Signals
and
Communication
Technology

Virendra P. Sinha

# Symmetries and Groups in Signal Processing

## An Introduction

Springer

Signals And Communication Technology

For other titles published in this series, go to
http://www.springer.com/series/4748

Virendra P. Sinha

# Symmetries and Groups in Signal Processing

## An Introduction

Springer

Prof. Virendra P. Sinha
Dhirubhai Ambani Institute
of Information and Comm. Tech.
Near Indroda Circle
382007 Gandhinagar, Gujarat, India
vp_sinha@daiict.ac.in

ISBN 978-90-481-9433-9 e-ISBN 978-90-481-9434-6
DOI 10.1007/978-90-481-9434-6
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2010931379

*Cover design*: SPi Publisher Services

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

*To my grandsons*
*Rohan, Ishaan, and Shloak,*
*for a whiff of algebra*

# Preface

The field of signal processing, as it stands today, abounds in varied generalizations of system theoretic concepts that can be said to rest on the notion of symmetry, and on group theoretic methods of exploiting symmetries.

A wide range of such generalizations and developments rely centrally on a transition from the classical Fourier theory to the modern theory of non-commutative harmonic analysis, with its roots in the representation theory of groups. In the framework that emerges through this transition, all the basic notions—transforms, convolutions, spectra, and so on, carry over in a form that allows a wide variety of interpretations, subsuming the old ones and admitting new ones.

This book is an introductory treatment of a selection of topics that together serve to provide in my view a background for a proper understanding of the theoretical developments within this framework. Addressed primarily to beginning graduate students in electrical and communication engineering, it is meant to serve as a bridge between what they know from their undergraduate years, and what lies ahead for them in their graduate studies, be it in the area of signal processing, or in related areas such as image processing and image understanding, coding theory, fault diagnostics, and the theory of algorithms and computation.

I assume that the reader is familiar with the theory of linear time–invariant continuous–time and discrete–time systems as it is generally taught in a basic undergraduate course on signals and systems. There are no mathematical prerequisites beyond what they would have learnt in their undergraduate years. Familiarity with rudiments of linear algebra would be helpful, but even that is not necessary; whatever of it is needed in the book, they can pick up on their own as they go along.

A point about pedagogy. In teaching mathematical concepts to engineering students, a plan of action that is commonly followed is to separate what is regarded as mathematics *per se* from its applications, and to introduce the two separately in alternation. Thus one first introduces them to differential equations, linear or modern algebra, or discrete mathematics, on abstract lines as they would appear in a mathematics text, and then one turns to their applications in solving engineering problems.

This plan works well, perhaps just about, when the students are fresh to their engineering studies. But at a stage when they have already had their first exposure to basic engineering principles, it has an inhibiting influence, both on their pace of learning and on their motivation for it. Faced with a new abstract concept at this stage, they instinctively begin to look for a pattern in which the new will fit in smoothly, and through analogies and metaphors, with what they already intuitively know of their main subjects. They look for the sort of experience that, for instance, they had at the time they learnt their elements of Euclidean geometry, when they saw how the theorem on triangle inequality, logically derived from the axioms, agreed with what they knew all along about triangles as they drew them on paper. It is the same experience which they had while learning elements of graph theory concurrently with network analysis. More generally, they look for a backdrop of intuition against which they would like the abstractions to be set and to unfold.

Study of new mathematical structures becomes, as result, an easier and more pleasant task for such students if the abstractions are presented seamlessly with their concrete engineering interpretations. I have tried to keep this point in mind in my presentation in this book.

The contents of the book are organized as follows. Chapter 1 is devoted to an overview of basic signal processing concepts in an algebraic setting. Very broadly, it is an invitation to the reader to revisit these concepts in a manner that places in view their algebraic and structural foundations. The specific question that I examine is the following: How should system theoretic concepts be formulated or characterized so that they are, in the first instance, independent of details such as whether the signals of interest to us are discrete, discrete finite, one–dimensional, or multi–dimensional. Implicit in this question is a finer question about representation of signals that I discuss first, focussing attention on the distinction between what signals are *physically*, and the models by which they are *represented*. Next I discuss those aspects of linearity, translation–invariance, causality, convolutions, and transforms, that are germane to their generalizations, in the context of discrete signals. Chapter 2 presents in a nutshell those basic algebraic concepts that are relied upon in a group theoretic interpretation of the concept of symmetry. In Chapter 3, the points made in Chapter 1 about the choice of mathematical models is taken up again. Chapter 4 is about symmetry and its algebraic formalization. Representation theory of finite groups is introduced in Chapter 5. Chapter 6 gives a final look at the role of group representation theory in signal processing.

## Acknowledgements

This book has grown out of notes written for a transition course offered to beginning graduate students in the Department of Electrical Engineering of IIT Kanpur, and also for a similar course at DA-IICT Gandhinagar. Creating these courses has been a very fruitful learning experience for me, and I have benefitted enormously from discussions and interactions with the participating students at both the places. I can not thank them enough for their active involvement. My special thanks are due to Ratnik Gandhi and Pratik Shah, both currently working for their doctoral degrees at DA-IICT, for their regular interactions and feedback. Ratnik has been my know-all

man for the subtleties of latex, and has acted as a sounding board for me at various stages of writing.

Writing has its lows, when one is held back by bouts of perfectionism. Constant nudging ‘to get on with the job’ is in such times a pragmatic antidote. My deep appreciation for that to Dr. A.P. Kudchadker, former Director of DA-IICT, and to Dr. S.C. Sahasrabudhe, the present Director.

Amongst colleagues, and former students, there are many who have directly and critically influenced my thought processes that have prompted this text. I gratefully acknowledge receiving constructive inputs from Drs. S.D. Agashe, S. Chatterji, S.K. Mullick, P. Ramakrishna Rao, K.R. Sarma, M.U. Siddiqi, V.R. Sule, and K.S. Venkatesh.

From the time I first put forth my book proposal to Springer in November 2008, it has been a pleasure interacting with Editor Mark de Jongh. My thanks go to him, and to Mrs. Cindy Zitter, his Senior Assistant, for benignly putting up with delays in my self-imposed deadlines, and for all the meticulous support.

Finally, to my wife, Meera, and daughters, Shubhra and Shalini, I am immeasurably grateful for being at one with me in negotiating the rhythms of academic life.

Gandhinagar
April, 2010

Virendra P. Sinha

# Contents

# Chapter 1

# Signals and Signal Spaces: A Structural Viewpoint

## 1.1 What Is a Signal?

The notion of a signal, like that of weight or temperature, is a two-sided one. We commonly think and speak of signals as functions of some sort, with numerical values both for their domain and for their range. Yet, signals to begin with have to do with what we perceive of objects and events around us through our senses. We could thus, to be more explicit, say that the term signal carries within it two connotations, one empirical and one formal. The former refers to the physical world in which statements about its objects and events are true or false in the sense that they are empirically observed to be so. The latter, on the other hand, refers to abstractions that are members of a formally defined mathematical world in which statements about a class of its members are true or false in the sense that they do or do not logically follow from the initial definitions and axioms governing that class.

To make the point clearer, consider the familiar notion of heaviness of an object for instance. How heavy an object is, we describe by means of its weight given as a number. The feeling of heaviness is, however, something empirical, and is assessed qualitatively. It is a matter of empirical verification that on every occasion that we check, we find two objects put together to be heavier than either of them individually, and therefore we inductively infer this to be a universal rule about all objects.

About numbers, on the other hand, it is a matter of deductive inference from the axioms of arithmetic that the sum of any two positive numbers is greater than either of the two. It makes sense to quantify the perception of heaviness by assigning numbers to objects as their weights because we have at our disposal a practical procedure, based on a device such as a pan balance, for assigning numbers to objects in a very special way. The resulting assignment is such that there is a match between what we find to be empirically true about objects regarding their heaviness, and what is logically true about their numerical weights on account of their arithmetic properties.

V.P. Sinha, *Symmetries and Groups in Signal Processing*, Signals and Communication Technology,
DOI 10.1007/978-90-481-9434-6_1, © Springer Science+Business Media B.V. 2010

This match between empirical attributes and their numerical representations may not, of course, be a total one. That is, the representations may be such that only some of the fundamental properties of numbers are put to use, the others having no meaningful role. Thus, both in the case of weight and temperature, real numbers are the means of representing empirical attributes—those of heaviness and hotness. But, whereas in the first case both the additive and ordering properties of numbers are relevant, only the ordering property is relevant in the second case and addition of numbers has no meaningful role.

Looked at in the same light, signals are primarily empirical entities, and based on their key attributes empirically determined, we choose for their representation a matching system of formal objects, like numbers, functions or relations, in a manner that may not always involve all attributes of these formal objects.[1]

To take a common example, take the case of audio signals, i.e., sounds as we hear them. Experience has shown us that sounds are characterized by air pressure variations with respect to time at the location of hearing, and that pressure and time are empirical entities that can be numerically represented through appropriate measurement procedures.

As a result of all this, we obtain for an audio signal a numerical representation in the form of a function $s : T \to P$, where $T$ is the set of reals assigned to time instants running from infinite past to infinite future, and $P$ is the set of reals as assigned to pressures corresponding to sounds. Since on values of incremental pressure, all possible arithmetic operations are meaningfully admissible, the full real number system figures in the representation of signal values in this case.

For the set $T$ of time instants, however, the situation is different. Since we perform only translation in time, at least in the situations we ordinarily encounter, the relevant operations on reals are addition and subtraction but not multiplication and division. That is, in the representation of time instants, we make use of the real number system only partially; we limit ourselves to a subsystem of it that consists of reals under addition, subtraction, and to accommodate the distinctions between past, present and future, also the relation of ordering.

So much for the general nature of signals. As far as signal processing theories are concerned, we do, of course, think of signals purely in terms of their formal representations, with the understanding that the right representations have been chosen and whatever conclusions we draw from a study of these representations, we can give them meaningful interpretations for the physical entities they represent.

In digital signal processing, these representations take the form of sequences or arrays of numbers. We are in general interested, for a class of such sequences or arrays, in the kind of relationships they bear amongst themselves, in the way they are affected when we perform certain kinds of numerical operations on them, and also in the design of such operations to produce certain desired effects. Further,

[1] For more on this issue from the point of view of modeling, see Rosen [29, Chapters 2, 3]. The book as a whole makes profound reading, touching on several fundamental issues.

amongst different classes of signals, there is a great deal that is common. We are interested in that too.

Two very familiar classes of signals that we come across in digital signal processing are those of one-dimensional (1-D) discrete signals consisting of sequences of reals, and two-dimensional (2-D) discrete signals consisting of 2-D arrays of reals. For signals of either kind, we talk of addition, scalar multiplication and convolution, and of their linear and shift-invariant processing. Suppressing the detailed formulae characterizing these actions for the two cases, we find that they obey the same algebraic rules of manipulation. For the effects of the actions that can be traced back solely to these rules, we need not then distinguish between the two classes of signals. We may in fact treat them, to start with, within one generalized framework.

Such a framework is provided by the notion of a signal space, as we shall see a little later. It is helpful first to understand what we mean here by the general terms "space" and "structure".

## 1.2 Spaces and Structures

Let us first take up space, a term we shall use here to mean what it is nowadays commonly understood to mean in mathematics: a set, together with certain relationships, usually of a geometric nature,[2] abstractly stipulated for its members, the so-called points of the space.

Why should such an abstraction be called a space? In the language of daily speech, by space we mean physical space, something that a chair, a table or any physical object occupies by virtue of its shape and size. Does this commonsense meaning have any links with the mathematical one? As we shall presently see, it indeed has. In naming abstract mathematical objects, the general practice is to see that the names are suggestive of the intuitive background from which the abstractions have evolved.[3] Such is the case here too.

Observe first of all that the notion of physical space is in itself an abstraction, one that must have already come into use, as surmised in Einstein [11], at a very early stage in the development of human thought. Distinct from physical objects, but inextricably tied up, nevertheless, with our sensory perception of them, this notion would have in all likelihood emerged as a natural next step in the transition from the conception of solid bodies to that of their spatial relations. But this intuitive conception of space did not quite acquire an explicit mathematical status until the times of Descartes (1596–1650) and Fermat (1601–1665).

[2] They may be of a geometric nature, either directly, as in the case of a *metric space*, or through implied possibilities, as in the case of a *vector space*.

[3] Desirable though this practice is, it is not a necessary requirement from a purely logical point of view. Thus in geometry axiomatically formulated on modern lines, nothing is in principle lost if, as Hilbert is said to have remarked, wherever we say "points, lines, and planes", we say instead, "tables, chairs and beer mugs". For more on this, and on the axiomatic method as understood in the modern sense, see Kennedy [18] and Meschkowski [23, Chapter 8, pp. 63–71].

The grounds for this had, of course, been well prepared in the geometry of the Greeks. Euclidean geometry—that is, geometry as laid down in Euclid's *Elements* (around 300 BC), however, made no explicit use of the notion of space as such; the mathematically idealized objects that it introduced were points, lines and planes considered on their own, and not as members of space as a continuum.[4] Space was still intuitively conceived as the container of all physical objects, and the intended role of Euclidean geometry was to provide a mathematical framework within which the spatial relationships of objects in physical space could be studied in an idealized manner.[5]

As it turned out, this geometry did far more than that. It launched a radically new trend in deductive reasoning—that of axiomatics, which was to have in due course a profound and pervasive influence on all subsequent mathematical and scientific thought. But, to start with, the axiomatic approach of Euclidean geometry, while it showed a new way of dealing with mathematical ideas in general, it also gave rise to serious questions about geometry itself. Was geometry a part of physics, and were its axioms and postulates laws of physical space? Was the fifth postulate concerning parallel lines really an independent postulate, or was it a theorem that could be derived from the other axioms and postulates? Sustained enquiries of this kind continued for over two thousand years since the first appearance of Euclid's work. It was finally in the mid-nineteenth century that the turning point came, when the independence of the fifth postulate was established, and it was discovered that there were geometries other than Euclidean, the so-called non-Euclidean geometries.[6]

In the new perspective that emerged, Euclidean geometry began to be seen in a sharper focus as just another mathematical system, and not as *the* geometry of *the* space, i.e., physical space. Every geometry was now seen to have its own space, the points and lines of which were defined by the axioms of that geometry.

This change was part of a major reform at a more fundamental level, which had to do with the nature and role of definitions, and with the axiomatic method in general. In the axiomatic system of Euclid, the starting point was a set of definitions

[4] Presumably, there was no need at this stage to consider space as a whole. How a set of straight lines, circles etc., were related to each other was for them a matter to be sorted out locally, without any reference to space in its entirety. The idea of coordinatizing space as a whole, and of referring to geometrical figures and curves in terms of coordinates of their points, originated in the works of Descartes and Fermat on what came to be known as analytic geometry. Of their independent works on the subject, the first to come out and to draw public attention to it was Descartes' *La Géométrie*, published in the year 1637 as an appendix to his *Discourses*. An English translation of this fascinating piece of work is available [30]. See Boyer [3] for a detailed historical account.

[5] See Einstein [11] for a very lucid account of this point and for a general discussion on the origins of the concept of space.

[6] Gauss (1777–1855) was ahead of all others in seeing all this, but he kept his ideas to himself, fearing that his radical views would be adversely received. Public attention was first drawn to these ideas through the independent works of Lobatchevsky (1793–1856) and Bolyai (1802–1860), and it was subsequently the unifying work of Riemann (1826–1866) on the foundations of geometry that brought them from the fringes to the mainstream of mathematical thinking of the time. See Boyer [4, Chapter 24, pp. 585–590] for more historical details.

of certain idealized mathematical objects (e.g., points and lines), followed by a set of axioms and postulates. In the modern axiomatic approach, which grew out of this reform and which is what is in vogue today, the idea of introducing idealized objects through definitions is abandoned. The objects (e.g., points and lines of geometry) are, instead, introduced simply as formal terms—the primitives, with no intuitive meanings attached to them, and they are considered to be *defined* by the axioms that follow.[7] The full and final transition to this modern interpretation of axiomatics, which was complete by 1899, is attributed to the works of Pasch, Peano and Hilbert.[8]

For mathematics on the whole, space concurrently became a general term, to be understood in the sense of modern axiomatics, as a set with members formally called points and defined by the axioms. There followed in rapid succession the notions of metric spaces, normed spaces and inner product spaces, each with its own axiom system.[9]

From the intuitive settings of classical geometry and its concerns with physical space, we have thus moved to the modern settings of abstract spaces, in which we talk of spaces of functions or of sequences in much the same way as we talk of the points of the Euclidean plane.

Besides this influence on the notion of space, the rise of the modern axiomatic approach had an equally profound influence on the classification of mathematics as a whole. While its topic-wise division into analysis, algebra, geometry, calculus etc., continued to hold, there came into being a more fundamental classification based on what is now commonly known as the structural viewpoint.

Culminating in the works of Bourbaki,[10] this viewpoint is based on the observation that no matter what mathematical constructs one considers, whether they belong to algebra, analysis or geometry, one can characterize and classify them in terms of the kinds of operations, relations, and neighbourhood notions that enter their axiomatic formalization.

[7] Ernst Mach (1838–1916), in his *Science of Mechanics* published in 1883, is credited to have explicitly articulated for the first time this modern viewpoint. See von Mises [34], and also Hempel [15] and Wilder [35], for details on this, and for very illuminating discussions on the axiomatic method as it is understood today.

[8] Hilbert's *Grundlagen der Geometrie*, published in 1899, brought it out in full glory in the axiomatization of geometry. Peano had, in 1888, already axiomatized the concept of vector spaces over the field of real numbers. See Kennedy [18]. Also see Boyer [4, Chapter 26, p. 659].

[9] Maurice Frechet, in his Ph.D. thesis in 1906, introduced the abstract concept of a metric space. See Bushaw [5, Chapter 1, pp. 1–3] and Boyer [4, Chapter 27, p. 667]. Also see Kline [19, Chapter 46, pp. 1076–1095] for a history of function spaces and functional analysis.

[10] Nicholas Bourbaki is, if one were to go by the works published in this name, one of the most outstanding mathematicians appearing on the scene in the twentieth century. In actual fact, however, the name is a collective pseudonym for a continually changing group of leading mathematicians, who have been at work since the 1930s on a unification drive for mathematics. As to the outcomes of their labours in this drive, these are perhaps too advanced for us to worry about here; being aware of their existence is just about enough. See Cartan [6] for more on the myth and reality concerning Bourbaki.

Consider for instance, the case of numbers under addition, and of rotations of the plane under composition. What is it that is common between them? Well, the rules of addition and the rules of combining rotations (e.g., commutativity and associativity), both are essentially the same, if we look at them in a properly codified symbolic form divested of the specific meanings attached to them in the two cases.

Such commonality as we observe in these instances is a part of a general pattern in which one mathematical construct appears to be the same as another if we abstractly consider only the way its form is "internally held together", visualizing it to have a *structure* somewhat in the manner that a building has a frame or a structure that holds it together.

At this level of abstraction, we think of an (abstract) structure, $\mathcal{S}$, as a triple, consisting of (i) a set, say $S$, (called its ground set), (ii) a list, $\mathcal{O}$, of symbols denoting names of operations, relations etc., over the members of $S$, and (iii) a set, $\mathcal{L}$, of rules or laws that axiomatically define the scope and behaviour of the members of the list $\mathcal{O}$. In notation, we write: $\mathcal{S} = (S, L, \mathcal{L})$. To simplify the notation, it is common practice to omit mentioning $\mathcal{L}$, assuming that it is there and it is given, and also to explicitly list the members of $\mathcal{O}$ if there are not too many of them. We further simplify the notation by using the same symbol for the structure as well as its ground set. We thus simply write $S = (S, +, \times)$ for a structure with two operations named "$+$" and "$\times$".

At the core of Bourbaki's classification, there are three basic types of structures, the so-called *mother-structures*, by whose amalgamation other more complex structures may be obtained.[11]

There are, to start with, the *algebraic structures*, in which the relationships that hold the members of the ground set together are defined by laws of composition. Thus, numbers under the operation of addition constitute an algebraic structure, in which the laws of addition are the pertinent laws of composition. It is an instance of an abstract algebraic structure called a group. There may, of course, be two or more operations that together define such a structure, as in the case of rings and fields.

Next, there are *relational structures*, in which ordering relations between elements of a set (e.g., the relation "$x$ is at most as large as $y$" for numbers) play the defining role. A set of sets under the containment relation is an instance of such a structure. So is a set of logical propositions under the relation of implication. Posets and lattices are two of the commonly encountered relational structures of the abstract kind.

Finally, there are the *topological structures*, which rest on an abstract axiomatic formulation of the intuitive notions of neighbourhood, distances, limits and continuity. Spaces of different kinds (e.g., metric spaces, normed spaces and inner product spaces) fall in this category.

Our interest here lies primarily in algebraic and relational structures, or rather, in studying the various ways in which they make their appearance in the theory

[11] See Bourbaki [2], Cartan [6], Yaglom [36, Chapter 3, pp. 63–79] and Piaget [26, Chapter 2, pp. 17–36] for a more detailed discussion.

of digital signal processing, and the manner in which they help systematize our understanding of signal processing concepts. In line with this interest, we now turn to the notion of signal spaces.

## 1.3 Signal Spaces and Systems

In the study of signals, there are usually many issues that are geometrical in character. It is commonly the case for instance that we need to have for signals some empirically meaningful measures of comparison. Thus, if the signals pertain to sounds then, by inspecting their representations, we need to be able to say whether one sound is louder than another. Likewise, we should be able to say how different they are from one another. We generally do this by introducing a distance function, a function that assigns to every pair of signals a non-negative number analogous in nature to the distances between points in physical space.

Again, when we talk of spectral components of signals, we essentially visualize them as points in a multi-dimensional space, like points of physical space in a three-dimensional coordinate system.

On the whole, it is thus appropriate to think of signals as points in a space, using the term in the sense that we have discussed in the previous section. It may well be that in a particular study, the geometrical features of signals do not explicitly come to the fore to start with. The theoretical framework to be used in the study must nevertheless allow for their inclusion.

So we generally refer to a class of signals as constituting a *signal space*, with its quintessential properties specified by assigning to it a structure, usually algebraic. As in the case of signals individually, for the signal space too, we work in the representation domain, assuming that a suitable representation for it has already been determined. Since our interest lies in what its structure implies for a space, the structure itself we call the space.

Now, with signals and signal spaces so visualized, what of (signal processing) systems? At the physical level, a system is intuitively thought of as an arrangement of devices or algorithms that act on a signal received as input to produce another signal as the output. The familiar diagram depicting this is shown in Figure 1.1.

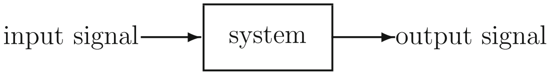

Figure 1.1 A system diagrammatically depicted

Stripped of all physical details, and reckoned in the representation domain solely in terms of the effects it has on signals, a *system* is then to be regarded as a set of ordered pairs of signals, i.e., as a relation on the ground set of the signal space.[12] Thus, if $X$ is the ground set, then any subset $H$ of $X \times X$ is a system, and for an ordered pair $(x, y) \in H$, $x$ is the *input* and $y$ is the *output*. It is ordinarily the case that for any input, there is just one output determined by the system so that, instead of being considered as a relation, it may be treated as a function or a mapping, $H : X \rightarrow X$, on the ground set. In what follows, we shall confine ourselves to such systems. We do not, however, prohibit them from being *many-to-one*, i.e., from being such that two different inputs produce the same output. Certain important classes of them are, of course, invariably *one-to-one*, producing different outputs for different inputs.

