|| < r$ for every $\mathbf{x} \in X$. Geometrically, this says that no point of *X* is farther than a distance *r* from the origin; that is, $X \subset N_r(\mathbf{0})$. For example, every neighborhood in $\mathbb{R}^n$ is bounded and infinite. Of course, every finite set is bounded.

If *X* is a bounded set of real numbers then, obviously, there exist numbers *r* and *s* such that $s \le x \le r \; \forall x \in X$. In this case, the numbers *r* and *s* are also called an *upper bound* and *lower bound* for *X*, respectively. Any set of real numbers that has an upper bound is said to be *bounded from above*. Similarly, any set of real numbers that has a lower bound is said to be *bounded from below*. Obviously, a bounded set of real numbers is just a set that is bounded from above and below. One of the most basic properties that characterizes the real numbers is the *least upper bound property*; namely, any nonempty set that is bounded from above has a smallest upper bound, and any nonempty set that is bounded from below has a greatest lower bound. To be more precise, suppose *X* is a bounded set of real numbers. Then $r \in \mathbb{R}$ is the *least upper bound* of *X*, denoted by lub or by *supX*, if and only if *r* is an upper bound of *X* and for any $t \in \mathbb{R}$ with $t < r$, *t* is not an upper bound for *X*. The *greatest lower bound* of *X*, denoted by glb or *infX*, is the number s with the property that s is a lower bound for *X* and if $t \in \mathbb{R}$ with $s < t$, then *t* is not a lower bound. For example, the set $(1, 2] = \{x \in \mathbb{R} : 1 < x \le 2\}$ has lub = 2, and glb = 1. Here $2 \in (1, 2]$ but $1 \notin (1, 2]$. As our next theorem shows, closed bounded sets have the important property that they contain their least upper bound and greatest lower bound.

2.10.14 Theorem. *If* $X \subset \mathbb{R}$ *is closed and bounded with* lub = *r and* glb = *s, then* $r \in X$ *and* $s \in X$.

Proof: Suppose $r \notin X$. For every $t > 0$, $\exists x \in X$ such that $r - t \le x \le r$, for otherwise $r - t$ would be an upper bound less than the lub of *X*. Thus every neighborhood of *r* contains a point *x* of *X* with $x \neq r$, since $r \notin X$. It follows that *r* is a limit point of *X* which is not in *X*. But this contradicts the fact that *X* is closed.

Q.E.D.

The notion of bounded sets can be extended to sets other than subsets of $\mathbb{R}^n$. Suppose *X* is an arbitrary nonempty set with metric *d* and $A \subset X$ with $A \neq \varnothing$. Then we say that *A* is *bounded* if and

only if the set $\{d(x,y) : x, y \in A\}$ is bounded. If *A* is bounded then we define the *diameter* of *A* by

$$d(A) = sup\{d(x,y) : x, y \in A\},$$

and $d(A) = \infty$ if *A* is not bounded.

The greatest lower bound can be used to define the distance between sets. In particular, for $x \in X$ we define the *distance* between *x* and *A* by

$$d(x,A) = inf\{d(x,a) : a \in A\},$$

and the *distance* between two nonempty subsets *A* and *B* of *X* by $d(A,B) = inf\{d(a,B) : a \in A\}$.

It is of interest to note that $d(A,B) \neq 0 \Rightarrow A \cap B = \varnothing$, but that the converse implication need not be true, even if both sets are closed sets in $\mathbb{R}$. As an example, consider the sets $A = \{n : n \in \mathbb{Z}^+\}$ and $B = \left\{n + \frac{1}{n} : n \in \mathbb{Z}^+\right\}$. Here $A \cap B = \varnothing$, while $d(A,B) = 0$.

2.11 Continuity and Compactness in $\mathbb{R}^n$

Closely associated with the notion of bounded sets is the concept of compactness. Compactness is usually defined in terms of open covers. By an *open cover* of a set *X* we mean a collection $\{Y_\lambda\}$ of open sets such that $X \subset \bigcup_\lambda Y_\lambda$.

2.11.1 Definition. A subset *X* of $\mathbb{R}^n$ is *compact* if every open cover contains a finite subcover.

More explicitly, the requirement is that if $\{Y_\lambda\}$ is an open cover of *X*, then there exists finitely many indices $\lambda_1, \ldots, \lambda_k$ such that $X \subset \bigcup_{i=1}^{k} Y_{\lambda_i}$.

It is clear that every finite set is compact. Recall that every finite set is also closed and bounded. The next theorem, known as the Heine-Borel Theorem, shows that this property is shared by all compact sets.

2.11.2 Theorem. (Heine-Borel) *$X \subset \mathbb{R}^n$ is compact if and only if X is closed and bounded.*

We omit the proof of this theorem which can be found in [49]. The notion of compactness is of great importance in connection with continuity. We recall the ϵ and δ definition of continuity.

2.11.3 Definition. Let $X \subset \mathbb{R}^n$ and $f : X \to \mathbb{R}^k$. Then *f* is said to be *continuous at a point* $\mathbf{x}_0 \in X$ if, given an arbitrary number $\epsilon > 0$, then there exists a number $\delta > 0$ (usually depending on ϵ) such that for every $\mathbf{x} \in N_\delta(\mathbf{x}_0) \cap X$ we have that $f(\mathbf{x}) \in N_\epsilon(f(\mathbf{x}_0))$. The function *f* is said to be *continuous on X* if it is continuous at every point of *X*.

Another way of saying that *f* is continuous at $\mathbf{x}_0$ is that given $\epsilon > 0$, then there exists $\delta > 0$ such that for $\mathbf{x} \in X$

$$\|\mathbf{x} - \mathbf{x}_0\| < \delta \Rightarrow \|f(\mathbf{x}) - f(\mathbf{x}_0)\| < \epsilon\,.$$

This is the usual definition of continuity taught at the beginning calculus level.

Continuous functions have a very useful characterization in terms of open sets:

2.11.4 **Theorem.** *Suppose $X \subset \mathbb{R}^n$ and $f : X \to \mathbb{R}^k$. Then f is continuous on X if and only if $f^{-1}(Y)$ is open in X for every open set Y in $\mathbb{R}^k$.*

Proof: Suppose *f* is continuous on *X* and *Y* is open in $\mathbb{R}^k$. We have to show that there exists an open set *W* in $\mathbb{R}^n$ such that $f^{-1}(Y) = W \cap X$. Let $\mathbf{x} \in f^{-1}(Y)$. Then $f(\mathbf{x}) \in Y$ and since *Y* is open, there exists $\epsilon > 0$ such that $N_\epsilon(f(\mathbf{x})) \subset Y$. Since *f* is continuous, there exists a number $\delta > 0$ such that for every $\mathbf{z} \in N_\delta(\mathbf{x}) \cap X$ we have that $f(\mathbf{z}) \in N_\epsilon(f(\mathbf{x})) \subset Y$. By the definition of $f^{-1}(Y)$, we have that $N_\delta(\mathbf{x}) \cap X \subset f^{-1}(Y)$. This shows that for every $\mathbf{x} \in f^{-1}(Y)$, we can find a neighborhood $N(\mathbf{x})$ such that $N(\mathbf{x}) \cap X \subset f^{-1}(Y)$. So let

$$W = \bigcup_{x \in f^{-1}(Y)} N(\mathbf{x}) \,,$$

where each neighborhood satisfies the property $N(\mathbf{x}) \cap X \subset f^{-1}(Y)$. Then $W \cap X \subset f^{-1}(Y)$. But clearly $f^{-1}(Y) \subset W \cap X$. Thus, $f^{-1}(Y) = W \cap X$.

Conversely, suppose $f^{-1}(Y)$ is open in *X* for every open set *Y* in $\mathbb{R}^k$. Fix $\mathbf{x} \in X$ and $\epsilon > 0$. Let $Y = N_\epsilon(f(\mathbf{x}))$. Then by **2.10.6**, *Y* is open. Hence $f^{-1}(Y)$ is open in *X*. Thus, there exists an open set *W* in $\mathbb{R}^n$ such that $f^{-1}(Y) = W \cap X$. Since *W* is open in $\mathbb{R}^n$, there exists $\delta > 0$ such that $N_\delta(\mathbf{x}) \subset W$. Therefore,

$$N_\delta(\mathbf{x}) \cap X \subset W \cap X = f^{-1}(Y) \,.$$

This shows that for every $\mathbf{z} \in N_\delta(\mathbf{x}) \cap X$, $f(\mathbf{z}) \in f\big(f^{-1}(Y)\big) = N_\epsilon(f(\mathbf{x}))$. Hence, *f* is continuous at $\mathbf{x} \in X$ and – since $\mathbf{x}$ was an arbitrarily chosen element of *X* – *f* is continuous on *X*.

Q.E.D.

Continuous functions have the important property of preserving compactness:

2.11.5 **Theorem.** *Suppose $X \subset \mathbb{R}^n$ is compact. If $f : X \to \mathbb{R}^k$ is continuous, then $f(X)$ is compact.*

Proof: Let $\{Y_\lambda\}$ be an open cover of $f(X)$. Since *f* is continuous, Theorem **2.11.4** shows that each of the sets $f^{-1}(Y_\lambda)$ is open. Since *X* is compact, there exists finitely many indices, say $\lambda_1, \ldots, \lambda_m$ such that

$$X \subset \bigcup_{i=1}^{m} f^{-1}(Y_{\lambda_i}) \,.$$

Since $f\big(f^{-1}(Y)\big) \subset Y$ for every $Y \subset \mathbb{R}^k$, (i) implies that

$$f(X) \subset \bigcup_{i=1}^{m} Y_{\lambda_i} \,.$$

This completes the proof.

Q.E.D.

A function $f : X \to \mathbb{R}^k$ is said to be *bounded* if there exists a real number *M* such that $||f(\mathbf{x})|| \leq M$ for all $\mathbf{x} \in X$. We now deduce some consequences of Theorem **2.11.5**.

2.11.6 Theorem. *Suppose $X \subset \mathbb{R}^n$ is compact. If $f : X \to \mathbb{R}^k$ is continuous, then $f(X)$is closed and bounded. In particular, f is bounded.*

This follows from **2.11.5** and the Heine-Borel Theorem. The result is particularly important when *f* is a real valued function.

2.11.7 Theorem. *Suppose $X \subset \mathbb{R}^n$ is compact. If $f : X \to \mathbb{R}$ is continuous and*

$$M = sup\{f(\mathbf{x}) : \mathbf{x} \in X\}, \qquad m = inf\{f(\mathbf{x}) : \mathbf{x} \in X\},$$

then there exist points **p** *and* **q** *in X such that $f(\mathbf{p}) = M$ and $f(\mathbf{q}) = m$.*

Proof: By Theorem **2.11.6** *f*(*X*) is a closed and bounded set of real numbers; hence *f*(*X*) contains its lub *M* and its glb *m* (Theorem **2.10.14**).

Q.E.D.

The conclusion of Theorem **2.11.7** may also be stated as follows: *There exist points* **p** *and* **q** *in X such that $f(\mathbf{q}) \leq f(\mathbf{x}) \leq f(\mathbf{p})$ for all $\mathbf{x} \in X$.* That is, *f* attains its maximum and minimum values at **p** and **q**, respectively.

We conclude this section with another important fact concerning continuous functions on compact sets. First we note that if *X* is a compact subset of $\mathbb{R}^n$, then since *X* is bounded there exists an *n*-dimensional rectangular box $B = \prod_{i=1}^{n} [a_i, b_i]$ such that $X \subset B$. Figure **2.11.1** illustrates this situation if $n = 2$.

Now given a continuous function $f : X \to \mathbb{R}$ we let $F : B \to \mathbb{R}$ be the extension of *f* defined by

$$F(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } \mathbf{x} \in X \\ 0 & \text{if } \mathbf{x} \in B \backslash X \end{cases}$$

The integral of *f* over *X* is then defined as

$$\int_X f(\mathbf{x})\, d\mathbf{x} = \int_B F(\mathbf{x})\, d\mathbf{x}\,.$$

By considering the usual Riemann sums of advanced calculus which approximate the integral $\int_B F(\mathbf{x})\, d\mathbf{x}$, it is easy to see that $\int_X f(\mathbf{x})\, d\mathbf{x}$ does not depend upon the choice of the *n*-dimensional box containing *X*. Furthermore, since *X* is compact and *f* is continuous on *X*, it follows from Theorem **2.11.7** and the definition of *F* that *F* is a bounded function on *B*. It now follows from the theory of integration (see [1] or [43]) that $\int_B F(\mathbf{x})\, d\mathbf{x}$ exists and, hence, so does $\int_X f(\mathbf{x})\, d\mathbf{x}$. The evaluation of the integral of the multivariable function *f* over *X* can be accomplished by evaluating *n* successive single variable integrals. This follows from Fubini's theorem [43], and [50]. We state these observations as a theorem:

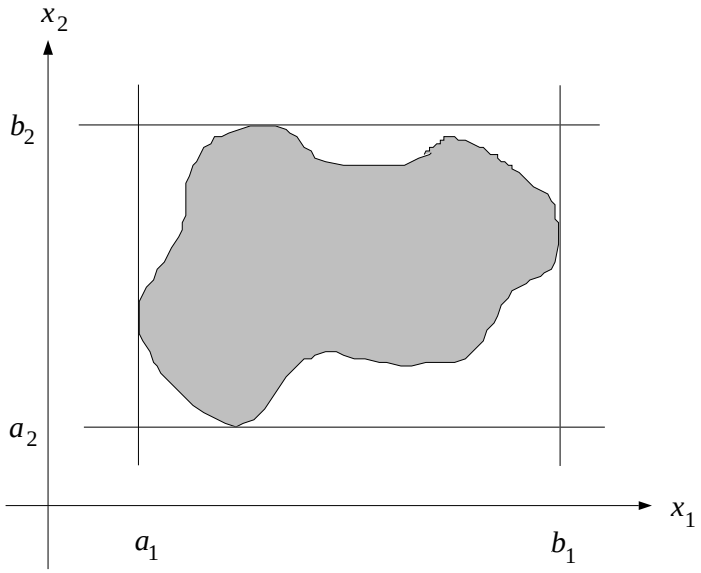


Figure 2.11.1 A compact set in $\mathbb{R}^2$ contained in a rectangle.

2.11.8 Theorem. *If $f : X \to \mathbb{R}$ is continuous and $X \subset \mathbb{R}^n$ is compact, then $\int_X f(\mathbf{x})\, d\mathbf{x}$ exists.*

Furthermore, if $X \subset \prod_{i=1}^{n} [a_i, b_i]$ and $F : B \to \mathbb{R}$ denotes the extension of f defined above, then

$$\int_X f(\mathbf{x})\, d\mathbf{x} = \int_{a_n}^{b_n} \left(\int_{a_{n-1}}^{b_{n-1}} \cdots \left(\int_{a_1}^{b_1} F(\mathbf{x}) dx_1 \right) \cdots dx_{n-1} \right) dx_n ,$$

where $\mathbf{x} = (x_1, \ldots, x_n)$.

Theorems **2.11.7** and **2.11.8** are essential in the definition of image algebra operations on continuous images.

2.12 Topological Spaces

In the previous section we discussed such notions as continuity, compactness, limit point, and boundary point. These notions are all *topological* concepts and a careful look at these concepts reveals that the basic ingredient of all them is the idea of an *open* set. Continuity of functions can be defined purely in terms of inverse images of open sets (Theorem **2.11.4**); closed sets are merely complements of open sets (**2.10.9**); the concept of compactness requires little more than the idea of open sets. However, open sets in $\mathbb{R}^n$ are really just an elementary example of open sets in more general spaces known as *topological spaces*.

2.12.1 Definition. Let X be a set. A set $T \subset 2^X$ is a *topology* on X if and only if T satisfies the following axioms:

O_1 *X* and ∅ are elements of *T*.

O_2 The union of any number of elements of *T* is an element of *T*.

O_3 The intersection of any finite number of elements of *T* is an element of *T*.

A pair (*X*,*T*) consisting of a set *X* and a topology *T* on *X* is called a *topological space*.

Whenever it is not necessary to specify *T* explicitly, we simply let *X* denote the topological space (*X*,*T*). Elements of topological spaces are called *points*. The members of *T* are called *open sets* of the topological space *X*. There is no preconceived idea of what "open" means, other than that sets called open in any discussion satisfy the axioms O_1, O_2, and O_3. Exactly what sets qualify as open sets depends entirely on the topology *T* on *X* — a set open with respect to one topology on *X* may be closed with respect to another topology defined on *X*.

2.12.2 Definition. Let (*X*,*T*) be a topological space and $x \in X$. By a *neighborhood* of *x*, denoted by *N*(*x*), we mean any open set (that is, any member of *T*) containing *x*.

The points of *N*(*x*) are *neighboring* points of *x*, sometimes called *N-close* to *x*. Thus, a topology *T* organizes *X* into chunks of nearby or neighboring points. In this way, topology provides us with a rigorous and general working definition of the concept of nearness: The topology of a space tells us when two points or two objects in the space are close to each other.

2.12.3 Examples:

(i) Let *X* be any set and let $T = \{\varnothing, X\}$. This topology, in which no set other than ∅ and *X* is open, is called the *indiscrete topology* on *X*. There are no "small" neighborhoods.

(ii) Let *X* be any set and let $T = 2^X$. This topology, in which every subset of *X* is an open set, is called the *discrete topology* on *X*, and *X* together with this topology is called a *discrete space*. Comparing this with example (i) above indicates the sense in which different topologies on a set *X* give different organizations of the points of *X*.

(iii) Recall from Section **2.10** that a set in $\mathbb{R}^n$ is an "open" set in $\mathbb{R}^n$ if and only if it is the union of neighborhoods; that is, a set $W \subset \mathbb{R}^n$ is open if and only if for each $\mathbf{x} \in W$ there is some $r > 0$ such that $N_r(\mathbf{x}) \subset W$. It is not difficult to verify that the collection of all sets satisfying this definition of "open in $\mathbb{R}^n$" determines a topology *T* on $\mathbb{R}^n$. Axiom O_1 is trivial. Axiom O_2 is also obvious, for if each member of $\{W_\lambda : \lambda \in \Lambda\}$ is "open in $\mathbb{R}^n$", then so is $\bigcup_{\lambda\in\Lambda} W_\lambda$, since

$$\mathbf{x} \in \bigcup_{\lambda\in\Lambda} W_\lambda \Rightarrow \exists\lambda \in \Lambda \text{ such that } \mathbf{x} \in W_\lambda \Rightarrow \exists r > 0 \text{ s.t. } N_r(\mathbf{x}) \subset W_\lambda \subset \bigcup_{\lambda\in\Lambda} W_\lambda .$$

We leave it to the reader to convince himself that Axiom O_3 holds. The topology *T* defined in this way is called the *Euclidean topology* on $\mathbb{R}^n$, and $\mathbb{R}^n$ together with *T* is called *Euclidean n-space*.

2.13 Basis for a Topology

The introduction of a basis in an abstract mathematical system allows us to reduce the number of objects we deal with in order to describe more easily typical elements of the system. The reader is already familiar with the utility of basis from his study of linear algebra: In an *n*-dimensional vector space the primitive objects we deal with are vectors and any vector can be expressed as the linear sum of a few (at most *n*) basis vectors. Similarly, the primitive objects we deal with in a topological space are open sets. Since the union of open sets is open, it makes sense to ask if there are classes of subsets *B* of a topology *T* such that any element of *T* can be expressed as the union of elements of *B*. Such classes do, in fact, exist and they are called *bases* for the topology *T*.

2.13.1 Definition. Let (*X*,*T*) be a topological space. A class *B* of open subsets of *X*, i.e. $B \subset T$, is a basis for the topology *T* if and only if

(i) every nonempty set $U \in T$ is a union of elements of *B*.

Equivalently, $B \subset T$ is a basis for *T* if and only if

(ii) for any point x belonging to an open set *U*, there exists $V \in B$ such that $x \in V \subset U$.

If *B* is a basis for a topology *T*, then we say that *B generates T*.

2.13.2 Examples:

(i) *T* is a basis for *T*.

(ii) Let *T* be the discrete topology on a set *X*. Then $B = \{\{x\} : x \in X\}$ is a basis for *T*.

(iii) The set $B_0 = \{N_r(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^{\mathbf{n}}, r \in \mathbb{R}^+\}$ is a basis for the Euclidean topology on $\mathbb{R}^n$.

(iv) Define $B_1 = \{N_r(\mathbf{x}) : \mathbf{x} \in \mathbb{Q}^{\mathbf{n}}, r \in \mathbb{Q}^+\}$. Then B_1 is a countable basis for the Euclidean topology on $\mathbb{R}^n$. For if $\mathbf{x} \in U \in T$, where *T* denotes the Euclidean topology, then $\exists\, r > 0$ such that $N_r(\mathbf{x}) \subset U$. Now if $\mathbf{x} \notin \mathbb{Q}^{\mathbf{n}}$ and $r \notin \mathbb{Q}^+$, pick $r' \in \mathbb{Q}^+$ such that $r' < \frac{1}{2}r$ and choose a point $\mathbf{y} \in \mathbb{Q}^{\mathbf{n}}$ such that $||\mathbf{x} - \mathbf{y}|| < r'$. Then $\mathbf{x} \in N_{r'}(\mathbf{y}) \subset N_r(\mathbf{x}) \subset U$. Figure **2.13** illustrates this situation. A slight modification of this argument will verify the case where $r \in \mathbb{Q}^+$, but $\mathbf{x} \notin \mathbb{Q}^{\mathbf{n}}$. The case where $\mathbf{x} \in \mathbb{Q}^{\mathbf{n}}$ and $r \notin \mathbb{Q}^+$ is trivial. Thus B_1 is a basis for *T* with $B_1 \subset B_0$. The fact that B_1 is countable follows from **2.7.2**(vii).

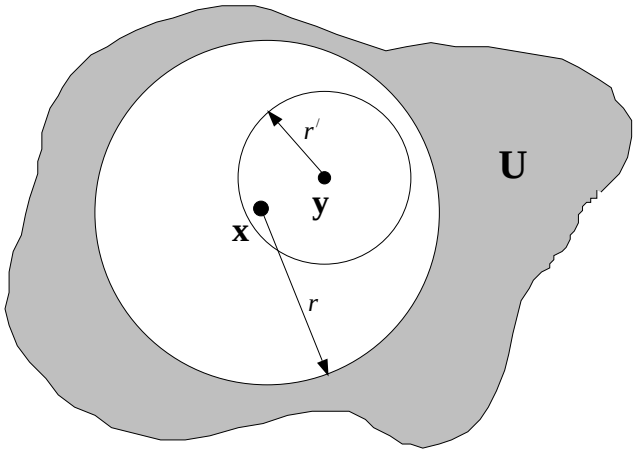


Figure 2.13.1 $\mathbf{x} \in N_{r'}(\mathbf{y}) \subset N_r(\mathbf{x}) \subset U$.

(v) Let *X* be a nonempty set and *d* a metric on *X*. Define the open sphere of radius $r \in \mathbb{R}^+$ about a point $x \in X$ by

$$N_r(x) = \{y \in X : d(x, y) < r\}.$$

Then $B = \{N_r(x) : x \in X, r \in \mathbb{R}^+\}$ is a basis for a topology on *X*. This topology is called the *metric topology* T_d induced by the metric *d*. The topological space (X, T_d) is called a *metric space*, and it is customary to use the simpler notation (*X*,*d*) for the space (X, T_d).

In view of these examples we see that a given topology may have many different bases that will generate it. This is analogous to the concept of a basis for a vector space: Different bases can generate the same vector space. Any linearly independent set of n vectors in $\mathbb{R}^n$ can be used as a basis for the vector space $\mathbb{R}^n$.

We now ask the following question: Given $B \subset 2^X$, when will *B* be a basis for *some* topology on *X*? Clearly, $X = \bigcup_{V \in B} V$ is necessary since *X* is open in every topology on *X*. The next example shows that other conditions are also needed.

2.13.3 Example: Let *X*={1,2,3}. The set *B*={{1,2},{2,3}} cannot be a basis for any topology on *X*. For otherwise the sets {1,2} and {2,3} would themselves be open and therefore their intersection $\{1, 2\} \cap \{2, 3\} = \{2\}$ would also be open; but {2} is not the union of elements of *B*.

The next theorem gives both necessary and sufficient conditions for a class of sets to be a basis for some topology.

2.13.4 Theorem. *Let B be a collection of subsets of X. Then B is a basis for some topology on X if and only if it possesses the following properties:*

(i) $X = \bigcup_{V \in B} V$.

(ii) *If for any* $U, V \in B$, $x \in U \cap V$, *then* $\exists W \in B$, *such that* $x \in W \subset U \cap V$, *or equivalently,* $U \cap V$ *is the union of elements of B.*

Proof: Suppose *B* is a basis for a topology *T* on *X*. Since *X* is open, *X* is the union of elements of *B*. Hence *X* is the union of all elements of *B*. This satisfies property (i). Now, if $U,\ V \in B$, then, in particular, *U* and *V* are open sets. Hence the intersection $U \cap V$ is also open; that is, $U \cap V \in T$. Since *B* is a basis for *T*, $U \cap V = \bigcup_{W \in B} W$. Thus, if $x \in U \cap V = \bigcup_{W \in B} W$, then $x \in W \subset U \cap V$ for some $W \in B$. This satisfies property (ii).

Conversely, suppose that *B* is a collection of subsets of *X* which satisfies properties (i) and (ii). Let $T(B) = \{U : U = \varnothing \text{ or } U \text{ is the union of elements of } B\}$, i.e., $T(B)$ is the collection of all possible subsets of *X* which can be formed from unions of elements of *B*. Then, obviously, $T(B)$ contains both X and $\varnothing$. Therefore, Axiom O_1 holds.

If $\{U_\lambda\}$ is a collection of elements of $T(B)$, then each U_λ is the union of elements of *B*; hence the union $\bigcup_\lambda U_\lambda$ is also the union of elements of *B*. Therefore $\bigcup_\lambda U_\lambda \in T(B)$. This shows that Axiom O_2 is also satisfied.

Lastly, suppose that $U, V \in T(B)$. We need to show that $U \cap V \in T(B)$. By definition of $T(B)$, $U = \bigcup_\lambda U_\lambda$ and $V = \bigcup_\gamma V_\gamma$, where each U_λ and V_γ is an element of *B*. By the distributive law, we have

$$U \cap V = \left(\bigcup_\lambda U_\lambda\right) \bigcap \left(\bigcup_\gamma V_\gamma\right) = \bigcup_{\lambda,\gamma} (U_\lambda \cap V_\gamma)\,.$$

But by (ii), each $U_\lambda \cap V_\gamma$ is the union of elements of *B*. Therefore $\bigcup_{\lambda,\gamma} (U_\lambda \cap V_\gamma)$ is the union of elements of *B* and so belongs to $T(B)$. This verifies Axiom O_3.

Q.E.D.

If *X* is a set and *B* a collection of subsets of *X* satisfying (i) and (ii) of Theorem **2.13.4**, then we say that $T(B)$ is the *topology on X generated by B*. If B_1 and B_2 are two bases for some topologies on *X*, then it is possible that $T(B_1) = T(B_2)$ even though $B_1 \neq B_2$. The two bases defined in Example **2.13.2** (iii) and (iv) illustrate this case. If $T(B_1) = T(B_2)$, then we say that the two bases B_1 and B_2 are *equivalent*. A necessary and sufficient condition that two bases B_1 and B_2 are equivalent is that both of the following two conditions hold:

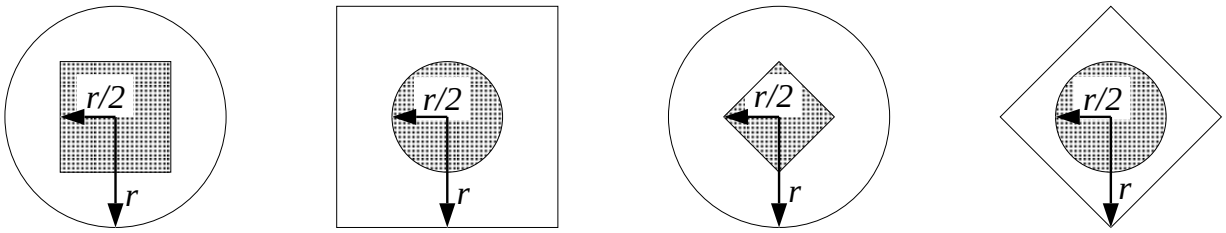
(1) For each $U \in B_1$ and each $x \in U$, there exists $V \in B_2$ such that $x \in V \subset U$.