To see what all this means in concrete terms, let us now take up a specific case. Consider a class of signals (commonly called 1-D signals) for which it is already settled that they can be represented by sequences of reals, and that the relevant operations on such representations are those of addition, scaling and convolution.

We begin with some basic notions, terminology and notation. For signals we admit all (one-sided) infinite sequences of reals indexed by the natural numbers. For such a sequence $x = \langle x_0, x_1, x_2, \ldots \rangle$, or in short $\langle x_i \rangle$, we call $x_k$ its *$k$-th sample*. Treating the sequence as a real-valued function of natural numbers, $x : \mathbb{N} \rightarrow \mathbb{R}$, we also at times write $x_k$ as $x(k)$, the value of $x$ at $k$, and the domain of $x$, in this case $\mathbb{N}$, we call the *index set* of the signals. We use $V$ to denote the set of all such sequences.

We say two sequences $x$ and $y$ are equal, and write $x = y$, if $x_n = y_n$ for all $n \in \mathbb{N}$. We define their *sum* to be a sequence $z$, $z = x + y$, obtained by component-wise, or point-wise, addition: $\langle z_0, z_1, z_2, \ldots \rangle = \langle x_0 + y_0, x_1 + y_1, x_2 + y_2, \ldots \rangle$. By *scalar multiplication* of $x$ by a real number $\alpha$, we mean the operation of producing the sequence $\langle \alpha x_0, \alpha x_1, \alpha x_2, \ldots \rangle$, which we write in short as $\alpha x$; for $\alpha = -1$, we write it simply as $-x$. Note that for any $x$, $x + (-x) = 0)$. Thus, as for numbers, subtraction as the inverse of addition is well-defined for sequences too; the sum $x + (-y)$, we shall in short write as $x - y$.

In order to formulate storage and recall operations on values of a sequence as operations on sequences, we bring in a *shift operator*, $D$, that produces from a sequence $x$ a shifted sequence $\langle 0, x_0, x_1, x_2, \ldots \rangle$, which we write as $Dx$. This unary operation is in effect a function $D : V \rightarrow V$ such that $Dx(0) = 0$ and $Dx(k) = x(k - 1)$ for $k \geq 1$.

From amongst the members of $V$, we identify certain special ones whose role will become clear as we go along. For any real number $\alpha$, we call the sequence $\langle \alpha, 0, 0, \ldots \rangle$ a *scalar sequence* $\alpha$. In particular, we have the *zero sequence*, $0 = \langle 0, 0, 0, \ldots \rangle$, and the *impulse*, $\langle 1, 0, 0, \ldots \rangle$, which we give the special name $\delta$. Then there is the *unit step*, $\sigma = \langle 1, 1, 1, \ldots \rangle$. Finally, there is the *shifted impulse*,

[12] We assume here for our purposes that the input and output signals are members of the same signal space. They may, in principle, belong to different spaces. We ignore this situation.

$\omega = \langle 0, 1, 0, \ldots \rangle$. In using $\alpha$ to denote numbers as well sequences, there is admittedly some abuse of notation, but it will always be clear from the context as to what is meant.

Given sequences $x$, $h$ and $y$, we say that $z$ is the *convolution* of $x$ and $h$, $y = x * h$, if their components are related by the formula[13]

$$y_k = \sum_{r=0}^{k} x_r h_{k-r} \tag{1.1}$$

The signal space in this case is then the algebraic structure $V = (V, \mathbb{R}, D, +, *)$, with $\mathbb{R}$ denoting here the set of unary operations on $V$ that consist of scalar multiplication by reals.

As to a system, to go by the general definition given earlier, it is in this case a function from $V$ to $V$. There is, however, a limitation. Recall that the common definition of a function is simply that it is a list or a table of ordered pairs of domain and range elements such that every domain element is associated with a unique range element. A rule is assumed to be given that enables us to identify for any domain element the corresponding range element from this table. But what is this rule? What if the rule says, “Look up the table”? It would make no practical sense if the domain and range sets are large or infinite.

We overcome this difficulty here by assuming that the system function takes the form of a computational algorithm. To be more specific, by processing an input sequence $x$ through a system we shall understand computations on the input values to produce a new sequence $y$ at the output, the computations organized in stages such that at the $k$-th stage of computing, $y_0, y_1, \ldots, y_k$, *its values up to index* $k$, become available. Further, the computations at any stage are to involve the usual real number arithmetic on a finite set of real numbers, some of them being values of $x$ and others preassigned and stored in memory independently of $x$.

The shift operator, $D$, or any (finite) power of $D$, is in this sense admissible as a system. Further, if in the convolutional formula (1.1) we treat $h$ as a given fixed sequence (with values $h_k$ given in a closed form as a function of $k$) then this formula too defines a function that is admissible as a system, which we will call a *convolutional system*.

Let us now examine certain structural issues related to our signal space. Recall that for an abstract structure, the ground set is an abstract set, with operations and relations for its members defined by a set of axioms that are explicitly laid down without any reference to what the members of the ground set stand for. In the present case, we have a concrete situation in which the ground set consists of a special kind of numerical objects—the sequences of reals. The operations on them have been defined in terms of operations of another structure, namely, the structure of real numbers.

These operations on sequences do, however, have properties whose statements are free from any explicit reference to the arithmetic of real numbers

[13] I assume that you are familiar with this convolutional formula and its basic properties.

(e.g., $x + y = y + x$). As we shall see, some of these properties are sufficiently deep to merit the role of axioms for an abstract structure of which our signal space $V$ is a particular instance. Let us, to start with, consider the following.

**P**1 $V$ is closed under every one of its operations, i.e., the sequences produced by them are also members of $V$.

**P**2 For any $x \in V$, $x = x$. Further, for $x, y, z \in V$, if $x = y$ then $y = x$. Also, if $x = y$ and $y = z$ then $x = z$. These are respectively the so-called reflexive, symmetric and transitive properties of the equality relation.

**P**3 $x + y = y + x$, i.e., addition of sequences is commutative.

**P**4 For any $x, y, z \in V$, $x + (y + z) = (x + y) + z$, i.e., addition is associative.

**P**5 $x + 0 = 0 + x = x$; we call 0 an identity element of addition.

**P**6 For any sequence $x$, there is the sequence $(-1)x$ such that $x + (-1)x = (-1)x + x = 0$; we say that for $x$, $(-1)x$ is its additive inverse.

**P**7 $\alpha(x + y) = (\alpha x) + (\alpha y)$ for any $x, y \in V$ and any $\alpha \in \mathbb{R}$.

**P**8 $x * y = y * x$, i.e., convolution is commutative.

**P**9 $x * (y * z) = (x * y) * z$, i.e., convolution is associative.

**P**10 $x * \delta = \delta * x = x$, i.e., there is an identity element for convolution.

**P**11 $x * (y + z) = (x * y) + (x * z)$, i.e., convolution distributes over addition.

**P**12 For any two sequences $x$ and $y$, if $x * y = 0$, then either $x = 0$, or $y = 0$ or both.

**P**13 $\alpha(x * y) = (\alpha x) * y = x * (\alpha y)$ for any $x, y \in V$ and any $\alpha \in \mathbb{R}$.

**P**14 $\alpha x = \alpha * x$, i.e., scalar multiplication by $\alpha$ is equivalent to convolution by the scalar sequence $\alpha$.

**P**15 $Dx = \omega * x$, i.e., the operation of shifting is equivalent to convolution by the shifted impulse.

Starting from the numerical definitions of sequences and operations on them, it is an elementary matter to check these properties; only for **P**12, and may be for **P**13–**P**15, we need to do some extra work.

Let us now suppress the fact that the members of $V$ are sequences, and also ignore the definitions of equality and the various operations in terms of components of sequences. All we know about them, let us say, is that they satisfy the conditions **P**1–**P**13, whatever the elements of $V$ may actually be.

Significantly, a great deal can be deduced about the members of $V$, and the operations on them, solely from these properties. Here is a simple example that gives a foretaste of their potential power.

**Example 1.3.1** As mentioned in **P**5, $0$ is an identity of addition, satisfying the condition $x + 0 = 0 + x = x$. Let us check whether there is another element, $u$, which also satisfies this condition.

One could check this for sequences by proceeding as follows. Assume that there is such a $u$. Then for any sequence $x$, $x_k + u_k = x_k$, and therefore from the properties of real numbers, $u_k = 0$ for all $k$. That is, there is no identity element of addition other than $0$.

But this could also be shown without bringing in numbers. Again suppose that besides an element $0$, there is another element $u$ that satisfies the condition $x + u = u + x = x \quad \text{for any} \quad x \in V$. We then have in particular, $0 + u = u + 0 = 0$. But, since $0$ is itself an identity, $u + 0 = u$. Combining the last two sets of equalities and invoking **P**2, we get $u = 0$. On the whole, we can then say that *an identity element for the operation $+$ on $V$, given that it exists, is unique.* □

Observe that in this alternative derivation of the result, the "internal" numerical nature of sequences does not figure at all. The result about the uniqueness of identities is thus a general result for an abstract algebraic structure whose axioms include **P**2 and **P**5.

Packaged in different combinations, **P**1–**P**13 serve as axioms for a variety of abstract structures, each of which is rich in properties derivable from its axioms. A selection of such abstract structures will be discussed in the next chapter. For now, we continue with the signal space of sequences.

One important point about this signal space, as we find from properties **P**14–**P**15, is that the operations defined on sequences are not quite independent of each other; scalar multiplication and the unary operation $D$ both can be replaced by convolution. Further, even if we confine ourselves to addition and scalar multiplication, the operator $D$ and convolution can be accommodated in the form of systems. This suggests that the signal space of sequences can be treated, without any loss of generality, in two alternative abbreviated forms: one, as the structure $(V, +, \mathbb{R})$ and two, as the structure $(V, +, *)$.[14] It is, however, the first form that is more in tune with the approach commonly adopted in the study of signals and systems. Let us therefore focus on that first, and examine further the nature of systems.

For the signal space $(V, +, \mathbb{R})$, there are three basic methods of combining systems to produce new systems—cascading, summation, and scaling.

***Cascade*** Given two systems $H$ and $G$, their *cascade*, $HG$, is the system defined in accordance with the usual composition law of functions: $(HG)x = H(Gx)$. For $HH$, we write $H^2$ in short, so that by $H^2(x)$ we mean $H(Hx)$.

[14] The first form makes it a vector space, and the second one a ring.

***Sum*** For two systems $H$ and $G$, their *sum*, $H+G$, is the system defined by the condition $(H+G)x = Hx + Gx$.

***Scalar Multiple*** For a system $H$, a *scalar multiple* of it, written $\alpha H$, is a system defined by the condition $(\alpha H)x = \alpha(Hx)$ for every $x \in V$, where $\alpha$ is a real number.

Consider now the set, $\mathcal{S}$, of all systems on the signal space $V$. Cascading and summation are binary operations on $\mathcal{S}$, and scalar multiples define for it a family of unary operations, which we denote by $\mathbb{R}$ as in the case of signals.[15] Closure of $\mathcal{S}$ under these operations is implicit in their definitions.

It is easily verified that both cascading and summation are associative operations. Further, summation is commutative, but cascading is in general not. Additional properties will come to light as we go along. The important point that follows from all this is the following.

**Proposition 1.3.1** *Like the signal space $V$, systems on $V$ also constitute collectively a structure, namely, the structure $(\mathcal{S}, +, \circ, \mathbb{R})$, where $\circ$ denotes cascading (composition).*[16]

This structure becomes all the more "lively" and interesting when we impose on systems additional constraints of practical significance.

## 1.4 Linearity, Shift–Invariance and Causality

We have already outlined certain computational constraints on the form of systems we are to admit in our study of the signal space $(V, +, \mathbb{R})$. But, even with these constraints, the choice we have for their form is enormously vast, in fact too vast to be of relevance for most practical purposes. Requiring them to be linear, shift–invariant and causal is one way of judiciously restricting this choice. We now look at these three properties one by one.

### 1.4.1 Linearity

Very broadly, a system is said to be linear if it satisfies the superposition principle, or alternatively, if it is additive and homogeneous. In our case it means that a system $H : V \to V$ is *linear* if, for any $x, u \in V$ and $\alpha \in \mathbb{R}$, it meets the conditions:[17]

$$\begin{array}{lrcl} \text{Additivity:} & H(x+u) & = & H(x) + H(u) \\ \text{Homogeneity:} & H(\alpha x) & = & \alpha(Hx) \end{array} \tag{1.2}$$

[15] While cascading is valid on any signal space whatever, summation and scaling are structure–specific; note the way the addition and scaling operations of the signal space enter their definitions.

[16] Such a structure is commonly referred to as an *algebra*.

[17] I should mention here as an aside that the two conditions together are equivalent to the single condition $H(\alpha x + \beta y) = \alpha(Hx) + \beta(Hy)$ for any real $\alpha$ and $\beta$. Further, homogeneity is, at least partially, implied by additivity. For, $H(1x) = 1(Hx)$, and assuming additivity,

**Example 1.4.1** It is easily checked that the shift operator, $D$, is a linear system. So is a convolutional system. □

Let us now see how the various operations on systems are disposed towards linearity.

**Proposition 1.4.1** *If systems $F$ and $G$ are linear then so are $FG$, $(F+G)$ and $\alpha F$, where $\alpha$ is any scalar multiplier. Further, for the set of all linear systems, the cascading operation distributes over summation.*

PROOF: Using additivity of $F$ and $G$, $FG(x+u) = F(G(x+u)) = F(Gx+Gu)) = FGx + FGu$, for any $x, u \in V$, i.e., $FG$ is additive. Its homogeneity follows in a similar manner, and so $FG$ is linear. Likewise we show for the sum and scalar multiple.

Consider now three systems $F$, $G$ and $H$, and suppose that $F$ is linear. Then $F(G+H)x = F(Gx+Hx)$ by definition, and from the linearity of $F$, finally, $F(G+H)x = (FG)x + (FH)x$ for any $x \in V$. Thus, cascading in this case distributes over summation. This is, however, not true for $G(F+H)$ or $H(G+F)$, unless we assume $G$ and $H$ also to be linear. That is, for linear systems, cascading distributes over summation. □

An instructive way to look at a linear system is to regard it as a *structure-preserving mapping*. Let me clarify this point with the help of an analogy. When you look at yourself in a mirror, your face in the mirror image has a likeness to your actual face, even though the depth information is lost in the image. Neighbouring points of your facial contours remain neighbouring points in the image, and so on. Indeed, the mirror establishes between an object plane in front of it and the corresponding image plane, not just a simple point-by-point mapping, but one that keeps intact many of the geometrical relationships amongst the points of the object plane.

Consider now a structure consisting of a ground set $V$ and a number of relations $R_1, R_2, \ldots, R_n$ on $V$. For a mapping $H : V \to V$ to have the mirror-like property, it needs to be such that it *preserves* the relations $R_i$, i.e., if any of the relations holds for a set of points of $V$ before applying the mapping then it should also hold after applying the mapping. More precisely, if one of the relations, say $R_1$, is a binary relation and $(x, y) \in R_1$ for $x, y \in V$ then $(H(x), H(x)) \in R_1$. Similarly, if $R_2$ is a ternary relation and $(x, y, z) \in R_2$ then $(H(x), H(y), H(z)) \in R_2$. Since a unary operation is a binary relation, and a binary operation is a ternary relation and so on, our stipulation for relational structures equally covers algebraic structures.[18]

$H(nx) = H[(n-1)x + x)] = H[(n-1)x] + H(x)$. If homogeneity is true for $(n-1)$, then by mathematical induction, $H(nx) = nH(x)$ for any positive integer $n$. With a little more work on the same lines, it can be shown that for any rational $\alpha$, additivity implies homogeneity. This cannot be said, however, for all reals. There are interesting deep set theoretic matters involved here. See Hrbacek and Jech [16, Chapter 9, pp. 177–178].

[18] A unary operation on a set $X$ maps a member $x \in X$ into another member $y \in X$, so that the operation can be looked upon as a binary relation consisting of a subset of $X \times X$; likewise, a binary operation can be looked upon as a ternary relation consisting of a subset of $X \times X \times X$.

Thus, *to say for our signal space $(V, +, \mathbb{R})$ that a system $H : V \to V$ satisfies the linearity conditions (1.2) is to say that it is a a structure–preserving mapping.* With this in mind, we are in a position to look at linearity in a broader light, relating it to what are called homomorphisms of structures. As an illustration of this point, consider the fact that a system is in general permitted to be a many-to-one mapping, i.e., for two distinct inputs it may produce the same output. Now, if it is linear, are there conditions under which it would be one-to-one, i.e., distinct inputs will produce distinct outputs? The following result, which is a special case of a property of homomorphisms in general, gives one such condition.

**Proposition 1.4.2** *An additive system $H : V \to V$ is one-to-one if and only if $Hx = 0$ implies that $x = 0$.*

PROOF: Let us first take note of some supportive details. To begin with, since $H$ is additive, $H(0) = H(0+0) = H(0)+H(0)$ so that $H(0)-H(0) = H(0)+H(0)-H(0)$, i.e., $H(0) = 0$. Consequently, for additive $H$, $H(x)+H(-x) = H(x-x) = H(0) = 0$, i.e., $H(-x) = -H(x)$ for any $x \in V$.

Now suppose that if $Hx = 0$ then $x = 0$. Consider $x, y \in V$ such that $Hx = Hy$. Then $H(x) - H(y) = H(y) + H(-y) = 0$. Then, since $H$ is additive, $H(x-y) = 0$, i.e., $x - y = 0$. Thus $H(x) = H(y)$ implies that $x = y$, i.e., $H$ is one-to-one.

Conversely, if $H$ is one-to-one, then for any $u \in V$, if $H(u) = 0$, since $H(0) = 0$, $u = 0$.[19] □

**Comment:** Consider alongside this result the case of linear simultaneous equations, $Ax = b$, where $A$ is a square matrix. For a given arbitrary vector $b$, how many solutions does it have? We know that if there is a nonzero $u$ such that $Au = 0$ then the equation has more than one solution. If, on the other hand, there is no such $u$ then it has a unique solution. Correspondingly, $A$ is invertible, and there is a one-to-one relationship between the given vectors and the solutions.

One important consequence of assuming linearity is that it significantly reduces our work in cataloguing system input-output pairs. Thus if the outputs $H(x)$ and $H(u)$ for inputs $x$ and $u$ respectively have already been determined, then we need not separately determine the output for any linear combination of $x$ and $u$. More generally, if our signal space is such that every signal is a linear combination of a fixed finite set of signals then a matching linear combination of the responses to these fixed signals gives the output for any other input. Further simplifications result if we invoke, in addition to linearity, the remaining two properties, namely, shift–invariance and causality. Let us now consider them one by one.

[19] The various intermediate steps might seem like fussing over what is trivially obvious. But that is because of our familiarity with the properties of numbers under addition. You will appreciate their merit better if you study algebraic structures in general.

## 1.4.2 Shift–Invariance

System theory was, in the early phases of its development, concerned mainly with the study of signals related to temporal events, and with their processing in “real time”. A system in that context came to be called *time–invariant* if its input-output relationship did not change with time.[20] More specifically, if no changes were made in the input except shifting it in time by a certain amount, then for such a system the corresponding output too would not undergo any changes except an equal amount of (time) shift or translation. Note that the role of time is incidental here. The crucial thing is that for functions representing signals, the domain should have a structure that allows us to properly define their translates or shifted versions. It does not matter whether the physical origin of the signals is temporal, spatial, or any other. For the property under consideration, a more appropriate name therefore is shift–invariance.

Now, for the signal space $(V, +, \mathbb{R})$, we have a natural way of defining shifts, and of describing them using the shift operator $D$. For a signal $x$, $Dx$ is its translate with a single shift, $D^2x$ its translate with two shifts, and so on.[21] We then say that a system $H : V \to V$ is *shift–invariant* (or *translation–invariant*) if it commutes with the shift operator $D$, i.e., if it satisfies the condition

$$HDx = DHx \quad \text{for any} \quad x \in V. \tag{1.3}$$

**Comment:** Note that instead of (1.3), textbooks usually stipulate the apparently more general requirement: $HD^n = D^nH$ for any $n$. But, if $HD = DH$, then $HD^2 = HDD = DHD = D^2H$, and likewise, $HD^n = D^nH$ for any $n$. Condition (1.3) is thus equally general for the present purposes. We will see, however, when we consider other kinds of signal spaces, that the condition needs to be stated for several (generalized) shift operators rather than just one. The needs in this respect are decided by the structure of the index set.

**Example 1.4.2** The system $D^n$, for any finite power $n$, is shift–invariant. So is a convolutional system. □

**Proposition 1.4.3** *The set of shift–invariant systems is closed under the operations of cascading, summing and scalar multiplication. In other words, for shift–invariant systems $G$ and $H$, their cascade $GH$, sum $G + H$ and scalar multiple $\alpha F$ are also shift–invariant.*

PROOF: For the cascade operation, this can be brought out as follows: $(GH)Dx = G(HD)x = (GD)(Hx) = (DG)(Hx) = D(GH)x \quad \text{for every} \quad x \in V$.

[20] In physical terms, an $RLC$ circuit, for instance, produces the same response no matter when you apply the input, whether today or tomorrow.

[21] Recall that $D^n$ denotes the composition of the operator $D$ applied $n$–times in cascade. For instance, by $D^2x$ we mean $D(D(x))$.

As to the sum, using linearity of $D$ in the last but one step, its shift–invariance is seen as follows: $((F+G)D)x = (F+G)(Dx) = F(Dx)+G(Dx) = (FD)x+(GD)x = (DF)x+(DG)x = D(Fx)+D(Gx) = D(Fx+Gx) = D(F+G)x.$

Similar manipulations, coupled with the fact that $D$ is linear, establish shift–invariance of the scalar multiple $\alpha F$. □

The set of shift–invariant systems is a proper subset of the set of all systems. Yet, in view of its closure under cascading, summing and scaling operations, it is a structure in its own right—a structure "within" the structure that we have mentioned earlier of all systems on $V$.[22]

### 1.4.3 Causality

We finally turn to causality. For signals and systems that relate to temporal events, its lay meaning is essentially that effects cannot appear before their causes. Unlike linearity and shift–invariance, causality is in this sense an intrinsic feature of systems for signals that represent temporal events. Whether we think of time as consisting of instants or epochs, there is the past, the present and the future, and ruling out for us the power to know what is to come in the future, all we can hope to achieve and work for is to condition our responses to what has already been up to now.

Carried over to the more technical setting of system theory, this essentially amounts to saying that if two temporal signals are identical up to a certain point in time then the outputs of a system corresponding to these signals are also identical up to that point in time.

Since over physical space, there is no such natural ordering of points as in the case of time, and no corresponding distinction like past, present and future, it follows that for signals of spatial origin, such as pictures and images, there is no compulsory requirement of causality for systems. All the same, we very often artificially impose this condition on systems even in more general settings not necessarily temporal.

In the case of the signal space $(V,+,\mathbb{R})$, we define a causal system as follows: Consider a system $H : V \to V$ and any two of its inputs $x$ and $u$. We say that the system $H$ is *causal* if the inputs are identical up to some index $n$, i.e., $x_k = u_k \quad \text{for} \quad 0 \le k \le n$, then the corresponding outputs are also likewise identical, i.e., $(Hx)_k = (Hu)_k \quad \text{for} \quad 0 \le k \le n.$

**Example 1.4.3** The operator $D^n$ is causal, and so is a convolutional system. □

Observe that this definition depends on the ordering relation "$\le$" that is available on the index set $\mathbb{N}$. In all then, to take care of the notions of shift–invariance

[22] Formally, it is an instance of a substructure, a term we shall properly examine in the next chapter.

and causality, we are pressing into use for the index set of signals the structure $(\mathbb{N}, +, \leq)$. Admittedly in this case, "+" and "$\leq$" are related.[23] But in more general situations, the two may be independently defined. In some situations, there may not be an order relation defined on the index set, and in that case we do without the notion of causality.

### 1.4.4 Characterization

In a technical sense, to characterize the members of a set defined in terms of a certain property is to give an explicit form, or an alternative expression, which the property entails for them in the light of some additional properties that the set has.

Thus, when we define linearity for functions, we essentially declare, to start with, that their domain and range have adequate structure to allow for additivity and homogeneity conditions to be stated for them, and then stipulate that the conditions are satisfied by the functions. If we now additionally stipulate for them the property that they are real-valued functions of a real variable, then for such a function, say $f : \mathbb{R} \to \mathbb{R}$, we can further claim that it is linear if and only if it has the form $f(x) = ax$, where $a = f(1)$. In so claiming, we have given a characterization of such linear functions. Of course, a characterization in general may not be as explicit as this. It may even consist of a necessary and sufficient condition that guarantees the defining property.[24]

Characterization of linear, shift–invariant and causal (LSIC) systems is what we examine next. We have, to start with, the following result that provides an alternative simpler form for the causality condition if we assume the system to be linear.

**Proposition 1.4.4** *A linear system on the signal space $V$ is causal if and only if, for any input whose values are zero up to an arbitrary index, the corresponding output is also zero up to that index.*

PROOF: Take the "only if" part first. For a linear causal system, $H : V \to V$, consider an arbitrary input $x$, and another input $u$ whose values are zero up to an index $n$ and arbitrary beyond $n$. Then inputs $x$ and $x + u$ are identical up to index $n$, and, since $H$ is causal, the outputs $Hx$ and $H(x + u)$ are also identical up to index $n$. But $H$ is linear, and so $H(x+u) = Hx+Hu$, and $Hu = H(x+u)-Hx$. Thus, for $k \leq n$, $(Hu)_k = 0$.

[23] For $a, b \in \mathbb{N}$, we say $a \leq b$ if and only if there is an $c \in \mathbb{N}$ such that $a + c = b$. The product $ab$ is similarly defined in terms of addition. Of course, for our present purposes, the product operation has no place in the role of $\mathbb{N}$ as an index set.

[24] A striking example of this is the so-called alternation theorem of minmax approximation theory. For the approximation of a real function by a polynomial of a fixed degree, it reads roughly as follows: the maximum deviation of the polynomial from the function is minimum over the interval if and only if the peaks of the error function alternate in sign as many times as the degree and are equal in magnitude. Filter designers make copious use of this result. See Rice [28, Chapter 3, p. 56] and Cheney [7, Chapter 3, p. 75] for the theorem, and Rabiner [27, Chapter 3, pp. 127–139] for filter applications.

Let us now argue in reverse for the “if” part. Suppose that the system is linear, and that if an input is zero up to index $n$ then the corresponding output is also zero up to index $n$. Let $x$ and $u$ be two inputs identical up to index $n$, so that $x - u$ is zero up to $n$. In that case, by hypothesis, $H(x - u)$ is also zero up to $n$. Then, invoking linearity of $H$, we conclude that $Hx$ and $Hu$ are identical up to $n$. □

**Exercise 1.4.1** Give an example of a system that is causal but nonlinear, and for which there is an input that is zero up to an index $n$, but the corresponding output is not zero up to the index $n$. (*Hint:* Examine the system for which the output $y$ is given in terms of the input $x$ by the formula $y_k = (x_k + 1)^2$.)

As already pointed out in Examples 1.4.1–1.4.3, if a system is convolutional, then it is also LSIC. As we shall presently see, the converse is also true. Putting these two facts together, we have, in all, the following result.

**Proposition 1.4.5** *An LSIC system, $H$, is characterized by a convolutional formula of the form (1.1), where $h$ is the sequence $H\delta$, the response of the system to an impulse. We thus have $Hx = x * h$ for any signal $x$.*

PROOF: Our full task is to show that a system is LSIC if and only if it is convolutional. The “if” part has already been verified by you.

Let us see the “only if” part. For any sequence $x$, let $x^{(n)}$ be the sequence obtained by truncating $x$ after $n$:

$$x_k^{(n)} = \begin{cases} x_k & : \quad 0 \le k \le n \\ 0 & : \quad k > n \end{cases}$$

Since $H$ is causal, clearly, $Hx$ and $Hx^{(n)}$ are identical up to the index $n$. For values of $Hx$ up to index $n$, we may therefore examine $Hx^{(n)}$ instead. Now, $x^{(n)}$ may be expressed, using the shifted impulses: $x^{(n)} = \sum_{r=0}^{n} x_r D^r \delta$.

Then, since $H$ is linear and shift-invariant, writing $h$ for $H\delta$, $Hx^{(n)} = \sum_{r=0}^{n} x_r D^r h$, and in terms of the sequence samples, $\left[Hx^{(n)}\right]_k = \sum_{r=0}^{n} x_r (D^r h)_k$. Now, by definition, $(D^r h)_k = 0$ for $r > k$ and $(D^r h)_k = h_{k-r}$ for $r \le k$, so that $\left[Hx^{(n)}\right]_k = \sum_{r=0}^{k} x_r h_{k-r}$.

Thus, finally, we have

$$(Hx)_k = \left[Hx^{(n)}\right]_k = \sum_{r=0}^{k} x_r h_{k-r} \quad \text{for} \quad k \le n.$$

Since $n$ is arbitrary, the convolutional formula holds in general for any output value of an LSIC system.[25] □

[25] The argument I have used here is not exactly one of mathematical induction, and it can be faulted for not being really rigorous. But it will do for the present.

**Comment:** Our use of truncates in the proof might raise some doubts. Why not just use the expansion $x = \sum_{r=0}^{\infty} x_r D^r \delta$, and for $Hx$ distribute $H$ over the infinite sum? Although this is very often done, we avoid this here. Addition, which is a binary operation, is defined, to start with, for two elements, and by successive applications it can properly be extended only to a finite number of elements. In order to talk of infinite sums, we have to bring in additional definitions incorporating notions of convergence. This is unnecessary here. Using truncates and invoking causality, we get a more appropriate and direct proof.

**Corollary 1.4.6** *Any two LSIC systems commute i.e., if $H$ and $G$ are two LSIC systems then $HG = GH$.*

PROOF: Being LSIC, $H$ and $G$ are convolutional, with impulse responses $h$ and $g$. Then for any $x$, since convolution is commutative and associative, $(HG)x = H(Gx) = H(x * g) = (x * g) * h = x * (g * h) = x * (h * g) = (x * h) * g = G(Hx) = (GH)x$. □

Proposition (1.4.5) explains why convolutional systems occupy such a central place in digital signal processing. Once we have determined the impulse response of an LSIC system, we can in principle compute its output for any input, using the convolutional formula. This result is certainly a powerful tool for the analysis of a system whose "inner details" we do not know except that it is LSIC. For purposes of design, however, it is not a very practical one. For, even though every output sample requires a finite number of computations in convolution, this number grows with the index.

It is therefore preferred in practice to confine only to those convolutional systems that can be *realized*, i.e., implemented in the form of an algorithm, such that for any output sample, there is required only a *fixed* finite number of computations. These are the so-called *non-recursive* and *recursive* realizations. They have the following forms.

$$\text{(1.4)} \qquad \text{Non–recursive:} \quad y = \sum_{i=0}^{m} a_i D^i x$$

$$\text{(1.5)} \qquad \text{Recursive:} \quad y = \sum_{i=0}^{m} a_i D^i x - \sum_{i=1}^{n} b_i D^i y,$$

where $m$ and $n$ are (fixed) given positive integers and $a_i$ and $b_i$ are given real numbers.

Although the non-recursive realization is a special case of the recursive one, it is common practice to consider it separately on account of its simpler properties. Looking at them together for the moment, (1.5) may be rewritten as

$$\text{(1.6)} \qquad b_n D^n y + b_{n-1} D^{n-1} y + \cdots + b_1 Dy + y = a_m D^m x + \cdots + a_0 x,$$

in the form of a difference equation, also called a recurrence relation, whose solution $y$ is sought for a given $x$. Expressed in terms of the samples of $y$ and $x$, it is

$$y_k + b_1 y_{k-1} + \cdots + b_n y_{k-n} = a_0 x_k + a_1 x_{k-1} + \cdots + a_m x_{k-m} \tag{1.7}$$

where, $x_r = y_r = 0$ for $r < 0$.

It is easily verified that both the realizations are LSIC, and are therefore characterized by their impulse responses. For the non-recursive type (1.4), the impulse response is the sequence

$$\langle a_0, a_1, \ldots, a_m, 0, 0, \ldots \rangle,$$

where the $a_i$'s are given coefficients.

An LSIC system whose impulse response has only a finite number of nonzero samples is commonly called a *finite impulse response* (FIR) system. Correspondingly, a LSIC system whose impulse response has infinitely many nonzero samples is called an *infinite impulse response* (IIR) system.

**Proposition 1.4.7** *An LSIC system is of the FIR type if and only if it has a non-recursive realization. A recursive system is of the IIR type, but not every IIR system has a recursive realization.*

PROOF: The first two parts are simple to verify. The last part calls for some work. Skipping the details, it should suffice for the present to mention that the impulse response of the recursive system (1.5) is the solution of the difference equation (1.7) for $x = \delta$. Such equations admit only certain kinds of sequences as solutions. It follows that not every sequence qualifies as the impulse response of a recursive (LSIC) system. An illustrative example is the sequence $\langle h_i | h_i = (0.5)^{i^2} \rangle$. □

## 1.5 Convolutional Algebra and the Z–Transform

So far we have looked at the signal space $V$ as the structure $(V, +, \mathbb{R})$. It is time for us to change tack and examine it as the structure $(V, +, *)$ instead. We shall presently see that it is an algebraic structure of the same species that integers under addition and multiplication are, and that we can embed it in a structure of fractions exactly in the same manner that from integers we obtain ordinary fractions. Such an embedding enables us to accomplish for sequences all of what is commonly accomplished with the help of the $Z$–transform.[26]

[26] A transform such as the Laplace or the $Z$–, with which I assume you are already familiar, maps the given functions into functions of a complex variable. The merit of this mapping is that, in addition to being linear, it replaces operations of differentiation, integration, shifting, and convolution of the original functions by ostensibly simpler algebraic operations in the transform domain. The idea of using addition and convolution as the basic operations on functions, and of using fractions of functions, was utilized by Jan Mikusiński [25] in the fifties to provide an alternative to the Laplace transform. For more on its applications to sequences, see Traub [33].

The structure $(V, +, *)$, which we call here a *convolutional algebra*, has its main properties already laid out in the set of statements **P**1–**P**15 on page 10. These properties are:

1. Closure under addition and convolution (**P**1)
2. Basic properties of equality (**P**2)
3. Basic properties of addition (**P**1–**P**6)
4. Basic properties of convolution (**P**8–**P**10 and **P**12)
5. Distributivity of convolution over addition (**P**11)

Reading multiplication for convolution, these are precisely the properties that govern the algebra of integers under addition and multiplication. We may thus say that our convolutional algebra is structurally the same as $(\mathbb{Z}, +, \cdot)$, the set $\mathbb{Z}$ of integers under addition and multiplication. Note in particular that, by **P**12, the usual cancellation law for multiplication of integers (i.e., for nonzero $a$, $b$ and $c$, $ab = ac$ implies that $b = c$.) has its equivalent for convolution of sequences. Bearing this structural similarity in mind, let us now refer to convolution as multiplication of sequences, and write "$x * y$" simply as "$xy$".[27]

One limitation that $\mathbb{Z}$ and $V$ have in common is that they are not equipped with multiplicative inverses, although they do possess multiplicative identities (1 for integers and $\delta$ for sequences). To be more precise, for any nonzero integer $a$ other than $1$, there is no integer $b$ such that $ab = 1$. It equivalently means that for given nonzero integers $a$ and $c$, there may not be an integer $b$ such that $ab = c$. Since this point is not as obvious in the case of sequences, let me give an illustrative example.

Consider a sequence $y$ for which $y_0 \neq 0$, and let $h$ denote its shifted version $\langle 0, y_0, y_1, y_2, \ldots \rangle$. You can see that $\omega y = h$. Is there a sequence $x$ such that $hx = y$? There is none. For, if there were, since $h_0 = 0$, its convolution with $h$ would give $y_0 = (hx)_0 = h_0 x_0 = 0$. But $y_0 \neq 0$ by hypothesis. So there is no such sequence $x$ in this case. Another way to put this fact is to say that for a sequence $h$ there may not be a sequence $g$ such that $h * g = \delta$. Thus convolution, like multiplication of integers, does not admit of an inverse operation.

The lack of multiplicative inverses is overcome in the case of integers by passing to fractions. For sequences too, the same strategy works. By a fraction $x/y$ let us now understand an ordered pair $(x, y)$ of sequences with $y \neq 0$, and let $\mathbb{V}$ denote the set of all such fractions. Consider then the new structure $(\mathbb{V}, +, \cdot)$, where, by

[27] Integers under addition and multiplication, and sequences under addition and convolution, are instances of a structure known as an *integral domain*.

analogy with fractions of integers, equality and its two operations, addition $(+)$, and product or multiplication $(\cdot)$, are defined by the following rules.[28]

1. We say $x/y = u/v$ if the sequences $x$, $y$, $u$ and $v$ satisfy the condition $xv = vy$. (Just as $2/3$, $4/6$ are equal fractions, although formed from different integers, so are apparently different fractions equal so long as they satisfy the stipulated condition. Thus $(xa)/(ya) = x/y$.)

2. The sum $(x/y) + (u/v)$ is any fraction equal to the fraction $(xv + uy)/(yv)$.

3. The product $(x/y)(u/v)$ is any fraction equal to the fraction $(xu)/(yv)$.

**Comment:** Notice that equality on $\mathbb{V}$, like that for ordinary fractions, partitions $\mathbb{V}$ into disjoint subsets of equal fractions; any two members of such a subset are interchangeable in computations with fractions. The idea will become clearer in the next chapter when we discuss equivalence relations.

With equality, addition, and product[29] so defined, fractions of sequences may be manipulated like ordinary fractions, and the structure $(\mathbb{V}, +, \cdot)$ behaves like the algebra of ordinary fractions.[30] Note in particular the following points. Using "1" to denote the unit sequence $\delta$, the fraction $0/1$, which we will write simply as $0$, is its additive identity, i.e., for any fraction $x/y$, $x/y + 0 = x/y$. The fraction $1/1$ (or simply $1$) is its multiplicative identity, i.e., $(x/y)1 = x/y$. For a fraction $x/y$, $x, y \neq 0$, $y/x$ is its multiplicative inverse, i.e., $(x/y)(y/x) = 1$. Every fraction other than $0$ has a multiplicative inverse. Finally, while a sequence $x$ by itself is not in $\mathbb{V}$, the corresponding fraction $x/1$, or any fraction equal to it, behaves virtually the same way. Thus, for sequences $x$ and $y$, $(x/1) + (y/1) = (x + y)/1$ and $(x/1)(y/1) = (xy)/1$. Sequences , implicitly treated as fractions with $1$ in the denominator, can therefore be regarded as members of $\mathbb{V}$.

Let us now take a look at certain special sequences expressed as fractions.

**Example 1.5.1** To express the unit step $\sigma = \langle 1, 1, 1, \ldots \rangle$ as a fraction, first expand it as

$$\sigma = \langle 1, 1, 1, \ldots \rangle = \langle 1, 0, 0, \ldots \rangle + \langle 0, 1, 1, 1, \ldots \rangle = 1 + \sigma\omega.$$

Then, rearranging and using the rules for fractions,

$$\sigma = \frac{1}{1 - \omega}$$

□

[28] Strictly speaking, we should use different symbols to designate equality and addition here, because they are not the same as those for sequences. But that would make the notation cumbersome. So we use the same symbols in both the cases.

[29] As in the case of the convolution sign, we shall omit the product sign and write $p \cdot q$ simply as $pq$ for fractions $p$ and $q$.

[30] Such a structure is called in algebra a *field*.

**Example 1.5.2** The sequence $x = \langle 1, a, a^2, a^3, \ldots \rangle$ may be expressed as

$$\begin{aligned} x &= \langle 1, 0, 0, \ldots \rangle + \langle a, 0, 0, \ldots \rangle \langle 0, 1, a, a^2, a^3, \ldots \rangle \\ &= 1 + a\omega x, \end{aligned}$$

where $a$ in the second step is a scalar sequence. The desired fractional form for it is then

$$x = \frac{1}{1 - a\omega}.$$

□

**Example 1.5.3** Consider an IIR system described by the recurrence relation $y = x + 0.5Dy$, with input samples $x_k = (0.8)^k$. Its output $y$, expressed in the fractional form, is then

$$y = \frac{1}{1 - 0.5\omega} x = \frac{1}{1 - 0.5\omega} \frac{1}{1 - 0.8\omega},$$

and by partial fraction expansion,

$$y = \frac{5/3}{1 - 0.5\omega} + \frac{25/24}{1 - 0.8\omega}.$$

Reverting to sequences, the output samples are then,

$$y_k = (5/3)(0.5)^k + (25/24)(0.6)^k$$

□

**Exercise 1.5.1** Show that the sequence $x$, $x_k = \sin(k\theta)$, has a fractional representation

$$x = \frac{\omega \sin\theta}{1 - 2\omega \cos\theta + \omega^2}$$

You should now be able to convince yourself that for a system described by the equation (1.6), the relationship between the input and output sequences can be expressed as a ratio of polynomials in $\omega$, analogous to the $Z$–transform transfer function:

$$\frac{y}{x} = \frac{a_0 + a_1\omega + a_2\omega^2 + \cdots + a_m\omega^m}{1 + b_1\omega + b_2\omega^2 + \cdots + b_n\omega^n}, \tag{1.8}$$

where $\omega^k$ means $\omega$ convolved $k$ times, and $a_i$'s and $b_i$'s are scalar sequences. Indeed, if in the fraction (1.8), $y$ and $x$ are replaced by their $Z$–transforms, and $\omega$ is replaced by $z^{-1}$ then it becomes the usual transfer function of the system in the $Z$–transform domain. Using the standard notation for the $Z$–transform, it is the function

$$\frac{Y(z)}{X(z)} = \frac{a_0 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_m z^{-m}}{1 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_n z^{-n}}, \tag{1.9}$$

where $X(z)$ and $Y(z)$ are the $Z$–transforms of the sequences $x$ and $y$, and $a_i$'s and $b_i$'s are the real numbers corresponding to the scalar sequences of (1.8).

It should be clear by now that the convolutional algebra has inherent in it all the arithmetic-like properties that we normally associate with signals and systems by going over to the transform domain. Are we to think that this algebra, or rather its extension to a structure of fractions, offers us an alternative superior to the $Z$–transform in any way? Not really; the two amount essentially to the same thing. It is, however, instructive on two counts.

First, it illustrates the general idea of creating new structures from old ones (fractions from integers or sequences). This idea consists of going through the following steps. Starting with a given structure (integers or sequences of reals), and using elements of its ground set, define a new ground set (e.g., ordered pairs of integers or sequences). Next, introduce for this new ground set a new set of relations and operations analogous in type to those of the older set, inheriting all their properties (as in the case of equality, addition and multiplication of fractions), and having some more of their own (like inverse of multiplication). The resulting new structure, consisting of the new ground set, taken together with the new relations and operations, is a richer structure than the older one, and at the same time it contains within it the older one (integers can also be interpreted as fractions).[31] We shall see more of this when we look at algebraic and relational structures more closely.

Secondly, it serves to illustrate the idea of treating signal spaces and systems in terms of their underlying structures. In the next section, we turn to yet another illustration of this idea, in which the focus is on transform domain techniques for signals and systems on finite index sets.

## 1.6 Shifts, Transforms and Spectra

The classical theory of linear time–invariant systems for continuous–time signals has given rise in its wake to analogous theories for various other classes of signals (e.g., 1-D, 2-D, discrete, and discrete finite). As we shall presently see, the connections between these theories go deeper than mere analogy. Observe, to start with, that all of them are centred on specific interpretations of two basic results: one, that a linear shift–invariant system is characterized by convolution, and two, that the transform of the convolution of two signals equals the product of their transforms (the convolution theorem).

Thus, in the case of discrete signals, the place of linear continuous–time convolution and Laplace and Fourier transforms is taken over in these results by discrete convolution and the $Z$–transform. The Discrete Fourier transform (DFT)

[31] Complex numbers, treated as ordered pairs of reals, provide another familiar illustration of this strategy.

acts the same way for discrete finite signals under cyclic convolution, and the Walsh–Hadamard transform[32] for dyadic convolution of discrete finite signals.[33]

A general theory, which covers all these variations as special cases, comes into view if we recognize that for the pertinent classes of signals, there are associated notions of translations or shifts over the index sets of these classes, analogous to that of translating signals in time.

Take for instance the case of images, or rather, signals that are represented by functions of two spatial coordinates. Their translations are composites of two 1-D translations, one along each of the two coordinates. With respect to these twin translations, one then has 2-D translation–invariant systems characterized by 2-D convolutions and corresponding 2-D transforms, both for the continuous and discrete cases.[34]

So, while signals belonging to the different classes differ in matters of detail—like whether they are continuous or discrete, 1-D or 2-D, they are identical in structure. To be more specific, they are functions defined on index sets that have identical structures, and their shifts or translates over their index sets are characterized by identical algebraic properties determined by the structure of the index sets. The system theoretic concepts of shift–invariance, convolutions and transforms rest in the main on these properties of shifts. Convolutional and transform domain characterizations of linear shift–invariant systems consequently admit of a unified treatment in which the versions specific to different classes of signals are particular interpretations.

Such a treatment highlights the important role that modern harmonic analysis and the representation theory of groups play in the study of symmetry, of which shift–invariance is an interesting instance.[35] In the rest of this section, we shall see its broad outlines, concentrating on the case of signals and systems defined on finite index sets.

[32] If you are not familiar with this, see Harmuth [14] to get an idea.