(2) For each $V \in B_2$ and each $X \in V$, there exists $U \in B_1$ such that $x \in U \subset V$.

2.13.5 Example: Let d_1, d_2, and d_3 denote the Euclidean, city block, and chessboard distance, respectively. Then the following three bases are all equivalent:

$$\begin{aligned} B_{d_1} &= \left\{N_{d_1,r}(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^2,\ r \in \mathbb{R}^+\right\}, \\ B_{d_2} &= \left\{N_{d_2,r}(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^2,\ r \in \mathbb{R}^+\right\}, \text{ and} \\ B_{d_3} &= \left\{N_{d_3,r}(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^2,\ r \in \mathbb{R}^+\right\}. \end{aligned}$$

The equivalence follows from the fact that $N_{d_i,\frac{1}{2}r}(\mathbf{x}) \subset N_{d_j,r}(\mathbf{x})$ for $1 \leq i,j \leq 3$. The four possible inclusions are illustrated in Figure **2.13.2**.



(a) (b) (c) (d)

Figure 2.13.2 Equivalence of three geometrically distinct bases.

The specification of a topology by giving a basis is generally accomplished by specifying for each $x \in X$ a family of neighborhoods $\{N_\lambda(x) : \lambda \in \Lambda(x)\}$, called a *neighborhood basis at x*, and verifying that the family $B = \{N_\lambda(x) : x \in X,\ \lambda \in \Lambda(x)\}$ satisfies the conditions of Theorem **2.13.4**. If the conditions of the theorem are met, then each member $N_\lambda(x)$ is called a *basic neighborhood* of *x*. For example, the set $\{N_r(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^2,\ r \in \mathbb{R}^+\}$ is a neighborhood basis for the Euclidean topology on $\mathbb{R}^2$ and each open disc $N_r(\mathbf{x})$ is a basic neighborhood of **x**. Similarly, if *X* is any set, then $B = \{N(x) : N(x) = \{x\},\ x \in X\}$ is a neighborhood basis for the discrete topology on *X*.

2.14 Point Sets in Topological Spaces

In this section we will emphasize topological concepts in terms of basic neighborhoods in order to retain as much of the geometric flavor of Section **2.10** as possible. We begin by giving some definitions which will have a familiar meaning when specialized to $\mathbb{R}^n$.

2.14.1 Definition. Let *X* be a topological space, $x \in X$, and $Y \subset X$. Then

(i) *x* is an *interior* point of *Y* if there exists a neighborhood $N(x)$ (i.e. an open set containing x) such that $N(x) \subset Y$;

(ii) *x* is an *exterior* point of *Y* if there exists a neighborhood $N(x)$ such that $N(x) \cap Y = \varnothing$;

(iii) *x* is a *boundary point* of *Y* if *x* is neither an exterior nor interior point of *Y*; and

(iv) *x* is a *limit point* of *Y* if for every neighborhood $N(x)$ of *x*, $N'(x) \cap Y \neq \varnothing$, where $N'(x)$ denotes the deleted neighborhood $N(x)\backslash\{x\}$.

The *closure* of *Y*, denoted by $\overline{Y}$, is defined as $\overline{Y} = \{p \in X : p \in Y, \text{ or } p \text{ is a limit point of } Y\}$. The set of all interior points of *Y* is called the *interior* of *Y* and is denoted by *intY*. The set of all boundary points of *Y* is called the *boundary* of *Y* and is denoted by ∂Y. In contrast to Euclidean spaces, an interior point of *Y* need not be a limit point of *Y*. For example, if *X* is a discrete space and $Y = \{x, y\} \subset X$, then $N(x) = \{x\}$ is an open neighborhood of an *x* in *X* with $N(x) \subset Y$ but $Y \cap N'(x) = Y \cap \varnothing = \varnothing$. Thus *x* is an interior point which is not a limit point.

2.14.2 Definition. Let *X* be a topological space and $Y \subset X$. Then *Y* is *closed* in *X* if and only if every limit point of *Y* is a point of *Y*.

Thus a set is closed in a topological space if and only if it contains all its limit points. This is in agreement with Euclidean spaces, where *Y* is closed if and only if $Y = \overline{Y}$. The next theorem characterizes closed sets in terms of open sets.

2.14.3 Theorem. *Let X be a topological space and* $Y \subset X$. *Then Y is closed if and only if* $Y' = X \backslash Y$ *is open.*

Proof: Suppose *Y* is closed. Then *Y* contains all its limit points and so for any $x \notin Y$ or, equivalently, $x \in Y'$, there exists some neighborhood $N(x)$ such that $N(x) \cap Y = \varnothing$ or, equivalently, $N(x) \subset Y'$. Thus, for each point $x \in Y'$ we can find a neighborhood $N(x) \subset Y'$. But then Y' can be written as the union of such neighborhoods, that is, $Y' = \bigcup_{x \in Y'} N(x)$. Since each $N(x)$ is an open set, Y' is open.

For the converse, suppose that Y' is open. Now, if *Y* were *not* closed, then there would have to be at least one limit point $x \in X$ of *Y* with $x \notin Y$. Thus, $x \in Y'$ and, since Y' is open, Y' is a neighborhood of x. But clearly, $Y \cap (Y' \backslash \{x\}) = \varnothing$, which contradicts our assumption that x is a limit point of *Y*.

Q.E.D.

This also proves statement **2.10.9** as it is a special case of this theorem. The next theorem is obvious and we dispense with its proof.

2.14.4 Theorem. *Let X be a topological space and* $Y \subset X$. *Then Y is open if and only if Y=intY.*

Let (X, T) be a topological space and $Y \subset X$. Then *T* induces a topology on *Y*, called the *induced* (or *relative* or *subspace*) *topology* on *Y*. Its importance lies in this: To determine what any concept defined on X becomes when the discussion is restricted to *Y*. We simply regard *Y* as a space with the induced topology and carry over the discussion verbatim.

2.14.5 Definition. Let (X, T) be a topological space and $Y \subset X$. Then the *induced* or *relative* topology T_Y on *Y* is defined as $T_Y = \{Y \cap U : U \in T\}$. The space (Y, T_Y) is called a *subspace* of *X*.

To verify that T_Y is actually a topology on *Y* is trivial. In fact, it is also straightforward to show that if *B* is a basis for *T*, then the set $\{Y \cap U : U \in B\}$ is a basis for T_Y.

2.14.6 Examples:

(i) Let $Y = (-1, 1] \cup \{2\} \subset \mathbb{R}^1$, where $\mathbb{R}^1$ has the Euclidean topology and $(-1, 1] = \{x \in \mathbb{R} : -1 < x \leq 1\}$. The induced topology on *Y* is generated by all sets of form $\{2\}$, all open intervals contained in $(-1, 1]$, and all intervals of form $(r, 1]$ with $-1 \leq r < 1$. These are exactly the sets one can obtain from $Y \cap N_r(x)$, where $r \in \mathbb{R}^+$ and $x \in \mathbb{R}^1$. Note that the proper subset $(-1, 1]$ of the space *Y* is both open and closed in *Y* while it is neither open or closed in the space $\mathbb{R}$.

(ii) The subspace $\mathbb{Z}^2$ of $\mathbb{R}^2$ is a discrete space: If $(m,n) \in \mathbb{Z}^2$, then $\mathbb{Z}^2 \cap N_{\frac{1}{2}}(m,n) = \{(m,n)\}$. Thus, each singleton set {(*m*,*n*)} is open in $\mathbb{Z}^2$.

(iii) Let $S^1 = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\| = 1\} \subset \mathbb{R}^2$. Then the relative topology on S^1 is generated by small open arcs of form $S^1 \cap N_r(\mathbf{x})$ as shown in Figure **2.14.1**. Again note that if $S^1 \cap N_r(\mathbf{x})$ is nonempty, then $S^1 \cap N_r(\mathbf{x})$ is open in S^1 but not in $\mathbb{R}^2$.

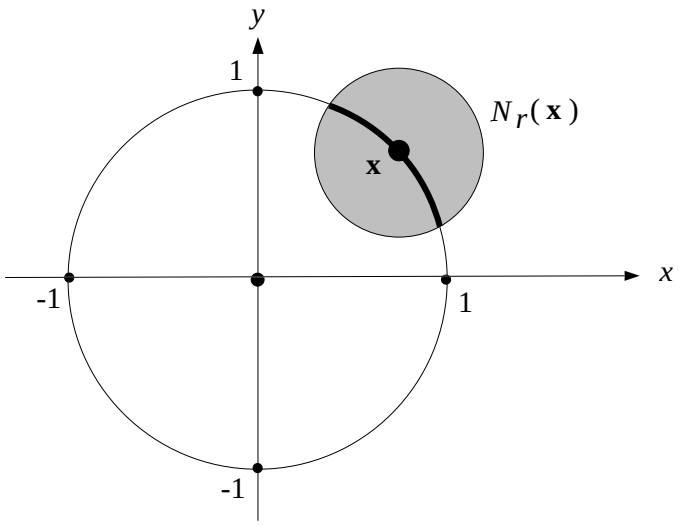


Figure 2.14.1 The intersection of S^1 with an open disk $N_r(\mathbf{x})$.

2.15 Continuity and Compactness in Topological Spaces

The generalization of the notions of compactness and continuity retain the geometric flavor of Section **2.11**.

Given a topological space *X* and $Y \subset X$, then an *open cover* of *Y* is any collection of open sets $\{Y_\lambda\}$ with the property $Y \subset \bigcup_\lambda Y_\lambda$.

2.15.1 Definition. If *X* is a topological space and $Y \subset X$, then *Y* is *compact* if and only if every open cover of *Y* contains a finite subcover. A space *X* is a *compact topological space* if and only if *X* is compact as a subset of *X*.

It follows from this definition that an indiscrete space and any finite space (i.e. the underlying set *X* is finite) is always compact.

2.15.2 Definition. Suppose *X* and *Y* are topological spaces and $f : X \to Y$. Then *f* is said to be *continuous at a point* $x_0 \in X$ if given any neighborhood $N(f(x_0)) \subset Y$ $\exists$ a neighborhood $N(x_0) \subset X$ such that for every $x \in N(x_0)$, $f(x) \in N(f(x_0))$. The function *f* is said to be *continuous on X* if it is continuous at every point of *X*.

The condition $f(x) \in N(f(x_0))$ for each $x \in N(x_0)$ simply means that $f(N(x_0)) \subset N(f(x_0))$. Note the similarity between this definition and **2.11.3**. Also, as before, continuous functions can be characterized in terms of open sets:

2.15.3 Theorem. *Suppose X and Y are topological spaces and $f : X \to Y$. Then f is continuous on X if and only if $f^{-1}(U)$ is open in X for every open set U in Y.*

Proof: Suppose $f^{-1}(U)$ is open in *X* for every open set *U* in *Y*. Let x_0 be an arbitrary point of *X* and let $N(f(x_0))$ be an arbitrary neighborhood of $f(x_0)$. Then, since $N(f(x_0))$ is open in *Y*, $f^{-1}(N(f(x_0)))$ is open in *X*. Obviously, $x \in f^{-1}[N(f(x_0))]$. Thus, letting $N(x_0) = f^{-1}[N(f(x_0))]$, we then have $f(N(x_0)) \subset N(f(x_0))$. Therefore, *f* is continuous at x_0, and since x_0 was arbitrary, *f* is continuous at every point of *X*.

To prove the converse, assume *f* is continuous and *U* is open in *Y*. We must show that $f^{-1}(U)$ is open in *X*. Let $x \in f^{-1}(U)$. Then $f(x) \in U$. Thus *U* is a neighborhood of $f(x)$ and, since *f* is continuous, there exists a neighborhood $N(x)$ such that $f(N(x)) \subset U$. But then $N(x) \subset f^{-1}(N(x)) \subset f^{-1}(U)$. This shows that x is an interior point of $f^{-1}(U)$. Since x was arbitrary, this means that every point of $f^{-1}(U)$ is an interior point. Therefore, $f^{-1}(U) = int\big[f^{-1}(U)\big]$ and, hence, by Theorem **2.14.4**, $f^{-1}(U)$ is open.

Q.E.D.

One of the most important concepts encountered in topology is that of a *homeomorphism*. Homeomorphisms tell us when two objects are *topologically the same*.

2.15.4 Definition. Suppose *X* and *Y* are topological spaces. A *homeomorphism* from *X* to *Y* is a continuous one-to-one and onto function $f : X \to Y$ such that $f^{-1} : Y \to X$ is continuous. If $f : X \to Y$ is a homeomorphism, then we say that *X* and *Y* are *homeomorphic*.

2.15.5 Example: Let *X* be the interval $X = (-1, 1)$ and let *B* be the set $B = \{N_r(x) : -1 < x < 1 \; and \; N_r(x) \subset (-1, 1)\}$. Then *B* is a basis for a topology on *X*. Define $f : X \to \mathbb{R}$ by $f(x) = tan\big(\frac{1}{2}\pi x\big)$. Then *f* is continuous, one-to-one, and onto. Furthermore, the inverse function f^{-1} is also continuous. Hence $\mathbb{R}$ and the open interval $(-1, 1)$ are homeomorphic.

A property *P* of sets is called a *topological property* or a *topological invariant* if whenever a topological space *X* has property *P*, then every space homeomorphic to *X* also has property *P*. As seen in the previous example, the real line $\mathbb{R}$ is homeomorphic to the open interval $X = (-1, 1)$. Hence *length* is not a topological property since $(-1, 1)$ has finite length while $\mathbb{R}$ is of infinite length. Similarly, boundedness is not a topological invariant since *X* is bounded but $\mathbb{R}$ is not.

2.16 Connected Sets

Much of topology concerns the investigation of consequences of certain topological properties such

as compactness, *connectedness*, and *Euler characteristic*. In fact, formally, topology is the study of topological invariants. Several of these invariants play an important role in image analysis.

Intuitively, a space or a subset of space is connected if it does not consist of two or more separate pieces. This simple idea is somewhat problematic in the analysis of computer images. Digital images are discrete objects from a signal processing point of view. They are discrete spaces when viewed as subspaces of Euclidean space. Thus, any object in a digital image consists of finitely many disjoint (disconnected) points. Yet the isolation and analysis of "connected" regions in digital images is a typical activity in computer vision. The reason that one is able to talk about connectivity in digital images in a rigorous sense stems from the fact that connectivity is a topological concept. Connectivity of a subset of a digital image depends on the topology defined on the image space.

2.16.1 Definition. A topological space *X* is *connected* if it is not the union of two nonempty disjoint open sets. A subset $Y \subset X$ is *connected* if it is connected as a subspace of *X*. A space or subset of a space is called *disconnected* if and only if it is not connected.

Observe that if *Y* is a subset of a topological space *X* then *Y* is disconnected if there exists open subsets *U* and *V* of *X* such that $(U \cap Y) \cap (V \cap Y) = \varnothing$, with $Y \subset U \cup V$ and $U \cap Y \neq \varnothing \neq V \cap Y$. Note also that $Y = (U \cap Y) \cap (V \cap Y) \iff Y \subset U \cup V$. The two sets $U \cap Y$ and $V \cap Y$ are called a *decomposition* of *Y*.

2.16.2 Examples:

(i) The set $\mathbb{Z}^2 \subset \mathbb{R}^2$ is disconnected. The sets $\left\{(x,y) : \frac{1}{2} > x\right\} \cap \mathbb{Z}^2$ and $\left\{(x,y) : \frac{1}{2} < x\right\} \cap \mathbb{Z}^2$ form a decomposition of $\mathbb{Z}^2$.

(ii) The rationals $\mathbb{Q} \subset \mathbb{R}$ are not connected since the sets $\left\{x \in \mathbb{R} : x > \sqrt{2}\right\} \cap \mathbb{Q}$ and $\left\{x \in \mathbb{R} : x < \sqrt{2}\right\} \cap \mathbb{Q}$ provide a decomposition.

2.16.3 Definition. Two subsets *A* and *B* of a topological space *X* are said to be *separated* if $A \cap \overline{B} = \varnothing = \overline{A} \cap B$.

If *A* and *B* are two nonempty separated sets in a topological space *X*, then $U = X \backslash \overline{A}$ and $V = X \backslash \overline{B}$ are open in *X*. Furthermore, $(A \cup B) \cap V = A$ and $(A \cup B) \cap U = B$ are nonempty sets whose union is $A \cup B$. Thus, we have a decomposition of $A \cup B$. This shows that if *A* and *B* are two nonempty separated sets, then $A \cup B$ is disconnected. The basic relationship between connectedness and separation is given by the next theorem.

2.16.4 Theorem. *A set A is connected if and only if it is not the union of two nonempty separated sets.*

Proof: We show, equivalently, that *A* is disconnected if and only if *A* is the union of two nonempty separated sets. We already know that the union of two nonempty separated sets is disconnected. In order to show the converse, assume that *A* is disconnected. Then *A* is the union of some decomposition $U \cap A$ and $V \cap A$. We claim that $U \cap A$ and $V \cap A$ are separated sets. Since $U \cap A$ and $V \cap A$ are disjoint, we need only show that each set contains no limit point of the other. Let *p* be a limit point of $U \cap A$ and assume that $p \in V \cap A$. Then *V* is an open set

containing *p* and so *V* contains a point of $U \cap A$ distinct from *p*; i.e. $(U \cap A) \cap V \neq \varnothing$. But by idempotency and associativity of set intersection we have the contradiction:

$$(U \cap A) \cap V = (U \cap A) \cap (V \cap A) = \varnothing\,.$$

Thus $p \notin V \cap A$. Similarly, if *p* is a limit point of $V \cap A$, then $p \notin U \cap A$. Therefore $U \cap A$ and $V \cap A$ are separated sets.

Q.E.D.

Theorem **2.16.4** can be used to show that the only connected subsets of $\mathbb{R}$ containing more than one point are $\mathbb{R}$ and the intervals (open, closed, or half-open). On the other hand, there exists a myriad of different connected sets in $\mathbb{R}^2$. For example, consider the set $Y = A \cup B$ shown in Figure **2.16.1**, where

$$A = \{(0, y) : -1 \leq y \leq 1\}$$
$$B = \{(x, y) : y = sin(1/x),\ 0 < x \leq 1\}\,.$$

Each point of *A* is a limit point of *B*; hence *A* and *B* are not separated sets.

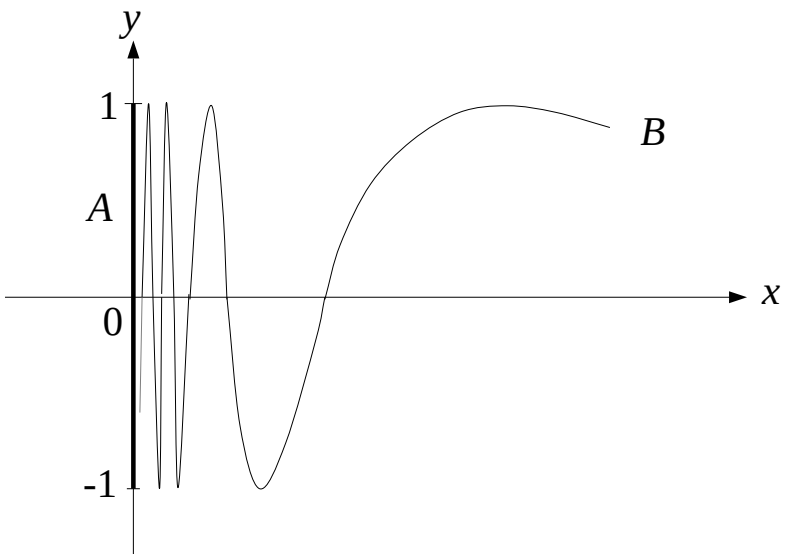


Figure 2.16.1 The connected set $Y = A \cup B$.

A useful notion of connectivity in digital image processing is that of *weak connectivity*.

2.16.5 Definition. A subset *A* of a topological space is *weakly connected* if and only if for each pair of open sets *U* and *V* satisfying

$$(1)\quad A \subset U \cup V \quad and \quad (2)\quad A \cap U \neq \varnothing \neq A \cap V\,,$$

it follows that $U \cap V \neq \varnothing$.

Observe that condition (1) implies that $A = (A \cap U) \cup (A \cap V)$. Thus, in a sense, weakly connected sets are *almost* connected since a set is weakly connected if and only if it is not the subset of a union of two nonempty disjoint open sets each intersecting it. This is very much like Theorem **2.16.4** with the open sets replacing the separated sets. In fact, in Euclidean spaces, the notions of connected and weakly connected are equivalent.

2.16.6 Theorem. *Suppose $X \subset \mathbb{R}^n$. Then X is connected if and only if X is weakly connected.*

Proof: Suppose *X* is connected and *U* and *V* are two open subsets satisfying conditions (1) and (2) of Definition **2.16.5**. Now if $U \cap V = \varnothing$, then since $(X \cap U) \cap (X \cap V) \subset U \cap V$, we have that $(X \cap U) \cap (X \cap V) = \varnothing$. Thus $X \cap U$ and $X \cap V$ is a decomposition of *X*. But this contradicts the hypothesis that *X* is connected. Therefore $U \cap V \neq \varnothing$.

Now suppose that *X* is weakly connected. If *X* is not connected, then according to Theorem **2.16.4**, *X* is the union of two separated sets *A* and *B*. Let

$$U = \{\mathbf{x} \in \mathbb{R}^n \,:\, d(\mathbf{x}, A) < d(\mathbf{x}, B)\} \quad \text{and} \quad V = \{\mathbf{x} \in \mathbb{R}^n \,:\, d(\mathbf{x}, B) < d(\mathbf{x}, A)\}\,.$$

Suppose $\mathbf{p} \in A$. Since *A* and *B* are separated sets, **p** is not a limit point of *B*. Thus there exists a number $\epsilon > 0$ such that $N_\epsilon(\mathbf{p}) \cap B = \varnothing$. Therefore $d(\mathbf{p}, B) \geq \epsilon > 0$ and $\mathbf{p} \in U$. Hence $A \subset U$. Similarly, $B \subset V$ and thus neither *U* nor *V* are empty. The two sets *U* and *V* are also disjoint because

$$d(\mathbf{x}, A) > d(\mathbf{x}, B) \quad \text{and} \quad d(\mathbf{x}, A) < d(\mathbf{x}, B)$$

cannot hold simultaneously. Let $\mathbf{x} \in U$, $\delta = d(\mathbf{x}, B) - d(\mathbf{x}, A)$, and $\mathbf{y} \in \mathbf{N}_{\delta/2}(\mathbf{x})$. Then by the triangle inequality

$$(1) \quad d(\mathbf{y}, A) < d(\mathbf{x}, A) + \delta/2\,.$$

Also,

$$d(\mathbf{y}, B) + \delta/2 > d(\mathbf{x}, B) \quad \text{or} \quad d(\mathbf{y}, B) + \delta/2 > d(\mathbf{x}, A) + \delta \quad \text{or}$$
$$(2) \quad d(\mathbf{x}, A) + \delta/2 > d(\mathbf{y}, B)\,.$$

But (1) and (2) imply that $d(\mathbf{y}, A) < d(\mathbf{y}, B)$. Therefore $\mathbf{y} \in U$ and, since **y** was arbitrary, $N_{\delta/2}(\mathbf{x}) \subset U$. It follows that **x** is an interior point of *U* and, since **x** was arbitrarily chosen, every point of *U* is an interior point of *U*. Thus, according to Theorem **2.14.4**, *U* is open. The proof that *V* is open is identical. This shows that there exist two disjoint open sets *U* and *V* with

$$X \subset U \cup V, \quad \text{and} \quad X \cap U = A \neq \varnothing \neq B = X \cap V\,.$$

But this means that *X* is not weakly connected which contradicts our hypothesis. Hence our assumption that *X* is not connected is false.

Q.E.D.

The first part of the proof actually shows that connectivity implies weak connectivity in any topological space; no use of special properties of $\mathbb{R}^n$ was made. The converse, however, does not hold in general topological spaces.

As mentioned earlier, connectedness is a topological property. In fact, even more is true, namely:

2.16.7 Theorem. *If X is connected and $f : X \to Y$ is continuous, then $f(X)$ is connected.*

Proof: We proceed contrapositively by proving that if $f(X)$ is not connected, then X is not connected. If $f(X)$ is not connected, then there exists a decomposition $f(X) = (f(X) \cap U) \cup (f(X) \cap V)$, where U and V are disjoint nonempty open sets in Y. But then, since f is continuous, $f^{-1}(U)$ and $f^{-1}(V)$ are open sets in X. Also, since U and V are disjoint, $f^{-1}(U)$ and $f^{-1}(V)$ are also disjoint. We now have

$$X = f^{-1}(f(X)) = \left[f^{-1}(f(X)) \cap f^{-1}(U)\right] \cup \left[f^{-1}(f(X)) \cap f^{-1}(V)\right],$$

and, therefore, the decomposition

$$X = \left(X \cap f^{-1}(U)\right) \cup \left(X \cap f^{-1}(V)\right).$$

Q.E.D.

As an application of Theorem **2.16.7** we obtain the following generalization of the intermediate value theorem of standard calculus:

2.16.8 Theorem. *Every continuous real valued function on a connected space X takes on all values between any two it assumes.*

Proof: Since $f : X \to \mathbb{R}$ is continuous, $f(X) \subset \mathbb{R}$ is connected according to Theorem **2.16.7**. By our observation following Theorem **2.16.4**, $f(X)$ is either a point, an interval, or equal to $\mathbb{R}$. If $f(X)$ is a point, then there is nothing to prove. But if $f(x) = a$ and $f(y) = b$, with say b greater than a, then we have $[a, b] \subset f(X)$. Now if c is any number with $a \leq c \leq b$ then for any $z \in f^{-1}(\{c\})$ we have that $f(z) = c$.

Q.E.D.

A *component* of topological space X is a maximal connected subset of X; that is, if C is a component of X, then C is connected and C is not a proper subset of any connected subset of X. Thus, if X is connected, then X has exactly one component, namely X itself. On the other hand, in a discrete space, every point is a component. If x is a point in a topological space, then the largest connected subset of X containing x is called the *component* of x and is denoted by C_x. It is intuitively clear and not difficult to prove that each point $x \in X$ belongs to a unique component C_x.

Components are closed sets. This follows from the fact that if Y is a set in a topological space and p is a limit point of Y, then the sets Y and $\{p\}$ are not separated sets. In particular, if Y is connected, then so is $\overline{Y}$. Thus, if C is a component, then $C = \overline{C}$ since C is a maximal connected subset. It also follows that if A and B are two distinct components, then $A \cap B = \varnothing$. If this were not the case, then A and B are not separated sets and, thus, $A \cap B$ is a connected set containing A (and B). But this contradicts the maximality of A. We summarize these comments as a theorem:

2.16.9 Theorem. *Every connected subset of a topological space X is contained in some component of X and the components form a partition of X.*

According to this theorem, a topological space can be decomposed uniquely into connected pieces, namely its components, and the number of components provides a rough indication of how disconnected a space is. A space *X* is called *totally disconnected* if the only components are points; i.e. if $C_x = \{x\}\ \forall x \in X$. Obviously, $\mathbb{Z}^n \subset \mathbb{R}^n$ is totally disconnected when viewed as a subspace, and so is any discrete space. On the other hand, the subspace $\mathbb{Q} \subset \mathbb{R}$ of rational numbers is a space that is not discrete but is totally disconnected.

2.17 Path-Connected Sets

For most purposes of analysis, the natural notion of connectedness is that two points can be joined by a path.

2.17.1 Definition. Let *X* be a topological space and $f : [0,1] \to X$ continuous. Then the image $Y = f([0,1])$ is called a *path* in *X*. The points $f(0)$ and $f(1)$ are called the *initial* and *terminal points* of the path *Y*, respectively, and *Y* is a path *from* $f(0)$ *to* $f(1)$. The initial and terminal points are also called the *end points* of the path. A point $y \in Y$ is a *multiple point* if the set $f^{-1}(y)$ contains more than one point. The image *Y* is an *arc* (or *simple curve* or *simple path*) if it contains no multiple points.

It follows from the definition that if *Y* is an arc, then $f : [0,1] \to Y$ is a homeomorphism. Also, if *Y* is a path from $f(0)$ to $f(1)$, then it is clear that the function

$$g : r \to f(1-r),\ r \in [0,1],$$

defines a path from $f(1)$ to $f(0)$.