[33] For all these transforms, there are also multi–dimensional generalizations. See Dudgeon and Mersereau [9]. Yet another generalization of recent origin is the so-called slope transform for morphological signals and systems. See Maragos [22].

[34] Of course, in the chain of developments along these lines, one soon finds that there are problems peculiar to the 2-D theory, requiring fresh thinking to deal with them. A case in point is the problem arising from the fact that for the so-called fundamental theorem of algebra—an $n$-th degree polynomial of a single variable has precisely $n$ roots in the complex plane, there is no counterpart in the two variable case. See Antoniou [20, p. 3] for more elaborate comments on this issue. To get an idea of the nuances of the theory of functions of several complex variables, see Grauert and Fritzsche [12].

[35] This point will come up for closer scrutiny in later chapters, but if you already feel curious enough about modern harmonic analysis, see Gross [13] for a brief but elegant treatment, and Mackey [21] for a masterly account of the key ideas set in a historical perspective.

### 1.6.1 Shift–Invariance on Finite Index Sets

Recall first of all that in the case of continuous–time ($-\infty < t < \infty$) signals and systems, the operation of translating or shifting a signal $x$ in time by $\tau$ is that of subjecting it to a shift operator $D_\tau$ to produce a signal $D_\tau x$, with values $(D_\tau x)(t) = x(t-\tau)$. We are implicitly assuming here that time, which is the index set in this case, has the structure of the real line under addition. More formally, the index set in this case is the structure $(\mathbb{R}, +)$, for which "$-$t" denotes the additive inverse of "t".

Now, for a finite index set, $I$, to analogously admit of translations, it needs to be a structure $(I, \oplus)$, where $\oplus$ is a commutative associative binary operation on $I$, with respect to which (*a*) there is an identity element, $0$, in $I$ i.e., for $a \in I$, $a \oplus 0 = a$, and (*b*) the elements of $I$ have inverses i.e., for $a \in I$, there is $a^{-1} \in I$ such that $a \oplus a^{-1} = 0$.[36] To have a notation matching that of subtraction for numbers, let us write $b \ominus a$ to mean $b \oplus a^{-1}$. For a signal, which we consider in this case to be a function $x : I \to \mathbb{R}$, we can then think of translation or shift on $I$ by any element $a \in I$ as applying to it a shift operator $D_a$, defined by the condition $(D_a x)(b) = x(b \oplus a^{-1}) = x(b \ominus a)$ for all $b \in I$. $D_a$ is clearly a linear operator.

In this context, we call a linear system, $H$, shift–invariant if for any input signal $x$, it satisfies the condition $HD_a(x) = D_a H(x)$ for all $a \in I$.

If we apply two shifts in succession, then for all $a \in I$, the resulting signal has values

$$\begin{aligned}(D_b D_c x)(a) &= (D_b(D_c x))(a)\\ &= (D_c x)(a \ominus b)\\ &= x(a \ominus b \ominus c)\\ &= x(a \ominus (b \oplus c))\\ &= D_{b\oplus c} x(a)\end{aligned}$$

i.e.,

$$D_b D_c = D_{b\oplus c}$$

In particular, $D_0 D_a = D_a D_0 = D_a$, and for any $D_a$ there is $D_{\ominus a}$ such that $D_a D_{\ominus a} = D_0$.

It can thus be seen that, besides being inherently associative, the shift operators are also closed, commutative and invertible under composition, and have $D_0$ as their identity element. Consider then the structure $\mathbb{D}$ consisting of the set of these operators under composition, and let $\rho : I \to \mathbb{D}$ be a map with values $\rho(a) = D_a$. Then clearly, $\rho$ is a structure preserving one–to–one onto map: $\rho(a \oplus b) = \rho(a)\rho(b)$. In other words, $(I, \oplus)$ and $(\mathbb{D}, \cdot)$ are *isomorphic structures*.

Consider now the special functions $\delta_a, a \in I$ defined by the condition

$$\delta_a(b) = \begin{cases} 1 & : \quad a = b \\ 0 & : \quad a \neq b \end{cases}$$

[36] See **P**1–**P**6 on page 10. In the language of algebra, $I$ is a finite abelian group.

You may verify by comparing values that $\delta_a$ is a shifted version of $\delta_0$: $D_a\delta_0 = \delta_a$.

Clearly, any signal $x$ can be written as a unique linear combination of these functions.[37] To be more specific, for any $x$, there is a unique expansion,

$$\begin{aligned} x &= \sum_{a\in I} x(a)\delta_a \\ &= \sum_{a\in I} x(a)D_a\delta_0 \end{aligned} \tag{1.10}$$

A finite set of functions that provides a unique expansion of this kind for any $x$, we shall refer to as a set of *basis functions*.[38]

Referring to $\delta_0$ as the impulse, let $h$ denote the impulse response of a system $H$, i.e., $h = H\delta_0$. Expressed in the form (1.10), $h$ has the expansion

$$h = \sum_{a\in I} h(a)D_a\delta_0$$

Further, if $H$ is linear and shift–invariant, then its output for any input $x$ has the expansion

$$\begin{aligned} y &= Hx \\ &= \sum_{a\in I} x(a)HD_a\delta_0 \\ &= \sum_{a\in I} x(a)D_a(H\delta_0) \\ &= \sum_{a\in I} x(a)D_a \sum_{b\in I} h(b)D_b\delta_0 \\ &= \sum_{a\in I} \left(\sum_{b\in I} h(b)D_b\right)(x(a)D_a\delta_0) \\ &= \left(\sum_{b\in I} h(b)D_b\right)x \end{aligned}$$

[37] We are treading on linear algebra here. The set of all signals under addition and scaling by reals has the structure of a finite dimensional vector space in this case. Saying that there is a unique expansion in terms of the $\delta$'s for any $x$ is equivalent to saying more formally that the $\delta$'s are linearly independent, and that they constitute a basis for the space. Although not essential, it would be worthwhile brushing up your linear algebra at this stage. For a compact introduction, see Artin [1].

[38] In linear algebra a basis is commonly introduced as a linearly independent set of vectors that spans the space. We mean essentially the same thing, except that for a finite dimensional vector space, we have put it differently in terms of the notion of uniqueness. This appears to be a more parsimonious way of introducing the concept at this stage. Uniqueness of expansion is equivalent to linear independence.

So we conclude that *a linear shift–invariant system, $H$, is a linear combination of the shift operators, i.e.,*

$$H = \sum_{a \in I} h(a) D_a, \tag{1.11}$$

*where the coefficients $h(a)$ are the values of its impulse response $h = H\delta_0$. Moreover, since the the shift operators commute, any two linear shift–invariant systems $H$ and $G$ also commute, i.e., $HG = GH$.*

Finally, by further expanding $x$, we get

$$\begin{aligned} y &= Hx \\ &= \left(\sum_{b\in I} h(b)D_b\right)\left(\sum_{a\in I} x(a)D_a\right)\delta_0 \\ &= \left(\sum_{a\in I}\sum_{b\in I} x(a)h(b)D_{a\oplus b}\right)\delta_0 \\ &= \sum_{a\in I} x(a)\left(\sum_{b\in I} h(b)D_{a\oplus b}\right)\delta_0 \\ &= \sum_{a\in I} x(a)\left(\sum_{c\in I} h(c\ominus a)D_c\right)\delta_0 \\ &= \sum_{c\in I}\left(\sum_{a\in I} x(a)h(c\ominus a)\right)D_c\delta_0 \\ &= \sum_{c\in I}\left(\sum_{a\in I} x(a)h(c\ominus a)\right)\delta_c \end{aligned}$$

Comparing this form of $y$ with its unique expansion, $y = \sum_{c\in I} y(c)\delta_c$, of the form (1.10), we thus find that*for a linear shift–invariant system $H$ on the index set $I$, the output $y$ is the convolution of the input $x$ and its impulse response $h$, i.e., the values of $y$ are related to those of $x$ and $h$ by the convolutional formula*

$$y(c) = \sum_{a\in I} x(a)h(c\ominus a), \quad \text{for } c \in I. \tag{1.12}$$

It is a straightforward matter to check the converse, namely, that a system whose input–output realtionship is given by the convolutional formula (1.12) is a linear shift–invariant system $H$ whose impulse response is $h$.

Having thus generalized the notions of shifts, shift–invariant systems, and their convolutional characterization, we need in addition to be sure that these notions are not vacuous for finite index sets.[39]

[39] Suppose there were no finite index sets of the stipulated kind (finite abelian groups). In that case there would be nothing that our derivations, even though correct, would refer to. At least one

To see that they are not, let us now examine, for a concrete illustration of these ideas, the case of finite discrete signals, consisting of $n$–tuples of reals. It would suffice to concentrate on the case $n = 4$.

**Example 1.6.1** For the index set, let us choose $\mathbb{Z}_4$, i.e., the set of integers $0, 1, 2, 3$ under modulo $4$ addition; "$\oplus$" and "$\ominus$" then respectively mean addition modulo $4$ ($+_4$) and subtraction modulo $4$ ($-_4$). You may check that it meets all requirements that we have placed on index sets.

For a signal, we have a 4–tuple of the form $x = (x_0, x_1, \ldots, x_3)$, or a function $x : \mathbb{Z}_4 \to \mathbb{R}$, with $x_i$ denoting $x(i)$, and all such signals under pointwise addition and scalar multiplication by reals constitute the signal space. A system is correspondingly a transformation on this space, taking 4–tuples into 4–tuples. The four shift operators $D_0, D_1, D_2, D_4$ are linear systems that produce shifts as shown in Table 1.1 below.

| $x(\cdot)$ | $x_0$ | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|---|
| $(D_0x)(\cdot)$ | $x_0$ | $x_1$ | $x_2$ | $x_3$ |
| $(D_1x)(\cdot)$ | $x_3$ | $x_0$ | $x_1$ | $x_2$ |
| $(D_2x)(\cdot)$ | $x_2$ | $x_3$ | $x_0$ | $x_1$ |
| $(D_3x)(\cdot)$ | $x_1$ | $x_2$ | $x_3$ | $x_0$ |

Table 1.1 Four shifted versions of a signal on $\mathbb{Z}_4$

It will suffice for us here to take a simple matrix approach, treating signals as $4 \times 1$ column vectors, and linear systems as $4 \times 4$ matrices. Thus for a system $H$ producing output $y = Hx$ from an input $x$, the matrix relationship in detail is,

$$\begin{bmatrix} h_{00} & h_{01} & h_{02} & h_{03} \\ h_{10} & h_{11} & h_{12} & h_{13} \\ h_{20} & h_{21} & h_{22} & h_{23} \\ h_{30} & h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} \tag{1.13}$$

The four shift operators are, in particular, the permutation matrices, $D_i$ shown in Table 1.2 below.

The matrices $D_i$ form a closed commutative set under multiplication, with $D_0$ as the identity element.[40] To be specific, it may be verified that

$$D_iD_j = D_jD_i = D_{(i+j)\ mod\ 4} = D_{(j+i)\ mod\ 4}.$$

such set, and preferably many such different sets, must therefore exist for the abstractions to have any significance. We will have more on this when we discuss abstract structures in general.

[40] Observe that $D_2$ and $D_3$ equal $D_1^2$ and $D_1^3$ respectively. The matrices form a special kind of abelian group–the so called cyclic group– of which one single member, raised to different powers, gives all other members.

$$D_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad D_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$D_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \qquad D_3 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Table 1.2 The four shift operators on $\mathbb{Z}_4$

The structure $\mathbb{D}$, which consists in this case of the set of matrices $D_i$ under multiplication, is isomorphic to $\mathbb{Z}_4$. Indeed, the map $\rho : I \longrightarrow \mathbb{D}$, with values $\rho(i) = D_i$, is one-to-one and onto, and $\rho(i +_4 j) = \rho(i)\rho(j)$.

For $H$ to be shift–invariant, it is required that

$$HD_i = D_i H, \quad i = 0, 1, 2, 3. \tag{1.14}$$

What does this constraint mean for the matrix $H$? Well, we know that it is of the form (1.11). But just as a cross check, let us work it out through sheer brute force by examining the $64$ identities given by (1.14). Containing a large number of redundancies, they can be checked to show that, with $h_{00}, h_{10}, h_{20}, h_{30}$ put respectively as $h_0, h_1, h_2, h_3$, $H$ turns out to be a cyclic matrix, with its first column producing the others through downward cyclic rotations:

$$H = \begin{bmatrix} h_0 & h_3 & h_2 & h_1 \\ h_1 & h_0 & h_3 & h_2 \\ h_2 & h_1 & h_0 & h_3 \\ h_3 & h_2 & h_1 & h_0 \end{bmatrix} \tag{1.15}$$

The cylic form of $H$ implies that it has the expansion

$$H = h_0 D_0 + h_1 D_1 + h_2 D_2 + h_3 D_3, \tag{1.16}$$

which is indeed the form (1.11).

Now, $h$, the first column of $H$, is its response to the impulse $\delta_0 = [1\,0\,0\,0]'$, i.e., $H\delta_0 = h$.[41] Further, it can be verified that, in view of (1.15), Eq. (1.13) reduces to the following input–output characterization for shift–invariant systems on $\mathbb{Z}_4$ in terms of cyclic convolution modulo $4$.

$$y_i = \sum_{k=0}^{3} x_k h_{(i-k) \ mod \ 4}, \quad i = 0, 1, 2, 3. \tag{1.17}$$

This checks with our general result (1.12). □

[41] For a matrix $A$, $A'$, $\bar{A}$, $A^*(\equiv \bar{A}')$ will denote its transpose, (complex) conjugate and conjugate transpose respectively.

**Exercise 1.6.1** For the set of integers $\{0, 1, 2, 3\}$, let the binary operation "$\oplus$" denote dyadic addition, which is defined as follows. For any $a$ in this set, let $a_1a_0$ denote its two–place binary equivalent. Then for $a$ and $b$, $a \oplus b$ is the integer satisfying the condition $(a \oplus b)_i = a_i +_2 b_i$. That is, this addition, or sum, is binary addition without carries. Verify that the integers $0, 1, 2, 3$ under dyadic addition are as good for an index set as $\mathbb{Z}_4$. Work out the counterparts of the results of Example 1.6.1.

The next natural step for us now is to look at a transform domain characterization that converts convolution (1.12) into point–wise product.

## 1.6.2 Transforms and Spectra

In obtaining a convolutional characterization of shift–invariant systems on the finite index set $I$, we relied on the expansion of a function in the form (1.10), in terms of the fixed set of basis functions $\{\delta_a | a \in I\}$. This expansion is local in the extreme in the sense that its coefficients are simply the values of the function at various points of the index set individually.

In a very broad sense, transforms have to do with alternative expansions in which each of the coefficients provides some global information about the function as a whole, rather than locally at each point of its index set.[42]

There are, admittedly, many different choices of basis functions and corresponding expansions, much the same as we have innumerably many different choices of coordinate systems for representing points in a plane. Which set we choose depends on what features of signals we want it to highlight, and what special relationships we want it to have with systems to be used for processing the signals.

For our present purposes, we require of the basis functions that they be for shift–invariant systems on the index set $I$ what (complex) exponentials are for linear time–invariant systems.

To be more explicit, suppose that there is a set of real– or complex–valued functions on $I$, say $\{\phi_a | a \in I\}$, in terms of which there is a unique expansion for any signal $x$, i.e.,

$$x = \sum_{a \in I} \hat{x}_a \phi_a, \tag{1.18}$$

where $\hat{x}_a$ are the coefficients of this expansion.

Treating the coefficients $\hat{x}_a$ as the values of a function $\hat{x}$ on $I$, and allowing both $x$ and the corresponding map $\hat{x}$ to be complex–valued, we can say that the expansion (1.18) defines a one-to-one onto linear map on the signal space, mapping $x$

[42] Seeing transforms from this angle has given rise to noteworthy developments in two areas. One is the area of digital logic and fault diagnostics. For details, see Hurst [17]. The other is of multiresolution signal representation using wavelet transforms. See Meyer [24] and Strang [32] for a good introduction.

into $\hat{x}$. Considering that it is one-to-one onto (i.e., invertible), it is more convenient, as we will see, to christen its inverse, for which we write $\mathbf{W}$:

$$\hat{x} = \mathbf{W}x. \tag{1.19}$$

In addition, we call $\hat{x}$ the *transform* of $x$, or the *spectrum* of $x$.

The property that we would like the $\phi_a$'s to have is simply this: *every one of them should be a common eigenfunction (eigenvector) of all shift operators $D_a$ (and thereby of all linear shift–invariant systems) on $I$.*

Imposing such a condition on the $\phi_a$'s has deep implications for them. Let us see. Consider such an eigenfunction $f$. Let us treat the corresponding eigenvalues of the shift operators as values of a function $\chi$ on $I$, writing the condition as $D_a f = \chi(a^{-1})f$.[43]

Then for any $a, b \in I$, since $(D_a f)(b) = f(b \oplus a^{-1})$, we have

$$f(b \oplus a) = \chi(a)f(b)\,. \tag{1.20}$$

In particular, for $b = 0$ (the identity element of $I$),

$$f(a) = f(0)\chi(a) \quad \text{for all } a \in I. \tag{1.21}$$

Thus, in view of (1.20), *the eigenfunction $f$ is simply a scaled version of the function $\chi$*. Furthermore, since the shift–operators under composition form an abelian group isomorphic to $I$, it follows that

$$D_{(b\oplus a)^{-1}} f = D_{a^{-1}} D_{b^{-1}} f\,.$$

Equivalently, in terms of the eigenfunction $f$, we conclude that

$$\begin{aligned} \chi(b \oplus a)f &= \chi(a \oplus b)f \\ &= \chi(a)\chi(b)f\,. \end{aligned}$$

That is, if $f \neq 0$, then the function $\chi$ satisfies the equation,

$$\chi(a \oplus b) = \chi(a)\chi(b). \tag{1.22}$$

For the sake of normalization, we put $f(0)$, the value of $f$ for the identity element of $I$, equal to 1. *The desired basis functions are then functions that satisfy the equations*

$$\begin{aligned} D_a\phi_c &= \phi_c(a^{-1})\phi_c && (1.23)\\ \phi_c(a \oplus b) &= \phi_c(a)\phi_c(b) \quad \text{for all} \quad a, b, c \in I. && (1.24) \end{aligned}$$

[43] Yes. The notation is rather puzzling at first glance. In using $\chi(a^{-1})$, and not $\chi(a)$, to denote the eigenvalue of $D_a$, the motive is to keep in line with the action of $D_a$ on $f$.

Stated in words, condition (1.23) means that the values of a basis function are the same as the associated eigenvalues of the shift–operators, to within a reordering! One simple function that satisfies the two conditions is the one whose value at every point of $I$ is 1. We shall refer to it as the *constant function*, and label it as $\phi_0$, the basis function indexed by the identity element of $I$. No matter which $I$ we consider, this is one of the desired basis functions.

There is yet another surprising and interesting consequence of the condition we have imposed on the basis functions. This has to do with orthogonality, and is brought out through an averaging procedure. For any function $f$ on $I$, consider the mean $M(f)$ defined by

$$M(f) = \frac{1}{|I|}\sum_{a\in I} f(a), \tag{1.25}$$

where $|I|$ denotes the cardinality of the set $I$. Then, since the values of $D_af$ are simply those of $f$ reshuffled, $M(f) = M(D_af)$ for any $a \in I$. On account of this invariance, it is called the *invariant mean*.[44]

For two functions $f$ and $g$, the shift of their product is the product of their individual shifts: $D_a(fg) = D_a(f)D_a(g)$. Further, if the two are eigenfunctions of the shift–operators, satisfying (1.23), then[45]

$$\begin{aligned} M(f\bar{g}) &= M\{D_a(f\bar{g})\} \\ &= M\{(D_af)(D_a\bar{g})\} \\ &= f(a^{-1})\bar{g}(a^{-1})M(f\bar{g}) \quad \text{for every} \quad a \in I. \end{aligned}$$

Thus,

$$\left\{1 - f(a^{-1})\bar{g}(a^{-1})\right\} M(f\bar{g}) = 0. \tag{1.26}$$

In particular, if $f = g \neq 0$ then, since $M(f\bar{f}) > 0$,

$$\begin{aligned} f(a^{-1})\bar{f}(a^{-1}) &= 1 \quad \text{for every} \quad a \in I && (1.27)\\ \text{and,} \qquad M(f\bar{f}) &= 1. && (1.28) \end{aligned}$$

*We thus see that the magnitude of any of the basis functions is every where equal to 1, and that the invariant mean of the square of its magnitude is* 1.[46]

In case $f \neq g$, then $\{1 - f(a^{-1})\bar{g}(a^{-1})\}$ can not be zero for every $a \in I$, and from (1.26), $M(f\bar{g}) = 0$, i.e.,

$$\frac{1}{|I|}\sum_{a\in I} f(a)\bar{g}(a) = 0. \tag{1.29}$$

[44] Invariant means and invariant integrals provide an important vantage point in modern harmonic analysis. For more on this, see Edwards [10, vol. 1, pp. 21–26].

[45] For a function $f$, $\bar{f}$ denotes its complex conjugate.

[46] The second part can also be said to mean that the basis functions are normalized with respect to the invariant mean as inner product.

In other words, under the condition we have imposed on them, *the basis functions are mutually orthogonal.*[47] A startling revelation indeed, and even more so because it is a natural consequence primarily of the fact that the index set has the structure of an abelian group.

An immediate corollary of orthogonality is that we have a simple inversion formula for the expansion (1.18), giving the coefficients $\hat{x}_a$:[48]

$$\hat{x}_b = M(x\bar{\phi}_b) \quad \text{for every} \quad b \in I. \tag{1.30}$$

Now, since $\phi_b(0) = 1$ for every $b \in I$, and the impulse $\delta_0$ is zero every where except at 0, $M(\delta_0\bar{\phi}_b) = 1$ for every $b \in I$. Thus the expansion for $\delta_0$ is,

$$\delta_0 = \sum_{a\in I} \phi_a \tag{1.31}$$

That is, *the transform of the impulse $\delta_0$ is the constant function, which is also the basis function $\phi_0$.* The impulse response, $h$, of a linear shift–invariant system $H$, for which the $\phi_a$'s are eigenfunctions, can in turn be seen to have the expansion

$$\begin{aligned} h &= H\delta_0 \\ &= \sum_{a\in I} H\phi_a \\ \text{or,} \qquad h &= \sum_{a\in I} \lambda_a\phi_a, \end{aligned}$$

where $\lambda_a$ is the eigenvalue of $H$ for the eigenvector $\phi_a$. Then, since the expansion for $h$ is unique, $\lambda_a = \hat{h}_a$, the value of the transform $\hat{h}$ of $h$ at $a$.

More generally, expanding an input $x$ of $H$, and also its output $y$ $(= Hx)$, we have

$$y = \sum_{a\in I} \hat{y}_a\phi_a, \tag{1.32}$$

and also,

$$\begin{aligned} y &= Hx \\ &= H\left(\sum_{a\in I} \hat{x}_a\phi_a\right) \\ &= \sum_{a\in I} \hat{h}_a\hat{x}_a\phi_a. \end{aligned} \tag{1.33}$$

[47] Take note of the fact that $M(f\bar{g})$ qualifies as what is called an inner product in linear algebra, making the signal space an inner product space.

[48] Use the familiar trick of Fourier series expansion. Multiply both sides by $\bar{\phi}_b$, take the invariant mean and invoke orthogonality. We are indeed doing a generalized fourier analysis here.

Then, invoking uniqueness of the expansion for $y$,

$$\hat{y}_a = \hat{h}_a \hat{x}_a, \quad a \in I. \tag{1.34}$$

But we also know that $y$ is the convolution of $x$ and $h$, as in (1.12). We have thus established the convolution theorem in this case: *the transform of the convolution of two functions (Eq. (1.12)) on $I$ is the pointwise product of their individual transforms.*

In line with conventional usage, it is appropriate to call $\hat{h}$ the spectrum of the system $H$, this being the function by which it scales the spectrum of its input to give the spectrum of its output.[49]

We may now ask whether there do actually exist, besides $\phi_0$, other basis functions of the kind we have discussed, with all their nice properties? Not only they do, they can be constructed in a systematic manner, using representation theory of groups, a topic we shall examine in Chapter 5.

For the present, we close the discussion with a concrete example.

**Example 1.6.2** We continue with the index set $\mathbb{Z}_4$ in the same spirit in which we have looked at convolution in Example 1.6.1, treating signals as column vectors and systems as matrices. In order to identify the transform in this case, we have to first check whether the shift operators $D_i$ have the right kind of common eigenvectors.

Using standard results of matrix theory, and the fact that they commute, it can be shown that the operators $D_i$ do have a common set of four linearly independent eigenvectors. You may check that the four (complex–valued) vectors shown in Table 1.3 are indeed these eigenvectors. Further, in view of the form (1.16), they are eigenvectors of $H$ as well.[50]

$$\phi_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad \phi_1 = \begin{bmatrix} 1 \\ j \\ -1 \\ -j \end{bmatrix} \qquad \phi_2 = \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} \qquad \phi_3 = \begin{bmatrix} 1 \\ -j \\ -1 \\ j \end{bmatrix}$$

Table 1.3 Eigenvectors of $D_i$ and $H$

[49] Having first called the transform of a function as spectrum, it might smack of ambiguity to use the term for an operator $H$. It is not really so because we are in effect referring to its impulse response $h$ and its transform. Incidentally, there is an interesting historical coincidence about the origin of the term "spectrum" that is worth mentioning here. Independently of physics, it was introduced in mathematics for operators in general, as an extension of the notion of eigenvalues, to mean the set of values of a scalar $\lambda$ such that $(\lambda I - T)$ is not invertible, where $I$ is the identity operator. It was later recognized that its use in physics (and subsequently in engineering), when formally interpreted, coincided with that in mathematics. See Steen [31]. Calling $\hat{h}$ the spectrum of $H$ is justified from this angle too.

[50] This simplistic approach of identifying the eigenvectors would admittedly be of little use for bigger index sets. As already mentioned, group representation theory provides us a general systematic technique for doing this. We will discuss it in Chapter 5.

You may also verify that they have the orthogonality property,

$$\frac{1}{4}(\phi_i^*\phi_j) = \begin{cases} 1 & : \quad i = j \\ 0 & : \quad i \neq j \end{cases} . \tag{1.35}$$

For any signal $x$, we then have the expansion

$$x = \frac{1}{4}\sum_{i=0}^{3} \hat{x}_i\phi_i, \tag{1.36}$$

where $\hat{x}_i$ are the coefficients of expansion given by the equality

$$\hat{x}_i = \phi_i^* x. \tag{1.37}$$

Let $\hat{x}$ denote the column vector of these coefficients, and $\mathbf{W}$ the $4 \times 4$ invertible matrix with $\phi_i^*$ as its $i$–th row:[51]

$$\mathbf{W} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -j & -1 & j \\ 1 & -1 & 1 & -1 \\ 1 & j & -1 & -j \end{bmatrix} \tag{1.38}$$

Then the relationships (1.36) and (1.37) between $x_i$ and $\hat{x}_i$ take the matrix forms,

$$x = \frac{1}{4}\mathbf{W}^*\hat{x}, \tag{1.39}$$

$$\hat{x} = \mathbf{W}x, \tag{1.40}$$

and,

$$\mathbf{W}^{-1} = \frac{1}{4}\mathbf{W}^*. \tag{1.41}$$

The matrix $\mathbf{W}$ is indeed the map defined in (1.19) that takes a signal $x$ into its transform $\hat{x}$. Cyclic convolution of $x$ and $h$ becomes pointwise product of their transforms $\hat{x}$ and $\hat{h}$ given by (1.40). Equivalently, the spectrum of the output of a linear shift–invariant system on $\mathbb{Z}_4$ is the spectrum of its input multiplied by the spectrum of the system.

□

**Exercise 1.6.2** For the index set of Excercise 1.6.1, determine the transform matrix $\mathbf{W}$.

[51] Observe that it is the 4–point DFT matrix.

The discussions and examples of this section should give a reasonable idea of the way the notions of shifts, convolutions, transforms, and spectra are nowadays being interpreted for signals and systems defined on finite index sets. As a first step towards introducing these interpretations, we have focussed on index sets with the structure of an abelian group. As we will see later, the interpretations are valid for finite groups in general, not necessarily abelian. Techniques of modern harmonic analysis and group representation theory provide the framework for this general treatment.

On the whole, this chapter would have introduced you to the rich connections that signal processing concepts have with algebra. Structures such as groups, rings, integral domains, fields, vector spaces, and algebras have come for specific mention in the course of our discussions. Modern signal processing relies heavily on the theories of such structures. My idea of mentioning them at appropriate places in this chapter has been essentially to indicate their relevance, and to motivate their study. In case you are already familiar with them, you may like to dig deeper into the history of modern algebraic concepts. A good source for that is Corry [8].

# References

1. Michael Artin. *Algebra*. Prentice–Hall, New York, 1991.

2. Nicholas Bourbaki. The architecture of mathematics. *Amer. Math. Month.*, pages 221–232, April 1950.

3. Carl B. Boyer. The invention of analytic geometry. *Scientific American*, pages 40–45, January 1949.

4. Carl B. Boyer. *A History of Mathematics*. Wiley, New York, 1968.

5. D. Bushaw. *Elements of General Topology*. Wiley, New York, 1963.

6. Henri Cartan. Nicolas Bourbaki and contemporary mathematics. *Mathematical Intelligencer*, 2(4):175–180, 1980.

7. E.W. Cheney. *Introduction to Approximation Theory*. McGraw-Hill, New York, 1966.

8. Leo Corry. *Modern Algebra and the Rise of Mathematical Structures*. Birkhäuser, Basel, 2004.

9. Dan E. Dudgeon and Russell M. Mersereau. *Multidimensional Digital Signal Processing*. Prentice–Hall, New York, 1984.

10. R.E. Edwards. *Fourier Series: A Modern Introduction*, volume 1. Springer-Verlag, New York, 1979.

11. Albert Einstein. The problem of space, ether and the world of physics. In Sonja Bargmann, editor, *Ideas and Opinions*, pages 276–285. Souvenir Press Ltd., London, 1973.

12. H. Grauert and K. Fritzsche. *Several Comples Variables*. Springer-Verlag, New York, 1976.

13. Kenneth I. Gross. On the evolution of noncommutative harmonic analysis. *Amer. Math. Month.*, 85:525–548, 1978.

14. Henning F. Harmuth. *Transmission of Information by Orthogonal Functions*. Springer-Verlag, New York, 1969.

15. Carl G. Hempel. Geometry and the empirical science. In James R. Newman, editor, *The World of Mathematics, vol 3*, pages 1609–1620. Tempus Books of Microsoft Press, Washington, 1988.

16. Karel Hrbacek and Thomas Jech. *Introduction to Set Theory*. Marcel Dekker, New York, 1984.

17. S.L. Hurst, D.M. Miller, and J.C. Muzio. *Spectral Techniques in Digital Logic*. Academic, London, 1985.

18. H.C. Kennedy. The origins of modern mathematics: Pasch to Peano. *Amer. Math. Month.*, 79:133–136, 1972.

19. Morris Kline. *Mathematical Thought from Ancient to Modern Times*. Oxford University Press, New York, 1972.

20. Wu-Sheng Lu and Andreas Antoniou. *Two–Dimensional Digital Filters*. Marcel Dekker, New York, 1992.

21. George W. Mackey. *The Scope and History of Commutative and Noncommutative Harmonic Analysis*, volume 5 of *History of Mathematics*. American Mathematical Society, 1992.

22. Petros Maragos. Morphological systems: slope transforms and max–min diference and differential equations. *Signal Processing*, 38:57–77, 1994.

23. H. Meschkowski. *Evolution of Mathematical Thought*. Holden-Day, San Fransisco, 1965.

24. Yves Meyer. *Wavelets: Algorithms and Applications*. Society for Industrial and Applied Mathematics, Philadelphia, 1993.

25. Jan Mikusiński. *Operational Calculus*. Pergamon, London, 1959.

26. Jean Piaget. *Structuralism*. Routledge/Kegan Paul, London, 1968.

27. Lawrence R. Rabiner and Bernard Gold. *Theory and Application of Digital Signal Processing*. Prentice–Hall (India), New Delhi, 1988.

28. John R. Rice. *The Approximation of Functions*, volume 1. Addison Wesley, Reading, MA, 1964.

29. Robert Rosen. *Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations*. Pergamon, London, 1985.

30. Transl. David Eugene Smith and Marcia L. Latham. *The Geometry of René Descartes*. Dover, New York, 1954.

31. L.A. Steen. Highlights in the history of spectral theory. *Amer. Math. Month.*, 80:359–381, 1973.

32. Gilbert Strang and Truong Nguyen. *Wavelets and Filter Banks*. Wellesley–Cambridge Press, Wellesley, 1996.

33. J.F. Traub. Generalized sequences with applications to the discrete calculus. *Mathematics of Computation*, 19:177–200, 1965.

34. Richard von Mises. Mathematical postulates and human understanding. In James R. Newman, editor, *The World of Mathematics, vol 3*, pages 1695–1724. Tempus Books of Microsoft Press, Washington, 1988.

35. R.L. Wilder. The axiomatic method. In James R. Newman, editor, *The World of Mathematics, vol 3*, pages 1621–1640. Tempus Books of Microsoft Press, Washington, 1988.

36. I.M. Yaglom. *Mathematical Structures and Mathematical Modelling*. Gordon and Breach, New York, 1986.

# Chapter 2

# Algebraic Preliminaries

*Ah! Abstractions!!*
*Those invisible shadows of reality!*

In Chapter 1 we examined some of the very basic signal processing concepts from a structural viewpoint, bringing out in the process their connections with the language of algebra in general. This chapter is devoted to those algebraic concepts that have a direct bearing on the study of symmetry. We use the standard set theoretic notation and conventions in discussing these concepts.

We begin by deliberating on the idea of a definition.

## 2.1 What's in a Definition?

Just as there are oranges that are sweet and memories that are haunting, there are circuits, resistors, capacitors, inductors, systems, functions, and operators, that are linear, or have symmetries, and there are structures that are homomorphic.

Whatever it be to be linear or symmetric or homomorphic, one thing is clear: it is a quality or property, the same way as being sweet is, or being haunting is. It is a property that is shared by several different objects but not necessarily by all possible objects. Indeed, if all objects had a particular property, there was no need to make an issue of that property.[1]

It follows that we are interested in situations in which the property in question is found in some objects but not in others. Given a set of objects, through this property we then have a partitioning of the set into two disjoint subsets, one of those objects that do have this property and the other of those that do not have it. Such a partitioning is the central task of definitions in a scientific discourse.

The partitions resulting from a definition must of course be suitably named. Usually, a name is first chosen for the objects with the stipulated property, and,

[1] Admittedly, there are subtle issues about what can properly qualify for being considered as a property, and about paradoxes. But we shall leave out all that from the present discussions.

V.P. Sinha, *Symmetries and Groups in Signal Processing*, Signals and Communication Technology,
DOI 10.1007/978-90-481-9434-6_2, © Springer Science+Business Media B.V. 2010

with a prefix such as 'non' or 'un' added to it, it becomes the name of the rest. Thus for systems, we have a partitioning into linear and nonlinear systems, stable and unstable systems, causal and noncausal systems.

This is, however, not always the case. A typical example is that of passive and active circuits, as in circuit theory. A circuit that fulfils a certain energy constraint at its ports is said to be passive; one that does not satisfy this constraint is commonly called active, rather than nonpassive. This is partly for historical or aesthetic reasons, and partly for being more suggestive of the motivation for introducing the definition.[2]

There is yet another point about definitions that is worth bringing up here. In any language, we work with a finite vocabulary; it may be very large, but it is still finite. So if we form a chain of words to explain the meaning of a word (by consulting a dictionary), it invariably loops back on a word already in the chain, thereby making the explanation circular. In ordinary language, we break this circularity by assuming at some point that the meaning beyond it is understood on the basis of intuition and shared experience. In the domain of mathematics and its applications, where the language used is that of set theory, such circularity is broken, or perhaps dodged, by proceeding axiomatically.

Hopefully, these comments will whet your appetite for knowing more about the nature of definitions and the inherent subtleties. You may, in that case, like to look up Copi [6, Chapter 5, pp. 120–159].[3] In closing, I might just add this quote from Halmos [12, p. 19]: "Nothing means something without a definition."

## 2.2 Set Theoretic Notation

I assume that the reader is familiar with basic set theoretic concepts and the related notation. There are, however, notational variations that are likely to cause confusion. To avoid that, I lay down below the conventions that I will follow in the rest of the book.

For sets in general, $\cup$ and $\cap$ denote union and intersection respectively, as is universally done. We write $X \subseteq Y$ to say that $X$ is a *subset* of $Y$, or that $Y$ contains $X$; in detail, this means that if $x$ is an element (or a member) of a set $X$ (in symbols,

[2] One needs to be careful in such cases; even though differently named, the terms passive and active are, as it were, complementary, and not independently defined. On the other hand, the mathematical notions of closed and open sets, as understood in topology, do not have such complementarity, even though we are apt to think of them to be so if we were to go by the day-to-day usage of the adjectives; as Bushaw [5, p. 23] very crisply points out, "Unlike doors, sets in a topological space may be both open and closed, or neither open nor closed."

[3] Also see Greenspan [11, Appendix A, pp. 276–285] for a very short but lucid discussion. Another book by the same author [10] is worth mentioning here; it contains radically new ideas about discrete modeling of physical phenomena, which do not seem to have received so far the kind of attention that they should in the area of signal processing. Written in the seventies, it may be hard to lay hands on a copy, but it is worth a try.

$x \in X$) then $x$ is also an element (member) of $Y$. We write $X = Y$ to mean that $X \subseteq Y$ and $Y \subseteq X$. To say that a set $X$ is a *proper subset* of $Y$, i.e., $X \subseteq Y$ but $X \neq Y$, we write $X \subset Y$. For a finite set $X$, $|X|$ denotes the number of its elements. The set with no elements, the *empty set*, is denoted by $\emptyset$. A set is called *nonempty* if it has at least one element.[4] For a set $X$, its *power set*, the set of all subsets of $X$, is denoted by $\mathcal{P}(X)$.

For a given set $X$, we frequently need to define another one, $Y$, by stipulating an additional condition on the elements of $X$. As is the common practice, we do this by writing

$$Y = \{x \in X \,|\, \text{a given condition holds for}\, x\}.$$

When a set is to be identified by listing its elements, we use curly brackets to enclose the list. Thus, for a set with three elements, $x$, $y$, and $z$, we say that the set is $\{x, y, z\}$; the order in which the elements are listed is irrelevant. When a set is large (has more than three elements), instead of using different letters for the names of its elements, we use the same letter appropriately subscripted, e.g., $x_1, x_2, x_3$ instead of $x, y, z$. The elements may still be too numerous to list individually, and in that case we make use of an ellipsis. Thus, for a set $X$ with $n$ elements, $x_1$ to $x_n$, we write $X = \{x_1, x_2, \ldots, x_n\}$. Furthermore, if the set is infinite (countably so), we stop at the ellipsis, as in $X = \{x_1, x_2, x_3, \ldots\}$.

Using numbers as subscripts in naming members of a set is an instance of what is called *indexing*. In general, we say that a set $X$ is *indexed by* a set $I$ if the members of $X$ are named by using members of $I$ as subscripts; the set $I$ is called the *index set* for $X$. All this is indicated by writing $X = \{x_i\}_{i \in I}$.

We reserve special symbols for sets of numbers that come up very frequently. Deviating a little from the standard practice, we write $\mathbb{N}^+$ for the set of natural numbers, and $\mathbb{N}$ for the set of nonnegative integers: $\mathbb{N}^+ = \{1, 2, 3 \ldots\}$, and $\mathbb{N} = \{0, 1, 2, \ldots\}$. The sets of integers is denoted by $\mathbb{Z}$, the set of rational numbers by $\mathbb{Q}$, the set of reals by $\mathbb{R}$, and the set of complex numbers by $\mathbb{C}$. Moreover, $\mathbb{R}^+$ stands for the set of nonnegative real numbers.

For an ordered pair of elements $x$ and $y$ of two (not necessarily different) sets, we write $(x, y)$. Likewise, for ordered $n$–tuples of elements $x_1, x_2, \ldots, x_n$, we write $(x_1, x_2, \ldots, x_n)$. For an infinite sequence, we write $(x_0, x_1, x_2, \ldots)$, or in short, $(x_i)_{i \in \mathbb{N}}$.[5]

[4] The empty set notion is highly counter–intuitive, and like that of zero, takes getting used to. As it happens, again like the role zero has in arithmetic, it has a crucial technical role in set theory. For now, we brush the related issues under the carpet. See Halmos [14, pp. 8–9] for a brief but engaging discussion. I might just add here that for an empty set, virtually any thing is true. Indeed, as Halmos argues, an assertion could be false for an element if there *is* an element to test for. But for an empty set? Well! That is why it is generally stipulated in definitions and results about sets that they are nonempty. For our purposes, this will be understood from the context, even if not explicitly stated.

[5] This overrides the notation of Chapter 1, where I have used angle brackets to indicate a sequence; parentheses were used there for several different purposes, and using them for sequences as well would have led to confusion.

Given (nonempty) sets $X$ and $Y$, $X \times Y$, the *cartesian product* of the two sets, is the set:

$$X \times Y = \{(x, y) | x \in X \,\text{and}\, y \in Y\}.$$

We talk of cartesian products of more than two sets in a similar fashion. For a cartesian product of the same set, say $X$, taken $n$ times, we use the abbreviation $X^n$:

$$X^n = \underbrace{X \times \cdots \times X}_{n\,times}.$$

Thus, we write $\mathbb{R}^2$ for $\mathbb{R} \times \mathbb{R}$, and $\mathbb{R}^3$ for $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$.

By a *relation*, unless otherwise mentioned, we mean a *binary relation*. Formally, a relation $R$ from a set $X$ to a set $Y$ is a subset of $X \times Y$: $R \subseteq X \times Y$. As is customary, most of the time we take the shortcut of writing this as $xRy$ to mean that $(x, y) \in R$. A relation *on* $X$ is a relation from $X$ to $X$ itself. Occasionally, we also need to bring in ternary and $n$–ary relations, the notation for which is on the same lines. In particular, an $n$–ary relation $S$ on a set $X$ is a subset of $X^n$: $S \subseteq X^n$.

In the engineering literature, the notation adopted for functions is so frequently ambiguous, not making a clear distinction between a function and its values. (It is a well established practice, for instance, to write $f(t)$ to denote a signal that is a function of continuous–time, or write $f(n)$ if it is a function of discrete–time.) This causes difficulties in talking about sets of signals and systems in an algebraic setting. We therefore do not fall in line, and follow, instead, the regular mathematical convention. Recall that a function $f$ from a set $X$ to a set $Y$ is a relation, $f \subseteq X \times Y$, such that for each $x \in X$ there is precisely one $y \in Y$ for which $(x, y) \in f$; we say that the *value* of $f$ at $x$ is $y$, or that $y$ is the *image* of $x$ under $f$. This is what is commonly expressed by saying that $y = f(x)$. As is the common practice, we too will frequently refer to a function as a *map* or a *mapping*.

We use the standard shorthand $f : X \rightarrow Y$ to mean that $f$ is a function from $X$ to $Y$; $X$ is called the *domain* of $f$, and $Y$ its *codomain*. It is often convenient to visualize a function as a processor, receiving a member of its domain as input, and producing a member of its codomain as output. The set $\{f(x) \mid x \in X\}$, which is a subset of the codomain of $f$, is referred to as the *range* of $f$. If the function $f$ is such that, for all $a, b \in X$, $f(a) = f(b)$ implies that $a = b$, we say that $f$ is *one–to–one*. If for every $y \in Y$, there is an $x \in X$ such that $f(x) = y$, we say that $f$ is an *onto* function.[6] Very frequently, it is convenient to treat sequences as functions. Thus, the sequence $(x_0, x_1, x_2, \ldots)$ of real numbers is a function $x : \mathbb{N} \rightarrow \mathbb{R}$, with $x(i) = x_i$ for $i \in \mathbb{N}$.

By an *operation* we usually mean a *binary operation*, and we adopt the standard infix notation for it. Thus, an operation $\circ$ on a set $X$ is a function from $X \times X$ to

[6] The terms 'one-to-one' and 'onto' are often frowned upon on grammatical grounds, and justly so, as pointed out for instance in Bloch [4, pp. 162–163]. But the alternative terms 'injective' and 'surjective', even though perhaps more fitting for the purpose, do not convey, at least to me, the ideas as vividly. Hence, with due deference to grammar, I avoid the hybrids, apparently introduced by Bourbaki.

$X$, and its value at $(x, y) \in X \times X$ is written $x \circ y$; we commonly visualize the operation as acting on an ordered pair of elements of $X$ to produce a unique third element of $X$ itself. An $n$–ary operation on a set $X$ is to be understood to be a function from $X^n$ to $X$; the number $n$ is referred to as the 'arity' of the operation.

An operation on a finite set is often specified by giving what is called its *Cayley table* (or *multiplication table*), which is analogous to the addition or multiplication tables of arithmetic. For an operation $\circ$ on a finite set $X$, $|X| = n$, the table consists of $n$ rows and $n$ columns, both indexed by the elements of $X$ in the same order, with the element $x \circ y$ as its entry in the row corresponding to the element $x$ and the column corresponding to the element $y$. For an operation on a set $\{a, b, c\}$ of three elements, the table given below is a sample. The table says for instance that $b \circ c = a$. Note that any arbitrary set of table entries from amongst the set elements will serve to define an operation on the set.

| $\circ$ | $a$ | $b$ | $c$ |
|---|---|---|---|
| $a$ | $a$ | $b$ | $b$ |
| $b$ | $b$ | $a$ | $a$ |
| $c$ | $c$ | $a$ | $b$ |

With this much on notation, we now turn to a selection of very basic structural concepts that we will need to rely on later.