2.17.2 Definition. A topological space is *path-connected* if for each pair of points *p* and *q* in the space there exists a path from *p* to *q*. A subset of a space is path-connected if and only if it is path-connected as a subspace.

2.17.3 Examples:

(i) $\mathbb{R}^n$ is path-connected and if $n > 1$, then for every countable set $X \subset \mathbb{R}^n$, $\mathbb{R}^n \backslash X$ is also path-connected.

(ii) If $Y \subset \mathbb{R}^n$ is an arc and $n > 1$, then $\mathbb{R}^n \backslash Y$ is path-connected.

(iii) A discrete space having more than one point is never path-connected. Every indiscrete space is path-connected.

A trivial but useful reformulation of **2.17.2** is provided by the next theorem.

2.17.4 Theorem. *Let X be a topological space and $p \in X$. Then X is path-connected if and only if for every $x \in X$, there is a path from p to x.*

Proof: If *X* is path-connected, then the condition holds automatically. Conversely, assume that the condition is satisfied and that $x, y \in X$. Let $f : [0,1] \to X$ define a path from *x* to *p* and $g : [0,1] \to X$ a path from *p* to *y*. Let $h : [0,1] \to X$ be defined by

$$h(r) = \begin{cases} f(2r) & if \ \ 0 \le r \le \frac{1}{2} \\ g(2r-1) & if \ \ \frac{1}{2} \le r \le 1 \end{cases} .$$

Then *h* is continuous since *h* is the continuous function *f* on the interval $\left[0, \frac{1}{2}\right]$, the continuous function *g* on the interval $\left[\frac{1}{2}, 1\right]$, and at $r = \frac{1}{2}$, $h(r) = f(1) = g(0)$. Thus, $h([0,1])$ is a path from *x* to *y*.

Q.E.D.

The next theorem establishes the general relation of connectedness and path-connectedness.

2.17.5 Theorem. *Each path-connected space X is connected.*

Proof: If *X* is not connected, then *X* is the union of two nonempty disjoint open sets *U* and *V*. Now let $u \in U$ and $v \in V$, and let $Y = f([0,1])$ be a path from *u* to *v*. Then $(U \cap Y) \cup (V \cap Y)$ is a decomposition of *Y*. But this contradicts Theorem **2.16.7** according to which $Y = f([0,1])$ is connected.

Q.E.D.

2.17.6 Example: A connected space need not be path-connected. Consider the example $X = A \cup B$, where $A = \{(0, y) : -1 \le y \le 1\}$ and $B = \{(x, y) : y = sin(1/x), 0 < x \le 1\}$ (Figure **2.16.1**). The space *X* is connected but not path connected: there is no path from $(0, 0)$ to $(1/\pi, 0)$. However, it is obvious that each of the sets *A* and *B* are path-connected. Also, as mentioned earlier, $\bar{B} = X$. Hence the closure of a path-connected set need not be path-connected.

In view of Theorem **2.17.4**, the union of any family of path-connected spaces having a point in common is again path-connected. Because of the property of unions, we can define a *path component* of a space as a maximal path-connected subset of the space. As before, the path components partition the space; indeed, from **2.17.5**, the path components partition the components. However, in contrast to components, path components need not be closed subsets of the space: in Example **2.17.6**, *B* is a path component of $\overline{B}$.

2.17.7 Theorem. *The following properties of a space X are equivalent:*

(1) *Each path component is open (and therefore also closed).*

(2) *Each point of X has a path−connected neighborhood.*

Proof: If each path component is open, then given $x \in X$, the path component containing x is a path-connected neighborhood of x. Thus, $(1) \Rightarrow (2)$.

To show that $(2) \Rightarrow (1)$, let *P* be any path component and $x \in P$. By hypothesis, *x* has a path-connected neighborhood *U*. However, *P* is a maximal path-connected set containing *x* and,

therefore, $U \subset P$. Thus every point of *p* is an interior point and therefore *U* is open (**2.14.4**). Noting that $P' = X \backslash P$ is the union of the remaining (open) path components, we have that P' is also open. Hence, by **2.14.3**, *P* is closed.

Q.E.D.

This theorem provides a tool for determining when path-connectedness and connectedness are equivalent.

2.17.8 Theorem. *A space X is path-connected if and only if it is connected and each $x \in X$ has a path-connected neighborhood.*

Proof: Since path-connectedness implies connectedness and *X* is a path-connected neighborhood of every point, only the converse requires proof. For this, we know from **2.17.7** that each path component is both open and closed in *X*; since *X* is connected, this path component must therefore be *X*.

Q.E.D.

This theorem has the following important consequence:

2.17.9 Corollary. *An open set in $\mathbb{R}^n$ is connected if and only if it is path-connected.*

Proof: Again, since path-connectedness implies connectedness, we only need to prove the converse. If $U \subset \mathbb{R}^n$ is open, then each point *x* of the space (U, T_U) has a neighborhood $N_\epsilon(x) \subset U$. But $N_\epsilon(x)$ is path-connected. Hence it follows from Theorem **2.17.8** that *U* is path-connected.

Q.E.D.

Of course, as Example **2.17.6** shows, non-open connected subsets of $\mathbb{R}^n$ need not be path-connected.

Simple paths have a particular useful and unique property that can be expressed in terms of *cut points*. A point *x* of a topological space *X* is a *cut point* of *X* provided that $X \backslash \{x\} = A \cup B$, where *A* and *B* are nonempty separated sets; otherwise *x* is a *non-cut point* of *X*.

2.17.10 Examples:

(i) Every point $r \in [0,1]$ with $0 < r < 1$ is a cut point of [0,1], and 0 and 1 are the only non-cut points of [0,1].

(ii) Every point of $\mathbb{R}^n$, where $n > 1$, is a non-cut point of $\mathbb{R}^n$.

In view of Example **2.17.10**(i) and the fact that arcs are homeomorphic images of the unit interval, the next theorem becomes intuitively obvious. However, its proof, which is given in [26], is far from trivial and is beyond the scope of this book.

2.17.11 Theorem. *If $X \subset \mathbb{R}^n$ is compact and connected with just two non-cut points, then X is an arc.*

A path $Y = f([0,1])$ in a space *X* is a *closed path* if $f(0) = f(1)$. A *simple closed path*, also called a *simple closed curve* or *Jordan curve*, is a closed path with exactly one multiple point $y \in f([0,1])$ such that $f^{-1}(y) = \{0,1\}$. An equivalent and more common way of defining a simple closed curve is as the homeomorphic image of the unit circle $S^1 = \{(x,y) : x^2 + y^2 = 1\}$. It is clear that the omission of any two distinct points from S^1 separates S^1 into two *open* arcs (objects homeomorphic to the interval (0,1)). It turns out that this property characterizes simple closed curves in $\mathbb{R}^n$.

2.17.12 Theorem. *If $X \subset \mathbb{R}^n$ is compact and connected and has the property that for any two points $\mathbf{x}, \mathbf{y} \in X$, $X \backslash \{\mathbf{x}, \mathbf{y}\}$ is not connected, then X is a simple closed curve.*

As in the case of Theorem **2.17.11**, the statement of this theorem (as well as that of the next theorem) is intuitively obvious, but its proof is nontrivial [26].

2.17.13 The Jordan Curve Theorem. *If $X \subset \mathbb{R}^2$ is a simple closed curve, then $\mathbb{R}^2 \backslash X$ has two components.*

There are various ways of proving the Jordan curve theorem; a geometric proof is given in [41] while [55] provides an algebraic version. According to this theorem, every simple closed curve in the plane $\mathbb{R}^2$ separates $\mathbb{R}^2$ into two components, each of which must be necessarily path-connected (**2.17.9**). In the next section we show that Theorems **2.17.11**, **2.17.12**, and **2.17.13** all have analogues in the discrete domain.

2.18 Digital Images

We now take a closer look at the set $\mathbb{Z}^n$. For $n = 1, 2$, and 3, this set plays an important role in discrete signal and image processing. Viewed as a subspace of $\mathbb{R}^n$, $\mathbb{Z}^n$ is a discrete space. The components are points; thus there are no interesting connected subsets of $\mathbb{Z}^n$ and no point is a limit point of any given subset of $\mathbb{Z}^n$. However, as mentioned previously, the isolation and analysis of connected regions in $\mathbb{Z}^n$ is a common activity in image analysis. Obviously, when talking about connected regions, there must be topologies on the set $\mathbb{Z}^n$ which provide for the connectivity of sets that contain more than one point. Before discussing different topologies on $\mathbb{Z}^n$, we provide an alternate geometric representation of $\mathbb{Z}^n$ and discuss the role of $\mathbb{Z}^n$ in image representation.

With each $\mathbf{p} = (p_1, p_2, \ldots, p_n) \in \mathbb{Z}^n$ we associate an *n*-dimensional cell, $c(\mathbf{p})$, with center $\mathbf{p}$ defined by

$$c(\mathbf{p}) = \left\{ \mathbf{x} : \mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n,\ |p_i - x_i| \leq \frac{1}{2} \text{ and } 1 \leq i \leq n \right\}.$$

The set of all *n*-dimensional cells is denoted by C^n; that is, $C^n = \{c(\mathbf{p}) : \mathbf{p} \in \mathbb{Z}^n\}$. Set theoretically, the sets $\mathbb{Z}^n$ and C^n are the same: The function $c : \mathbb{Z}^n \to C^n$ defined by $c : \mathbf{p} \to c(\mathbf{p})$ is one-to-one and onto. Figure **2.18.1** illustrates this relationship for $n = 2$. The set C^n is also referred to as the *dual representation* of $\mathbb{Z}^n$.

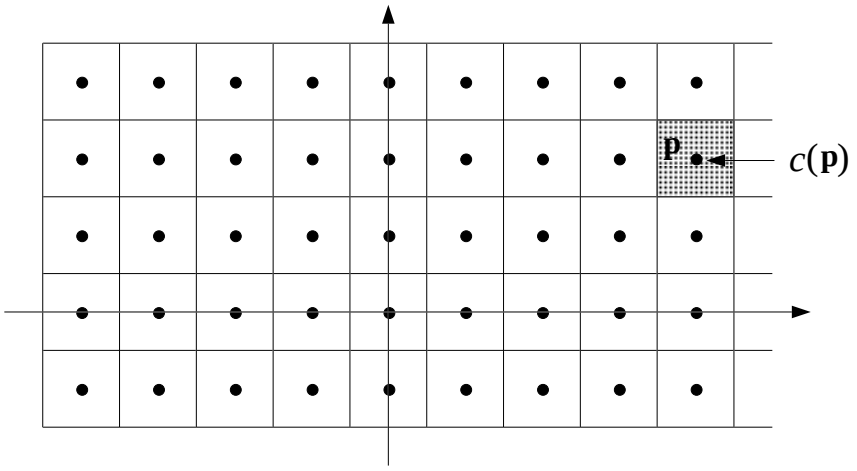


Figure 2.18.1 The set $\mathbb{Z}^2$ and its dual C^2.

One reason for using the representation C^n of $\mathbb{Z}^n$ has to do with sampling of continuous images and the display of sampled images. In order to process a picture (continuous image) or any other signal by computer, we must convert it into a finite set of numbers. Sampling is the selection of a set of discrete points from a compact time and/or spatial domain. Only the values of the signal at those points will usually be used in further processing. In the one-dimensional case the fundamental mathematical result is Shannon's sampling theorem [42]. It shows that any continuous signal over any duration *T* but band-limited in frequency to ω cycles per second can be completely specified (i.e. reconstructed) if we sample its amplitude at intervals less than $1/2\omega$ seconds. It follows that all we need to do is to sample the signal at *k* greater than $2\omega T$ equidistant points during the duration *T* in order to encompass the total signal.

Shannon's theorem does not suggest a way for reconstructing the continuous signal from its discrete samples; it only says that it is possible. In fact, it is necessary to use fairly sophisticated techniques to reconstruct a signal when it is sampled at the minimum frequency. In addition, the choice of algorithms for reconstruction is usually severely limited in image processing. Pictorial data must usually be sampled at a much higher rate, about 160 times as often, than what one might expect from the trivial extension of Shannon's theorem to two dimensions. To illustrate the problem, consider the high resolution image (512×512) shown in Figure **2.18.2**(a), which to the human eye is indistinguishable from a continuous image such as a photograph. The two images in Figures **2.18.2**(b) and **2.18.2**(c) have been obtained from the former by skipping samples, so that Figure **2.18.2**(b) consists of 64×64 samples and Figure **2.18.2**(c) of 32×32. They are displayed on a larger grid by repeating the values of each sample 8 and 16 times. The quality of their appearance is not due to undersampling alone. Their quality improves when squinting one's eyes or looking at them from a distance; this is due to the fact that most of the information is still there. It is our method of piecewise constant reconstruction, i.e. the simple replication of values, that introduces the distortions. Even a simple linear interpolation between samples, instead of replication, would have greatly improved the quality of the images.

Figure 2.18.2 Effects of reducing sampling grid size.

In the case of two-dimensional images, such as shown in Figure **2.18.2**, the image is usually viewed as being derived from a continuous image function $f : \mathbb{R}^2 \to \mathbb{R}^+ \cup \{0\}$ by taking a finite number of samples. The value $f(\mathbf{x})$ represents the intensity, photographic density, or some desired parameter of the physical image at the point $\mathbf{x} \in \mathbb{R}^2$. In a perfect image sampling system, spatial samples of the ideal image *f* would, in effect, be obtained by multiplying *f* by a spatial sampling function *s* composed of an infinite array of Dirac delta functions arranged in a grid of spacing $\triangle\mathbf{x} = (\triangle x_1, \triangle x_2)$. The sampled image f_s is then given by $f_s(\mathbf{x}) = f(\mathbf{x}) \cdot s(\mathbf{x})$; that is

$$f_s(\mathbf{x}) = f(\mathbf{x}) \cdot \sum_{\mathbf{z}\in\mathbb{Z}^2} \delta(\mathbf{x} - \mathbf{z} \cdot \triangle\mathbf{x}) = \sum_{\mathbf{z}\in\mathbb{Z}^2} f(\mathbf{z} \cdot \triangle\mathbf{x}) \cdot \delta(\mathbf{x} - \mathbf{z} \cdot \triangle\mathbf{x}) \,. \tag{2.18.1}$$

In this equation, vector subtraction and multiplication are defined componentwise; e.g. if $\mathbf{x} = (x_1, x_2)$ and $\mathbf{z} = (z_1, z_2)$, then

$$\mathbf{x} - \mathbf{z} = (x_1 - z_1, x_2 - z_2) \text{ and } \mathbf{z} \cdot \mathbf{x} = (z_1 \cdot x_1, z_2 \cdot x_2) \,.$$

A continuous image function may be obtained from the sampled image f_s by linear interpolation or by linear spatial filtering. If *r* represents the impulse response of an interpolating filter, then a continuous image function f_r is obtained by convolving f_s with *r*; i.e. $f_r = f_s * r$ where $*$ denotes the convolution product. However, substituting f_s from equation 2.18.1 and performing the convolution yields

$$f_r(\mathbf{x}) = \sum_{\mathbf{z}\in\mathbb{Z}^2} f_s(\mathbf{z} \cdot \triangle\mathbf{x}) \cdot r(\mathbf{x} - \mathbf{z} \cdot \triangle\mathbf{x}) \,. \tag{2.18.2}$$

This shows that the impulse response function *r* acts as a two-dimensional interpolation waveform for the sampled image f_s.

Of course, Equation 2.18.1 represents an idealized description of f_s. It is physically impossible to obtain measurements at a point. The evaluation of f_s at a point **x** represents the measured intensity over a small convex area centered at **x**. The dimension of the sampling areas are approximately equal to their spacing (Figure **2.18.3**). Another physical restriction is that *f* can only be sampled at a finite number of places. Thus, the union of all convex sampling areas forms a compact subset $X \subset \mathbb{R}^2$. These comments can be formally expressed and are the rationale for the next definition.

2.18.1 **Definition.** Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is continuous. A *sampling cell* associated with a point $\mathbf{p} \in \mathbb{R}^2$ is a compact and convex subset of $\mathbb{R}^2$ with barycenter $\mathbf{p}$ over which the value of a sample $f_s(\mathbf{p})$ of f is calculated. The union of all the centers of the sampling cells is called the *sampling grid*. The pair $(\mathbf{p}, f_s(\mathbf{p}))$ is called a *picture element* or *pixel* and $f_s(\mathbf{p})$ is called the *pixel value*.

Figure **2.18.3** illustrates the idea behind this definition. In general, the spatially sampled or *spatially quantized* image consists of an $n \times m$ array of equally distributed samples and can therefore be viewed as points in the discrete plane $\mathbb{Z}^2$ arranged in rectangular form. We also need to point out that the sampling cells can be disjoint, although in most sampling devices they overlap as illustrated.

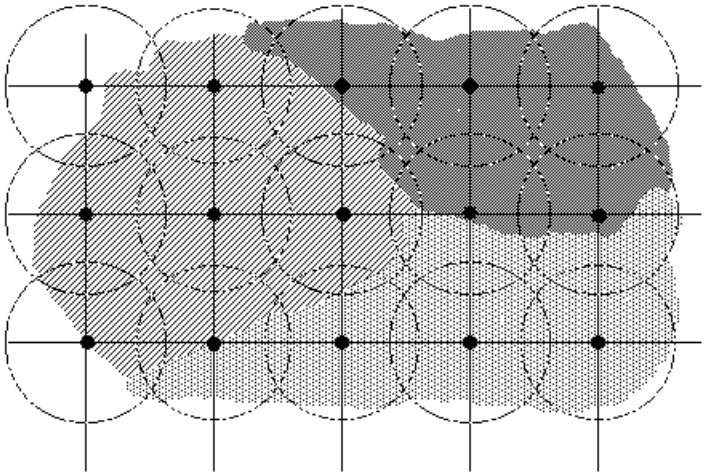
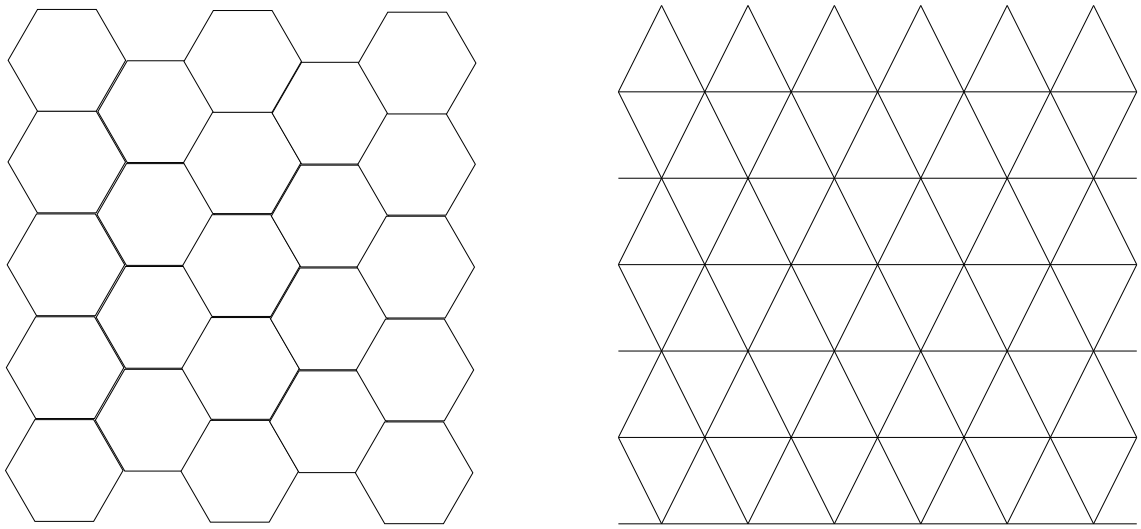


Figure 2.18.3 Sampling grid with sampling cells; the different shadings represent different values of the image function *f*.

The most common grid used in image processing is the rectangular grid. Hexagonal and triangular grids (Figure **2.18.4**) are often discussed in the literature but are rarely implemented.



(a) Hexagonal grid (b) Triangular grid

Figure 2.18.4 (a) The hexagonal grid and (b) the triangular grid.

Although the spatial samples $f_s(\mathbf{p})$ can be represented as points, it is often intuitively more satisfying and closer to the sensing process to use the dual representation of $\mathbb{Z}^2$ and view the samples as cells. In addition, this view corresponds to the actual display of sensed images on a variety of display devices. Television frames, for example, might be quantized into 450 lines of 560 cells each.

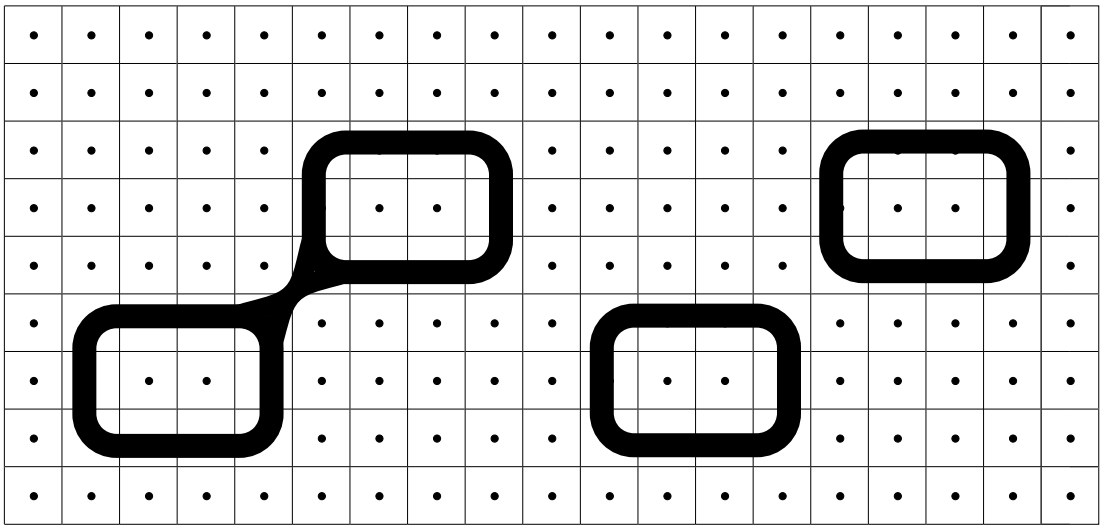
2.18.2 **Definition.** The set $D(f_s) = \{c(\mathbf{p}) : \mathbf{p} \in domain(f_s)\}$ is called the *display grid* of *f* and the function $f_d : D(f_s) \to \mathbb{R}$ defined by $f_d[c(\mathbf{p})] = f_s(\mathbf{p})$ is called the *display image*. The pair $[c(\mathbf{p}), f_d(c(\mathbf{p}))]$ is called a *display pixel* or simply a *pixel*.

In case one desires a larger display or a more dense set of display cells than provided for by $D(f_s)$, we could define the display function in terms of an interpolating function f_r by setting $f_d[c(\mathbf{p})] = f_r(\mathbf{p})$.

In order for the sampled image to be processed by a digital computer, the function *f* must also be sampled in amplitude; i.e., each real number $f(\mathbf{p})$ must be assigned a binary code. This process is called *amplitude quantization* and can be considered as a mapping from the real numbers into either the integers or into $\mathbb{Z}_{2^k}$ as each binary code is of finite length. The corresponding integers are called the *gray levels* or *gray values* of the image. It is common practice in image processing to have the discrete gray levels equally spaced between 0 and some maximum number L of form $L = 2^k - 1$. These comments form the basis of the following definition:

2.18.3 **Definition:** A *digital image* is a spatially-quantized and amplitude-quantized image and the full range of amplitude quantization levels available for a particular image is called a *gray scale*. The process of obtaining a digital image from a continuous image is called *digitization*.

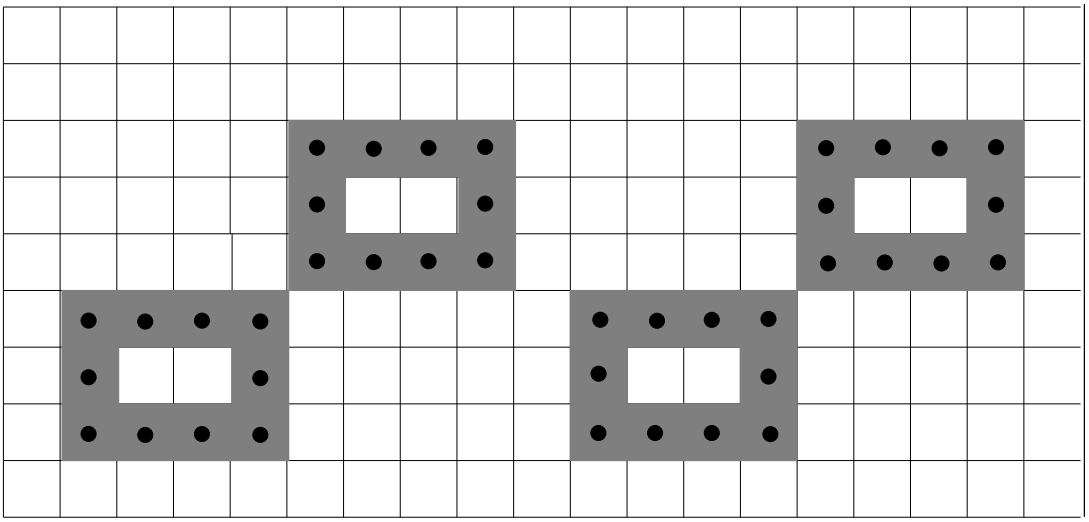
Digital images constitute a special class of *computer* images, a topic discussed in Chapter **4**. Here we are only interested in exploring some useful topologies on digital images. The first problem to consider is that of connectedness. Consider the digitization of the continuous curves shown in Figure **2.18.5** (a) and (b) using the gray scale {0,1}.



(a) (b)

Figure 2.18.5 Continuous curves with sampling grid; the dots indicate the centers of the sampling cells.

The continuous curve in **2.18.5**(a) is connected, while the object in **2.18.5**(b) consists of two separate continuous curves that are spatially close to each other. Due to the limitations of the spatial sampling grid, the resulting digital images obtained from (a) and (b) are identical (Figure **2.18.6**). The digital representation in Figure **2.18.6** of either set of curves as subsets of $\mathbb{Z}^2$ appears totally disconnected while the dual representation appears connected and seems a good representative of either **2.18.5**(a) of **2.18.5**(b). Thus, the question arises as to which curve **2.18.6**(b) represents; i.e. does it represent a continuous figure eight curve or two separate simple closed curves that are spatially close? The answer to this question depends, of course, on the choice of the topology. Keeping the subspace topology for $\mathbb{Z}^2 \subset \mathbb{R}^2$ is useless in the analysis of connectivity since the digitized curves are then totally disconnected. However, we may choose topologies that provide for the desired connectivity of the curve shown in **2.18.6**(b). In the next sections we shall take a closer look at these topologies.



(a) (d)

Figure 2.18.6 Illustration of the dual representation of digitized curves.

2.19 Digital Topology

Any topology on $\mathbb{Z}^n$ or C^n is called a *digital topology*. Topologies on C^n are also referred to as *cellular topologies*. The set $\mathbb{Z}^n$ or C^n together with a topology is called a *digital space* or *cellular space*. Topologies other than the discrete topology can be defined in terms of the coordinates of points of $\mathbb{Z}^n$. One of the most popular digital topologies uses the concept of even and odd points. In particular, a point $\mathbf{p} = (p_1, p_2, \ldots, p_n) \in \mathbb{Z}^n$ is called *even* if and only if $\sum_{i=1}^{n} p_i$ is even. If $\mathbf{p}$ is not even, then $\mathbf{p}$ is said to be *odd*. Let $J = \{-1, 0, 1\}$, $\mathbf{p} = (p_1, p_2, \ldots, p_n) \in \mathbb{Z}^n$, and define a basic neighborhood $N(\mathbf{p})$ of $\mathbf{p}$ by

$$N(\mathbf{p}) = \begin{cases} \{\mathbf{p}\} \text{ if } \mathbf{p} \text{ is odd} \\ \{(p_1, \ldots, p_i + j, \ldots, p_n) : 1 \le i \le n,\ j \in J\} \text{ if } \mathbf{p} \text{ is even.} \end{cases}$$

It is easy to verify that the collection $B = \{N(\mathbf{p}) : \mathbf{p} \in \mathbb{Z}^{\mathbf{n}}\}$ satisfies the conditions of Theorem **2.13.4**. Therefore *B* is a basis for a topology on $\mathbb{Z}^n$ and the topology thus derived is called the *von Neumann topology*. This topology was first described by A. Rosenfeld for the case $n = 2$ [48]. Rosenfeld's pioneering work in digital topology has provided a variety of useful tools and a rigorous foundation for many image processing operations [29].

Defining basic neighborhoods $N(c(\mathbf{p}))$ for points $c(\mathbf{p}) \in C^n$ by $N(c(\mathbf{p})) = c(N(\mathbf{p}))$ provides for an equivalent topology on C^n. In particular, $N(c(\mathbf{p})) = \{c(\mathbf{p})\}$ if $\mathbf{p}$ is odd and $N(c(\mathbf{p})) = \{c(p_1, \ldots, p_{i-1}, p_i + j, p_{i+1}, \ldots, p_n) : 1 \le i \le n,\ j \in J\}$ if $\mathbf{p}$ is even. If *n*=2 or 3, then the possible neighborhoods $N(c(\mathbf{p}))$ of $\mathbf{p}$ are shown in Figure **2.19.1**(a) and (b), respectively.

As an easy consequence of the neighborhood definition we have

2.19.1 Theorem. *The basic neighborhoods for the von Neumann topology are path-connected.*

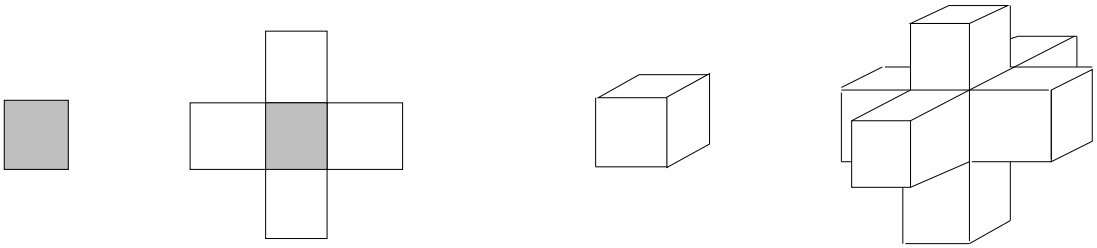


Figure 2.19.1 The von Neumann basis (a) if $n = 2$, and (b) if $n = 3$.

Proof: The proof is trivial if **p** is odd, for then $N(\mathbf{p})$ is a point. If **p** is even, let **q** and **r** be two distinct points in $N(\mathbf{p})$. Then at least one of **q** or **r** must be odd, say **q**. Now either **r** is even or odd. If **r** is even, then $\mathbf{r} = \mathbf{p}$. In this case, we define a path $f : [0,1] \to N(\mathbf{p})$ with initial point **q** and end point **r** by

$$f(x) = \mathbf{q} \text{ if } 0 \le x < \frac{1}{2} \text{ and } f(x) = \mathbf{p} \text{ if } \frac{1}{2} \le x \le 1 .$$

Since $\{\mathbf{q}\} = N(\mathbf{q}) \subset N(\mathbf{p})$, we have $f^{-1}(N(\mathbf{q})) = [0, \frac{1}{2})$ and $f^{-1}(N(\mathbf{p})) = [0,1]$. This shows that the inverse images of open sets are open in [0,1]. Hence *f* is continuous.
If the point **r** is odd, then we define $f : [0,1] \to N(\mathbf{p})$ by

$$f(x) = \mathbf{q} \text{ if } 0 \le x < \frac{1}{2} , \quad f\left(\frac{1}{2}\right) = \mathbf{p}, \text{ and } f(x) = \mathbf{r} \text{ if } \frac{1}{2} < x \le 1 .$$

In this case we have $f^{-1}(N(\mathbf{q})) = [0, \frac{1}{2})$, $f^{-1}(N(\mathbf{p})) = [0,1]$, and $f^{-1}(N(\mathbf{r})) = (\frac{1}{2}, 1]$. Thus, again, inverse images of open sets are open and, therefore, *f* is continuous.

Q.E.D.

2.20 Path-Connected Sets in Digital Spaces

There are many metrics that can be defined on $\mathbb{Z}^n$. Two commonly used metrics are the city block metric $d_1(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{n} |p_i - q_i|$ and the chessboard metric $d_2(\mathbf{p},\mathbf{q}) = max\{|p_i - q_i| : 1 \le i \le n\}$, where $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_n)$. Given a point $\mathbf{p} \in \mathbb{Z}^{\mathbf{n}}$, then its *von Neumann neighborhood* $F(\mathbf{p})$ is the set $F(\mathbf{p}) = \{\mathbf{q} : d_1(\mathbf{p},\mathbf{q}) \le 1\}$. In the case $n = 2$, $F(\mathbf{p})$ is also known as the *4–neighborhood* of **p** since it consists of the point **p** and its directly adjacent horizontal and vertical neighbors. Furthermore, if **p** is even, then $F(\mathbf{p}) = N(\mathbf{p})$.

The *Moore neighborhood* of **p** is denoted by $E(\mathbf{p})$ and defined as $E(\mathbf{p}) = \{\mathbf{q} : d_2(\mathbf{p},\mathbf{q}) \le 1\}$. For *n*=2, $E(\mathbf{p})$ is also known as the *8–neighborhood* of **p**.

2.20.1 **Definition.** A sequence of points $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k\} \subset \mathbb{Z}^n$ is called a d_1*-path* if $\mathbf{p}_{i+1} \in F(\mathbf{p_i})$ for $1 \le i \le k-1$, and a d_2*-path* if $\mathbf{p}_{i+1} \in E(\mathbf{p_i})$ for $1 \le i \le k-1$.
A set $S \subset \mathbb{Z}^n$ is said to be d_1*-connected* (d_2*-connected*) if for each pair of points $\mathbf{p}, \mathbf{q} \in S$ there exists a d_1-path (d_2-path) $\mathbf{p} = \mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k = \mathbf{q}$ from **p** to **q** such that $\mathbf{p}_i \in S$ for $1 \le i \le k$.

We call a d_i-path ($i = 1$ or 2) a *digital path* if it is clear from the discussion which type of path is meant. Unless otherwise specified, in the von Neumann space a digital path shall always mean a d_1-path and a *digital path-connected* set a d_1-connected set. Also, in the case *n*=2, d_1-paths and d_1-connectivity are usually referred to as *4–paths* and *4–connectivity*, while d_2-paths and d_2-connectivity are called *8–paths* and *8–connectivity*.

It follows from the definition that every d_1-path is a d_2-path and every d_1-connected set is also d_2-connected. The converse is obviously false; for *n*=2, the set $S = \{(x, y), (x + 1, y + 1)\}$ is d_2-connected but not d_1-connected. The relationship between connectivity, d_1-connectivity, and path-connectivity is given by

2.20.2 **Theorem.** *Let $\mathbb{Z}^n$ be the digital space with the von Neumann topology and $S \subset \mathbb{Z}^n$. Then the following are equivalent:*

(1) *S is connected.*

(2) *S is digital path-connected.*

(3) *S is path-connected.*

Proof: The equivalence of connected and path-connected follows from theorems **2.17.5**, **2.17.8**, and **2.19.1**.

To show the equivalence of connectivity and digital path-connectivity, assume first that *S* is connected. Let $\mathbf{p} \in S$ and set $A_{\mathbf{p}} = \{\mathbf{q} \in S : \text{there is a } d_1 - \text{path from } \mathbf{p} \text{ to } \mathbf{q} \text{ in } S\}$. If $A_{\mathbf{p}} = S$, there is nothing to prove. If $A_{\mathbf{p}} \neq S$, let $B_{\mathbf{p}} = S \backslash A_{\mathbf{p}}$. Then $A_{\mathbf{p}} \cup B_{\mathbf{p}} = S$ and $A_{\mathbf{p}} \cap B_{\mathbf{p}} = \varnothing$. Now if $\mathbf{q} \in A_{\mathbf{p}}$, then $N(\mathbf{q}) \cap S \subset A_{\mathbf{p}}$. This shows that $A_{\mathbf{p}}$ is open in *S*. Similarly, if $\mathbf{q} \in B_{\mathbf{p}}$, then $N(\mathbf{q}) \cap S \subset B_{\mathbf{p}}$; for otherwise there exists a path from $\mathbf{p}$ to $\mathbf{q}$, a condition that violates the definition of $B_{\mathbf{p}}$. Thus *S* is the union of two relatively open disjoint sets, namely $A_{\mathbf{p}}$ and $B_{\mathbf{p}}$. But this contradicts the assumption that *S* is connected. Therefore $A_{\mathbf{p}} = S$.

If, on the other hand, *S* is digital path-connected but not connected, then *S* is the union of two separated sets *A* and *B*. In this case let $\mathbf{p} \in A$, $\mathbf{q} \in B$, and $\mathbf{p} = \mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k = \mathbf{q}$ be a d_1-path in *S*. Then for some *j*, $\mathbf{p}_j \in A$ and $\mathbf{p}_{j+1} \in B$. One of $\mathbf{p}_j$ or $\mathbf{p}_{j+1}$ must be even. If $\mathbf{p}_j$ is even, then $\mathbf{p}_{j+1} \in F(\mathbf{p}_j) = N(\mathbf{p}_j)$ and, hence, $N'(\mathbf{p}_j) \cap B \neq \varnothing$. Thus, $\mathbf{p}_j \in A$ is a limit point of *B*; i.e. $\overline{B} \cap A \neq \varnothing$. But this contradicts the assumption that *A* and *B* are separated sets. A similar argument holds if $\mathbf{p}_{j+1}$ is even.

Q.E.D.

According to the theorem, the connected sets in $\mathbb{Z}^n$ are exactly the path-connected sets. Comparing this with Corollary **2.17.9**, we see that $\mathbb{Z}^n$ with the von Neumann topology enjoys some of the properties of $\mathbb{R}^n$. Subsequent theorems will show further similarities between $\mathbb{R}^n$ and $\mathbb{Z}^n$.

For the remainder of this section we shall assume, unless otherwise specified, that $\mathbb{Z}^n$ is the von Neumann space. The next lemma is a main ingredient in proving several interesting properties of digital curves.

2.20.3 **Lemma.** *Suppose $S \subset \mathbb{Z}^n$ and $\mathbf{p} \in S$. If S is connected and $S\backslash\{\mathbf{p}\} = A \cup B$ a separation, then each $A \cup \{\mathbf{p}\}$ and $B \cup \{\mathbf{p}\}$ are connected sets.*

Proof: We show that $A \cup \{\mathbf{p}\}$ is connected. By interchanging *A* with *B*, this also proves that $B \cup \{\mathbf{p}\}$ is connected.

Define $f : S \to A \cup \{\mathbf{p}\}$ by

$$f(\mathbf{x}) = \begin{cases} \mathbf{x} & \text{if } \mathbf{x} \in A \cup \{\mathbf{p}\} \\ \mathbf{p} & \text{if } \quad \mathbf{x} \in B \end{cases}.$$

We shall show that *f* is continuous. The result then follows from Theorem **2.16.7**.

Let $\mathbf{x} \in S$, $\mathbf{y} = f(\mathbf{x})$, and $W(\mathbf{y})$ a basic open neighborhood of $\mathbf{y}$ in $A \cup \{\mathbf{p}\}$. We show that $f^{-1}(W(\mathbf{y}))$ is open in *S*. Since $W(\mathbf{y})$ is a basic open neighborhood in $A \cup \{\mathbf{p}\}$, $W(\mathbf{y})$ is of form

$$W(y) = N(\mathbf{y}) \cap (A \cup \{\mathbf{p}\}).$$

There are two cases to consider, namely $\mathbf{y} \neq \mathbf{p}$ and $\mathbf{y} = \mathbf{p}$. Suppose $\mathbf{y} \neq \mathbf{p}$. In this case $\mathbf{y} \in A$ and $\mathbf{y} = \mathbf{x}$. Since *A* and *B* are separated, $\mathbf{y}$ cannot be a limit point of *B*. Thus $N(\mathbf{y}) \cap B = \varnothing$. If, in addition, $\mathbf{p} \notin N(\mathbf{y})$, then $W(y) = N(\mathbf{y}) \cap A$ and $f^{-1}(W(\mathbf{y})) = N(\mathbf{x}) \cap S$. Therefore, $f^{-1}(W(\mathbf{y}))$ is a relatively open set in *S* containing $\mathbf{x}$.

If $\mathbf{p} \in N(\mathbf{y})$, then $\mathbf{p}$ is odd and

$$f^{-1}(W(\mathbf{y})) = (N(\mathbf{x}) \cap A) \cup \{\mathbf{p}\} \cup B.$$

Since $N(\mathbf{x}) \cap B = \varnothing$ and $\mathbf{p} \in N(\mathbf{x})$,

$$(N(\mathbf{x}) \cap A) \cup \{\mathbf{p}\} = N(\mathbf{x}) \cap (A \cup \{\mathbf{p}\}) = N(\mathbf{x}) \cap S,$$

which is open in *S*. However, $B \cup \{\mathbf{p}\}$ is also open in *S*. This can be ascertained from the fact that since $\mathbf{p}$ is odd, we have for each $\mathbf{z} \in B \cup \{\mathbf{p}\}$ that $N(\mathbf{z}) \cap A = \varnothing$ and, therefore,

$$B \cup \{p\} = \bigcup_{\mathbf{z} \in B \cup \{\mathbf{p}\}} (N(\mathbf{z}) \cap S).$$

Thus, $f^{-1}(W(\mathbf{y}))$ is the union of two relatively open sets $(N(\mathbf{x}) \cap A) \cup \{\mathbf{p}\}$ and $B \cup \{\mathbf{p}\}$. This shows that $f^{-1}(W(\mathbf{y}))$ is open in *S* as long as $\mathbf{y} \neq \mathbf{p}$.

Next suppose that $\mathbf{y} = \mathbf{p}$. In this case,

$$f^{-1}(W(\mathbf{y})) = (N(\mathbf{p}) \cap A) \cup \{\mathbf{p}\} \cup B.$$

Now if $\mathbf{p}$ is odd, then $N(\mathbf{p}) = \{\mathbf{p}\}$ and, therefore, $f^{-1}(W(\mathbf{y})) = B \cup \{\mathbf{p}\}$. By the above argument, $B \cup \{\mathbf{p}\}$ is open in *S*. If, on the other hand, $\mathbf{p}$ is even, let $U = N(\mathbf{p}) \cap A \cup \{\mathbf{p}\} \cup B$. Then if $\mathbf{z} \in U$ we must have either $\mathbf{z} \in N(\mathbf{p}) \cap A$ with $\mathbf{z} \neq \mathbf{p}$, or $\mathbf{z} = \mathbf{p}$, or $\mathbf{z} \in B$.

If $\mathbf{z} \in N(\mathbf{p}) \cap A$ with $\mathbf{z} \neq \mathbf{p}$, then $\mathbf{z}$ is odd and $\{\mathbf{z}\}$ is open. Hence $\mathbf{z} \in \{\mathbf{z}\} \subset U$ is an interior point of *U*.

If $\mathbf{z} = \mathbf{p}$, then

$$\begin{aligned} N(\mathbf{z}) \cap U &= N(\mathbf{p}) \cap A \cup \{\mathbf{p}\} \cup N(\mathbf{p}) \cap B \\ &= N(\mathbf{p}) \cap (A \cup \{\mathbf{p}\} \cup B) \ = N(\mathbf{p}) \cap S. \end{aligned}$$

which is open in *S*. Therefore $\mathbf{p} \in N(\mathbf{p}) \cap S \subset U$ is an interior point of *U*.

Finally, if $\mathbf{z} \in B$, then $N(\mathbf{z}) \cap A = \varnothing$ and

$$N(\mathbf{z}) \cap U = N(\mathbf{z}) \cap (B \cup \{\mathbf{p}\}) = N(\mathbf{z}) \cap S$$

which is open in *S*. Thus, in each case **z** is an interior point of *U*. Since **z** was arbitrary, this shows that $intU = U$. This completes the proof that $f^{-1}(W(\mathbf{y}))$ is open in *S*.
Let $V = f^{-1}(W(\mathbf{y}))$ and $\mathbf{q} \in V$. Then $f(\mathbf{q}) \in f(V) = f\big[f^{-1}(W(\mathbf{y}))\big] = W(\mathbf{y})$. This shows that given a point $\mathbf{x} \in S$ and a basic open neighborhood $W(f(\mathbf{x})) \subset A \cup \{\mathbf{p}\}$, then there exists an open set *V* containing **x**, namely $V = f^{-1}(W(f(\mathbf{x})))$, such that for each $\mathbf{q} \in V$, $f(\mathbf{q}) \in W(f(\mathbf{x}))$. Therefore, *f* is continuous at **x**, and since **x** was arbitrary, *f* is continuous on *S*.

Q.E.D.

Lemma **2.20.3** is needed to prove the following surprising theorem:

2.20.4 **Theorem.** *Suppose $S \subset \mathbb{Z}^n$ and $card(S) \geq 2$. If S is compact and connected, then S has at least two non-cut points.*

Proof: Let *N* be the set of all non-cut points of *S* and suppose to the contrary that $card(N) < 2$. Let $\mathbf{p} \in S \backslash N$. Then $S\backslash\{\mathbf{p}\} = A \cup B$, where *A* and *B* are separated sets with *N* contained in one of *A* or *B*. Suppose without loss of generality that $N \subset B$. For each point $\mathbf{q} \in A$, let $S\backslash\{\mathbf{q}\} = A_{\mathbf{q}} \cup B_{\mathbf{q}}$, where $A_{\mathbf{q}}$ and $B_{\mathbf{q}}$ are separated sets with $\mathbf{p} \in B_{\mathbf{q}}$. Since $\mathbf{p} \in B_{\mathbf{q}}$ and by Lemma **2.20.3** $A_{\mathbf{q}} \cup \{\mathbf{q}\}$ is connected, we must have $A_{\mathbf{q}} \cup \{\mathbf{q}\} \subset A$.
Partially order the collection $\mathcal{A} = \{A_{\mathbf{q}}\}_{\mathbf{q} \in A}$ by subset inclusion. Since *S* is compact, *S* is finite. Therefore $\mathcal{A}$ must be finite. Thus we can select a maximal simply ordered subcollection $\mathcal{B} = \{A_{\mathbf{q}_i}\}_{i=1}^{k}$ of $\mathcal{A}$ such that $A_{\mathbf{q}_i} \subset A_{\mathbf{q}_j}$ whenever $i > j$. Consider $A_{\mathbf{q}_k} = \bigcap_{i=1}^{k} A_{\mathbf{q}_i}$. Since the sets $A_{\mathbf{q}_k}$, $B_{\mathbf{q}_k}$ form a separation of $S\backslash\{\mathbf{q}_k\}$, $A_{\mathbf{q}_k} \neq \varnothing$. Let $\mathbf{q} \in A_{\mathbf{q}_k}$. Then since $A_{\mathbf{q}_k} \cup \{\mathbf{q}_k\} \subset A$, **q** must be an element of *A* and therefore a cut point of *S*. Now consider the set $A_{\mathbf{q}}$. Since $A_{\mathbf{q}} \cup \{\mathbf{q}\}$ is a connected subset of $S\backslash\{\mathbf{q}\}$, it must either be a subset of $A_{\mathbf{q}_k}$ or of $B_{\mathbf{q}_k}$. Since $\mathbf{q} \in A_{\mathbf{q}_k}$, $A_{\mathbf{q}} \cup \{\mathbf{q}\} \subset A_{\mathbf{q}_k}$ and hence $A_{\mathbf{q}} \subset A_{\mathbf{q}_k}$. Thus the collection $\mathcal{B}$ is a strict subset of the simply ordered collection $\{A_{\mathbf{q}_1}, A_{\mathbf{q}_2}, \ldots, A_{\mathbf{q}_k}, A_{\mathbf{q}}\}$. This contradicts the maximality of $\mathcal{B}$ and proves that *N* must contain more than one point.

Q.E.D.

The reason that **2.20.4** is a surprising theorem is that it is true for Euclidean space $\mathbb{R}^n$ but not for general topological spaces. In particular, **2.20.4** holds for a class of spaces known as T_1-spaces but not for a class known as T_0-spaces. A topological space *X* is a T_0-*space* if given any two points of *X*, then at least one of them is contained in an open set not containing the other. Obviously, both $\mathbb{Z}^n$ and $\mathbb{R}^n$ are T_0-spaces. However, Euclidean spaces also satisfy several stronger conditions. One of these is the T_1-*hypothesis*. A space *X* is called a T_1-*space* if given any two points of *X*, then each of them lies in an open set not containing the other. For example, if $\mathbf{p}, \mathbf{q} \in \mathbb{R}^n$ and $d(\mathbf{p}, \mathbf{q}) = r$, then $\mathbf{q} \notin N_r(\mathbf{p})$ and $\mathbf{p} \notin N_r(\mathbf{q})$. Thus, $\mathbb{R}^n$ is a T_1-space. On the other hand, if $\mathbf{p} \in \mathbb{Z}^n$ is even and $\mathbf{q} \in \mathbb{Z}^n$ with $d(\mathbf{p}, \mathbf{q}) = 1$, then $\mathbf{q} \in N(\mathbf{p})$. It follows that **q** is in every open set containing **p** and $\mathbf{p} \notin N(\mathbf{q}) = \{\mathbf{q}\}$. Hence $\mathbb{Z}^n$ is

a T_0-space but not a T_1-space. The reason that **2.20.4** holds for $\mathbb{Z}^n$, even though it is not a T_1-space, is due to the von Neumann topology which was heavily employed in the proof of the theorem.

2.21 Digital Arcs and Curves

There are many other properties shared by $\mathbb{Z}^n$ and $\mathbb{R}^n$, and one of these is the similarity between digital arcs in $\mathbb{Z}^n$ and arcs in $\mathbb{R}^n$. In the definition of digital arcs given below, the set $F'(\mathbf{p})$ denotes the deleted neighborhood $F(\mathbf{p})\backslash\{\mathbf{p}\}$.

2.21.1 **Definition.** A *digital arc* is a d_1-path $\{\mathbf{p}_1, \ldots, \mathbf{p}_k\} \subset \mathbb{Z}^n$ such that, for all $1 \leq i, j \leq k$,

(1) $\mathbf{p}_i = \mathbf{p}_j \iff i = j$, and

(2) $\mathbf{p}_i \in F'(\mathbf{p}_j) \iff i = j \pm 1$.

If the arc consists of one point $(k = 1)$, then it is also called a *degenerate arc*. A *digital simple closed curve* is a d_1-path $\{\mathbf{p}_1, \ldots, \mathbf{p}_k\}$ such that $k > 4$ and for all $1 \leq i, j \leq k$,

$(1')$ $\mathbf{p}_i = \mathbf{p}_j \iff i = j$, and

$(2')$ $\mathbf{p}_i \in F'(\mathbf{p}_j) \iff i = j \pm 1 \; mod \; k$.

A digital arc satisfying the additional condition

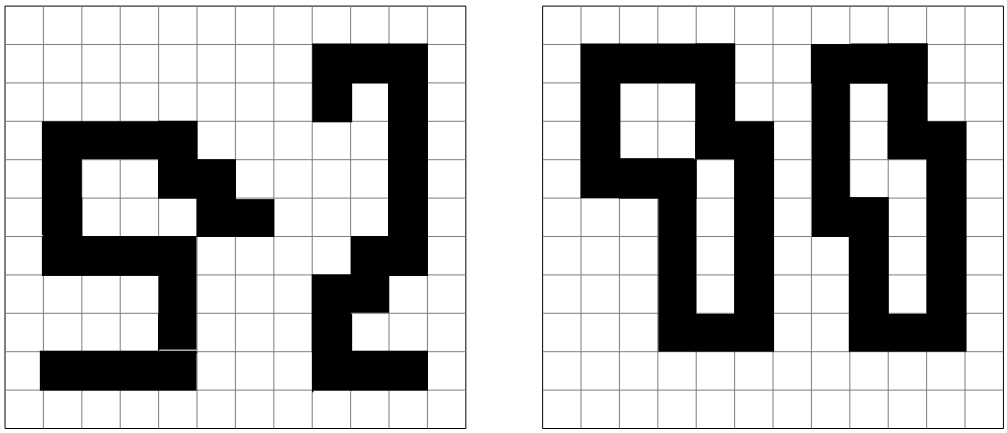
(3) $\mathbf{p}_i \notin F(\mathbf{p}_j)$ and $F(\mathbf{p}_i) \cap F(\mathbf{p}_j) \neq \varnothing \iff i = j \pm 2$.

is called a *digital Jordan arc.* A *digital Jordan curve* is a digital simple closed curve satisfying condition

$(3')$ $\mathbf{p}_i \notin F(\mathbf{p}_j)$ and $F(\mathbf{p}_i) \cap F(\mathbf{p}_j) \neq \varnothing \iff i = j \pm 2 \; mod \; k$.

Note that a digital arc is a digital path which, because of condition (1), cannot double back on itself or cross itself. Condition (2) implies that a digital arc does touch itself; a point $\mathbf{p}_j$ with $j > i + 1$ or $j < i - 1$ must be a (Euclidean) distance greater than one from the point $\mathbf{p}_i$. For Jordan arcs, condition (3) means that the path cannot even double back within the distance $\sqrt{2}$ of another point of the path. In higher dimensions it allows arcs to double back to within the Moore neighborhood of a previous point $\mathbf{p}_i$, but at a distance greater than $\sqrt{2}$.

The requirement $k > 4$ for digital simple closed curves rules out one point (degenerate) closed curves, two point closed curves, and four point closed curves that form a 2×2 block of cells. In fact, it is not hard to see that these are the only d_1-paths with $k < 8$ that satisfy $(1')$ and $(2')$. Hence a digital simple closed curve must contain at least eight points. Figure **2.21.1** provides examples of digital arcs and digital simple closed curves.



(a) (b)

Figure 2.21.1 (a) A digital arc (left) and a digital Jordan arc (right). (b) A digital simple closed curve (left) and a digital Jordan curve (right).

Digital arcs and digital simple closed curves as defined in **2.21.1** are d_1-connected sets. Obviously, by using $E(\mathbf{p})$ instead of $F(\mathbf{p})$ one can just as well define the concepts of a digital arc and digital simple closed curve in terms of d_2-connectivity.

The points of a digital arc $\mathbf{p}_1, \ldots, \mathbf{p}_k$ are all distinct, and $\mathbf{p}_1$ and $\mathbf{p}_k$ are called the *end points* of the arc. It is also obvious from the definition that the end points are non-cut points of the arc and that any other point $\mathbf{p}_i$ is a cut point. This is one property digital arcs share with topological arcs (Example **2.17.10**(i)). Before discussing further common properties we need to establish the following result:

2.21.2 **Lemma.** *If $S \subset \mathbb{Z}^n$ is connected, then for any two given points $\mathbf{p}, \mathbf{q} \in S$ there exists a digital arc from $\mathbf{p}$ to $\mathbf{q}$ in S.*

Proof: By Theorem **2.20.2**, there exists a digital path from $\mathbf{p}$ to $\mathbf{q}$ in *S*. Let $P = \{\mathbf{p}_1, \ldots, \mathbf{p}_k\}$ be a digital path from $\mathbf{p}$ to $\mathbf{q}$ in *S* of minimal length with $\mathbf{p} = \mathbf{p}_1$ and $\mathbf{q} = \mathbf{p}_k$. Then $\mathbf{p}_i \neq \mathbf{p}_j$ if $i < j$; otherwise, $\{\mathbf{p}_1, \ldots, \mathbf{p}_i, \mathbf{p}_{j+1}, \ldots, \mathbf{p}_k\}$ would be a shorter path than *P*.

If $i < j - 1$, then $\mathbf{p}_i \notin F(\mathbf{p}_j)$, for otherwise $\{\mathbf{p}_1, \ldots, \mathbf{p}_i, \mathbf{p}_j, \ldots, \mathbf{p}_k\}$ is shorter than *P*. Similarly, we cannot have $\mathbf{p}_i \in F(\mathbf{p}_j)$ for $i > j+1$, for otherwise the path $\{\mathbf{p}_1, \ldots, \mathbf{p}_j, \mathbf{p}_i, \ldots, \mathbf{p}_k\}$ is shorter than *P*. It follows that *P* is a digital arc from $\mathbf{p}$ to $\mathbf{q}$.

Q.E.D.

We are now able to prove the digital version of Theorem **2.17.11**.