## 2.3 Relations and Operations

We all know from our school days what relations and operations are in a concrete sort of way. In arithmetic we freely make use of the relations of equality ($=$) and order ($\leq$), and likewise, the operations of addition and multiplication. As we progress in our studies, we learn to use them in more general settings such as those of sets, matrices, and functions. The final stage in this progression is reached when we conceptualize them abstractly in terms of set theoretic properties. It is appropriate at this point to review some of these properties, and the relations and operations they characterize.

### 2.3.1 Equivalence Relations and Partitions

Consider, to begin with, those properties that characterize the notion of equivalence. Intuitively speaking, two objects are considered to be equivalent if they can serve for each other equally well in some sense. There are two alternative ways to formally capture the basic idea, one through the concept of equivalence relations, and another through the notion of set partitions.

A relation $R$ on a set $X$ is said to be an *equivalence relation* if it simultaneously fulfills the following three conditions for all $x, y, z \in X$.

1. *Reflexivity*: For any $x \in X$, $xRx$
2. *Symmetry*: If $xRy$ then $yRx$
3. *Transitivity*: If $xRy$ and $yRz$, then $xRz$

One may get the impression that there is redundancy here. Surely, if $xRy$ then $yRx$ by symmetry, and then, by transitivity (with $z = x$), $xRx$. Thus symmetry and transitivity imply reflexivity. Not quite. Observe that, while the reflexivity condition is for all elements, the symmetry and transitivity conditions start with an 'if'. For a particular $x$ there may not be any $y$ such that $xRy$, and in that case the argument falls through. Perhaps, an example will help visualize this situation. Consider the relation $R = \{(a, b), (b, a), (a, a), (b, b)\}$ on the set $X = \{a, b, c\}$. You may check that $R$ is symmetric and transitive. But it is not reflexive, since $(c, c)$ is not in it. It is worth noting, on the whole, that an arbitrary relation may satisfy one or two of the three conditions but not the rest.

For an equivalence relation $R$ on $X$, and for any $x \in X$, the set $\{y \in X | xRy\}$ is called the *equivalence class* of $x$ with respect to $R$, denoted commonly by $[x]$; if $xRy$, we say that $x$ and $y$ are *equivalent*.

A very important aspect of an equivalence relation is that it is essentially synonymous with the idea of partitioning a set into a union of disjoint subsets.

A *partition* of a (nonempty) set $X$ consists of a collection of its nonempty subsets $\{X_1, X_2, \ldots, X_n\}$, called *cells* or *blocks* of the partition, such that

$$X_i \cap X_j = \emptyset \quad \text{for} \quad i \neq j\,, \text{ and } \quad X_1 \cup X_2 \cup \cdots \cup X_n = X\,.$$

Let $P$ be such a partition of a set $X$, and let $\sim$ be the relation on $X$ defined by the condition that, for $x, y \in X$, $x \sim y$ if and only if $x$ and $y$ are in the same cell of $P$. You can check that the relation $\sim$ is an equivalence relation, i.e., it is reflexive, symmetric, and transitive. We say that the partition $P$ of $X$ *induces* the equivalence relation $\sim$ on $X$.

Conversely, an equivalence relation $R$ on a set $X$ induces a partition of $X$, and the equivalence classes with respect to $R$ are the cells of this partition.[7] This correspondence between equivalence relations and partitions is put to frequent use in algebra and its applications.

### 2.3.2 Operations

Turning now to operations, notice first of all that the output of an operation belongs to the same set to which the input elements belong. As is commonly done, we refer to this property of an operation as *closure*, and we say that the set is *closed* under the operation.

[7] This calls for a proof, which I leave the reader to figure out.

The closure property is part of the definition of an operation; it is common to all operations. There is no need therefore to mention it separately. Yet, in many circumstances, we can not help bringing it up explicitly. Consider for instance an operation $\circ$ on a set $X$, and let $S$ be a subset of $X$. When can we say that it is also an operation on $S$? Well, if we check and find that $S$ is closed under the mapping $\circ$, then this mapping defines an operation on $S$ as well, called the *induced operation* on $S$. Strictly speaking, we should use a different symbol for the induced operation. But, to keep the notation simple, we commonly use the same symbol for it.

We are concerned most of the time with operations that have some special properties. To be more specific, for an operation $\circ$ on a set $X$, these properties are the following.

1. *Associativity*: For all $x, y, z \in X$, $(x \circ y) \circ z = x \circ (y \circ z)$; if this condition holds then we say that the operation $\circ$ is *associative*, or that it satisfies the *associative law*.

2. *Existence of an Identity*: The operation $\circ$ satisfies the *identity law*, i.e., there exists and element $e \in X$ such that $e \circ x = x = x \circ e$ for all $x \in X$. The element $e$ is called an *identity element* (or simply, an *identity*) for the operation $\circ$.

3. *Existence of Inverses*: Let $e$ be an identity element for $\circ$. For any $x \in X$, if there exists an element $y \in X$, called an *inverse* of $x$, such that $x \circ y = e = y \circ x$, then we say that the operation $\circ$ satisfies the *law of inverses*. If an element $x$ has a unique inverse, it is commonly denoted by $x^{-1}$.

4. *Commutativity*: For any $x, y \in X$, if $x \circ y = y \circ x$ then the operation $\circ$ is called *commutative*. Alternatively, we say that the operation satisfies the *commutative law*.

Two points related to the associative law are worth mentioning here. The first is about the use of parentheses in general, and the second is about the use of the law for more than three elements.

All of us are so familiar with the use of parentheses that nothing about the way we use them seems to require a proof for its correctness. Yet, there are matters that do. Can we really be sure that for a given string of properly paired parentheses in an expression, there can not be an alternative proper pairing? Next, are we entitled, on the strength of the associative law, to drop the parentheses altogether for any finite number of elements? The answer to both questions is: Yes. For more on these questions, in case you have not thought about them so far, see Appendix A.

From operations and relations, we now move on to structures centred around them. An algebraic structure, you will recall, consists, in general, of a set together with a finite set of operations (not necessarily binary, but of different arities). In the study of symmetry and its implications, there are two structures that play a central role. One is that of a group, and the other is that of a vector space (linear algebra). It is to these two that the next two sections are devoted. We present only the bare essentials, mainly to indicate what kinds of notions and results we shall later draw upon. For a more detailed treatment, the reader will need to refer to standard introductory literature on abstract and linear algebra.

## 2.4 Groups

In abstract axiomatic terms, a *group* is an algebraic structure $(G, \circ)$, where $G$ is a set, and $\circ$ is an operation on $G$, which satisfies the associative law, the identity law, and the inverse law (the group axioms). We shall assume throughout that the set $G$ is nonempty, and we shall use the same symbol to denote the set as well as the group. The number of elements in $G$ is the *order* of the group $G$, denoted by $|G|$. A group is *finite* if its order is finite; otherwise it is an *infinite* group. A group that also satisfies the commutative law is an *abelian* (or *commutative*) group.

Let us look at some of the elementary but important consequences of the group axioms. To start with, note that *for a group* $(G, \circ)$*, its identity element is unique.* For, suppose that there are two elements $e, f \in G$ such that $e \circ x = x = x \circ e$, and also $f \circ x = x = x \circ f$, for every $x \in G$. Then, $e$ being an identity, we have $e \circ f = f$. Similarly, since $f$ is an identity, $e \circ f = e$. Combining the last two equalities, we see that $e = f$.

A similar result holds for inverses: *Every element of a group $G$ has a unique inverse.* Let us say that an element $a$ of a group $G$ has two inverses, $x$ and $y$, i.e., $a \circ x = x \circ a = e$ and $a \circ y = y \circ a = e$, where $e$ is the unique identity of $G$. Then, combining the two sets of equalities and using associativity, we see that $x = y$:

$$\begin{aligned} x &= x \circ e \\ &= x \circ (a \circ y) \\ &= (x \circ a) \circ y \\ &= e \circ y \\ &= y \end{aligned}$$

Next, there are the *cancellation laws* for a group: *Let* $(G, \circ)$ *be a group. Then for any* $x, y, z \in G$*, if* $x \circ y = x \circ z$ *then* $y = z$*. Likewise, if* $x \circ y = z \circ y$ *then* $x = z$*.* It follows that each element of a (finite) group appears precisely once in each row and each column of the Cayley table of the group.

In discussing groups, it is frequently convenient to omit the operation symbol, writing $x \circ y$ simply as $xy$, and calling $xy$ the "product" of $x$ and $y$. As an additional shorthand, for any group element $x$, one makes use of the exponential notation as follows: with $x^0 = e$ (the identity element), $x^n$ is recursively defined as $x^n = x^{n-1}x$, for any positive integer $n \geq 1$; furthermore, $x^{-n}$ stands for $(x^{-1})^n$, where $x^{-1}$ is the inverse of $x$.

## 2.4.1 Groups Within Groups

For a set with no additional structure specified, all its subsets have the same standing, so to say. But for a set with additional structure, there are subsets that are special, the ones that inherit the additional structure. In the case of a group, these special subsets are its subgroups.

More precisely, let $(G, \circ)$ be a group, and let $S$ be a subset of $G$ such that $S$ is closed under $\circ$. If $(S, \circ)$ is a group then we say that $S$ is a *subgroup* of $G$. Checking for $S$ to be a group is, of course, a shade easier than having to check for all the three group axioms separately. Assuming that $S$ is closed under the operation, associativity for $G$ implies associativity for $S$ as well. Furthermore, if $x \in S$ implies $x^{-1} \in S$, then, by closure, $x \circ x^{-1} = e$ is also in $S$, i.e., the inverse law implies the identity law in this case.

Going strictly by the definition, every group is a subgroup of itself. Moreover, the set consisting of the identity element of a group is also a subgroup. A group has, in general, many other more interesting subgroups, and there is a wide variety of interesting and powerful results that relate groups to their subgroups. To get the flavour, let us look at one particular result, which shows that the orders of finite groups and their subgroups are very tightly related.

Let $H$ be a subgroup of a finite group $G$, with $e$ denoting its identity element. We want to see if the order of $H$, is related in some way to the order of $G$. For this purpose, for every $a \in G$, introduce the function $f_a : H \rightarrow G$, $f_a(x) = ax$, with range $T_a = \{ax | x \in H\}$.[8]

Focussing on $T_a$ as a subset of $G$, we shall presently see that all such subsets together form a partition of $G$. Note that $T_e$ is one of these subsets, and it consists precisely of the elements of $H$. Moreover, for any $a \in G$, $f_a(e) = a$, i.e., $a \in T_a$. In other words, *every element of G is contained in at least one of these subsets.*

Next, for $a \in G$ , let us check on the number of elements in $T_a$, the range of $f_a$. Suppose that $f_a(x) = f_a(y)$ for $x, y \in H$, i.e., $ax = ay$. Since $a, x, y$ are all in $G$, we have $x = a^{-1}ay = y$. That is, *the function $f_a$ is one–to–one, i.e., its domain and range have the same number of elements: $|T_a| = |H|$.*

Now, what about common elements? Well, suppose that, for $a, b \in G$ and $a \neq b$, $T_a$ and $T_b$ have a common element. That is, there are $x, y \in H$ such that $f_a(x) = f_b(y)$, or $ax = by$. In that case, $a = b(yx^{-1})$. Since both $x, y \in H$, $w = yx^{-1} \in H$. So $a = bw$ for some $w \in H$, i.e., $f_b(w) = a$ and $a \in T_b$. Next, since $a \in T_b$, $f_a(v) = av = bwv = b(wv)$, for $v \in H$. That is, for any $v \in H$, $f_a(v) \in T_b$. Thus, $T_a \subseteq T_b$. Likewise, by symmetry, $T_b \subseteq T_a$. Thus $T_a = T_b$. In other words, *if any two subsets $T_a$ and $T_b$ have a common element, then they are identical.*

In summary then, *the sets $T_a$, for $a \in G$, form the cells of a partitioning of G; the number of elements in every cell is the same, and is equal to the order of $H$.* This

[8] We are here using the product notation for groups. The subsets $T_a$ are known as *right cosets* of $H$. The notion of a coset is an important one in many ways, but we have avoided using the term here, because we do not put it to use in later discussions.

means that $|G| = m|H|$, where $m$ is the number of cells in the partition. As all this is true for any subgroup of a finite group, it follows that*for a finite group, the order of a subgroup divides the order of the group.* This is the celebrated **Langrange's Theorem.**

There is yet another way of partitioning a finite group that is going to be important for us later. This rests on the idea of conjugate elements. An element $y$ of a group $G$ is said to be a *conjugate* of another element $x$ of $G$ if there is an element $h$ in $G$ such that $y = h^{-1}xh$. It is easily verified that "conjugate of" is an equivalence relation on $G$. It follows that this equivalence relation induces a partitioning of $G$, in which the cells are the equivalence classes of conjugate elements. In the context of groups, these equivalence classes are commonly called *conjugacy classes*, or simply *classes* of a group.

**Remark 1** Recall at this point the idea of similarity transformation of matrices. Given two (square) matrices, $A$ and $B$ of the same size, $B$ is said to be similar to $A$ if there is an invertible matrix $P$ such that $B = P^{-1}AP$. Clearly, the relation 'similar to' is an equivalence relation, its equivalence classes consisting of similar matrices. Matrices that are similar in this sense have the same eigenvalues, and have the same traces. Thus, if our interest is in the eigenvalues, or traces of matrices, any one member of an equivalences class will serve the purpose. Classes of groups have an analogous role.

♠

### 2.4.2 Group Morphisms

Equivalence relations have to do with the 'sameness' of members of a set in general. There is a corresponding idea of sameness for structures, such as groups, which is formally captured in terms of homomorphisms and isomorphisms, together referred to as morphisms.

For two groups, $(G, \circ)$ and $(H, *)$, let $f : G \rightarrow H$ be a map such that, for any $x, y \in G$, $f(x \circ y) = f(x) * f(y)$. Then the map $f$ is called a *homomorphism.* If $f$ is, in addition, a one–to–one and onto map, then it is called an *isomorphism*, and the groups $G$ and $H$ are called *isomorphic*. Finally, if $f$ is an isomorphism from $G$ to $G$ itself, it is called an *automorphism* of $G$.

The following simple but significant result about morphisms is an immediate consequence of the definition: *For groups G and H, let $f : G \rightarrow H$ be a homomorphism, and let $e_G$ and $e_H$ be the identity elements of the groups respectively. Then $f(e_G) = e_H$, and $f(x^{-1}) = (f(x))^{-1}$for all $x \in G$.*

It is worth mentioning here that the concept of a morphism extends in a natural way to algebraic and relational structures in general. As already pointed out in Chapter 1, an intuitive way of perceiving a morphism is to think of it as a map that preserves structure. At a later point, while discussing symmetry, we shall have an occasion to put it to use for relational structures.

### 2.4.3 Groups and Geometry

We have so far discussed the notion of a group in the abstract. It is appropriate at this point to reflect on how it connects, in a rather concrete manner, with geometry.

Recall that in Euclid's geometry, there are no coordinates. There are geometrical objects, such as points, lines, circles, triangles and so on, and there are rigid motions (actions) applied to these objects. One can, no doubt, introduce cartesian coordinates, and describe these objects, and actions on them, with reference to a coordinate system in a space of points.

But while this may be convenient for computations, it is far more convenient and natural in many ways to visualize objects, and actions on them, purely by themselves, and not in terms of their locations with respect to some coordinate system. When you look at a chair, for instance, all that initially matters is the mutual relationships of the points that constitute the chair, not how the points are individually situated in a cartesian coordinate system. Moreover, when you move the chair, what you observe is that the set of points that make up the chair have preserved their mutual relationships; it is still the same chair. More specifically, transformations (rigid motions, like translations) have taken place under which every point has been moved to replace precisely one other point, and the distance between any two points has remained invariant.

To look at the basic idea algebraically, let us confine ourselves to geometry in the plane. Let $P$ denote the set of all points of the plane, and for any $x, y \in P$, let $d(x, y)$ denote the distance between $x$ and $y$, in the sense commonly understood. The intuitive idea of a rigid motion of the plane corresponds to an onto function $f : P \to P$ such that $d(f(x), f(y)) = d(x, y)$. Commonly referred to as an *isometry*, such a function is also one-to-one, as can be easily checked.[9] Significantly, the set of all isometries of $P$ constitute a group under composition.

This connection of geometrical properties (such as distances and angles) with groups of transformations was what led Felix Klein to enunciate, in his classic work, known as *Erlangen Programme*, a very profound and general rendering of the idea of geometry. Very roughly, it can be summed up by saying that *geometry is the study of those properties of a space that remain invariant under a group of transformations on that space.*[10]

Klein's ideas have had a profound influence on the evolution of geometry and algebra, and of mathematics in general. Their implications in the study of symmetry is what is of concern to us here. For the reader interested in knowing more about these ideas and their broad ramifications, there are several very engaging books to turn to. See, for instance, Mumford [22], Ash [2], Beardon [3], and Solomon [24]. To go with the original paper of Klein's, there is, of course, Klein's three volume treatise [19].

[9] Recall that $d(x, y) = d(y, x)$, and also that $d(x, y) = 0$ if and only if $x = y$. Thus, if $f(x) = f(y)$ then, since $d(f(x), f(y)) = d(x, y) = 0$, it follows that $x = y$. In other words, $f$ is one-to-one.

[10] Klein's original notes for a lecture, written in 1872, were in German. For an English translation, see [18].

## 2.5 Vector Spaces

The theory of vector spaces has long occupied a central place in the study of signals and systems. It is reasonable to assume therefore that the reader is already familiar with the basic concepts of the theory. There is extensive introductory literature on the subject, and I take the liberty of using the basic terms, such as those of bases, subspaces, dimension, linear transformations, and so on, without explicitly laying down their definitions here. Considering that the related terminology and notation is highly standardized, this should not cause any difficulty.

Our concern here is with finite-dimensional vector spaces over the field of real or complex numbers ($\mathbb{R}$ or $\mathbb{C}$). There are essentially two main ideas about them that we shall make use of in our treatment of symmetry. One of them has to with matrices of vectors and linear transformations. The other one is related to decomposition of vector spaces into subspaces.

### 2.5.1 Matrices of Vectors and Linear Transformations

With respect to a particular ordered basis for a vector space $V$ of dimension $n$, a vector as well as a linear transformation has a unique matrix representation. For vectors, this follows from the fact that, for a particular basis, say, $\nu = (\nu_1, \nu_2, \ldots, \nu_n)$ for $V$, any vector $x$ is expressible as a unique linear combination of the basis vectors:

$$x = a_1\nu_1 + a_2\nu_2 + \cdots + a_n\nu_n\,, \tag{2.1}$$

where $a_1, a_2, \ldots, a_n$, the *coordinates* of $x$ with respect to the basis $\nu$, are scalars from the field $\mathbb{F}$ of the vector space.

The *matrix* of $x$ with respect to the basis $\nu$, which we shall write as $[x]_\nu$, is the $n \times 1$ column matrix formed by the coordinates of $x$ with respect to $\nu$:

$$[x]_\nu = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

Most of the time, the basis in use is clear from the context, or it is understood to be same throughout a part of a discussion. At such places, we shall omit the subscript and simply write $[x]$ for the matrix of $x$.

Between vectors in $V$ and their matrices (with respect a particular basis), there is one-to-one correspondence. Moreover, for $x, y \in V$ and $\lambda \in \mathbb{F}$,

$$[x + y] = [x] + [y]\,, [\lambda x] = \lambda[x]\,.$$

In short, the vector space $V$ is isomorphic to the vector space of $n \times 1$ matrices over the field $\mathbb{F}$.

Now, using the expansion (2.1), we see that the action of a linear transformation $H : V \to V$ on a vector $x$ is also expressible in matrix form. Observe that, with respect to the basis $\nu$,

$$[Hx] = a_1[H\nu_1] + a_2[H\nu_2] + \cdots + a_n[H\nu_n]\,, \tag{2.2}$$

The *matrix* of the linear transformation $H$, which we write as $[H]$, is the $n \times n$ matrix with $[H\nu_i]$ as its columns:

$$[H] := \Big[\; [H\nu_1] \quad [H\nu_2] \quad \cdots \quad [H\nu_n] \;\Big] \tag{2.3}$$

We may then recast (2.2) as the matrix equality,

$$[Hx] = [H][x] \tag{2.4}$$

where it is understood that the same basis $\nu$ is in use throughout.

**Exercise 2.5.1** *If $G$ and $H$ are two linear transformations on* $\mathbf{V}$*, then the matrix of their composition is the product of their individual matrices: $[GH] = [G][H]$. Likewise, $[G + H] = [G] + [H]$, and $[\lambda H] = \lambda[H]$. Thus, for a fixed basis, the algebra of linear transformations turns into algebra of their matrices.*

Although all basis sets serve essentially the same purpose—that of providing alternative coordinate systems, different ones may be preferred in different situations. This is something like what you do when you describe the location of a place to some one. You may direct him for instance to start from a point known to both of you, move 4 m to the east, and then from there move 53 m to the north. To reach the same final point, you may instead ask him to choose the same starting point, but to go 35 m from there at 30° north of east, take a left turn at right angles and then proceed for 40 m. Either way, you have used a particular coordinate system, the east-west-north-south (EWNS) system with your chosen spot as the origin, for reference. In the first case you give your directions in this coordinate system itself, and in the second you describe with respect to it a new coordinate system, and give directions for movement in the new system. Note that there is nothing sacrosanct about the EWNS coordinates; you could have chosen any other coordinate system in its place. But you prefer to use this rather than any other, at least to start with, because every one knows how to go about using this.

For bases in an $n$-dimensional vector space $V$, we have a similar situation. In this case too, there is in general no preferred basis, and we keep changing from one to another, depending upon what is convenient in a particular work.

It is important to remember at this point that matrices of vectors and linear transformations are tied to a particular basis; if we change the basis, the matrices also change. To see what form these changes take, consider on $V$ a new basis $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_n)$, in addition to the old one $\nu = (\nu_1, \nu_2, \ldots, \nu_n)$. Using (2.1), the expansion of vector $x$ in terms of the old basis, the matrix of $x$ with respect to the new basis is then expressible as

$$[x]_\varphi = a_1[\nu_1]_\varphi + \cdots + a_n[\nu_n]_\varphi \,, \tag{2.5}$$

and, considering that the $a_i$ are the coordinates of $x$ with respect to the old basis, we have

$$[x]_\varphi = [M]_{\nu\varphi}[x]_\nu \,, \tag{2.6}$$

where $M_{\nu\varphi}$ denotes the so called *change of basis matrix* from the basis $\nu$ to basis $\varphi$, whose columns are the matrices of the old basis vectors $\nu_i$ with respect to the new basis $\varphi$:

$$[M]_{\nu\varphi} := \begin{bmatrix} [\nu_1]_\varphi & [\nu_2]_\varphi & \cdots & [\nu_n]_\varphi \end{bmatrix} \tag{2.7}$$

Note that the matrix $[M]_{\nu\varphi}$ is invertible. For, by interchanging the roles of the two bases, instead of (2.6), we get

$$[x]_\nu = [M]_{\varphi\nu}[x]_\varphi \,, \tag{2.8}$$

where $[M]_{\varphi\nu}$ is the change of basis matrix from $\varphi$ to $\nu$:

$$[M]_{\varphi\nu} := \begin{bmatrix} [\varphi_1]_\nu & [\varphi_2]_\nu & \cdots & [\varphi_n]_\nu \end{bmatrix} \tag{2.9}$$

Combining (2.6) and (2.8), we conclude that the two change of basis matrices are inverses of each other, and indeed they are invertible.