2.21.3 **Theorem.** *If $S \subset \mathbb{Z}^n$ is compact and connected with just two non-cut points, then S is a digital arc.*

Proof: Let $\mathbf{p}, \mathbf{q} \in S$ be the two non-cut points of *S*. By Lemma **2.21.2** there exists a digital arc *P* from **p** to **q** in *S*. If $P \neq S$, then there exists a point $\mathbf{x} \in S \backslash P$. Since **x** is a cut point of *S*, $S\backslash\{\mathbf{x}\} = A \cup B$, where *A* and *B* are separated sets. Since *P* is connected, *P* must be contained in one of *A* or *B*, say *A*. By Lemma **2.20.3**, $B \cup \{\mathbf{x}\}$ is connected. Since $B \neq \varnothing$, $card(B \cup \{\mathbf{x}\}) \geq 2$. Hence, by Theorem **2.20.4**, $B \cup \{\mathbf{x}\}$ has at least two non-cut points, one of which, call it **y**, is not the point **x**. We now have two connected sets, namely $(B \cup \{\mathbf{x}\})\backslash\{\mathbf{y}\}$ and $A \cup \{\mathbf{x}\}$, and these sets have the point **x** in common. Thus

$$S\backslash\{\mathbf{y}\} = (A \cup \{\mathbf{x}\}) \cup [(B \cup \{\mathbf{x}\})\backslash\{\mathbf{y}\}]$$

is connected. But this means that **y** is a non-cut point of *S* that is not in *P*, a contradiction.

Q.E.D.

A pair of points $\mathbf{p}, \mathbf{q} \in \mathbb{Z}^n$ are called *von Neumann neighbors* or simply *neighboring points* if $\mathbf{p} \in F(\mathbf{q})$. If **p** and **q** are not von Neumann neighbors, then they are referred to as *non-neighboring points*.

2.21.4 Theorem. *If $S \subset \mathbb{Z}^n$ is compact and connected with the property that for any pair of non-neighboring points $\mathbf{x}, \mathbf{y} \in S$, $S\backslash\{\mathbf{x}, \mathbf{y}\}$ is not connected, then S is a digital simple closed curve.*

Proof: We divide the proof into five parts.

(1) We first prove that *S* contains no cut points. Suppose to the contrary that **p** is a cut point of *S*. Then $S\backslash\{\mathbf{p}\} = A \cup B$, where *A* and *B* are separated sets. By Lemma **2.20.3**, $A \cup \{\mathbf{p}\}$ and $B \cup \{\mathbf{p}\}$ are both compact connected sets. According to Theorem **2.20.4**, there exist points $\mathbf{x} \in A$ and $\mathbf{y} \in B$ such that **x** does not separate $A \cup \{\mathbf{p}\}$ and **y** does not separate $B \cup \{\mathbf{p}\}$. Since *A* and *B* are separated sets, **x** and **y** are not neighboring points. But then

$$S\backslash\{\mathbf{x}, \mathbf{y}\} = [(A \cup \{\mathbf{p}\})\backslash\{\mathbf{x}\}] \cup [(B \cup \{\mathbf{p}\})\backslash\{\mathbf{y}\}]$$

is the union of two connected sets that have the point **p** in common. Thus $S\backslash\{\mathbf{x}, \mathbf{y}\}$ is connected, contrary to the theorem's hypothesis.

(2) Next, suppose that $S\backslash\{\mathbf{x}, \mathbf{y}\} = A \cup B$, where *A* and *B* are separated sets. Then $A \cup \{\mathbf{x}, \mathbf{y}\}$ and $B \cup \{\mathbf{x}, \mathbf{y}\}$ are both connected sets. For suppose to the contrary that $A \cup \{\mathbf{x}, \mathbf{y}\} = U \cup V$, where *U* and *V* are separated sets. If *U* contains both **x** and **y**, let $\mathbf{z} \in V$. Since $V \subset A$, $V\backslash\{\mathbf{z}\}$ and B are separated sets. Thus $S\backslash\{\mathbf{z}\} = (V\backslash\{\mathbf{z}\}) \cup (B \cup U)$ is a separation, contrary to part (1) of the proof. If *U* contains only one of the points **x** and **y**, say **x**, then $S\backslash\{\mathbf{x}\} = (U\backslash\{\mathbf{x}\}) \cup (B \cup V)$ is a separation. Thus again we have a contradiction to part (1).

(3) At least one of $A \cup \{\mathbf{x}, \mathbf{y}\}$ or $B \cup \{\mathbf{x}, \mathbf{y}\}$ is a digital arc. For if not, then it follows from Theorems **2.20.4** and **2.21.3** that each of the sets contains a non-cut point, say $\mathbf{p} \in A \cup \{\mathbf{x}, \mathbf{y}\}$ and $\mathbf{q} \in B \cup \{\mathbf{x}, \mathbf{y}\}$, distinct from **x** and **y**. But then

$$S\backslash\{\mathbf{x}, \mathbf{y}\} = [(A \cup \{\mathbf{x}, \mathbf{y}\})\backslash\{\mathbf{p}\}] \cup [(B \cup \{\mathbf{x}, \mathbf{y}\})\backslash\{\mathbf{q}\}]$$

is a connected set since it is the union of two connected sets having the points **x** and **y** in common. In addition, **p** and **q** are non-neighboring points since $\mathbf{p} \in A$ and $\mathbf{q} \in B$. This contradicts the hypothesis.

(4) Both $A \cup \{\mathbf{x}, \mathbf{y}\}$ and $B \cup \{\mathbf{x}, \mathbf{y}\}$ are digital arcs. By (3), at least one of these two sets is a digital arc. Let $B \cup \{\mathbf{x}, \mathbf{y}\}$ be the digital arc guaranteed by (3). If $A \cup \{\mathbf{x}, \mathbf{y}\}$ is not a digital arc, then it must contain a non-cut point $\mathbf{p} \notin \{\mathbf{x}, \mathbf{y}\}$. Since $B \cup \{\mathbf{x}, \mathbf{y}\}$ is a digital arc and **x** and **y** are not neighboring points, there exists a point $\mathbf{q} \in B$ which separates **x** from **y**; that is, $(B \cup \{\mathbf{x}, \mathbf{y}\}) \backslash \{\mathbf{q}\} = U \cup V$ is a separation with $\mathbf{x} \in U,\ \mathbf{y} \in V$, and each *U* and *V* is a connected set. But then

$$S \backslash \{\mathbf{x}, \mathbf{y}\} = [(A \cup \{\mathbf{x}, \mathbf{y}\}) \backslash \{\mathbf{p}\}] \cup (U \cup V)$$

is a connected set, with **p** and **q** non-neighboring points. This again contradicts the hypothesis.

(5) The points **x** and **y** are end points of $A \cup \{\mathbf{x}, \mathbf{y}\}$ and of $B \cup \{\mathbf{x}, \mathbf{y}\}$. If neither **x** nor **y** is an end point of $A \cup \{\mathbf{x}, \mathbf{y}\}$, let $\mathbf{p}, \mathbf{q} \in A$ denote the end points of $A \cup \{\mathbf{x}, \mathbf{y}\}$. Then

$$S \backslash \{\mathbf{p}, \mathbf{q}\} = [(A \cup \{\mathbf{x}, \mathbf{y}\}) \backslash \{\mathbf{p}, \mathbf{q}\}] \cup (B \cup \{\mathbf{x}, \mathbf{y}\})$$

is the union of two connected sets having the points **x** and **y** in common. Thus $S \backslash \{\mathbf{p}, \mathbf{q}\}$ is connected. Since **p** and **q** are end points and $card(A \cup \{\mathbf{x}, \mathbf{y}\}) > 2$, they are non-neighboring points. This contradicts the hypothesis. Thus $\{\mathbf{x}, \mathbf{y}\}$ contains at least one end point of *A* and, for analogous reasons, of $B \cup \{\mathbf{x}, \mathbf{y}\}$.

Suppose **y** is the end point of $A \cup \{\mathbf{x}, \mathbf{y}\}$ but **x** is not. If $\mathbf{p} \in A$ denotes the other end point, then **p** and **y** cannot be neighbors. Also, both **x** and **y** must be end points of $B \cup \{\mathbf{x}, \mathbf{y}\}$. For if not, let $\mathbf{q} \in B$ denote the other endpoint. Then

$$S \backslash \{\mathbf{p}, \mathbf{q}\} = [(A \cup \{\mathbf{x}, \mathbf{y}\}) \backslash \{\mathbf{p}\}] \cup [(B \cup \{\mathbf{x}, \mathbf{y}\}) \backslash \{\mathbf{q}\}]$$

is the union of two connected sets having the points **x** and **y** in common. Thus $S \backslash \{\mathbf{p}, \mathbf{q}\}$ is connected, where **p** and **q** are non-neighboring points. But this contradicts the hypothesis.

We now have that **x** and **y** are the end points of $B \cup \{\mathbf{x}, \mathbf{y}\}$ and **p** and **y** are end points of $A \cup \{\mathbf{x}, \mathbf{y}\}$. Then $(A \cup \{\mathbf{x}\}) \backslash \{\mathbf{p}\}$ and $B \cup \{\mathbf{x}\}$ are connected sets having the point **x** in common. Therefore,

$$S \backslash \{\mathbf{p}, \mathbf{y}\} = [(A \cup \{\mathbf{x}\}) \backslash \{\mathbf{p}\}] \cup (B \cup \{\mathbf{x}\})$$

is connected. This again contradicts the hypothesis.

Thus *S* is the union of two digital arcs having only their end points in common. Furthermore, since *A* and *B* are separated sets, we have that $\mathbf{p} \notin F(\mathbf{q})\ \forall \mathbf{p} \in A$ and $\mathbf{q} \in B$. It now follows that *S* has at least eight points and satisfies conditions $(1')$ and $(2')$ of Definition **2.21.1**.

Q.E.D.

If $A \subset \mathbb{Z}^n$, then a d_2-*component* of *A* is a maximal d_2-connected subset of *A*. In particular, if $n = 2$, then an *8–component* is a maximal 8–connected subset of *A*. In [47] and [48], Rosenfeld proved the following digital version of the Jordan Curve Theorem (**2.17.13**):

2.21.5 **Theorem.** *If $S \subset \mathbb{Z}^2$ is a digital simple closed curve, then $\mathbb{Z}^2 \backslash S$ has two 8–components.*

It can be seen from Figure **2.21.1**(b) that the 8–components of $\mathbb{Z}^2 \backslash S$ need not be connected sets in the von Neumann topology. However, the exact analogue of the Jordan Curve theorem follows as an easy corollary of Rosenfeld's theorem.

2.21.6 **Corollary.** *If $S \subset \mathbb{Z}^2$ is a digital Jordan curve, then $\mathbb{Z}^2 \backslash S$ has two components.*

Proof: Let *C* be one of the 8–components of $\mathbb{Z}^2 \backslash S$ guaranteed by **2.21.5** and $\mathbf{p} \in C$. Let $C_{\mathbf{p}}$ denote the component of *C* containing **p**. We shall show that $C_{\mathbf{p}} = C$.

Since $\mathbf{p} \in C_{\mathbf{p}}$, $C_{\mathbf{p}} \neq \varnothing$. Suppose to the contrary that $C_{\mathbf{p}} \neq C$. Then there is a point $\mathbf{q} \in C \backslash C_{\mathbf{p}}$ and, since *C* is 8–connected, and 8–path $P = \{\mathbf{p}_1, \ldots, \mathbf{p}_k\}$ from **p** to **q** in *C*. Since $\mathbf{q} \notin C_{\mathbf{p}}$, there exists an n with $1 \leq n < k$ such that $\mathbf{p}_n \in C_{\mathbf{p}}$ and $\mathbf{p}_{n+1} \in C \backslash C_{\mathbf{p}}$. Let $(i, j) = \mathbf{p}_n$. Then $\mathbf{p}_{n+1} \notin F(\mathbf{p}_n)$, for otherwise $C_{\mathbf{p}} \cup \{\mathbf{p}_{n+1}\}$ is a connected subset of *C* larger than $C_{\mathbf{p}}$. Thus, $\mathbf{p}_{n+1}$ must be one of the four diagonally adjacent points $(i \pm 1, j \pm 1)$. Suppose, without loss of generality, that $\mathbf{p}_{n+1} = (i + 1, j + 1)$. Now consider the points $\mathbf{x} = (i + 1, j)$ and $\mathbf{y} = (i, j + 1)$. Since *S* is a Jordan curve, it follows from **2.21.1** $(3')$, that at least one of the points, say **x**, is not an element of *S*. But then $C_{\mathbf{p}} \cup \{\mathbf{x}\}$ is a connected subset of *C* larger than $C_{\mathbf{p}}$. Therefore $C_{\mathbf{p}} = C$.

Q.E.D.

The following digital version of Example **2.17.3**(ii) was also proven by Rosenfeld [47].

2.21.7 **Theorem.** *If $S \subset \mathbb{Z}^2$ is a digital arc, then $\mathbb{Z}^2 \backslash S$ is 8–connected.*

In view of Figure **2.21.1**(a), it is obvious that $\mathbb{Z}^2 \backslash S$ need not be connected. The following digital analogue of **2.17.3**(ii) follows as an easy corollary to **2.21.7**:

2.21.8 **Corollary.** *If $S \subset \mathbb{Z}^2$ is a digital Jordan arc, then $\mathbb{Z}^2 \backslash S$ is connected.*

The proof of this corollary is identical to that of Corollary **2.21.6**.

A topological invariant that provides a method for counting objects in the digital plane $\mathbb{Z}^2$ is the Euler number.

2.21.9 **Definition.** If $S \subset \mathbb{Z}^2$ is compact, then the *Euler number* of *S*, *E*(*S*), is defined as

$$E(S) = m(S) - n(S)\,,$$

where *m*(*S*) denotes the number of components of *S* and *n*(*S*) the number of bounded components of $\mathbb{Z}^2 \backslash S$. The bounded components of $\mathbb{Z}^2 \backslash S$ are also called the *holes* of *S*.

If one of the numbers *m*(*S*) or *n*(*S*) is known, then the Euler number provides a means for obtaining the other. This observation has direct practical applications. For example, there are various "hole filling" algorithms whose output are connected objects without holes. The Euler number can then be applied to find the number of objects present. Conversely, if we know that we are dealing with one connected object, then the Euler number provides us with the number of holes in the object.

2.22 Weakly Connected Sets and d_2-Connectivity

In the preceding two sections we formulated most definitions and theorems in terms of d_1-connectivity and observed that most of the theorems have analogous interpretations for d_2-connectivity. This holds for the Euler number as well. For 8–connectivity we define

$$E_8(S) = m_8(S) - n_8(S)\,,$$

where $m_8(S)$ denotes the number of 8–components of *S* and $n_8(S)$ the number of bounded 8–components of $\mathbb{Z}^2 \backslash S$. Obviously, in most cases $E_8(S) \neq E(S)$.

One reason we preferred using d_1-connectivity is Theorem **2.20.2**. It is well known that there does not exist a topology on $\mathbb{Z}^n$ for which connectivity is equivalent to d_2-connectivity [4]. However, there exist topologies on $\mathbb{Z}^n$ in which every d_2-connected set is weakly connected. Note that if $\mathbf{p} = (i, j)$ is an even point in the von Neumann space $\mathbb{Z}^2$ and $\mathbf{q}$ is any one of the diagonally adjacent neighboring points $(i \pm 1, j \pm 1)$, then the set $\{\mathbf{p}, \mathbf{q}\}$ is weakly connected but not connected. This is in contrast to $\mathbb{R}^n$, where weak connectivity and connectivity are equivalent notions (Theorem **2.16.6**). It does not mean, however, that 8–connected sets are weakly connected; the set $\{(i, j+1), (i+1, j)\}$ is not weakly connected.

In the von Neumann topology on $\mathbb{Z}$ a basic open set $N(p)$ is of form $N(p) = \{p\}$ if $p \in \mathbb{Z}$ is odd and $N(p) = \{p-1, p, p+1\}$ if *p* is even. The Cartesian product of these basic neighborhoods will be used to define a basis for a topology on $\mathbb{Z}^n$ which is different from the von Neumann topology. Specifically, with each $\mathbf{p} = (p_1, p_2, \ldots, p_n) \in \mathbb{Z}^n$ we associate a basic neighborhood $N(\mathbf{p})$ defined by

$$N(\mathbf{p}) = (N(p_1), \ldots, N(p_n)) = \prod_{i=1}^{n} N(p_i)\,,$$

where each $N(p_i)$ is a basic neighborhood of $p_i \in \mathbb{Z}$ of the von Neumann space $\mathbb{Z}$. It is not difficult to show that the collection $B = \{N(\mathbf{p}) : \mathbf{p} \in \mathbb{Z}^{\mathbf{n}}\}$ is a basis for a topology on $\mathbb{Z}^n$. Since there are 2^n possible neighborhood configurations, this topology is appropriately called the 2^n-*topology* on $\mathbb{Z}^n$. The set $\mathbb{Z}^n$ together with this topology is also referred to as the *product space* of the von Neumann space $\mathbb{Z}$.

For illustrative purposes we again consider the dual space C^n obtained by substituting $c(\mathbf{p})$ for $\mathbf{p}$ and $N(c(\mathbf{p}))$ for $N(\mathbf{p})$. The four different neighborhood configurations for the product space C^2 are shown in Figure **2.22.1**. The shaded cell represents the cell $c(i, j) \in N(c(i, j))$. Here the left-most neighborhood results when *i* and *j* are both even. Proceeding from left to right, the next neighborhood pictured represents the case *i* even and *j* odd followed by *i* even and *j* odd, and *i* and *j* both odd, respectively.

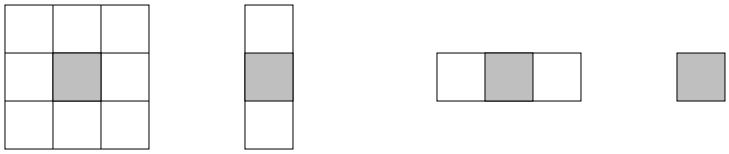


Figure 2.22.1 The four possible basic neighborhoods in the product space C^2.

Product spaces and von Neumann spaces share several topological properties. In particular, we have the following analogues of Theorems **2.19.1** and **2.20.2**:

2.22.1 Theorem. *The basic neighborhoods for the 2^n-topology are path-connected.*

The proof of **2.22.1** is analogous to that of **2.19.1**. Given a neighborhood $N(\mathbf{p})$ and $\mathbf{p}, \mathbf{q} \in N(\mathbf{p})$, we define the path $P = \{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\}$, where $\mathbf{p}_1 = \mathbf{q}$, $\mathbf{p}_2 = \mathbf{p}$, and $\mathbf{p}_3 = \mathbf{r}$, and define a continuous path $f : [0,1] \to N(\mathbf{p})$ from **q** to **r** according to the types of coordinates of the points **p**, **q**, and **r**.

2.22.2 Theorem. *Let $\mathbb{Z}^n$ be the digital space with the 2^n-topology and $S \subset \mathbb{Z}^n$. Then S is connected $\iff$ S is path-connected.*

As in the case of **2.20.2**, the equivalence follows from Theorems **2.17.5**, **2.17.8**, and **2.22.1**. In contrast to Theorem **2.20.2**, connectivity in the 2^n-topology is not equivalent to digital path-connectivity for either d_1-paths or d_2-paths. It is true that every d_1-path is connected in the product space $\mathbb{Z}^n$. However, connected sets need not be d_1-path connected. For example, if $(i,j) \in \mathbb{Z}^2$ with *i* and *j* both even integers, then $\{(i,j),(i+1,j+1)\}$ is connected but not d_1-connected. Although $\{(i,j),(i+1,j+1)\}$ is d_2-connected, it does not mean that all d_2-connected sets are connected; the set $\{(i+1,j),(i,j+1)\}$ is d_2-connected but not connected. However, $\{(i+1,j),(i,j+1)\}$ is weakly connected, and so is every d_2-connected set in the product space $\mathbb{Z}^n$. These observations are relevant in connection with the topological notion of *weak path-connectivity*.

2.22.3 Definition. Let $\mathbb{Z}^n$ be a digital space. A sequence of points $\{\mathbf{p}_1, \ldots, \mathbf{p}_k\} \in \mathbb{Z}^n$ is called a *weak path* if $\{\mathbf{p}_i, \mathbf{p}_{i+1}\}$ is weakly connected for $1 \leq i \leq k-1$.

A set $S \subset \mathbb{Z}^n$ is *weakly path-connected* if for each pair of points $\mathbf{p}, \mathbf{q} \in S$, there exists a weak path $P = \{\mathbf{p} = \mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k = \mathbf{q}\} \subset S$. The set *P* is called a *weak path from* **p** *to* **q**.

As a consequence of this definition we have the following:

2.22.4 Theorem. *If P is a weak path in a digital space, then P is weakly connected.*

Proof: Suppose to the contrary that some weak path $P = \{\mathbf{p}_1, \ldots, \mathbf{p}_k\} \subset \mathbb{Z}^n$ is not weakly connected. Then there exist open sets *U* and *V* such that $P \subset U \cup V$, $P \cap U \neq \varnothing \neq P \cap V$, and $U \cap V = \varnothing$. Assume without loss of generality that $\mathbf{p}_k \in P \cap V$. Let $j = max\{i : \mathbf{p}_i \in P \cap U\}$. Since $j+1 \leq k$ we have by definition of *j* that $\mathbf{p}_{j+1} \in P \cap V$, while $\mathbf{p}_j \in P \cap U$. Thus $\{\mathbf{p}_j, \mathbf{p}_{j+1}\}$ is not weakly connected, contrary to the hypothesis that *P* is a weak path.

Q.E.D.

It is not difficult to verify that connectivity implies weak path-connectivity in the von Neumann topology as well as the 2^n-topology. Also, every d_1-path is a topological path and, hence, a weak path in these topologies. However, as we noted earlier, diagonally adjacent points in the von Neumann space $\mathbb{Z}^2$ need not be weakly path-connected. Hence d_2-paths in von Neumann spaces need not be weakly path-connected. On the other hand, d_2-paths are weak paths in the 2^n-topology. This follows from the fact that if $\mathbf{p}_{i+1} \in E(\mathbf{p}_i)$, then $N(\mathbf{p}_{i+1}) \cap N(\mathbf{p}_i) \neq \varnothing$ for the basic neighborhoods in the 2^n-topology.

In analogy to path connectivity, we also have that weak path-connectivity implies weak connectivity in digital spaces. For suppose that $S \subset \mathbb{Z}^2$ is weakly path-connected but not weakly connected. Then there exist open sets *U* and *V* such that $S \subset U \cup V$, $S \cap U \neq \varnothing \neq S \cap V$, and $U \cap V = \varnothing$. Let $\mathbf{p} \in S \cap U$, $\mathbf{q} \in S \cap V$, and *P* a weak path from **p** to **q** in *S*. Then $P \subset U \cup V$, $P \cap U \neq \varnothing \neq P \cap V$, and $U \cap V = \varnothing$. Thus *P* is not weakly connected, contrary to Theorem **2.22.4**. The next theorem shows that the converse also holds in the von Neumann as well as the 2^n-topology.

2.22.5 **Theorem.** *Let $\mathbb{Z}^n$ be the digital space with either the von Neumann or the 2^n-topology. If $S \subset \mathbb{Z}^2$, then S is weakly connected $\iff$ S is weakly path-connected.*

Proof: We already know that weak path-connectivity implies weak connectivity in digital spaces. To prove the converse, suppose that *S* is weakly connected. Let $\mathbf{p} \in S$ and $A_{\mathbf{p}} = \{\mathbf{q} \in S : \text{there is a weak path from } \mathbf{p} \text{ to } \mathbf{q} \text{ in } S\}$. If $A_{\mathbf{p}} = S$, then there is nothing to prove. So suppose that $A_{\mathbf{p}} \neq S$. Let $B_{\mathbf{p}} = S \backslash A_{\mathbf{p}}$,

$$U = \bigcup_{\mathbf{q} \in A_{\mathbf{p}}} N(\mathbf{q}), \text{ and } V = \bigcup_{\mathbf{x} \in B_{\mathbf{p}}} N(\mathbf{x}),$$

where $N(\mathbf{q})$ and $N(\mathbf{x})$ denote basic neighborhoods.

If $\mathbf{x} \in B_{\mathbf{p}}$, then $N(\mathbf{x}) \cap N(\mathbf{q}) = \varnothing$ for every $\mathbf{q} \in A_{\mathbf{p}}$. For if $N(\mathbf{x}) \cap N(\mathbf{q}) \neq \varnothing$ for some $\mathbf{q} \in A_{\mathbf{p}}$, let *P* be a weak path from **p** to **q** in *S*. Then $P \cup \{\mathbf{x}\}$ is a weak path from **p** to **x** in *S* and, therefore, $\mathbf{x} \in A_{\mathbf{p}}$, contrary to the fact that $A_{\mathbf{p}} \cap B_{\mathbf{p}} = \varnothing$. It follows that $U \cap V = \varnothing$, $U \cap S = U \cap A_{\mathbf{p}} \neq \varnothing$, $V \cap S = V \cap B_{\mathbf{p}} \neq \varnothing$, and $S \subset U \cup V$. but this contradicts the hypothesis that *S* is weakly connected. Therefore $A_{\mathbf{p}} = S$.

Q.E.D.

Although d_2-connectivity implies weak path-connectivity and, hence, weak connectivity, the converse does not hold in either the von Neumann topology or the 2^n-topology. For example, if $(i, j) \in \mathbb{Z}^2$ with *i* and *j* both even integers, then $P = \{(i, j), (i + 2, j)\}$ is weakly connected and hence weakly path-connected but not d_2-connected in the 2^n-topology. Note, however, that if we shift the set one unit in the diagonal direction, then the shifted set $\{(i + 1, j + 1), (i + 3, j + 1)\}$ is not weakly path-connected or weakly connected. This is due to the fact that shifts are not continuous transformations in this topology. As we shall show, d_2-paths are the only weak paths that are shift invariant weak paths in the 2^n-topology.

If $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_n)$ are points in $\mathbb{Z}^n$, and $S \subset \mathbb{Z}^n$, then we define

$$\mathbf{p} + \mathbf{q} = (p_1 + q_1, \ldots, p_n + q_n) \quad \text{and} \quad S + \mathbf{p} = \{\mathbf{p} + \mathbf{q} : \mathbf{q} \in S\}. \tag{2.22.1}$$

A weak path $P = \{\mathbf{p}_1, \ldots, \mathbf{p}_k\}$ is a *translation invariant* or *shift invariant weak path* if $P + \mathbf{p} = \{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$, where $\mathbf{q}_i = \mathbf{p}_i + \mathbf{p}$ for $1 \leq i \leq k$, is a weak path for every $\mathbf{p} \in \mathbb{Z}^{\mathbf{n}}$.

2.22.6 **Theorem.** *Let $\mathbb{Z}^n$ be the digital space with the 2^n-topology and $S \subset \mathbb{Z}^n$. Then S is d_2-connected $\iff$ for each pair of points $\mathbf{p}, \mathbf{q} \in S$ there exists a translation invariant weak path P from* **p** *to* **q** *in S.*

In particular, P is a d_2-path $\iff$ P is a translation invariant weak path.

Proof: Suppose *S* is d_2-connected. Let $\mathbf{p}, \mathbf{q} \in S$, $P = \{\mathbf{p_1}, \ldots, \mathbf{p}_k\}$ a d_2-path from $\mathbf{p}$ to $\mathbf{q}$ in *S*, and $\mathbf{x} \in \mathbb{Z}^n$. Since $\mathbf{p}_{i+1} + \mathbf{x} \in E(\mathbf{p}_i) + \mathbf{x} = E(\mathbf{p}_i + \mathbf{x})$, $P + \mathbf{x}$ is a d_2-path in $S + \mathbf{x}$. Since d_2-paths are weak paths, *P* is a translation invariant weak path.