Equations (2.6) and (2.8) tell us how matrices of vectors change with change of basis. Let us now see how matrices of transformations change with change of basis. Abbreviating $[M]_{\varphi\nu}$ as $[M]$, and using Eq. (2.8) together with Eq. (2.4), we arrive at the relationship

$$[H]_\varphi[x]_\varphi = [M]^{-1}[H]_\nu[M][x]_\varphi \tag{2.10}$$

for every $x$. Thus,

$$[H]_\varphi = [M]^{-1}[H]_\nu[M] \tag{2.11}$$

Put in words, Eq. (2.11) means that the new matrix of transformation is obtained from the old one through a similarity transformation under the matrix $[M]$. Recall that for matrices $A$, $B$ of the same size, if there is a matrix $P$ such that $B = P^{-1}A\,P$, then we say $B$ is *similar* to $A$, and that $B$ is obtained from $A$ through a *similarity transformation* under $P$. Note that if $B$ is similar to $A$ then $A$ is similar to $B$, $A$ is similar to $A$, and if $A$ is similar to $B$ and $B$ is similar to $C$ then $A$ is similar to $C$. Thus, for matrices of the same size, similarity transformations define an equivalence relation “similar to” that partitions them into equivalence classes.

**Remark 2** We have used square brackets to distinguish vectors and transformations from their matrices. Most of the time, we shall drop the brackets and rely on the context to clarify what we have in mind. Likewise, we shall omit the subscripts for indicating the bases in use, unless there is a chance for confusion. ♠

## 2.5.2 Direct Sums of Subspaces

The notions of complements and direct sums of subspaces often present some difficulty of visualization to students. One way to overcome this difficulty is to start with the notions of complements of sets and their (disjoint) partitioning, and then to show how they lead to analogous notions for vector spaces.

Recall that for a set $S$, there is its *power set* $\mathcal{P}(S)$, the set of all its subsets; the null set, $\emptyset$, is included as one of its members, and so is $S$ itself. On $\mathcal{P}(S)$, we have the binary operations of *union* and *intersection* ($\cup$ and $\cap$ respectively). Given a subset $X$ of a set $S$, its *complement* is another subset $X'$ of $S$ such that $X$ and $X'$ are disjoint, i.e., $X \cap X' = \emptyset$, and $X \cup X' = S$. If $X_1, X_2, \cdots, X_n$ are pairwise disjoint subsets of a set $S$, i.e., $X_i \cap X_j = \emptyset$ for $i \neq j$, and if $X_1 \cup X_2 \cdots \cup X_n = S$ then $X_1, X_2, \cdots, X_n$ constitute a disjoint partitioning of $S$.

For a vector space $W$, consider two subspaces $U$ and $V$. Can the two ever be disjoint in the sense of sets? No, because the element 0 is present in both of them. But if that is the only common element, ie, $U \cap V = \{0\}$, then we do have a condition here that serves essentially the same purpose as disjointness of sets, and so we say that subspaces $U$ and $V$ are *disjoint* when such is the case.

Consider now the set theoretic intersection and union of the two subspaces. Their intersection $U \cap V$ is also a subspace. Indeed, if $x$ and $y$ are both in $U \cap V$ then $x + y$ and $\alpha x$ are also $U$ as well as in $V$, making $U \cap V$ a subspace.

Their union is, however, not a subspace in general.[11] For, suppose that, as sets, neither of the subspaces contains the other one. That is, there are elements $x$ and $y$ of $W$ such that $x \in U$ but $x \notin V$, and $y \in V$ but $y \notin U$. (Consider the geometrical case of two planes through the origin in 3-D Euclidean space.) Both $x$ and $y$ are in $U \cup V$, but what about the element $x + y$? Suppose this too is in the union, i.e., it is either in $U$ or $V$, or in both. If it is in $U$ then so is $y = (x + y) - x$. But this contradicts the hypothesis, and so it is not in $U$, and from symmetry of the argument, it is not in $V$ either. Thus $x + y$ is not in the union, and therefore the union is not a subspace.

But then, can we find a binary operation for subspaces that serves virtually the same purpose that union does for sets? Let us try out their *sum* defined as $U + V = \{x + y | x \in U \text{ and } y \in V\}$.[12] You may verify that $U + V$ is also a subspace of $W$. For the set of all subspaces, this operation is indeed a natural analogue of set theoretic union.

[11] The qualifier 'in general' points to the fact that there are situations in which it is. Indeed, if one of them is contained in the other, say, $U \subseteq V$, then certainly $U \cup V$ $(= V)$ is a subspace. In fact, the union is a subspace only in such a situation. To see this, let us say that $U \cup V$ is a subspace. Then pick any $x \in U$ and $y \in V$. Since both $x$ and $y$ are in $U \cup V$, $x + y$ is also in $U \cup V$, by hypothesis. That is, $x + y$ is in $U$, or in $V$, or in both. If $x + y \in U$, then $(x + y) - x = y$ is in $U$ for any $y \in V$. Therefore, $V \subseteq U$. Likewise, if $x + y \in V$, then $U \subseteq V$. Finally, if both hold then $U = V$.

[12] The '+' sign is used twice here; on the left it denotes the newly defined binary operation on subspaces, and on the right it denotes an operation on vectors. What it actually denotes in a particular situation is decided by the context.

If the subspaces are disjoint, i.e., $U \cap V = \{0\}$, then the sum $W = U + V$ is called a *direct sum* of $U$ and $V$, written as $U \oplus V$. Putting it the other way round, $U \oplus V$ is a *direct sum decomposition* of $W$.

Clearly, since vector addition is commutative and associative, the binary operations of sum and direct sum of subspaces are also commutative and associative. Moreover, their associativity implies that their repeated application to any finite number of subspaces yields unique final sums, independent of the sequence in which the summands are associated in pairs. So for any finite number of subspaces, say $U_1, U_2, \ldots, U_n$, we leave out the associating brackets and write the sums simply as $U_1 + U_2 + \cdots + U_n$ and $U_1 \oplus U_2 \oplus \cdots \oplus U_n$ (as we do for finite unions of subsets, and in fact for any associative binary operation in general).

As in the case of sets, if a vector space $W$ has subspaces $U$ and $V$ such that $W = U \oplus V$ then we say that $V$ is a *complement* of $U$ (or the other way round).

Observe that while a subset's complement is unique, the complement of a subspace is not. As a pictorial illustration of this, consider the case of the plane, with the usual $x$- and $y$-axes centred on a specified origin, as a model of $\mathbb{R}^2$. Then the $x$-axis represents a subspace, say $X$, of $\mathbb{R}^2$. Likewise, the $y$-axis represents another subspace, say $Y$, of $\mathbb{R}^2$. Clearly, $X \cap Y = \{0\}$, and $X + Y = \mathbb{R}^2$, so that we have $X \oplus Y = \mathbb{R}^2$. Consider now the points on a straight line through the origin at some arbitrary (nonzero) angle to the $x$-axis. These points too represent a subspace of $\mathbb{R}^2$, say $W$, such that $X + W = \mathbb{R}^2$ and $X \cap W = \{0\}$, i.e., $X \oplus W = \mathbb{R}^2$. Thus, $W$ is another complement of $X$, which is different from $Y$.[13]

Now, suppose that a vector space $W$ has a decomposition $W = U + V$. Then any vector $x \in W$ is, by definition, expressible as a sum: $x = u + v$, where $u \in U$ and $v \in V$. What can we say about the uniqueness of such a vector decomposition? Well, we can say the following.

**Theorem 2.5.1** *Let two subspaces $U$ and $V$ of a vector space $W$ be such that $W = U + V$. If $U \cap V = \{0\}$, so that $W = U \oplus V$, then any vector $x \in W$ has a unique decomposition $x = u + v$ with $u \in U$ and $v \in V$. Conversely, if every vector $x \in W$ has a unique decomposition $x = u + v$ with $u \in U$ and $v \in V$, then $U \cap V = \{0\}$.*

PROOF: Suppose that the subspaces are disjoint, i.e., $U \cap V = \{0\}$. Further suppose that we also have $x = u' + v'$ with $u' \in U$ and $v' \in V$. In that case, $0 = x - x = (u + v) - (u' + v') = (u - u') + (v - v')$. That is, $u - u' = v' - v$. But $u - u' \in U$ and $v' - v \in V$, i.e., $(u - u'), (v' - v) \in U \cap V$. Therefore, since $U \cap V = \{0\}$ by hypothesis, we have $u - u' = v - v' = 0$. In other words, if the subspaces $U$ and $V$ are disjoint then the decomposition $x = u + v$ is unique.

[13] My arguments here are essentially the same as those of Mansfield [21, Chapter 2, Section 6].

Next, suppose there is a nonzero vector $z$ that is in $U$ as well as in $V$. Then, with $u' = u - z$ and $v' = v + z$, $x = u' + v'$, where $u' \in U$ and $v' \in V$ and, since $z \neq 0$, we have another decomposition for $x$. Thus if the subspaces $U$ and $V$ are not disjoint then the decomposition $x = u + v$ is not unique.

The two sets of arguments together complete the proof. □

Associated with the notion of direct sum decompositions, there is the important notion of projections, or projection operators. ( For a physical analogy, consider the space of all audio signals. An audio signal, $x$ can be thought of as a sum of two parts, one consisting of its low frequency components, say up to 1 kHz, and the other consisting of its high frequency components, those above 1 kHz. Call these parts $x_L$ and $x_H$ respectively. Then, with $x$ as input to an ideal low–pass filter with cut-off at 1 kHz, $x_L$ is its output. Likewise, if we put $x$ through a high–pass filter, we get $x_H$. A projection is like these filters.)

For any vector space $W$, consider a decomposition $W = U \oplus V$. Since every vector $x \in W$ is uniquely expressible as $x = u + v$, with $u \in U$ and $v \in V$, this decomposition of $W$ determines a transformation $H : W \to W$ defined by $Hx = u$.[14] There is just one and only one such transformation. We say that $H$ is a *projection*, or, if we want to spell out the details, the projection *of* $W$ *on* $U$ *along* $V$. Likewise, the decomposition $U \oplus V$ also determines, as a companion to $H$, another transformation $G : W \to W$ defined by $Gx = v$; it is the projection of $W$ on $V$ along $U$.

Note that $U$ and $V$ are respectively the range and null space of $H$. Indeed, for any $x \in W$, $Hx$ is in $U$ by definition. For any $u \in U$, there is $x = u + 0$ such that $Hx = u$. Thus $U$ is the range of $H$. Now, if $x \in V$ then $Hx = H(0 + x) = 0$, and if $Hx = 0$ then $x = 0 + v$ for some $v \in V$, and so $x = 0 + v = v \in V$. Thus $V$ is the null space of $H$.

It is easily checked that a projection is a linear transformation. Continuing with the projection $H$, suppose $x_1 = u_1 + v_1$ and $x_2 = u_2 + v_2$, with the $u_1, u_2 \in U$ and $v_1, v_2 \in V$. Then $x_1 + x_2 = (u_1 + u_2) + (v_1 + v_2)$, where $(u_1 + u_2) \in U$ and $(v_1 + v_2) \in V$. Consequently, $H(x_1 + x_2) = u_1 + u_2 = Hx_1 + Hx_2$. Likewise for scaling. Thus $H$ is linear, and so is the projection $G$ on $V$ along $U$. Incidentally, if $I$ denotes the identity operator on $W$ ($Ix = x$ for any $x \in W$), then $H + G = I$, i.e., for any $x \in W$, $(G + H)x = x$.[15]

There is another important point about a projection $H$. For any vector $x = u+v$, with $u$ in $U$, $v$ in $V$, it follows that $H^2x = HHx = Hu = H(u + 0) = u = Hx$, i.e., $H^2 = H$.[16] (Passing a signal through an ideal filter twice is only as good as passing it once.)

[14] Many texts treat it as a transformation $H : W \to U$. While it is quite in order to proceed this way, this will not allow us to talk in the usual manner of the sum $H + G$, where $H$ and $G$ are two such transformations, taking $W$ to different subspaces.

[15] Recall that for two linear transformations $H$ and $G$ on $W$, their sum $H + G$ is defined by $(H + G)x = Hx + Gx$ for $x \in W$.

[16] In words, we say that $H$ is *idempotent*.

Conversely, if $H$ is a linear transformation on $W$, and if $H^2 = H$, then $H$ is a projection of $W$. To see this, let $V$ be the null space of $H$, and $U$ its range. For a vector $u \in U$, there is $x \in W$ such that $u = Hx$, and so $Hu = H^2x = Hx = u$. On the other hand, if $Hu = u$ then $u \in U$. Thus a vector $u$ is in the range of $H$ if and only if $Hu = u$. Now, if $u$ is also in $V$, i.e., $Hu = 0$, then $u = 0$. In other words, $U \cap V = \{0\}$. Finally, for an arbitrary vector $x \in W$, $Hx$ is in $U$, and, since $H(x - Hx) = Hx - H^2x = 0$, $(x - Hx)$ is in $V$. We thus get a decomposition $x = Hx + (x - Hx)$ with $Hx$ in $U$ and $(x - Hx)$ in $V$. Coupled with the fact that $U$ and $V$ are disjoint, this means that $W = U \oplus V$, and that $H$ is the projection on $U$ along $V$. To sum up, we have the following result.

**Theorem 2.5.2** *For a transformation $H$ on a vector space $W$, the following two statements are equivalent.*

1. *$H$ is a projection of $W$ on $U$ along $V$.*
2. *$H$ is a linear transformation with $U$ and $V$ as its range and null space respectively, and $H$ is idempotent, i.e., $H^2 = H$.*[17]

In the light of this result, one is justified in defining a projection not as we have chosen to do here, but as a linear idempotent transformation instead.[18]

For direct sum decompositions into more than two subspaces, there is a natural extension of the notion of projection operators. You may consult a text such as Hoffman and Kunze [15] or Roman [23] for details.

With this much on vector spaces, we now turn to a class of structures that have two faces, one relational and another algebraic. Partially ordered sets, lattices and Boolean algebras are what I have in mind. In the rest of this chapter, I give a very brief outline of the basic ideas related to them. You have, I am sure, already met these structures in one context or another, but perhaps as isolated entities. Boolean algebra, for instance, you may associate exclusively with logic circuits and their design, or perhaps with probability theory. Similarly, you may be familiar with Venn diagrams and their use in a pictorial visualization of the algebra of power sets (under containment, union, and intersection). My purpose here is to highlight the underlying unity.

[17] Note that linearity has been stipulated separately in this statement. Idempotence does not imply linearity. Consider for instance $E : V \to V$, $E(x) = a$ for any $x \in V$ and a fixed $a \in V$. $E$ is idempotent but not linear.

[18] Many texts in fact do that. See, for example, Hoffman and Kunze [15, p. 211]. Dunford and Schwartz also adopt this definition in their treatise [8, p. 37]. We have followed the line adopted in Halmos [13, pp. 73–74].

# 2.6 Posets, Lattices, and Boolean Algebras

Like sameness or equality, the notion of order is a very basic one in our daily activities of comparing objects. Assessments such as “longer than”, “heavier than”, “at least as heavy as”, “contained in”, are all instances of ascribing to objects some kind of an ordering relation. In arithmetic, we have for numbers the property of being “less than” or “less than or equal to”. In the theory of sets, we talk of one set being a subset of another. For purposes of mathematical formalization, the relation that has acquired a place of central importance in this context is that of a partial order.

For a precise definition of a partial order, it is helpful first to introduce one more property of relations, in addition to the ones introduced earlier (namely, reflexivity, symmetry, and transitivity). Consider a set $S$ equipped with an equality (equivalence) relation. Using $=$ to denote this equality, let $\preceq$ be another relation on $S$ such that, if $x \preceq y$ and $y \preceq x$, then $x = y$. We then say that the relation $\preceq$ is *antisymmetric*. A relation $\preceq$ on a set $S$ is called a *partial order* if it is reflexive, transitive and antisymmetric. The structure $(S, \preceq)$ is correspondingly called a *partially ordered set*, or in short a *poset*.

Note that the relation $\leq$ on $\mathbb{R}$ is antisymmetric. So is the relation $\subseteq$ on the power set of a set. But there is a difference. It is always the case for $\leq$ that any two real numbers $x$ and $y$ are comparable, i.e., either $x \leq y$ or $y \leq x$. For the relation $\subseteq$, however, this may not always be so; it may well be that for two subsets $X$ and $Y$ of a set $S$, neither $X \subseteq Y$ nor $Y \subseteq X$. The definition of antisymmetry accommodates this difference.

Also note that, in addition to being antisymmetric, the relation $\leq$ is reflexive and transitive as well. Thus the structure $(\mathbb{R}, \leq)$ is a poset. Likewise, for a set $S$, the structure $(\mathcal{P}(S), \subseteq)$ is a poset too.

## 2.6.1 From Posets to Lattices

Built up on the notion of a poset, there is the important notion of a lattice, which is a special kind of a poset. In order to see what it is, it is perhaps best to first examine closely the structure of the power set of a set.

Recall that for a set $S$, there are two ways of looking at the structure of its power set $\mathcal{P}(S)$: one as a relational structure $(\mathcal{P}(S), \subseteq)$, and two, as an algebraic structure $(\mathcal{P}(S), \cup, \cap)$. Furthermore, the two views are very tightly interlinked. To be specific, we know that, for two subsets $X$ and $Y$ of $S$, $X \subseteq Y$ if and only if $X \cup Y = Y$. Equivalently, $X \subseteq Y$ if and only if $X \cap Y = X$. Can we say some thing similar about posets in general? Given a partial order, can we associate with it two binary operations (like $\cup$ and $\cap$) in a similar fashion? The answer is a qualified yes. This is made possible by the fact that for posets in general one can talk of their smallest and largest elements, and of upper and lower bounds for their subsets. Analogous to what we mean by these terms in the case of sets and their subsets, they are defined for posets as follows.

For a poset $(P, \preceq)$, if there exists an element $s \in P$ such that $s \preceq a$ for every $a \in P$ then, by antisymmetry, $s$ is unique. Such an element $s$ is called the *least* (or *smallest*) element of the poset. Similarly, if there exists an element $s \in P$ such that $a \preceq s$ for every $a \in P$ then $s$ is the *greatest* ( or the *largest*) element of the poset. Next, let $X \subseteq P$. An element $s \in P$ is an *upper bound* of $X$ if $x \preceq s$ for every $x \in X$. If $s$ is an upper bound of $X$ such that $s \preceq a$ for any other upper bound $a$ of $X$, then $s$ is a *least upper bound* (or *supremum*) of $X$, written in short as $s = \sup(X)$. The element $s$ is a *lower bound* of $X$ if $s \preceq x$ for every $x \in X$; $s$ is a *greatest lower bound* (or *infimum*) of $X$ if $a \preceq s$ for any other lower bound $a$ of $X$, and we write $s = \inf(X)$ in this case. Again, it follows from antisymmetry that both $\sup$ and $\inf$, if they exist, are unique.

In the special case in which the set $X$ consists of a pair of elements, $X = \{a, b\}$, and for which $\sup$ and $\inf$ exist, we respectively write them as $\sup(a, b)$ and $\inf(a, b)$. Furthermore, if they exist for every pair of elements of the poset $S$, then the two qualify as binary operations on $P$. In that case, a common practice is to write $\sup(a, b)$ as $a \vee b$, and $\inf(a, b)$ as $a \wedge b$. At places, we refer to the operations $\vee$ and $\wedge$ as $\sup$ and $\inf$ respectively.

**Remark 3** Observe that for the poset $(\mathcal{P}(S), \subseteq)$, the operations $\cup$ and $\cap$ are special cases of the operations $\vee$ and $\wedge$ respectively.[19] (Recall that the union of two subsets is the smallest subset containing both, and the intersection of any two subsets is the largest subset common to both.) ♠

We are now ready to define a lattice: A poset $(P, \preceq)$ is a *lattice* if each pair of elements of $P$ has a $\sup$ and an $\inf$.

For any set $S$, the poset $(\mathcal{P}(S), \subseteq)$ is a lattice. However, not every lattice is as rich in structural properties as a power set. All the same, we can bring in additional constraints on lattices, with the power set lattice in mind as template. To be specific, we can talk of complements, and of distributivity. But before that, let us take stock of some of the properties of a lattice that are reminiscent of what we know for power sets. It can be verified that the following claims are true for any three elements $x, y, z$ of a lattice $(P, \preceq)$:

| | | | |
|---|---|---|---|
| (2.12) | Claim 1 | $x \vee y = y \vee x\,,$ | $x \wedge y = y \wedge x$ |
| (2.13) | Claim 2 | $x \vee (y \vee z) = (x \vee y) \vee z\,,$ | $x \wedge (y \wedge z) = (x \wedge y) \wedge z$ |
| (2.14) | Claim 3 | $x \vee (x \wedge y) = x\,,$ | $x \wedge (x \vee y) = x$ |
| (2.15) | Claim 4 | $x \vee x = x\,,$ | $x \wedge x = x$ |

Claims 1 to 4 are respectively called the *commutative*, *associative*, *absorptive*, and *idempotence* properties of a lattice; these hold for all lattices. Observe that, with $\vee$ and $\wedge$ replaced by $\cup$ and $\cap$ respectively, these become for power sets their four

[19] It may perhaps be more appropriate to say in reverse that $\vee$ and $\wedge$ are generalizations of $\cup$ and $\cap$ respectively.

familiar properties with the same names. Furthermore, as for power sets, these four properties uniquely characterize a partial order.

In order to see this, let us go backwards. Let us say that for an algebraic structure $(P, \vee, \wedge)$ with two binary operations, all we know is that the operations satisfy the conditions (2.12)–(2.15). Next, let us define a relation $\preceq$ on $P$ by the condition,

$$x \preceq y \ \text{ if and only if } \ x \vee y = y. \tag{2.16}$$

It can be shown that *the relation $\preceq$ is a unique partial order on $P$ that makes $(P, \preceq)$ a lattice, with $\vee$ and $\wedge$ serving as its* $\sup$ *and* $\inf$ *respectively.*[20] Moreover, the condition (2.16) is equivalent to the condition,

$$x \preceq y \ \text{ if and only if } \ x \wedge y = x. \tag{2.17}$$

We thus come a full circle, starting off with a relational definition of a lattice, passing on to an algebraic structure with its two binary operations satisfying the conditions (2.12)–(2.15), and then back.

### 2.6.2 Complemented and Distributive Lattices

For a power set lattice $(\mathcal{P}(S), \subseteq)$, we have the familiar idea of the complement of a subset: for a subset $X$ of $S$, its complement is a unique subset $X'$ of $S$ such that $X \cup X' = S$ and $X \cap X' = \emptyset$.

This suggests an analogous notion of a complement for a lattice in general, which reduces to the familiar one for power sets. Let $(P, \preceq)$ be a lattice with greatest element $U$ and smallest element $E$. Then we say that an element $y$ of $P$ is a *complement* of an element $x$ of $P$ if the two together satisfy the condition

$$x \vee y = U \quad \text{and} \quad x \wedge y = E\,, \tag{2.18}$$

where $\vee$ and $\wedge$ respectively denote $\sup$ and $\inf$ for the lattice. We say that the lattice is a *complemented lattice* if it has a greatest element and a smallest element and, in addition, every element of it has at least one complement.