To prove the converse, suppose that for each pair $\mathbf{p}, \mathbf{q} \in S$, there exists a translation invariant weak path from $\mathbf{p}$ to $\mathbf{q}$ in *S*. If *S* is not d_2-connected, then for some pair of points $\mathbf{p}, \mathbf{q} \in S$, there does not exist a d_2-path from $\mathbf{p}$ to $\mathbf{q}$ in *S*. Let $P = \{\mathbf{p}_1, \ldots, \mathbf{p}_k\}$ be a translation invariant weak path from $\mathbf{p}$ to $\mathbf{q}$ in *S*. Since *P* is not a d_2-path, we have that for some i $\mathbf{p}_{i+1} \notin E(\mathbf{p}_i)$, where $1 \leq i \leq k-1$. Let $\mathbf{p}_{i+1} = (p_1, \ldots, p_n)$ and define $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{Z}^n$ by

$$x_j = \begin{cases} 0 & \text{if } p_j \text{ is odd} \\ 1 & \text{if } p_j \text{ is even,} \end{cases}$$

where $1 \leq j \leq n$. Then the coordinates of the point $\mathbf{p}_{i+1} + \mathbf{x}$ are all odd and, therefore, $N(\mathbf{p}_{i+1} + \mathbf{x}) = \{\mathbf{p}_{i+1} + \mathbf{x}\}$. But then

$$N(\mathbf{p}_{i+1}+\mathbf{x}) \cap N(\mathbf{p}_i+\mathbf{x}) \subset N(\mathbf{p}_{i+1}+\mathbf{x}) \cap E(\mathbf{p}_i+\mathbf{x}) = \{\mathbf{p}_{i+1}+\mathbf{x}\} \cap N(\mathbf{p}_i+\mathbf{x}) = \varnothing$$

since $\mathbf{p}_{i+1}+\mathbf{x} \notin E(\mathbf{p}_i)+\mathbf{x} = E(\mathbf{p}_i+\mathbf{x})$. This contradicts the hypothesis that *P* is a translation invariant weak path.

Q.E.D.

A *shift* by a vector $\mathbf{x} \in \mathbb{Z}^\mathbf{n}$ is a function $f_\mathbf{x} : \mathbb{Z}^n \to \mathbb{Z}^n$ defined by $f_\mathbf{x}(\mathbf{p}) = \mathbf{p} + \mathbf{x}$ $\forall$ $\mathbf{p} \in \mathbb{Z}^\mathbf{n}$. The fundamental problem with both the von Neumann and the 2^n-topology is that they are not shift invariant; a shift $f_\mathbf{x}$ is not necessarily continuous. Shifts are important operations in image processing, while continuous functions preserve such important features as connectedness. Therefore it would be desirable to have shifts represented as continuous functions. There are topologies on $\mathbb{Z}^n$ that provide for continuity of shifts. The discrete and indiscrete topologies are two such examples. However, these topologies do not provide the useful properties we associate with $\mathbb{R}^n$ such as the classification of arcs, the Jordan Curve Theorem, and the various properties of surfaces embedded in $\mathbb{R}^n$. Digital images are usually viewed as discrete representations of regions in $\mathbb{R}^n$. Thus we would like properties of $\mathbb{R}^n$ to carry over into the discrete domain. As we have seen, the von Neumann and 2^n-topologies preserve many of these useful properties. In addition, even though shifts are not continuous functions in these topologies, shifts do preserve connectivity in the von Neumann topology and weak connectivity of d_2-connected sets in the 2^n-topology.

There are many other topics in digital topology that provide for a theoretical foundation of many important image processing operations. We refer the reader interested in this subject to [29] and [44] for further references.

Bibliography

[1] T.M. Apostol. *Mathematical Analysis*. Addison-Wesley, Reading, MA, 1964.

[2] K.E. Batcher. Design of a massively parallel processor. *IEEE Transactions on Computers*, 29(9):836–840, 1980.

[3] G. Birkhoff and J. Lipson. Heterogeneous algebras. *Journal of Combinatorial Theory*, 8:115–133, 1970.

[4] J. Chassery. Connectivity and consecutivity in digital pictures. *Computer Vision, Graphics, and Image Processing*, 9(3):294–300, 1979.

[5] E.L. Cloud. The geometric arithmetic parallel processor. In *Proceedings Frontiers of Massively Parallel Processing*. George Mason University, October 1988.

[6] E.L. Cloud. Geometric arithmetic parallel processor: Architecture and implementation. In V.K. Prasanna, editor, *Parallel Architectures and Algorithms for Image Understanding*, Boston, MA., 1991. Academic Press, Inc.

[7] E.L. Cloud and W. Holsztynski. Higher efficiency for parallel processors. In *Proceedings IEEE Southcon 84*, pages 416–422, Orlando, FL, March 1984.

[8] T.R. Crimmins and W.M. Brown. Image algebra and automatic shape recognition. *IEEE Transactions on Aerospace and Electronic Systems*, AES-21(1):60–69, January 1985.

[9] D. Crookes, P.J. Morrow, and P.J. McParland. An algebra-based language for image processing on transputers. *IEE Third Annual Conference on Image Processing and its Applications*, 307:457–461, July 1989.

[10] D. Crookes, P.J. Morrow, and P.J. McParland. An implementation of image algebra on transputers. Technical report, Department of Computer Science, Queen's University of Belfast, Northern Ireland, 1990.

[11] J.L. Davidson. *Lattice Structures in the Image Algebra and Applications to Image Processing*. PhD thesis, University of Florida, Gainesville, FL, 1989.

[12] J.L. Davidson. Classification of lattice transformations in image processing. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 57(3):283–306, May 1993.

[13] J.L. Davidson and F. Hummer. Morphology neural networks: An introduction with applications. *IEEE Systems Signal Processing*, 12(2):177–210, 1993.

[14] J.L. Davidson and A. Talukder. Template identification using simulated annealing in morphology neural networks. In *Second Annual Midwest Electro-Technology Conference*, pages 64–67, Ames, IA, April 1993. IEEE Central Iowa Section.

[15] E.R. Dougherty. Unification of nonlinear filtering in the context of binary logical calculus, part ii: Gray-scale filters. *Journal of Mathematical Imaging and Vision*, 2(2/3):185–192, November 1992.

[16] M. J. B. Duff. Review of CLIP image processing system. In *Proceedings of the National Computing Conference*, pages 1055–1060. AFIPS, 1978.

[17] M.J.B. Duff. Clip4. In K.S. Fu and T. Ichikawa, editors, *Special Computer Architectures for Pattern Processing*, chapter 4, pages 65–86. CRC Press, Boca Raton, FL, 1982.

[18] M.J.B. Duff, D.M. Watson, T.J. Fountain, and G.K. Shaw. A cellular logic array for image processing. *Pattern Recognition*, 5(3):229–247, June 1973.

[19] G.R. Fischer and M.R. Rowlee. Computation of disparity in stereo images on the Connection Machine. In *Image Algebra and Morphological Image Processing*, volume 1350 of *Proceedings of SPIE*, pages 328–334, 1990.

[20] T.J. Fountain, K.N. Matthews, and M.J.B. Duff. The CLIP7A image processor. *IEEE Pattern Analysis and Machine Intelligence*, 10(3):310–319, 1988.

[21] J. Goutsias. On the morphological analysis of discrete random shapes. *Journal of Mathematical Imaging and Vision*, 2(2/3):193–216, November 1992.

[22] H. Hadwiger. *Vorlesungen Über Inhalt, Oberflæche und Isoperimetrie*. Springer-Verlag, Berlin, 1957.

[23] R.M. Haralick, L. Shapiro, and J. Lee. Morphological edge detection. *IEEE Journal of Robotics and Automation*, RA-3(1):142–157, April 1987.

[24] R.M. Haralick, S.R. Sternberg, and X. Zhuang. Image analysis using mathematical morphology: Part I. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(4):532–550, July 1987.

[25] W.D. Hillis. *The Connection Machine*. The MIT Press, Cambridge, MA, 1985.

[26] J.G. Hocking and G.S. Young. *Topology*. Addison-Wesley, Reading, MA, 1961.

[27] E. Kamke. *Theory of Sets*. Dover Publications, Inc., New York, 1950.

[28] J.C. Klein and J. Serra. The texture analyzer. *Journal of Microscopy*, 95, 1972.

[29] T.Y. Kong and A. Rosenfeld. Digital topology: Introduction and survey. *Computer Vision, Graphics, and Image Processing*, 48(3):357–393, 1989.

[30] L. Koskinen and Jaako Astola. Asymptotic behaviour of morphological filters. *Journal of Mathematical Imaging and Vision*, 2(2/3):117–136, November 1992.

[31] R.M. Lougheed. A high speed recirculating neighborhood processing architecture. In *Architectures and Algorithms for Digital Image Processing II*, volume 534 of *Proceedings of SPIE*, pages 22–33, 1985.

[32] R.M. Lougheed and D.L. McCubbrey. The cytocomputer: A practical pipelined image processor. In *Proceedings of the Seventh International Symposium on Computer Architecture*, pages 411–418, 1980.

[33] P. Maragos. *A Unified Theory of Translation-Invariant Systems with Applications to Morphological Analysis and Coding of Images*. Ph.D. dissertation, Georgia Institute of Technology, Atlanta, 1985.

[34] P. Maragos and R.W. Schafer. Morphological skeleton representation and coding of binary images. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(5):1228–1244, October 1986.

[35] P. Maragos and R.W. Schafer. Morphological filters Part I: Their set-theoretic analysis and relations to linear shift-invariant filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-35:1153–1169, August 1987.

[36] P. Maragos and R.W. Schafer. Morphological filters Part II : Their relations to median, order-statistic, and stack filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-35:1170–1184, August 1987.

[37] G. Matheron. *Random Sets and Integral Geometry*. Wiley, New York, 1975.

[38] P.E. Miller. Development of a mathematical structure for image processing. Optical division tech. report, Perkin-Elmer, 1983.

[39] H. Minkowski. Volumen und oberflache. *Mathematische Annalen*, 57:447–495, 1903.

[40] H. Minkowski. *Gesammelte Abhandlungen*. Teubner Verlag, Leipzig-Berlin, 1911.

[41] M.H. Newman. *Elements of the Topology of Plane Point Sets*. Cambridge University Press, Cambridge, England, 1961.

[42] A.V. Oppenheim and R.W. Schafer. *Discrete-Time Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1989.

[43] E.R. Phillips. *An Introduction to Analysis and Integration Theory*. Intext Educational Publishers, Scanton, PA, 1971.

[44] G.X. Ritter. Topology of computer vision. In *Topology Proceedings*, volume 12, pages 117–158, Auburn University Press, 1987.

[45] G.X. Ritter. Image algebra with applications. Unpublished manuscript, available via anonymous ftp from `ftp://ftp.cise.ufl.edu/pub/src/ia/documents`, 1994.

[46] G.X. Ritter and J.N. Wilson. *Handbook of Computer Vision Algorithms in Image Algebra*. CRC Press, Boca Raton, 1996.

[47] A. Rosenfeld. Arcs and curves in digital pictures. *Journal of the ACM*, 20, 1976.

[48] A. Rosenfeld. Digital topology. *American Math Monthly*, 86, 1979.

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

[50] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, New York, NY, 1974.

[51] D. Schonfeld and J. Goutsias. Optimal morphological pattern restoration from noisy binary images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(1):14–29, January 1991.

[52] J. Serra. Introduction a la morphologie mathematique. Booklet no. 3, Cahiers du Centre de Morphologie Mathematique, Fontainebleau, France, 1969.

[53] J. Serra. Morphologie pour les fonctions “a peu pres en tout ou rien”. Technical report, Cahiers du Centre de Morphologie Mathematique, Fontainebleau, France, 1975.

[54] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, London, 1982.

[55] E. Spanier. *Algebraic Topology*. McGraw-Hill, New York, 1966.

[56] S. R. Sternberg. Language and architecture for parallel image processing. In *Proceedings of the Conference on Pattern Recognition in Practice*, Amsterdam, May 1980.

[57] S.R. Sternberg. Biomedical image processing. *Computer*, 16(1), January 1983.

[58] S.R. Sternberg. Overview of image algebra and related issues. In S. Levialdi, editor, *Integrated Technology for Parallel Image Processing*. Academic Press, London, 1985.

[59] L. Uhr. Pyramid multi-computer structures, and augmented pyramids. In M.J.B. Duff, editor, *Computing Structures for Image Processing*, pages 95–112. Academic Press, London, 1983.

[60] L. Uhr. *Algorithm-Structured Computer Arrays and Networks*. Academic Press, New York, NY, 1984.

[61] S.H. Unger. A computer oriented toward spatial problems. *Proceedings of the IRE*, 46:1144–1750, 1958.

[62] University of Florida Center for Computer Vision and Visualization. *Software User Manual for the `iac++` Class Library*, version 1.0 edition, 1994. Avialable via anonymous ftp from `ftp.cise.ufl.edu` in `/pub/ia/documents`.

[63] J. von Neumann. The general logical theory of automata. In *Cerebral Mechanism in Behavior: The Hixon Symposium*. John Wiley & Sons, New York, NY, 1951.

[64] J. von Neumann. *Theory of Self-Reproducing Automata*. University of Illinois Press, Urbana, IL, 1966.

[65] J.N. Wilson. An introduction to image algebra Ada. In *Image Algebra and Morphological Image Processing II*, volume 1568 of *Proceedings of SPIE*, pages 101–112, San Diego, CA, July 1991.

[66] J.N. Wilson. Supporting image algebra in the C++ language. In *Image Algebra and Morphological Image Processing IV*, volume 2030 of *Proceedings of SPIE*, pages 315–326, San Diego, CA, July 1993.

[67] J.N. Wilson, G.R. Fischer, and G.X. Ritter. Implementation and use of an image processing algebra for programming massively parallel computers. In *Frontiers '88: The Second Symposium on the Frontiers of Massively Parallel Computation*, pages 587–594, Fairfax, VA, 1988.

[68] J.N. Wilson, D.C. Wilson, G.X. Ritter, and D. Langhorne. Image algebra FORTRAN language, version 3.0. Technical Report TR-89–03, University of Florida CIS Department, Gainesville, 1989.

CHAPTER 3
ELEMENTS OF ABSTRACT ALGEBRA

If one surveys the subjects of arithmetic, elementary algebra, or matrix theory, certain features stand out. One notes that these subjects deal with some given or derived set of objects, usually numbers or symbolic expressions, and with rules for combining these objects. Examples of these are the set of real numbers, the set of real valued functions on a set X, and the set of complex valued $n \times n$ square matrices with the usual rules of addition, subtraction, and multiplication. Moreover, one finds that there are some properties which these combining operations have in common: e.g. adding zero to any real number, adding the zero function to a function, or adding the zero matrix to a matrix does not change the value of the real number, the function, or the matrix, respectively. Other properties, such as commutativity, do not always hold. Multiplication of square matrices is, in general, not a commutative operation.

Abstract algebra aims at providing a fuller understanding of these subjects through a systematic study of typical mathematical structures. Such a study has the advantage of economy in that many superficially distinct structures are found to be basically the same, and hence open to a unified treatment.

Image algebra has an analogous goal in that it aims at providing a deeper understanding of image processing through a systematic study of image processing operations. Various structures in the image algebra are equivalent to those studied in abstract algebra. Familiarity with some of these structures is, therefore, essential to the understanding of image algebra.

3.1 Relations and Operations on Sets

A binary relation $\mathcal{R}$ on a set X is, intuitively, a proposition such that for each ordered pair (x, y) of elements of X , one can determine whether $x\mathcal{R}y$ is or is not true. Here, $x\mathcal{R}y$ means that "x is related (by the relation $\mathcal{R}$) to y." For example, if L is the set of all lines in a plane, then "is parallel to" or "is perpendicular to" are binary relations on L.

The notion of a binary relation on a set can be rigorously defined by stating it formally in terms of the set concept.

3.1.1 Definition: A *binary relation* $\mathcal{R}$ on a set X is a subset $\mathcal{R} \subset X \times X$.

Thus, any subset $\mathcal{R}$ of $X \times X$ is a binary relation on X and if such a subset is being used to define a relation on X, then it is customary to write $x\mathcal{R}y$ for $(x, y) \in \mathcal{R}$.

3.1.2 Examples:

(i) Set inclusion is a relation on any power set. In particular, let X be any set and

$$\mathcal{R} = \{(A, B) : A \subset B,\ A, B \in 2^X\}.$$

Then $\mathcal{R}$ is a binary relation on 2^X.

(ii) The relation of *less or equal*, $\leq$, between real numbers is the set $\{(x, y) : x \leq y\} \subset \mathbb{R} \times \mathbb{R}$.

(iii) For any set X, the diagonal $\Delta = \{(x, x) : x \in X\}$ is the relation of equality.

(iv) The *inverse relation* of $\mathcal{R}$, denoted by $\mathcal{R}^{-1}$, is the relation $\mathcal{R}^{-1} = \{(y,x) : (x,y) \in \mathcal{R}\}$. Thus, the inverse relation of $\leq$ in (ii) above is the relation of *greater or equal* $\geq$.

Note that in binary relations, each pair of elements need not be related. For instance, in (iii) above, if $x, y \in X$ and $x \neq y$, then neither (x,y) nor (y,x) are in Δ.

An obvious generalization of Definition **3.1.1** is to define any subset of $X \times Y$ to be a binary relation between the elements of X and those of Y; thus, a function $f : X \to Y$ is a special type of a binary relation between X and Y . Operations between images and templates as defined in Chapter **4** provide examples of binary relations between elements of different sets that are pertinent to image processing.

Certain relations on a set allow elements of that set to be arranged in some order. For example, when a child arranges a set of sticks in order, from longest to shortest, he has an intuitive grasp of the relation "is longer than." From this example we can see that there are at least two properties which a relation $\mathcal{R}$ must have if it is to order a set. Specifically:

$\mathcal{R}$ must be antisymmetric. That is, given two sticks, one of them must be longer than the other. Otherwise, they could not be given relative positions in the order.

$\mathcal{R}$ must be transitive. That is, given three sticks x, y, and z, with x longer than y and y longer than z, then x must be longer than z.

We collect these ideas in a definition.

3.1.3 **Definition.** A relation $\preceq$ on a set X is called a *partial order* on X if and only if for every $x, y, z \in X$ the following three conditions are satisfied:

(i) $x \preceq x \quad (\text{reflexive})$

(ii) $x \preceq y \text{ and } y \preceq x \Rightarrow x = y \quad (\text{anti} - \text{symmetric})$

(iii) $x \preceq y \text{ and } y \preceq z \Rightarrow x \preceq z \quad (\text{transitive})$

The relations defined in Example **3.1.2** are all partial order relations. The relation of less or equal given in Example **3.1.2** (ii) is also called the *natural order* on $\mathbb{R}$.

A set X together with the partial order $\preceq$, i.e. the pair $(X, \preceq)$, is called a *partially ordered set*. If $x \preceq y$ in a partially ordered set, then we say that x *precedes* y or that x is *smaller* than y and that y *follows* or is *larger* than x.

If $\mathcal{R} = \preceq$ is a partial order on X, then it is easy to see that the inverse relation $\mathcal{R}^{-1}$, denoted by $\succeq$, is also a partial order on X. The inverse partial order relation $\succeq$ is also called the *dual* of $\preceq$ and gives rise to the following definition:

3.1.4 **Definition.** The *dual* of a partially ordered set X is that partially ordered set X^* defined by the inverse partial order relation on the same elements.

Since $(X^*)^* = X$, this terminology is legitimate.

Note that Definition **3.1.3** does *not* imply that given $x, y \in X$, then either $x \preceq y$ or $y \preceq x$; that is, in a partially ordered set not every pair of elements need to be related. A partially ordered set in which every pair of elements is related under the order relation is called a *totally* (or *linearly*) ordered set. The set $\mathbb{R}$ together with the natural order of $\leq$ is an example of a totally ordered set. On the other hand, the

relation of set inclusion (Example **3.1.2**(i)) is a partial order which is not a total order. An extremely useful special case of a linear order is provided by the next example.

3.1.5 **Example:** Let X and Y be totally ordered. Then the product set $X \times Y$ can be totally ordered as follows:

$$(x, y) \preceq (x', y') \text{ if } x \preceq x' \text{ or if } x = x' \text{ and } y \preceq y' .$$

This order is called the *lexicographical order* on $X \times Y$ as it is similar to the way words are arranged in a dictionary. For example, suppose that $\mathbf{X} = X \times Y$, where $X = \{1, 2, 3\}$ and $Y = \{1, 2, 3, 4\}$ (see also Example **2.3.1**). If the integers in X and Y are considered ordered by the natural order of less or equal, then $\mathbf{X}$ is totally ordered by the order relation $(i, j) < (i', j')$ defined above. Thus, if we rename the elements of $\mathbf{X}$ by $\mathbf{x}_k = (i, j) \in \mathbf{X}$, where $k = 4(i - 1) + j$, then $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{12}\}$ and $\mathbf{x}_k < \mathbf{x}_h$ if and only if $k < h$.

If we view the elements (i, j) of $\mathbf{X}$ as the usual ith row and jth column location of a matrix array, then this order is also known as the *common scanning order* of the matrix array $\mathbf{X}$. This corresponds to the usual way a computer reads (scans) the entries of a matrix; namely, row by row from left to right, starting with the top first row then the second, and so on, until the last or bottom row is read.

If $X \subset Y$, then $X \times X \subset Y \times Y$. Thus, if $\mathcal{R}$ is a binary relation on Y, then $\mathcal{R} \cap (X \times X)$ is a binary relation on X. We call the relation $\mathcal{R} \cap (X \times X)$ *the relation induced by* $\mathcal{R}$ *on* X. In particular, a binary relation on Y induces a definite binary relation on every subset of Y. For example, the natural order relation $\mathcal{R} = \leq$ on $Y = \mathbb{R}$ induces the natural order $\leq$ on the set of integers $X = \mathbb{Z}$.

One of the most fundamental relations between elements of a set is that of equivalence. Equivalence relations are used in practically all fields of mathematics; they arise whenever one desires to regard all those members of set that have some preassigned characteristic as a single entity.

3.1.6 **Definition.** A relation $\mathcal{R}$ on a set X is called an *equivalence relation* if it satisfies the following three conditions:

(1) $x \in X \Rightarrow x \mathcal{R} x$ (reflexive)

(2) $x \mathcal{R} y \Rightarrow y \mathcal{R} x$ (symmetric)

(3) $x \mathcal{R} y$ and $y \mathcal{R} z \Rightarrow x \mathcal{R} z$ (transitive)

If $\mathcal{R}$ an equivalence relation and $x \mathcal{R} y$, then we say that x and y are *equivalent*.

3.1.7 **Examples:**

(i) Consider the relation $\mathcal{R}$ of set inclusion (Example **3.1.2**(i)). For each $A \subset 2^X$, $A \subset A$ and if $A \subset B$ and $B \subset C$, then $A \subset C$. Hence, the relation $\mathcal{R} = \subset$ is both reflexive and transitive. On the other hand, $A \subset B$ and $A \neq B \Rightarrow B \not\subset A$. Accordingly, $\subset$ is not symmetric and hence not an equivalence relation.

(ii) The diagonal relation Δ of Example **3.1.2**(iii) is an equivalence relation.

(iii) Let $f : X \to Y$ be a function and $\mathcal{R} = \{(x, x') : f(x) = f(x')\}$. Then $\mathcal{R}$ is an equivalence relation on X.

If $\mathcal{R}$ is an equivalence relation on X, then the *equivalence class* of any element $x \in X$, denoted by $[x]$, is the set

$$[x] = \{y : y\mathcal{R}x,\ y \in X\} .$$

The collection of equivalence classes of X, denoted by $X/\mathcal{R}$, is called the *quotient set* of X by $\mathcal{R}$. Thus,

$$X/\mathcal{R} = \{[x] : x \in X\}$$

The quotient set $X/\mathcal{R}$ possesses the following properties:

3.1.8 **Theorem.** *Let $\mathcal{R}$ be an equivalence relation on X. Then*

(1) *For every $x \in X$, $x \in [x]$*

(2) $[x] = [y] \Leftrightarrow x\mathcal{R}y$

(3) $[x] \neq [y] \Leftrightarrow [x] \cap [y] = \varnothing$

Proof: (1) Since $\mathcal{R}$ is reflexive, we have $x\mathcal{R}x$ and, hence, $x \in [x]$.

(2) Suppose $[x] = [y]$. By part (1), $x \in [x] = [y]$. Thus, $x\mathcal{R}y$. To prove the converse, let $z \in [y]$. Then $z\mathcal{R}y$. By symmetry, $y\mathcal{R}z$. We now have $y\mathcal{R}z$ and, by hypothesis, $x\mathcal{R}y$. Hence, by transitivity, $x\mathcal{R}z$. Again, by symmetry, $z\mathcal{R}x$ and, therefore, $z \in [x]$, which shows that $[x] \subset [y]$. Arguing in a similar fashion, we can show that $[y] \subset [x]$ and, hence, $[x] = [y]$.

(3) Suppose the conclusion is false, i.e. $[x] \cap [y] \neq \varnothing$. Then $\exists z \in X$ with $z \in [x] \cap [y]$. Hence, $z\mathcal{R}x$ and $z\mathcal{R}y$. By symmetry, $x\mathcal{R}z$. Since $x\mathcal{R}z$ and $z\mathcal{R}y$, we have by transitivity that $z\mathcal{R}y$. It now follows from part (2) that $[x] = [y]$ which contradicts the hypothesis. The converse argument is just as easy.

Q.E.D.

A collection $\{A_\lambda\}_{\lambda\in\Lambda}$ of subsets of X is called a *partition* of X if the following conditions are satisfied:

$$(1)\ \bigcup_{\lambda\in\Lambda} A_\lambda = X \text{ and}$$

$$(2)\ A_\lambda \cap A_\gamma = \varnothing \text{ whenever } \lambda \neq \gamma\ \ (\lambda, \gamma \in \Lambda).$$

The following fundamental theorem of equivalence relations is a consequence of Theorem **3.1.8**:

3.1.9 **Theorem:** *If $\mathcal{R}$ is an equivalence relation on X, then X/R is a partition of X.*

Proof: Obviously, $\bigcup_{x\in X} [x] \subset X$. If $y \in X$, then by Theorem **3.1.8**(1) $y \in [y] \subset \bigcup_{x\in X} [x]$. Hence, $X \subset \bigcup_{x\in X} [x]$. The remainder of the proof follows immediately from Theorem **3.1.8**(3).

Q.E.D.

Each element y of an equivalence class $[x]$ (i.e., each $y \in [x]$) is called a *representative* of $[x]$. Note that if y is a representative of $[x]$, then $[y] = [x]$.

3.1.10 Examples:

(i) Two integers are said to have the same *parity* if either both are even or both are odd. The relation "has the same parity as" on $\mathbb{Z}$ is an equivalence relation and partitions $\mathbb{Z}$ into two equivalence classes. In particular, if $n \in \mathbb{Z}$, then

$$[2n] = \{k \in \mathbb{Z} : k \text{ is even}\} \text{ and } [2n+1] = \{k \in \mathbb{Z} : k \text{ is odd}\}$$

Clearly, $[2n] \cup [2n+1] = \mathbb{Z}$ and $[2n] \cap [2n+1] = \varnothing$.

(ii) Each equivalence class resulting from the relation Δ of Example **3.1.2**(iii) contains exactly one element.

(iii) In Example **3.1.7**(iii), the equivalence classes are the sets $\{f^{-1}(y) : y \in f(X)\}$.

3.1.11 Definition. For arbitrary integers m, n and r, we say that m *is congruent to* r *modulo* n (or m *is congruent to* $r \ mod \ n$), and write $m \equiv r \ (mod \, n)$, if the difference $m - r$ is an integral multiple of n; that is, if $m = nk + r$ for some integer k.

It is easily verified that *congruence modulo* n, i.e. the set $(n) = \{(m, r) : m \equiv r \ (mod \, n)\}$, is an equivalence relation on $\mathbb{Z}$. For the relation of congruence modulo n is obviously reflexive and symmetric. The transitivity also follows easily: If $m = nk + r$ and $r = nj + q$ for some integers k and j, then $m = n(k + j) + q$, so that $m \equiv q(mod \, n)$.

If $\mathcal{R} = (n)$ denotes the relation on $\mathbb{Z}$ defined by "m is congruent to $r \ mod \ n$", then the quotient $\mathbb{Z}/(n)$ is called *the set of integers mod n.*

3.1.12 Example: Let $\mathcal{R} = (5)$ be the relation on $\mathbb{Z}$ defined by $m \equiv r(mod \, 5)$. Then there are exactly five distinct equivalence classes in $\mathbb{Z}/(5)$:

$$A_0 = \{\ldots, -10, -5, 0, 5, 10, \ldots\} = \cdots = [-10] = [-5] = [0] = [5] = [10] = \cdots$$
$$A_1 = \{\ldots, -9, -4, 1, 6, 11, \ldots\} = \cdots = [-9] = [-4] = [1] = [6] = [11] = \cdots$$
$$A_2 = \{\ldots, -8, -3, 2, 7, 12, \ldots\} = \cdots = [-8] = [-3] = [2] = [7] = [12] = \cdots$$
$$A_3 = \{\ldots, -7, -2, 3, 8, 13, \ldots\} = \cdots = [-7] = [-2] = [3] = [8] = [13] = \cdots$$
$$A_4 = \{\ldots, -6, -1, 4, 9, 14, \ldots\} = \cdots = [-6] = [-1] = [4] = [9] = [14] = \cdots$$

Observe that each integer m is uniquely expressible in the form $m = 5n + r$, where $0 \leq r < 5$ and $r \in A_r$ is the remainder. Clearly, $\mathbb{Z} = \bigcup_{i=0}^{4} A_i$ and $A_i \cap A_j = \varnothing$ whenever $i \neq j$.

In Example **3.1.10**(i), there are exactly two equivalent classes, the set of even integers and the set of odd integers, and 0 and 1 are representatives of these classes. Although 2 and 3 are also representatives of these classes, it is customary to let 0 and 1 *represent* these classes; i.e., to identify the set $\mathbb{Z}_2 = \{0, 1\}$ with the set $\mathbb{Z}/(2) = \{[2n], [2n-1]\}$ by identifying 0 with $[2n]$ and 1 with $[2n+1]$. In general, it is customary to let $\mathbb{Z}_n = \{0, 1, 2, \ldots, n-1\}$ denote the set $\mathbb{Z}/(n)$ as these two sets are in one-to-one correspondence under the function $i \to [i] \in \mathbb{Z}/(n)$.

It is easy to show that for any given $k \in \mathbb{Z}^+$, the set $\mathbb{Z}/\left(2^k\right) \approx \mathbb{Z}_{2^k}$ is in one-to-one correspondence with the set $\prod_{i=1}^{k} \mathbb{Z}_2 = (\mathbb{Z}_2)^k$ (we leave it to the reader to convince himself of this fact). Thus, the elements

of $\mathbb{Z}_{2^k}$ can be uniquely identified with the elements of $\prod_{i=1}^{k} \{0,1\}$; i.e. with binary numbers of fixed length k. This corresponds to the usual representation of digital image values in digital image processing by computers (Section **2.18**).

3.1.13 **Example:** The set of integers mod 8 consists of eight equivalence classes; namely $\mathbb{Z}_{2^3} = \{[0],[1],[2],[3],[4],[5],[6],[7]\}$. Identifying [0] with (0,0,0), [1] with (0,0,1), [2] with (0,1,0), [3] with (0,1,1), [4] with (1,0,0), [5] with (1,0,1), [6] with (1,1,0), [7] with (1,1,1) provides for a unique one-to-one correspondence between $\mathbb{Z}_{2^3}$ and $\{0,1\} \times \{0,1\} \times \{0,1\} = (\mathbb{Z}_2)^3$.

A binary operation is a relation between sets which provides a rule for combining two arbitrary elements of one or two sets. The precise definition is as follows:

3.1.14 **Definition.** Let X, Y, and Z be three (not necessarily distinct) sets. A *binary operation* $\bigcirc$ *between X and* Y *with resultant* in Z is a function $\bigcirc : X \times Y \to Z$. If $X = Y = Z$, then $\bigcirc$ is simply called a *binary operation on* X.

The evaluation $\bigcirc(x,y)$ is commonly denoted by $x \bigcirc y$ and is called the *resultant* of the operation. Thus, if $(x,y) \in X \times Y$, then $x \bigcirc y = z \in Z$. Binary operations between sets play an important role in image processing. In this chapter, however, we will deal mostly with binary operations on a set.

Addition, multiplication, and division are examples of binary operations on $\mathbb{R}^+$. Addition and multiplication are also binary operations on $\mathbb{R}$. Due to the fact that for any pair of numbers of form $(r,0)$, $r/0$ is undefined, division is not a binary operation on $\mathbb{R}$. However, it is a binary operation on $\mathbb{R}\backslash\{0\}$.

Some binary operations may satisfy special properties. Commutativity and associativity are the most important of these special properties. A binary operation $\bigcirc$ on a set X is called *commutative* whenever $x \bigcirc y = y \bigcirc x \ \ \forall\, x,y \in X$, and *associative* whenever $(x \bigcirc y) \bigcirc z = x \bigcirc (y \bigcirc z) \ \ \forall\, x,y,z \in X$.

3.1.15 **Example:** Addition and multiplication are commutative and associative binary operations on $\mathbb{R}$. Division is not commutative on $\mathbb{R}^+$. Defining $\bigcirc$ on $\mathbb{R}$ by

$$m \bigcirc n = m + 2n \ \ \forall m,n \in \mathbb{R}$$

then

$$(m \bigcirc n) \bigcirc k = (m+2n) \bigcirc k = (m+2n) + 2k$$

and

$$m \bigcirc (n \bigcirc k) = m \bigcirc (n+2k) = m + 2(n+2k) = m + 2n + 4k.$$

Thus, the operation $\bigcirc$ is not associative. Furthermore, $\bigcirc$ is not commutative since $m \bigcirc n = m + 2n \neq n + 2m = n \bigcirc m$.

A set X is said to have an *identity* element with respect to a binary operation $\bigcirc$ on X if there exists an element $e \in X$ with the property

$$x \bigcirc e = e \bigcirc x = x \quad \forall\, x \in X.$$

The identity element of $\mathbb{R}$ with respect to addition is 0 since $0 + x = x + 0 = x \ \ \forall\, x \in \mathbb{R}$; the identity element of $\mathbb{R}^+$ with respect to multiplication is 1 since $x \cdot 1 = 1 \cdot x = x \ \ \forall\, x \in \mathbb{R}^+$. Observe that $\mathbb{R}^+$ has no identity element with respect to addition. The next theorem shows that identity elements are unique.

3.1.16 **Theorem.** *An identity element, if one exists, of a set X with respect to a binary operation $\bigcirc$ on X is unique.*

Proof: Assume the contrary; that is, assume e_1 and e_2 are two distinct identity elements of X. Then $e_1 \bigcirc e_2 = e_2$ since e_1 is an identity element. Similarly, $e_2 \bigcirc e_1 = e_1$. Therefore, $e_1 = e_1 \bigcirc e_2 = e_2$, which contradicts the fact that e_1 and e_2 are distinct.

Q.E.D.

If a set X has an identity element e with respect to a binary operation $\bigcirc$, then an element $y \in X$ is called an inverse of $x \in X$ provided that $x \bigcirc y = y \bigcirc x = e$. The inverse with respect to addition (also called *additive inverse*) of $x \in \mathbb{R}$ is $-x$ since $x + (-x) = 0$. The inverse with respect to multiplication (also called *multiplicative inverse*) of $x \in \mathbb{R}\backslash\{0\}$ is x^{-1} since $x \cdot x^{-1} = 1$. Note that the set of all $n \times n$ square matrices under matrix multiplication has a multiplicative identity, namely the $n \times n$ identity matrix. However, not every $n \times n$ matrix has a multiplicative inverse.

The proof of the next theorem is similar to the proof of Theorem **3.1.16** and is left as an exercise for the reader.

3.1.17 **Theorem.** *Let $\bigcirc$ be a binary operation on set X. The inverse with respect to $\bigcirc$ of $x \in X$, if it exists, is unique.*

Although not every binary operation on a set X provides for inverse elements, many operations provide for elements that behave almost like inverses. Obviously, if y is the inverse of $x \in X$ with respect to the operation $\bigcirc$, then $x \bigcirc y \bigcirc x = e \bigcirc x = x$ and $y \bigcirc x \bigcirc y = e \bigcirc y = y$. Any element y satisfying the two conditions

$$x \bigcirc y \bigcirc x = x \text{ and } y \bigcirc x \bigcirc y = y \tag{3.1.0}$$

is called a *pseudo inverse* of $x \in X$. Thus, every inverse is a pseudo inverse. However, in Section **4.4** we shall see that the converse does not necessarily hold.

Suppose X is a set with two binary operations $\bigcirc$ and $\bigcirc'$. The operation $\bigcirc$ is said to be *left distributive* with respect to $\bigcirc'$ if

$$x \bigcirc \left(y \bigcirc' z\right) = (x \bigcirc y) \bigcirc' (x \bigcirc z) \ \ \forall\, x, y, z \in X\,. \tag{3.1.1}$$

and right distributive if

$$\left(y \bigcirc' z\right) \bigcirc x = (y \bigcirc x) \bigcirc' (z \bigcirc x) \ \ \forall\, x, y, z \in X\,. \tag{3.1.2}$$

When both 3.1.1 and 3.1.2 hold, we simply say that $\bigcirc$ is distributive with respect to $\bigcirc'$. Note that the right members of 3.1.1 and 3.1.2 are equal whenever $\bigcirc$ is commutative. Obviously, on $\mathbb{R}$, multiplication is distributive with respect to addition. However, division on $\mathbb{R}^+$ is not left distributive over addition. That is, $(y + z)/x = (y/x) + (z/x)$ but $x/(y + z) \neq (x/y) + (x/z)$.

3.2 Groups and Semigroups

We begin our short survey of algebraic structures by listing characteristic features of special abstract algebraic systems that are important in the study of image algebra.

3.2.1 **Definition.** A *groupoid* is any set X together with a binary operation on X. A groupoid whose binary operation is associative is called a *semigroup*.

To be completely precise in denoting a semigroup, we should use some symbolism such as $(X, \bigcirc, =)$, which specifies the set of elements, the binary relation, and the equality relation used to specify the equality of elements, e.g., $x \bigcirc (y \bigcirc z) = (x \bigcirc y) \bigcirc z$. However, it is customary to use either the pair $(X, \bigcirc)$ or simply the letter designation of the set of elements, in this case X, as a designation of the groupoid or semigroup, provided there is no danger of confusion as to the notation being used for binary composition. Also, algebraists as a rule do not use a special symbol "$\bigcirc$" to denote a binary operation different from the usual addition and multiplication. They stick with the conventional additive or multiplicative notation and even call these operations *addition* or *multiplication*, depending on the symbol used. The symbol for addition is of course "+", and for multiplication "$\cdot$". Thus, in place of the notation "$x \bigcirc y$", we shall be using either "$x + y$" or "$x \cdot y$". There is also a sort of gentlemen's agreement that the symbol "0" is used to denote an additive identity and the symbol "1" to denote a multiplicative identity, even though they may not be actually denoting the integers 0 and 1. Of course, if a person is also talking about numbers at the same time, other symbols are used to denote these identities in order to avoid confusion.

To the uninitiated, semigroups may seem too poor in properties to be of much interest. However, the set of all $n \times n$ square matrices under matrix multiplication forms a semigroup. Anyone who has had experience with matrix theory is well aware that this system, far from being too poor in properties to be of interest, is, indeed, extremely rich in properties. Research into the fascinating ramifications of matrix theory has provided the stimulus to a great deal of mathematical development and is an active and growing branch of mathematics.

The set of $n \times n$ square matrices under matrix multiplication has the additional property of having a multiplicative identity. This leads us to the next definition:

3.2.2 **Definition**. A *monoid* is a semigroup with identity.

3.2.3 **Examples:**

(i) Let Y be a set and $X = 2^Y$. Then X together with the operation of union is a monoid. By the laws of set operations (**2.2.1**), union is an associative operation with identity $\varnothing$.

(ii) The set of positive integers $\mathbb{Z}^+$ together with the operation + is not a monoid. There is no identity for + in $\mathbb{Z}^+$. However, $(\mathbb{Z}^+, +)$ is a semigroup.

(iii) The system $(\mathbb{Z}^+, \cdot)$ is a monoid with identity the integer 1.

Of the various possible algebraic systems having a single associative operation, the type known as a *group* has been by far the most extensively studied. Also, the theory of groups is one of the oldest parts of abstract algebra, as well as one particularly rich in applications.

3.2.4 Definition. A *group* is a monoid with the property that each element has an inverse.

It is customary to denote the inverse of an element x in a group X by "x^{-1}" if multiplicative notation is used, and by "$-x$" if additive notation is used.

Recalling the definition of a monoid, we may define a group alternatively as a set X together with a binary operation, say $(X, \cdot)$, such that :

(1) The operation $\cdot$ is associative, i.e., $\forall\, x, y, z \in X,\ x \cdot (y \cdot z) = (x \cdot y) \cdot z$.

(2) There is an identity element $1 \in X$ such that $\forall\, x \in X,\ x \cdot 1 = 1 \cdot x = x$.

(3) $\forall\, x \in X,\ \exists$ an inverse element $x^{-1} \in X$ such that $x \cdot x^{-1} = x^{-1} \cdot x = 1$.

If in addition to these three properties the operation is commutative, then the group X is called an *abelian* group.

3.2.5 Examples:

(i) The set $\mathbb{N}$ with the operation + is not a group. There is an identity element 0, but no inverses for integers greater than 0.

(ii) The set of integers $\mathbb{Z}$ with the operation + is a group. This group is abelian.

(iii) On $\mathbb{Z}_n$ we define a binary operation (which we shall again write as +, although it is certainly not ordinary addition) by $[j] + [k] = [j + k]$. Here j and k are any elements of the respective sets $[j]$ and $[k]$ of $\mathbb{Z}_n$, and the sum $j + k$ is the ordinary sum of j and k. In order to show that we actually have defined an operation, i.e., that the function (operation) $([j], [k]) \to [j] + [k]$ is *well-defined*, we must show that the image element of the pair $([j], [k])$ is uniquely determined by $[j]$ and $[k]$ alone, and does not depend in any way upon the representative elements j of $[j]$ and k of $[k]$ which we happen to choose. So, suppose that i and h h are also arbitrary elements of the sets $[j]$ and $[k]$, respectively. We then have that

$$i = j + an \quad \text{and} \quad h = k + bn$$

for some integers a and b. But then

$$i + h = (j + an) + (k + bn) = (j + k) + (a + b)n$$

by virtue of the associativity and commutativity of addition, and the distributivity of multiplication over addition for the integers. Thus, we have that $[i + h] = [j + k]$, and our operation is well-defined, independent of the choice of representatives of the respective equivalence classes. That this operation of *addition modulo* n is associative follows from the associativity of ordinary integer addition. The identity element is [0] and the inverse of $[k]$ is $[-k]$ (we leave the verification of these two facts to the reader). Thus, $\mathbb{Z}_n$ with this operation of addition forms a group. It also follows from the commutativity of addition of ordinary integers that the group $\mathbb{Z}_n$ is abelian. This group is known as the *group of integers modulo* n.

An important property inherent to all groups is the cancellation law provided by the following theorem:

3.2.6 **Theorem.** *If X is a group with binary operation $\cdot$, then the left and right cancellation law holds in X, that is, $x \cdot y = x \cdot z$ implies $y = z$, and $y \cdot x = z \cdot x$ implies $y = z$, $\forall\, x, y, z \in X$*

Proof: Suppose $x \cdot y = x \cdot z$. Then, multiplying by x^{-1}, the inverse of x, we obtain

$$x^{-1} \cdot (x \cdot y) = x^{-1} \cdot (x \cdot z)\,.$$

By the associative law,

$$\left(x^{-1} \cdot x\right) \cdot y = \left(x^{-1} \cdot x\right) \cdot z\,.$$

By definition of an inverse, $x^{-1} \cdot x = 1$ and, hence,

$$1 \cdot y = 1 \cdot z, \text{ and by definition of identity, } y = z\,.$$

Similarly, from $y \cdot x = z \cdot x$ one can deduce that $y = z$.

Q.E.D.

Note that we had to use the definition of a group in order to prove this theorem.

A common activity among scientists and engineers is to solve problems. Often these problems lead to equations involving some unknown number or quantity x which is to be determined. The simplest equations are the linear ones of the forms $a + x = b$ for the operation of addition, and $a \cdot x = b$ for multiplication. Equations of form $a \cdot x = b$ are in general not solvable in the monoid $(\mathbb{Z}^+, \cdot)$. For instance, $2 \cdot x = 3$ has a solution $x = 3/2$, which is not an integer. However, equations of form $a \cdot x = b$ are always solvable in the structure $(\mathbb{R}^+, \cdot)$. The reason for this is that the structure $(\mathbb{R}^+, \cdot)$ is a group. As the next theorem shows, the properties necessary to solve linear equations within a system are precisely the properties of a group.

3.2.7 **Theorem.** *If X is a group with binary operation $\cdot$, and if a and b are elements of X, then the linear equations $a \cdot x = b$ and $y \cdot a = b$ have unique solutions in X.*

Proof: Note that

$$\begin{aligned} a \cdot \left(a^{-1} \cdot b\right) &= \left(a \cdot a^{-1}\right) \cdot b, && (\text{associative law}) \\ &= 1 \cdot b, && \left(\text{definition of } a^{-1}\right) \\ &= b, && (\text{property of } 1). \end{aligned}$$

Thus, $x = a^{-1} \cdot b$ is a solution of $a \cdot x = b$. In a similar fashion, $y = b \cdot a^{-1}$ is a solution of $y \cdot a = b$.

To show that y is unique, suppose that $y \cdot a = b$ and $y_1 \cdot a = b$. Then $y \cdot a = y_1 \cdot a$, and by Theorem **3.2.6**, $y = y_1$. The uniqueness of x follows similarly.

Q.E.D.

It is important to note that $x = a^{-1} \cdot b$ and $y = b \cdot a^{-1}$ need not be the same unless $\cdot$ is commutative.

Following common mathematical convention, to indicate the composite of an element x of a group with itself n times, where n is a positive integer, we shall write $x^n = x \cdot x \cdot \ldots \cdot x$ (n factors of x) whenever we use multiplicative notation, or, when using additive notation, $nx = x + x + \ldots + x$ (n summands of x).

Similarly, if n is any positive integer, $x^{-n} = \left(x^{-1}\right)^n$ and $-nx = n(-x)$, where $-x$ is, of course, the inverse of x in the additive notation. Still following customary notation, $x^0 = 1$ and $0 \cdot x = 0$, where 1 and 0 represent the identity element in the multiplicative and additive notations, respectively.

The usual rules of power follow at once:

$$x^n \cdot x^m = x^{n+m},$$
$$(x^n)^m = x^{nm},$$
$$nx + mx = (n+m)x\,,$$
$$n(mx) = (nm)x\,.$$

In particular, it follows that $\left(x^{-1}\right)^{-1} = x$.

3.2.8 **Definition.** A group $(X, \cdot)$ is called *cyclic* if for some $g \in X$, every $x \in X$ is of the form $x = g^n$, where $n \in \mathbb{Z}$. The element g is called a *generator* of X.

Clearly, every cyclic group is abelian.

3.2.9 **Examples:**

(i) Let $X = \{\omega_0, \omega_1, \ldots, \omega_5\}$, where

$$\omega_k = cos\frac{2\pi k}{6} + i \cdot sin\frac{2\pi k}{6} = e^{2k\pi i/6}\,,\ k = 0, 1, \ldots, 5 \text{ and } i = \sqrt{-1}\,.$$

Thus, X is the set of solutions of the equation $z^6 = 1$, where $z \in \mathbb{C}$. Each ω_k is called a *sixth root of unity*. X together with the operation of complex multiplication is a cyclic group with generators ω_1 and ω_5. For example, $\omega_4 = \omega_1^4$ and $\omega_4 = \omega_5^2$.

(ii) $(\mathbb{Z}, +)$ is cyclic with generator 1, since in the additive notation we have for every $n \in \mathbb{Z}$, $n = n \cdot 1$. Note that -1 is also a generator for this group since any integer $k$can be expressed as $k = n \cdot (-1)$, where $n = -k$.

(iii) The group of integers modulo n is cyclic with generator [1], since for any $[k] \in \mathbb{Z}_n$, $[k] = k[1]$.

3.3 Permutations

3.3.1 **Definition.** A *permutation of a set* X is a function from X to X which is both one-to-one and onto.

Suppose X is a finite set of n elements, say $X = \{1, 2, \ldots, n\}$, and ρ a permutation of X. No significance is to be given to the fact that X consists of the first n natural numbers, it is only a matter of notational convenience. It is customary to use the notation

$$\rho = \begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ x_1 & x_2 & x_3 & \cdots & x_n \end{pmatrix}$$

$\cdot$	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6
ρ_1	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6
ρ_2	ρ_2	ρ_3	ρ_1	ρ_5	ρ_6	ρ_4
ρ_3	ρ_3	ρ_1	ρ_2	ρ_6	ρ_4	ρ_5
ρ_4	ρ_4	ρ_6	ρ_5	ρ_1	ρ_3	ρ_2
ρ_5	ρ_5	ρ_4	ρ_6	ρ_2	ρ_1	ρ_3
ρ_6	ρ_6	ρ_5	ρ_4	ρ_3	ρ_2	ρ_1

Figure 3.3.1 Products of permutations.

to describe the permutation ρ, where $x_i = \rho(i)$ for $i = 1, \ldots, n$.

Consider, for example, the case where X consists of three elements, say $X = \{1, 2, 3\}$. In this case we have the following six possible permutations:

$$\rho_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \qquad \rho_4 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix},$$
$$\rho_2 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \qquad \rho_5 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix},$$
$$\rho_3 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, \qquad \rho_6 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}.$$

The inverse of a permutation is simply the reverse mapping. For example,

$$\rho_2^{-1} = \begin{pmatrix} 2 & 3 & 1 \\ 1 & 2 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} = \rho_3 .$$

Thus, $\rho_2^{-1}(2) = 1$, $\rho_2^{-1}(3) = 2$, and so forth. Similarly,

$$\rho_4^{-1} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} = \rho_4 ,$$
$$\rho_3^{-1} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} = \rho_2 ,$$

etc. A product of two permutations is simply the composition of the two permutation functions. Since the composition of two one-to-one and onto functions is again a one-to-one and onto function (Theorem **2.5.12**), the product of two permutations is again a permutation. For example,

$$\rho_3 \cdot \rho_4 \equiv \rho_3 \circ \rho_4 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} = \rho_5 .$$

The product defines how two elements permute or interchange by repeated application of permutations. For instance, $(\rho_3 \cdot \rho_4)(1) = (\rho_3 \circ \rho_4)(1) = \rho_3(\rho_4(1)) = \rho_3(1) = 3$, while $(\rho_3 \cdot \rho_4)(3) = \rho_3(\rho_4(3)) = \rho_3(2) = 1$. The set of all possible products is given by the multiplication table in Figure **3.3.1**.

It follows from the multiplication table that the set of all permutations on X together with the operation of permutation multiplication is a group with identity ρ_1. Note that this group is not abelian

$(\rho_3 \cdot \rho_4 = \rho_5 \neq \rho_6 = \rho_4 \cdot \rho_3)$. It is the smallest possible example of a nonabelian group as any group with fewer than six elements must be abelian [16].

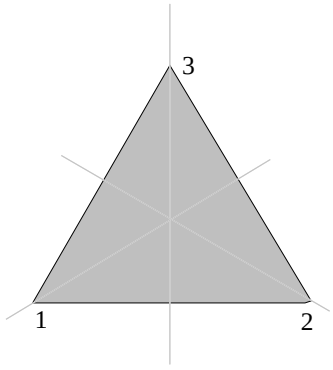


Figure 3.3.2 Symmetries of an equilateral triangle.

There is a natural correspondence between the elements of this group and the symmetries of the equilateral triangle shown in Figure **3.3.2**. The permutations ρ_i for $i = 1, 2, 3$, represent the *rotations* of the triangle onto itself about its barycenter, with ρ_0 representing the "no move" rotation. The permutations ρ_i, for $i = 4, 5, 6$, represent the *mirror images* across the bisectors of the angles. These symmetries are also nicely reflected in the four quadrants of the multiplication table (Fig. **3.3.1**).

We now show that the collection of all permutations of any nonempty set X forms a group under permutation multiplication.

3.3.2 **Theorem.** *Let X be a nonempty set, and let S_X be the collection of all permutations of X. Then S_X is a group under permutation multiplication.*

Proof: We have three axioms to check. Since permutations are functions, in order to show for permutations ρ, σ, and τ that

$$(\rho \cdot \sigma) \cdot \tau = \rho \cdot (\sigma \cdot \tau),$$

we have to show that each composite function maps each $x \in X$ onto the same image in X. That is, we must show that

$$[(\rho \cdot \sigma) \cdot \tau](x) = [\rho \cdot (\sigma \cdot \tau)](x) \ \ \forall\, x \in X\,.$$

We have

$$[(\rho \cdot \sigma) \cdot \tau](x) = (\rho \cdot \sigma)(\tau(x)) = \rho(\sigma(\tau(x))) = \rho((\sigma \cdot \tau)(x)) = [\rho \cdot (\sigma \cdot \tau)](x)\,.$$

Thus, $(\rho \cdot \sigma) \cdot \tau$ and $\rho \cdot (\sigma \cdot \tau)$ map each $x \in X$ into the same element in X. This satisfies the associativity axiom for groups.

Obviously, the identity function 1_X acts as the multiplicative identity. This satisfies the second group axiom.

As we remarked earlier, the inverse of a permutation ρ is simply defined to be the permutation which reverses the direction of the function ρ. More precisely, since ρ is one-to-one and

onto, each $y \in X$ is the image of some unique $x \in X$ and we simply define, for each $y \in X$, $\rho^{-1}(y) = x$ such that $\rho(x) = y$. It follows from this definition of ρ^{-1} that $\rho \cdot \rho^{-1} = \rho^{-1} \cdot \rho = 1_X$. This proves the existence of inverses, which satisfies the third group axiom.

Q.E.D.

Groups of permutations play an important role in geometry. For example, consider the set of all translations of the coordinate plane $\mathbb{R}^2$, that is, mappings $\rho : \mathbb{R}^2 \to \mathbb{R}^2$, defined by

$$\rho(x, y) = \big(x', y'\big), \text{ where } x' = x + s,\ y' = y + t\,.$$

Here s and t are some fixed real numbers and $(x, y) \in \mathbb{R}^2$ is arbitrary. The reader should verify that such a mapping ρ is a permutation of $\mathbb{R}^2$. It is also easy to verify that the set of all translations of this type forms a group under the operation of permutation multiplication. The reader will no doubt think of many other groups of permutations of the plane, or $\mathbb{R}^n$, which are of geometric interest. Indeed, one of the most famous approaches to geometry, known as the "Erlanger Programm," is by means of the determination of the geometric properties which remain invariant under a particular group of transformations [15].

There is nothing in our definition of a permutation that requires the set X to be finite. Our last example with $X = \mathbb{R}^2$ is a case in point. However, most of our examples of permutation groups will be concerned with permutations of finite sets. Clearly, if X and Y both have the same number of elements, then the group of all permutations of X has the same structure as the group of all permutations of Y; i.e., one group can be obtained from the other by just renaming the elements. This is the concept of *isomorphic* structures of which more will be said in Section **3.4**.

3.3.3 **Definition**. If X is the finite set $\{1, 2, \ldots, n\}$, then the group of all permutations of X is called the *symmetric group on n letters*, and is denoted by S_n.

Note that S_n has $n!$ elements, where $n! = n(n-1)(n-2)\cdots(3)(2)(1)$.

There is another standard notation for a permutation which is often used. The permutation

$$\rho_2 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$$

of the set $X = \{1, 2, 3\}$ can be written in *cyclic* notation as $\rho_2 = (1, 2, 3)$, where the $cycle(1, 2, 3)$ is interpreted to mean: 1 is replaced by 2, 2 is replaced by 3, and 3 is replaced by 1. The permutation

$$\rho_5 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$$

can be written as (1,3), where the cycle (1,3) is interpreted as: 1 is replaced by 3, 3 by 1, and the missing symbol 2 remains unchanged.

Not every permutation can be written as a cycle. Consider the permutation

$$\rho = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 3 & 2 & 5 & 4 \end{pmatrix}$$

on the set $X = \{1, 2, 3, 4, 5\}$. There is no consistent way of writing ρ as a cycle. However, we can write ρ as (2,3)(4,5). The interpretation is clear: 1 is unchanged; 2 is replaced by 3 and 3 is replaced

by 2; 4 is replaced by 5 and 5 by 4. Note that ρ corresponds to the product of the two permutations $\tau = (2,3)$ and $\sigma = (4,5)$ on X. For this reason we call (2,3)(4,5) the product of cycles. These two cycles are also disjoint, i.e., they have no symbol in common. Thus, in cyclic notation we shall expect a permutation on n symbols to consist of a single cycle or the product of two or more mutually disjoint cycles. This fact is expressed by the next theorem.

3.3.4 **Theorem.** *Every permutation ρ on a finite set X is a product of disjoint cycles.*

Proof: We assume, without loss of generality, that $X = \{1, 2, \ldots, n\}$. Consider the elements

$$1,\ \rho(1),\ \rho^2(1),\ \rho^3(1), \ldots$$

Since X is finite, these elements cannot all be distinct. Let $\rho^k(1)$ be the first term in the sequence which has appeared previously. Then $\rho^k(1) = 1$, for if $\rho^k(1) = \rho^j(1)$, with $0 < j < k$, we would have $\rho^{k-j}(1) = 1$, with $k - j < k$, contradicting our choice of k. Let

$$\sigma_1 = \left(1, \rho(1), \rho^2(1), \rho^3(1), \ldots, \rho^{k-1}(1)\right).$$

It is easy to see that σ_1 has the same effect as ρ on all elements of X appearing in this cyclic notation for σ_1.

Let i be the first element of X not appearing in this cyclic notation for σ_1. Repeating the above argument with the sequence

$$i, \rho(i), \rho^2(i), \rho^3(i), \ldots$$

we arrive at a cycle σ_2. Now σ_2 and σ_1 are disjoint, for if they had any element m of X in common, they would be identical, since each cycle could be constructed by repeated application of the permutation ρ starting at *m*.

Continuing, we pick the first element in X not appearing in the cyclic notations of either σ_1 or σ_2 and construct σ_3, etc. Since X is finite, this process must terminate with some σ_p. The product

$$\sigma_1 \cdot \sigma_2 \cdot \ldots \cdot \sigma_p$$

then clearly has the same effect on each element of X as ρ does. Therefore,

$$\rho = \sigma_1 \cdot \sigma_2 \cdot \ldots \cdot \sigma_p .$$

Q.E.D.

The reader can easily convince himself that the representation of a permutation as a product of disjoint cycles, none of which is the identity permutation, is unique up to the orders of the factors.

A cycle of length 2 is called a *transposition*. Thus, a transposition leaves all but two elements fixed, and maps each of these onto the other. A computation shows that

$$(1,2,3,4,5) = (1,2) \cdot (1,3) \cdot (1,4) \cdot (1,5)$$

and, in general,

$$(a_1, a_2, \ldots, a_n) = (a_1, a_2) \cdot (a_1, a_3) \cdot \ldots \cdot (a_1, a_n) .$$

Therefore, any cycle is a product of transpositions. We have the following corollary of Theorem **3.3.4**.

3.3.5 Corollary. *Any permutation of a finite set of at least two elements is a product of transpositions.*

The cycle (1,3,5) can be written as the following product of transpositions:

$$(1,3,5) = (1,3)\cdot(1,5) = (1,5)\cdot(3,5)\,,$$

that is, as the product of two different transpositions. Similarly, in the previous two examples, the cycles (1,2,3,4,5) and $(a_1, a_2, \ldots, a_n)$ were written as products of four and $n-1$ transpositions, respectively. This illustrates the fact that a cycle of length n can always be written as a product of $n-1$ transpositions. Thus, if n is even, then the number of transpositions is odd, and if n is odd, then the number of transpositions is even. For a permutation which is not necessarily a cycle, the following theorem holds.

3.3.6 Theorem. *If ρ is a permutation on n symbols expressed as the product of k transpositions and also as a product of j transpositions, then k and j are either both even or both odd.*

This is quite an important fact, the usual proof of which may seem a bit artificial and can be found in [17]. A permutation will be called *even* or *odd* according to whether it can be expressed as the product of an even or odd number of transpositions, respectively.

Permutation of data is one important reason for studying permutations and permutation groups in signal processing and computer science. Matrix versions of the Fast Fourier Transform (FFT) involve sophisticated shuffling of data which is accomplished with the use of permutation matrices.

Let S_n denote the group of permutations on $\{0, 1, \ldots, n-1\}$ and A' the transpose of the matrix A.

3.3.7 Definition. Let $\sigma \in S_n$ and define the $n \times n$ matrix P_σ by

$$P_\sigma = (p_{ij}) \text{ where } p_{ij} = \begin{cases} 1 & \text{if } j = \sigma(i) \\ 0 & \text{otherwise} \end{cases}.$$

P_σ is called a *permutation matrix*.

It is well known that permutation matrices are invertible and that $P_\sigma^{-1} = P'_\sigma = P_{\sigma^{-1}}$. Using this fact, it is easy to see the set of all $n \times n$ permutation matrices forms a group under matrix multiplication which has the same structure (i.e. is isomorphic to) as S_n. Note that if $A = (a_{ij})$ is an $n \times n$ matrix, then multiplication on the left by P_σ permutes the rows of A by σ^{-1} and multiplication on the right by $P_{\sigma^{-1}} = P'_\sigma$ permutes the columns of A by σ^{-1}. Hence, we can write $P_\sigma A P'_\sigma = \left(a_{\sigma(i),\sigma(j)}\right)$.

3.4 Isomorphisms

Throughout this book, we deal with various kinds of abstract mathematical systems. The name "abstract mathematical system" is used to describe any well-defined collection of mathematical objects consisting, for example, of a set together with relations and operations on the set, and a collection of postulates, definitions, and theorems describing various properties of the structure.

It is a fundamentally important fact that even when systems have very little structure, such as semigroup or groups, it is often possible to classify them according to whether or not they are mathematically similar or equivalent. These notions are made mathematically precise by the *morphism* relationship between abstract systems.

3.4.1 **Definition.** Let $G = (G, \bigcirc)$ and $G' = (G', \bigcirc')$ denote two systems. A *homomorphism* from G to G' is a function $\psi : G \to G'$ such that for each $g, h \in G$,

$$\psi(g \bigcirc h) = \psi(g) \bigcirc' \psi(h) .$$

Thus, a homomorphism is required to preserve the operations of the systems; i.e., performing the operation $g \bigcirc h$ in G and then applying the function ψ to the result is the same as first applying ψ to each g and h and then applying the operation $\psi(g) \bigcirc' \psi(h)$ in G'. If such a function exists, then the two systems are said to be *homomorphic*.

By definition, a homomorphism need not be a one-to-one correspondence between the elements of G and G'. One-to-one and onto functions that preserve the mathematical structures of systems lead to the extremely important concept of an *isomorphism*.

3.4.2 **Definition.** Let $G = (G, \bigcirc)$ and $G' = (G', \bigcirc')$ denote two systems. An *isomorphism* of G into G' is a homomorphism $\psi : G \to G'$ which is both one-to-one and onto.

If such a homomorphism exists, then we say that the two systems are *isomorphic*. Hence the idea that the two systems G and G' are isomorphic means that they are identical except for the names of the elements and operations. That is, we can obtain G' from G by renaming an element g in G with the name of a certain element g' in G', namely $g' = \psi(g)$, and by renaming the operation $\bigcirc$ as $\bigcirc'$. Then the counterpart of $g \bigcirc h$ will be $g' \bigcirc' h'$. The next theorem we prove is very obvious if we consider an isomorphism to be a renaming of one system so that it is just like another.

3.4.3 **Theorem.** *Let G and G' be two groups and suppose* e *is the identity of G. If $\psi : G \to G'$ is an isomorphism, then $\psi(e)$ is the identity of G. Moreover,*

$$\psi\left(g^{-1}\right) = [\psi(g)]^{-1} \quad \forall\, g \in G .$$

Proof: Let $g' \in G'$. Since ψ is onto, $\exists g \in G$ such that $\psi(g) = g'$. Then

$$g' = \psi(g) = \psi(e \cdot g) = \psi(e) \cdot \psi(g) = \psi(e) \cdot g' .$$

Similarly,

$$g' = \psi(g) = \psi(g \cdot e) = \psi(g) \cdot \psi(e) = g' \cdot \psi(e) .$$

Thus, for every $g' \in G'$ we have

$$\psi(e) \cdot g' = g' = g' \cdot \psi(e) .$$

Therefore, $\psi(e)$ is the identity of G.

Moreover, for $g \in G$ we have

$$\psi(e) = \psi\left(g^{-1} \cdot g\right) = \psi\left(g^{-1}\right) \cdot \psi(g)$$

and

$$\psi(e) = \psi\left(g \cdot g^{-1}\right) = \psi(g) \cdot \psi\left(g^{-1}\right) .$$

Thus, $\psi\left(g^{-1}\right) = [\psi(g)]^{-1}$.

Q.E.D.

The essence of the theorem is that isomorphisms map identities onto identities and inverses onto inverses.

It is immediate from our discussion that every system is isomorphic to itself; we simply let ψ be the identity function. To show whether or not two different systems are isomorphic can be a difficult task. Proceeding from the definition, the following algorithm can be used to show that two systems $G = (G, \bigcirc)$ and $G' = (G', \bigcirc')$ are isomorphic:

STEP 1. Define the function ψ from G to G' which is proposed as a candidate for isomorphism.

STEP 2. Show that ψ is a one-to-one function.

STEP 3. Show that ψ is an onto function.

STEP 4. Show that $\psi(g \bigcirc h) = \psi(g) \bigcirc' \psi(h)$.

Step 4 is usually just a question of computation. One computes both sides of the equation and checks out whether or not they are the same. We illustrate this procedure with an example.

3.4.4 **Example:** We want to show that $(\mathbb{R}, +)$ is isomorphic to $(\mathbb{R}^+, \cdot)$.

STEP 1. Define the function $\psi : \mathbb{R} \to \mathbb{R}^+$ by $\psi(x) = e^x \quad \forall\, r \in \mathbb{R}$.

STEP 2. If $\psi(x) = \psi(y)$, then $e^x = e^y$, and taking the natural log we obtain that $x = y$. Thus, ψ is one-to-one.

STEP 3. If $x \in \mathbb{R}^+$, then $\psi(lnx) = e^{lnx} = x$. Thus, for every $x \in \mathbb{R}^+$, $\exists\, y \in \mathbb{R}$, namely $y = lnx$, such that $\psi(y) = x$. Therefore, ψ is onto.

STEP 4. For $x, y \in \mathbb{R}$, we have

$$\psi(x + y) = e^{x+y} = e^x \cdot e^y = \psi(x) \cdot \psi(y)\,.$$

Another example of two isomorphic groups was mentioned in the previous section. There we noted that the symmetric group S_n is isomorphic to the group of $n \times n$ permutation matrices. This fact signifies the importance of the symmetric group in applications. In the theory of groups, S_n plays an even more central role; it can be shown that any finite group is isomorphic to some subgroup of S_n for some n [17]. However, finding the right candidates for isomorphisms in order to establish this fact is a nontrivial task.

To show that two systems are not isomorphic means that there cannot exist a one-to-one correspondence which preserves the algebraic structure of the systems. This is a trivial problem whenever the two systems have a different number of elements. For example, $\mathbb{Z}_4$ and S_6 are not isomorphic as there cannot exist a one-to-one correspondence between their elements. Similarly, since $\mathbb{Z}$ is countable and $\mathbb{R}$ is uncountable, they can never be isomorphic as algebraic structures.

3.5 Rings and Fields

The systems we have considered thus far have been concerned with sets on which a single binary operation has been defined. Our earliest experience with arithmetic, however, has taught us the use of two distinct binary operations on sets of numbers, namely addition and multiplication. This early and important experience should indicate that a study of sets on which two binary operations have been defined is of great importance. Modeling our definition on properties common to these number systems, as well as such structures as the set of all $n \times n$ matrices with elements in one of the number systems, or the set of all polynomials with coefficients in, say, the set of all integers, we now define a type of algebraic structure known as a ring.

3.5.1 **Definition.** A *ring* $(R, +, \cdot)$ is a set R together with two binary operations + and $\cdot$ of addition and multiplication defined on R such that the following axioms are satisfied:

R_1 $(R, +)$ is an abelian group.

R_2 $(R, \cdot)$ is a semigroup.

R_3 $\forall\, a, b, c \in R,\ a \cdot (b + c) = (a \cdot b) + (a \cdot c)\ and\ (a + b) \cdot c = (a \cdot c) + (b \cdot c)$.

If axiom R_1 is weakened to

R_1' $(R, +)$ is a commutative semigroup

then R is called a *semiring*.

In subsequent chapters it will become apparent that the theory of rings and semirings plays an important role in the analysis and application of image algebra. Of the many examples of rings which come readily to mind from experience with common systems of elementary mathematics, the most natural is, perhaps, the ring $\mathbb{Z}$ of integers with the usual addition and multiplication. However, if we examine the properties of the ring of integers, we note that it has properties not enjoyed by rings in general. Among these properties are:

(i) The existence of a multiplicative identity element, which must be unique, called the *unit* element, and which is usually designated by the number 1.

(ii) The commutativity of multiplication.

(iii) The nonexistence of an element $a \neq 0$ such that for some positive integer n, $na \equiv a + a + \ldots + a = 0$ (where na is defined to be the sum of n a's).

On the other hand, the integers themselves fail to possess a most useful property, namely that:

(iv) For every nonzero $a \in R$ there is an element in R, denoted by a^{-1}, such that $a \cdot a^{-1} = a^{-1} \cdot a = 1$, i.e., $(R, \cdot)$ is a group.

When it does exist in a particular ring, the element a^{-1} is called the *inverse* of a. In fact, $(\mathbb{Z}, +, \cdot)$ also fails to have the slightly weaker property:

(v) For every nonzero $a \in R$ there is an element in R, denoted by $\tilde{a}$, such that $a \cdot \tilde{a} \cdot a = a$ and $\tilde{a} \cdot a \cdot \tilde{a} = \tilde{a}$; i.e., every nonzero element has a pseudo inverse.

These properties lead us to some further definitions.

3.5.2 **Definition:** If a ring satisfies (i), it is called a *ring with unity*. If a ring satisfies (ii), it is called *commutative* or *abelian*. If a ring satisfies (iii) it is said to have *characteristic zero*. If it satisfies (iv) it is called a *division ring* or a *quasi-field*. A *field* is a commutative division ring. A ring which satisfies (v) is called a *von Neumann ring*.

As we noted, $(\mathbb{Z}, +, \cdot)$ is a commutative ring but not a division ring. On the other hand, $(\mathbb{R}, +, \cdot)$ is an example of a commutative division ring and, hence, a field. It is also an example of a ring with unity and characteristic zero. Other well known examples of commutative rings are $(\mathbb{Q}, +, \cdot)$ and $(\mathbb{C}, +, \cdot)$.

In order to give some indication of the generality of the concept of a ring, we turn to some less familiar examples of rings.

3.5.3 **Examples:**

(i) The set of real valued functions on a set X together with the operation of function addition and multiplication $(\mathbb{R}^X, +, \cdot)$ is a commutative ring with unity. We know from Theorem **2.8.2** that $(\mathbb{R}^X, +)$ is an abelian group. That $(\mathbb{R}^X, \cdot)$ is a commutative semigroup follows from Definition **2.8.1** (iv) of multiplication of real valued functions. For example,

$$[(f \cdot g) \cdot h](x) = [(f \cdot g)(x)]h(x) = [f(x)g(x)]h(x)$$
$$= f(x)[g(x)h(x)] = f(x)[(g \cdot h)(x)] = [f \cdot (g \cdot h)](x)\,.$$

Thus,

$$(f \cdot g) \cdot h = f \cdot (g \cdot h)\,.$$

Commutativity and distributivity of multiplication over addition can be demonstrated in a similar fashion. The multiplicative identity is, of course, the constant function $1(x) = 1\ \forall\, x \in X$. For any function f with the property $f(x) \neq 0\ \ \forall x \in X$ we may define a multiplicative inverse $(f)^{-1}$ by $(f)^{-1}(x) = 1/f(x)$. However, since there are functions $f \in \mathbb{R}^X$ with $f \neq 0$ but $f(x) = 0$ for some $x \in X$, $(\mathbb{R}^X, +, \cdot)$ — in contrast to $(\mathbb{R}, +, \cdot)$ — is not a division ring.

(ii) Consider the cyclic group $(\mathbb{Z}_n, +)$. For $i, j \in \mathbb{Z}_n$ we define the product $i \cdot j$ to be the remainder of the usual product of the integers i and j when divided by n. For example, in $\mathbb{Z}_5$ we have $3 \cdot 4 = 2$. This operation on $\mathbb{Z}_n$ is *multiplication modulo* n. We leave it to the reader to check that the system $(\mathbb{Z}_n, +, \cdot)$ satisfies the ring axioms.

An important observation concerning example (i) is the fact that the operations on $\mathbb{R}^X$ are *induced* by the operations on $\mathbb{R}$. That is, the addition of two functions $f + g$ is defined in terms of addition of real numbers, e.g. $(f + g)(x) = f(x) + g(x)$, and multiplication of two functions is defined in terms of multiplication of real numbers. In view of this observation, it should be clear that the set $\mathbb{R}$ can be replaced with any field $\mathbb{F}$ in these two examples and that the result would be a commutative ring with unity $(\mathbb{F}^X, +, \cdot)$, where the addition $f + g$ on $\mathbb{F}^X$ is defined in terms of the addition on $\mathbb{F}$, e.g., $(f + g)(x) = f(x) + g(x)$, and likewise for multiplication. In addition, due to the fact that the operations on $\mathbb{F}^X$ are induced by the operations on $\mathbb{F}$, the ring $(\mathbb{F}^{\mathbf{X}}, +, \cdot)$ behaves very much like the ring $(\mathbb{F}, +, \cdot)$. The only missing ingredient is the lack of multiplicative inverses for $\mathbb{F}^X$.

Hopefully the reader is beginning to realize that in the study of any sort of mathematical structure, an idea of basic importance is the concept of two systems being structurally alike or identical, i.e., one being similar to the other or one being exactly like the other except for its name and the names of its

elements. In algebra, the concept of being identical is always called "isomorphism." The concept of two rings being just alike except for names of elements leads us, just as it did for systems with one operation, to the following definition.

3.5.4 **Definition.** An *isomorphism* ψ of a ring R with a ring R' is a one-to-one function mapping R onto R' such that $\forall\, r, s \in R$

$$(1) \quad \psi(r+s) = \psi(r) + \psi(s)\,,$$
$$(2) \quad \psi(r \cdot s) = \psi(r) \cdot \psi(s)\,.$$

If such a function exists, then we say that the two rings are *isomorphic*.

3.5.5 **Example:** Consider the ring $(\mathbb{R}^n, +, \cdot)$, where addition corresponds to vector addition and multiplication is defined by multiplying the corresponding vector components, i.e.,

$$(x_1, x_2, \ldots, x_n) + (y_1, y_2, \ldots, y_n) = (x_1 + y_1,\, x_2 + y_2\,, \ldots,\, x_n + y_n)$$

and *Hadamard multiplication*

$$(x_1, x_2, \ldots, x_n) \cdot (y_1, y_2, \ldots, y_n) = (x_1 y_1,\, x_2 y_2\,, \ldots,\, x_n y_n)\,.$$

We leave it to the reader to convince himself that $(\mathbb{R}^n, +, \cdot)$ is a commutative ring with unity. Suppose that X is a finite set with n elements, say $X = \{1, 2, \ldots, n\}$. Let $\nu : \mathbb{R}^X \to \mathbb{R}^n$ be the function defined by

$$\nu(f) = (f(1),\, f(2), \ldots,\, f(n))\,.$$

We know from Example **2.8.3** that ν is one-to-one and onto. Furthermore,

$$\begin{aligned}\nu(f) + \nu(g) &= (f(1),\, f(2), \ldots, f(n)) + (g(1),\, g(2), \ldots, g(n)) \\ &= (f(1) + g(1),\, f(2) + g(2), \ldots,\, f(n) + g(n)) \\ &= ((f+g)(1), (f+g)(2), \ldots,\, (f+g)(n)) = \nu(f+g).\end{aligned}$$

An analogous argument shows that $\nu(f \cdot g) = \nu(f) \cdot \nu(g)$. This proves that the rings $(\mathbb{R}^n, +, \cdot)$ and $\left(\mathbb{R}^X, +, \cdot\right)$ are isomorphic. Of course, by arguing in an analogous fashion, we can prove that for any field $\mathbb{F}$ the corresponding rings $(\mathbb{F}^n, +, \cdot)$ and $\left(\mathbb{F}^X, +, \cdot\right)$ are isomorphic.

Thus far, all our examples have dealt with commutative rings. However, noncommutative rings play an important role in the structure of the image algebra which is the central theme of this treatise. We present the most pertinent example of such a ring.

3.5.6 **Example:** Let $\mathbb{F}$ be any field, say $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$, and consider the set $M_{2\times 2}(\mathbb{F})$ of all 2×2 matrices of form

$$(a_{ij}) = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

where the a_{ij}'s are all in $\mathbb{F}$. The set $M_{n\times n}(\mathbb{F})$ of all $n \times n$ matrices over $\mathbb{F}$ is similarly defined.

Matrix addition on $M_{2\times 2}(\mathbb{F})$ is defined by

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} \\ a_{21}+b_{21} & a_{22}+b_{22} \end{pmatrix},$$

that is, by adding corresponding entries using addition in $\mathbb{F}$. After a few moments of thought, it is clear from the axioms of a field that $\big(M_{2\times 2}(\mathbb{F}),+\big)$ is an abelian group with additive identity

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

and with additive inverse

$$-\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} -a_{11} & -a_{12} \\ -a_{21} & -a_{22} \end{pmatrix}.$$

Matrix multiplication on $M_{2\times 2}(\mathbb{F})$ is defined by

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11}+a_{12}b_{21} & a_{11}b_{12}+a_{12}b_{22} \\ a_{21}b_{11}+a_{22}b_{21} & a_{21}b_{12}+a_{22}b_{22} \end{pmatrix}.$$

If $\mathbb{F}$ equals $\mathbb{R}$ or $\mathbb{C}$, then this multiplication corresponds of course to the regular matrix product over these fields and can best be remembered by

$$(a_{ij})(b_{ij}) = (c_{ij}),$$

where

$$c_{ij} = \sum_{k=1}^{2} a_{ik}b_{kj}\ .$$

Of course, the analogous definition holds for matrix multiplication in which the sum goes from $k=1$ to n. In short, everything we said about the system $\big(M_{2\times 2}(\mathbb{F}),+,\times\big)$ is also valid for the system $\big(M_{n\times n}(\mathbb{F}),+,\times\big)$.

To show that $\big(M_{n\times n}(\mathbb{F}),+,\times\big)$ is a ring, it remains to prove the associative and distributive laws. Using the field properties of $\mathbb{F}$ and the definition of matrix multiplication in $\big(M_{n\times n}(\mathbb{F}),+,\times\big)$, then if d_{rs} denotes the entry in the rth row and sth column of $(a_{ij})[(b_{ij})(c_{ij})]$, we have

$$d_{rs} = \sum_{k=1}^{n} a_{rk}\Big(\sum_{j=1}^{n} b_{kj}c_{js}\Big) = \sum_{j=1}^{n}\Big(\sum_{k=1}^{n} a_{rk}b_{kj}\Big)c_{js} = e_{rs}\ ,$$

where e_{rs} is the entry in the rth row and sth column of $[(a_{ij})(b_{ij})](c_{ij})$. The distributive property is proved in a similar fashion.

The last example proves the following theorem:

3.5.7 **Theorem.** *If $\mathbb{F}$ is a field, then the set $M_{n\times n}(\mathbb{F})$ of all $n \times n$ matrices with entries from $\mathbb{F}$ forms a ring under matrix addition and matrix multiplication.*

The rings of matrices over a field $\mathbb{F}$ are an important tool in the theory and practice of image transformations. In this context, they can be viewed as corresponding to certain functions called *templates*, and matrix multiplication, when viewed in this light, can be shown to correspond to template convolutions (Chapter **4**). This provides an elegant demonstration of the associative law for template convolutions.

On the down side, we need to point out that the ring $\left(M_{n\times n}(\mathbb{F}), +, \times\right)$ lacks some important algebraic properties. Since matrix multiplication is not commutative, $M_{n\times n}(\mathbb{F})$ is not a commutative ring. Also, one of the most important properties of the real number system is that the product of two numbers can only be zero if at least one of the factors is zero. The working engineer or scientist uses this fact constantly, perhaps without realizing it. Suppose for example, one needs to solve the equation

$$2x^2 + 9x - 5 = 0\,.$$

The first thing to do is to factor the left side:

$$2x^2 + 9x - 5 = (2x - 1)(x + 5)\,.$$

One then concludes that the only possible values for x are $\frac{1}{2}$ and –5. Why? Because the resulting product is zero if and only if one of the factors $2x - 1$ or $x + 5$ is zero.

The property that if a product equals zero then at least one of the product factors must also equal zero, does not hold for rings in general. For instance, the definition of matrix product in $M_{2\times 2}(\mathbb{F})$ shows that

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\,.$$

Similarly, the Hadamard product of the two nonzero vectors (1,0,0,...,0) and (0,0,...,0,1) in $\mathbb{R}^n$ is the zero vector (0,0,...,0).

These ideas are of such importance that we formalize them in a definition.

3.5.8 **Definition.** If r and s are two nonzero elements of a ring R such that $r \cdot s = 0$, then r and s are *divisors of zero* or *zero divisors*. In particular, r is a *left zero divisor* and s is a *right zero divisor* of the product $r \cdot s$.

An important consequence of the concept of zero divisors is provided by Theorem **3.5.10** below. Let R be a ring and let $r, s, t \in R$. We say that the *cancellation laws* hold in R if $r \cdot s = r \cdot t$, with $r \neq 0$, implies $s = t$ and $s \cdot r = t \cdot r$, implies $s = t$. These are multiplicative cancellation laws. Additive cancellation laws hold since $(R, +)$ is a group.

3.5.9 **Definition.** An *integral domain* is a commutative ring with unity containing no zero divisors.

3.5.10 **Theorem.** *The cancellation laws hold in integral domains.*

Proof: Suppose that D is an integral domain and that $ab = ac$ with $a \neq 0$. Then

$$ab - ac = a(b - c) = 0\ .$$

Since $a \neq 0$, and since D has no right divisor of zero, we must have $b - c = 0$. Thus, $b = c$. A similar argument shows that $ba = ca$, with $a \neq 0$, implies that $b = c$.

Q.E.D.

Suppose that R is a division ring and $r \cdot s = 0$ with $r \neq 0$. Then

$$0 = r^{-1} \cdot 0 = r^{-1} \cdot (r \cdot s) = \left(r^{-1} \cdot r\right) \cdot s = 1 \cdot s = s\,.$$

Similarly, if $s \cdot r = 0$, then multiplication on the right by r^{-1} leads to the conclusion that s must be zero. This demonstrates that a division ring contains no divisors of zero. Thus a division ring lacks only commutativity of being an integral domain. However, since commutativity was not used in proving Theorem **3.5.10**, we see that the theorem also holds for division rings.

3.6 Polynomial Rings

An important consequence of Theorem **3.5.10** is that we can solve polynomial equations in which the polynomials can be factored into linear factors in the usual fashion by setting each factor equal to zero, as long as we are dealing with polynomials with coefficients from an integral domain or division ring. This leads us directly into the topic of polynomials with coefficients in a ring.

3.6.1 **Definition.** A *polynomial $p(x)$ with coefficients in a ring R in the indeterminate x* is an informal sum

$$\sum_{i=0}^{\infty} a_i x^i = a_0 + a_1 x + \cdots + a_n x^n + \cdots\,,$$

where $a_i \in R$ are called the *coefficients of $p(x)$*, and $a_i = 0$ for all but a finite number of values i. If for some $i > 0$ $a_i \neq 0$, then the largest such value of i is the *degree of $p(x)$*. If no such $i > 0$ exists, then $p(x)$ is of *degree zero*.

We also use the notation

$$p(x) = a_0 + a_1 x + \cdots + a_n x^n$$

whenever $a_i = 0$ for all $i > n$. Any element of R is a *constant polynomial*. The most important constant polynomials are the *zero polynomial* $0 \in R$ and, if R has unity, the *unit polynomial* 1.

Addition and multiplication of polynomials with coefficients in a ring R are defined in a way formally familiar to the reader. If

$$p(x) = a_o + a_1 x + \cdots + a_n x^n + \cdots$$

and

$$q(x) = b_0 + b_1 x + \cdots + b_n x^n + \cdots\,,$$

then for polynomial addition, we have

$$p(x) + q(x) = c_0 + c_1 x + \cdots + c_n x^n + \cdots\,,$$