As we shall see in the next subsection, complements need not be unique in this general setting. Recall that complements, as defined for power sets, are unique. This is ensured by the distributivity property of union and intersection. We can analogously define distributivity for lattices as follows.

[20] Details of the proof can be found in Arnold [1, Chapter 4]. Lidl [20, Chapter 1] and Bloch [4, Chapter 7] are also worth looking at in this connection.

Let $(P, \preceq)$ be a lattice, with $\vee$ and $\wedge$ respectively denoting its $\sup$ and $\inf$. Then we say that it is a *distributive lattice* if any three elements $x, y, z \in P$ meet the condition,

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z). \tag{2.19}$$

Again, just as condition (2.16) is equivalent to condition (2.17), condition (2.19) is equivalent to the condition,

$$x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z). \tag{2.20}$$

Now, suppose $(P, \preceq)$ is a complemented distributive lattice. Then, by definition, any element $x \in P$ has at least one complement, say $y$. Let us say that $x$ has another complement $z$. Clearly, if $U$ and $E$ are the greatest and least elements of the lattice, then $y = y \wedge U = y \wedge (x \vee z)$. Invoking distributivity and properties of a complement, we get $y = (y \wedge x) \vee (y \wedge z) = E \vee (y \wedge z) = (y \wedge z)$. Similarly, by symmetry, $z = (z \wedge y) = (y \wedge z)$. In other words, $y = z$. In summary, we find that *every element of a complemented distributive lattice has a unique complement.* Thus, a complemented distributive lattice is, in this respect, put at par with a power set lattice.

In the mathematical literature, a common practice nowadays is to call a complemented distributive lattice *a Boolean algebra*. This apparently differs from the common perception that an engineering student has of Boolean algebra. For him (her) Boolean algebra is what you use in the design and simplification of digital circuits made up of binary switches. It is generally presented in this context as an algebraic structure with two binary operations satisfying conditions (2.12)–(2.15), together with condition (2.19), and one unary operation (complementation) satisfying condition (2.18). Based on the preceding discussions, we can see that there is no conflict between the two perceptions of a Boolean algebra. They yield two equivalent characterizations of the same structural entity. Both characterizations are very tightly connected with the so-called **Stone's Representation Theorem**, which reads in the finite case as follows: *Every finite Boolean algebra is isomorphic to the Boolean algebra of the power set of some set.* It is this result that justifies the use of the familiar Venn diagrams in depicting logical expressions.

The theory of lattices in general, and of Boolean algebras, is very rich and vast, as you can see from Grätzer [9]. For its modern applications in signal processing, spectral domain design of digital devices, and in cryptography, you may like to look up Stanović [25], Karpovsky [17], and Cusick [7].

### 2.6.3 Lattice of Subspaces of a Vector Space

We have seen that the power set of a set forms a complemented distributive lattice (Boolean algebra). As already mentioned, not every lattice is distributive. As an example of a non-distributive lattice, let us consider the set $\mathcal{S}(V)$ of all subspaces of a vector space $V$. We will first show that it is a lattice under set inclusion ($\subseteq$).

Note, first of all, that $\mathcal{S}(V)$ is a poset under set inclusion, i.e., set inclusion, $\subseteq$, is a reflexive, antisymmetric, and transitive relation on $\mathcal{S}(V)$. This follows directly from set theoretic arguments.

Recall, further, that for $X, Y \in \mathcal{S}(V)$, the set theoretic intersection $X \cap Y$ is also in $\mathcal{S}(V)$; in fact it is the largest subspace contained in both $X$ and $Y$. More formally,

$$\inf\{X, Y\} = X \cap Y.$$

At the same time, the sum $X + Y$, of $X$ and $Y$,

$$X + Y = \{x + y | x \in X, y \in Y\}$$

is the smallest subspace containing both $X$ and $Y$, i.e.,

$$\sup\{X, Y\} = X + Y.$$

Thus the set $\mathcal{S}(V)$, under set inclusion, is a lattice, i.e., it is a poset with a $\sup$ and an $\inf$ for any two of its elements.

But it is not a distributive lattice. We show this with the help of a counterexample for $\mathbb{R}^2$. Let $V_1$ and $V_2$ be its subspaces spanned by the standard basis vectors $\delta_1$ and $\delta_2$ respectively. Then $V_1 + V_2 = \mathbb{R}^2$. Let $W$ denote the subspace spanned by the vector $\delta_1 + \delta_2$. Then

$$W \cap (V_1 + V_2) = W \cap \mathbb{R}^2 = W.$$

But $W \cap V_1 = W \cap V_2 = \{0\}$, the zero subspace of $\mathbb{R}^2$, so that

$$(W \cap V_1) + (W \cap V_2) = \{0\}.$$

Thus, in this case,

$$W \cap (V_1 + V_2) \neq (W \cap V_1) + (W \cap V_2),$$

violating distributivity.[21]

What about complements in this case? Well, for $\mathcal{S}(V)$, the least element is the subspace $\{0\}$ and the greatest element is $V$ itself. Treating it as a lattice, we then find that the condition (2.18), which defines complements for lattices, reduces to the following condition:

$$X + Y = V \quad \text{and} \quad X \cap Y = \{0\}\,, \tag{2.21}$$

where $X$ and $Y$ are subspaces of $V$. Condition (2.21) is equivalent to saying that $X \oplus Y = V$. Thus the notion of a complement, as defined for lattices, coincides in this case with that of a complement of a subspace as defined earlier in Section 2.5.2. Moreover, as explained there, complements of subspaces are in general not unique.

In all, *in the lattice of subspaces of a vector space, we have an example of a complemented non-distributive lattice for which complements are not necessarily unique.*

[21] The discussions given here are based on the treatment in Roman [23, pp. 30–31] and Jauch [16].

## 2.7 Closing Remarks

The brief narrative of this chapter is not a substitute for a detailed exposition of any of the topics. It is essentially meant to give the reader a working start for immediate use in the later chapters. For a fuller account, the references given at various points of the narrative may be consulted. There has been a steady stream of very good texts on algebra and its applications, some of these specially meant for engineers and scientists.

# References

1. B.H. Arnold. *Logic and Boolean Algebra*. Prentice–Hall, New Jersey, 1962.
2. Avner Ash and Robert Gross. *Fearless Symmetry: Exposing the Hidden Patterns of Numbers*. Princeton University Press, Princeton, 2006.
3. Alan F. Beardon. *Algebra and Geometry*. Cambridge University Press, Cambridge, 2005.
4. Ethan D. Bloch. *Proofs and Fundamentals*. Birkhäuser, Basel, 2000.
5. D. Bushaw. *Elements of General Topology*. Wiley, New York, 1963.
6. Irving M. Copi and Carl Cohen. *Introduction to Logic*. Pearson Education, Singapore, 1998.
7. Thomas W. Cusick and Pantelimon Stănică. *Cryptographic Boolean Functions and Applications*. Academic, London, 2009.
8. Nelson Dunford and Jacob T. Schwartz. *Linear Operators; Part I; General Theory*. Interscience Publishers, New York, 1964.
9. George Grätzer. *General Lattice Theory*. Birkhäuser, Basel, 2003.
10. Donald Greenspan. *Discrete Models*. Addison Wesley, Reading, MA, 1973.
11. Donald Greenspan. *Discrete Numerical Methods in Physics and Engineering*. Academic, London, 1974.
12. Paul Halmos and Steven Givant. *Logic as Algebra*. The Mathematical Association of America, USA, 1998.
13. Paul R. Halmos. *Finite–Dimensional Vector Spaces*. D. Van Nostrand, Princeton, 1958.
14. Paul R. Halmos. *Naive Set Theory*. Springer, New York, 1974.
15. Kenneth Hoffman and Ray Kunze. *Linear Algebra*. Prentice–Hall (India), New Delhi, 1972.

16. Joseph M. Jauch. *Foundations of Quantum Mechanics*. Addison Wesley, Reading, MA, 1968.

17. Mark G. Karpovsky, Radomir S. Stanković, and Jaakko T. Astola. *Spectral Logic and Its Applications for the Design of Digital Devices*. Wiley, New Jersey, 2008.

18. F.C. Klein. A comparative review of recent researches in geometry. *Arxiv preprint arXiv:0807.3161*, 2008. http://arxiv.org/abs/0807.3161

19. Felix Klein. *Elementary Mathematics from an Advanced Standpoint: Geometry, Vol. 2*. Dover, New York, 2004.

20. Rudolf Lidl and Günter Pilz. *Applied Abstract Algebra*. Springer-Verlag, New York, 1998.

21. Larry E. Mansfield. *Linear Algebra with Geometric Applications*. Marcel Dekker, New York, 1976.

22. David Mumford, Caroline Series, and David Wright. *Indra's Pearls: The Vision of Felix Klein*. Cambridge University Press, Cambridge, 2002.

23. Steven Roman. *Advanced Linear Algebra*. Springer-Verlag, New York, 1992.

24. Roland Solomon. *Abstract Algebra*. Thomson, Australia, 2003.

25. Radomir S. Stanković, Claudio Moraga, and Jaakko T. Astola. *Fourier Analysis on Finite Groups with Applications in Signal Processing and System Design*. Interscience Publishers, New York, 2005.

# Chapter 3

# Measurement, Modeling, and Metaphors

From a review of mathematical structures in the abstract, we now turn to their representational role. What prompts us to think of signals in terms of vector spaces? What justification is there for us to think of long or short rods in terms of their lengths (in some unit) given as real numbers? Or for that matter, what justifies the practice of reckoning time in terms of numbers shown by a clock? There is a common thread running through all such questions. It is this thread that we wish to examine in this chapter. Our deliberations here run on two overlapping lines of thought. One of them is the line adopted in what has come to be known as the representational theory of measurement, as outlined in such works as Stevens [18], Suppes and Zinnes [19], Scott and Suppes [17], Krantz [5], Roberts [14], Narens [7,8]. The other one is what forms the core of the notion of modeling as outlined by Rosen [15, 16].

## 3.1 Archimedes and the Tortoise

Recall, to begin with, the popular story about the hare and the tortoise, which is generally understood to conclude with the message that slow and steady wins the race. Significantly, alongside the message, the story also has a hidden assumption: *In walking from a point $A$ to another point $B$ at a finite distance from $A$, no matter what your step size, you will be able to cover the finite distance in a finite number steps.*[1] Let us call this assumption $\mathbf{HA}$.

There is a numerical counterpart of this observation, known as the *Archimedean Principle*: If $a$ and $b$ are real numbers, and $a > 0$, then there is a positive integer $n$ such that $b < na$. Note that there is a close connection between this principle and the $\mathbf{HA}$. Assign the number $a$ to the step size, and the number $b$ to the distance between points $A$ and $B$. Then the Archimedean Principle becomes the $\mathbf{HA}$.

[1] What is finite distance, one might ask. But we will ignore such questions here.

V.P. Sinha, *Symmetries and Groups in Signal Processing*, Signals and Communication Technology,
DOI 10.1007/978-90-481-9434-6_3, © Springer Science+Business Media B.V. 2010

But there is a catch. In the axiomatic theory of numbers, the Archimedean Principle is an unavoidable consequence of the axioms. The $\mathbf{HA}$, on the other hand, is at best a hypothesis about real-life situations, borne out of experience. Imagine yourself walking on perfect ice, and you will have a scenario in which $\mathbf{HA}$ may not hold true. In that case, assigning numbers to step sizes (or to distances in physical space in general), and then using the laws of arithmetic to draw conclusions from these numbers, is liable to lead to major errors of judgement about real-life situations.

So, a very pertinent question to ask in this connection is the following: What conditions must hold so that a study of certain empirically verifiable qualitative attributes of physical objects or processes can be carried out quantitatively by mapping them into a system of numbers? Let us see how this question is addressed in the representational theory of measurement.

## 3.2 The Representational Approach

Measurement, in its widest interpretation, is a matter of assigning numbers to certain attributes of objects or events such that, for purposes of analysis, the numbers so assigned can serve for the attributes by proxy. It is a matter of representing a structure of physically observed relations, e.g., those between lengths, weights, preferences, or utilities, by a structure of abstract mathematical objects such as numbers or functions.

In the representational view of measurement, this idea is made more precise by characterizing objects and their attributes in terms of relational structures. Such a characterization is perhaps best explained with the help of an example. We choose for this the case of measurement, or comparison, of lengths of a given set of rods.[2]

### 3.2.1 Measuring Lengths

Let $X$ denote a set of rods, idealized as line segments, and let us say that we are interested in comparing them in terms of what we might call their "longness" attribute. Such a comparison may consist of figuring out for any two rods whether one is longer than the other, or whether the two are equally long. This may be ascertained in the usual way by physically placing them side by side, flush with each other at one end. Results of such pairwise comparisons together define an ordering relation "at least as long as" on $X$, which we will denote here by the symbol $\succeq$.

Besides comparing two rods, we also join them together to make a new rod. In an idealized sense, this amounts to placing the two rods adjacent to each other along a straight line, leaving no gap in between. Assuming that the resulting new rod is

[2] Measurement of weight, or of temperature, could have equally well served the purpose of bringing out the basic idea. But the case of length is perhaps the simplest.

also in the set $X$, this defines a binary operation (a ternary relation), the so called *concatenation* operation, on the set $X$ of rods, which we will denote here by $\circ$.

Our study of the rods in this context is essentially a study of the relational structure $\mathcal{X} = (X, \succeq, \circ)$.[3] Following common practice in the theory of measurement, we shall call such a structure an *empirical structure* to suggest that it is determined through empirical evidence. This is to be contrasted with a numerical structure such as $\mathcal{R} = (\mathbb{R}^+, \geq, +)$, which is a *formal structure*, i.e., one that is an axiomatically defined mathematical construct.

Rather than using direct physical comparisons to determine the empirical structure $\mathcal{X} = (X, \succeq, \circ)$, we commonly take the help of a measuring tape to read off numbers that give the lengths of the individual rods, and then compare the lengths instead. In doing this, we intuitively appeal to the fact that the relation $\geq$ mirrors for the numbers assigned as lengths the same relation that $\succeq$ does for the rods. Furthermore, we also take for granted that the lengths of rods $x$ and $y$ add up to the length of the rod $(x \circ y)$.

To make all this more precise using the language of algebra, we can say that the process of length measurement involves three things here: the empirical structure $\mathcal{X} = (X, \succeq, \circ)$, the formal structure $\mathcal{R} = (\mathbb{R}^+, \geq, +)$, and a mapping $\phi : X \rightarrow \mathbb{R}^+$. The mapping $\phi$, which can be looked upon as an abstraction of the measuring tape, is a structure preserving mapping, a homomorphism. It is a homomorphism from $(X, \succeq, \circ)$ into $(\mathbb{R}^+, \geq, +)$, i.e., for any $x, y \in X$, it satisfies the conditions:

$$x \succeq y \quad \text{if and only if} \quad \phi(x) \geq \phi(y)\,, \tag{3.1}$$

$$\phi(x \circ y) \quad = \quad \phi(x) + \phi(y). \tag{3.2}$$

Such a homomorphism is said to provide *a representation* of the structure $(X, \succeq, \circ)$ by the structure $(\mathbb{R}^+, \geq, +)$. On the lines of representational measurement theory, the basic question posed towards the end of the previous section takes here the following form: What empirically verifiable conditions must be placed on the structure $\mathcal{X}$ so that the structure $\mathcal{R}$ serves as a representation for it?

Consider for instance the following four conditions.

1. The relation $\succeq$ is transitive, i.e., for $x, y, z \in X$, if $x \succeq y$ and $y \succeq z$, then $x \succeq z$.
2. For any $x, y \in X$, either $x \succeq y$ or $y \succeq x$ or both.
3. The operation $\circ$ is associative and commutative.
4. Let us assume that for $x \in X$ we can have as many 'perfect copies' as we like, and let $x_n$ denote the rod obtained by concatenating $n$ perfect copies of $x$. *Then, for* $x, y \in X$*, there is a positive integer* $n$ *such that* $x_n \succeq y$ *but* $y \not\succeq x_n$.[4]

[3] Note that a binary operation is a ternary relation. So, in principle, we can treat $\circ$ as a relation.

[4] This is brought in to fall in line with the Archimedean principle for numbers.

Intuitively, we will all agree—perhaps after some reflection, that these conditions must necessarily be met by $\mathcal{R}$ for the measuring tape procedure to work. Are these conditions really necessary? Are they sufficient as well? Is the representation unique?[5] If it is not unique then what kind of structural relationship the various possible representations have? How can we formalize the notion of scales of measurement?

Representational measurement theory is concerned with a host of such issues underlying the concept of measurement in general. The reader is referred to Roberts [14], Narens [7], and Krantz [5], and Narens [8], for a comprehensive discussion. Narens [8] also discusses connections of theories of measurement with symmetry. For a recent critique of the theory, see Domotor [3].

Our concern here is not with the details of the theory, but with the spirit of it, and with the connections it has with the notion of modeling.

Notice that measurement is typically concerned with numerical representation of a specific attribute of a certain class of objects, events, or processes. The notion of modeling, as it is to be understood here, has to do with a formal representation of the class as a whole, one that mirrors all its attributes of interest.[6] Let me go back to the case of sounds, or audio signals, as an illustration.

### 3.2.2 From Measurement to Modeling

For a sound, a microphone enables us to convert it into a voltage at a terminal pair as a function of time. Every sound has such a map. And through a complementary transducer, the loud speaker, this map is invertible. Note that the set of sounds has additional structure, in the sense that there are natural operations that we perform on sounds. The act of mixing of two different sounds produces another sound. There is also the operation of raising or lowering the volume of a sound. Then there are delays. The functions into which sounds map should also have the equivalent of these operations, and the map from sounds to functions should preserve these operations. Thus mixing corresponds to adding functions, volume control corresponds to scalar multiplication of functions, and operations of time delay correspond to time translation of functions.

Then there are more intricate issues about sounds. We think of them in terms of certain common features. Thus apart from volume, there is the notion of pitch. We talk of low–pitched and high–pitched voices. Likewise, we distinguish between the drone of a plane and a whistle, attributing the difference to their basic pitches. More generally, we may say that we categorize sounds in terms of some of

[5] Clearly, it is not unique. We know that the numbers depend on what scale we choose—centimeter or inch, for instance. But we also know that there is a proportionality relationship between numbers we get on these two scales.

[6] Both in mathematics and the sciences, the word ‘model’ has been used in a variety of apparently different ways. A critique of these can be found in Suppes [20], where he takes stock of some of these different ways, and presents arguments to show their essential unity.

their discernable features and attributes. Such features and attributes translate into properties of functions into which they map, e.g., how these functions change from one moment to the next. For functions of time, for instance, such features and attributes are related to their various derivatives.

Moving over to the set of functions that encodes the set of all sounds, we now can see a more elaborate structure for it. If the set of such admissible functions is $X$, then it has prescribed on it the following operations:

1. Adding
2. Scaling
3. Delaying, or shifting in time
4. Differentiating, which gives an idea as to how rapidly the function changes from moment to moment
5. Storage in and retrieval from memory on demand
6. Making copies

**Remark 4** I use the term "differentiation" in a general sense here. If time is envisaged as continuous then it the usual operation of taking the derivative. In case time consists of a countable set of moments then it involves first and higher order differences. ♠

These can be regarded as a set of elementary operations on functions, with the understanding that they have their physical realizations as systems. Thus, if sounds $s_1$ and $s_2$ are mapped by a transducer into voltages $V_1$ and $V_2$, physically mixing $s_1$ and $s_2$ carries over into summing the voltages $V_1$ and $V_2$, a task that is carried out by a circuitry. Likewise, making $s_1$ louder by a certain amount carries over into magnifying or amplifying the voltage $V_1$ by a matching factor or gain, again through a circuitry. At the same time, the voltages encoding the sounds in turn map, through a process of measurement (recording values at every time instant with the help of meters or digital oscilloscopes), into mathematical functions $f_1$, $f_2$ and so on.

The transducers, together with the measuring system, enable us to make a transition from the physical world of aural perceptions and sounds to the world of numbers and functions. This transition is of crucial importance because it enables us to do things to sounds that we cannot do with our natural means (like storing sounds for listening to them later, or for sending them over to distant places). By encoding sounds into voltages, we are in a position to do a lot of processing that is not naturally possible, e.g., adding and subtracting, filtering and smoothing, through clever circuitry, and then converting back to sounds through complementary transducers.[7]

[7] This is where technology comes in. Rather, it is this possibility of increasing our power through synthetic means that is the essence of technology.

Given the elementary operations and their physical realizations, we can then ask ourselves what sort of derived operations we can build from them. The class of all such derived operations becomes then an interesting subject of study for the design of processing systems. The elementary operations are the building blocks from which other more complicated derived systems are built so to say.

In summary, the process of modeling audio signals, in the sense just explained, involves four things: (*a*) an empirical structure, $\mathcal{S}$, consisting of the set of sounds, together with all the elementary operations, (*b*) a formal structure, $\mathcal{F}$, consisting of a set of mathematical functions of a certain kind, together with mathematical equivalents of the elementary operations, (*c*) an encoding procedure $\phi : \mathcal{S} \rightarrow \mathcal{F}$ for mapping propositions in $\mathcal{S}$ into corresponding propositions in $\mathcal{F}$ and, (*d*) a decoding procedure $\psi : \mathcal{F} \rightarrow \mathcal{S}$ for going from the $\mathcal{F}$ back to $\mathcal{S}$. The formal structure $\mathcal{F}$ serves as a *model* of the empirical structure $\mathcal{S}$.[8]

In order to further illustrate the idea of modeling, let us examine the way we perceive of time and space in the context of signals and systems.

### 3.2.3 Time and Space

Both in our conception of time and space, it is considered natural to turn to the real line as a model, treating time as consisting of *moments* and space as consisting of *locations*. Yet, if a distinction is to be made between what we learn through our sense perceptions and what we conceptualize through contemplation alone, the practice of identifying time or space with the real line ceases to be that natural. There is a need here to show that the empirical structure of time, or of space, has the right correspondence with the formal structure of numbers.

Let us take up the case of time as consisting of discrete moments. Let $T$ denote the set of such moments (or instants), with one of them $t_0$ chosen as the starting moment. Furthermore, let $S$ denote a unary operation on $T$, with $St$ denoting the moment "immediately after" the moment $t$. Note that our empirical temporal experience supplies the interpretation for the expression 'immediately after', as also for the notion of an initial time instant. In this sense, no arithmetic like operation is yet involved. Again, based on empirical evidence, we can say that the structure $(T, t_0, S)$ meets the following conditions in which "=" stands for the relation "same as":

N1 For any $t \in T$, $St \neq t_0$.

N2 For all $r, t \in T$, if $r \neq t$ then $Sr \neq St$.

[8] Notice that, following Rosen [15], I have called $\phi$ and $\psi$ encoding and decoding procedures, respectively, rather than morphisms; morphisms are maps from one formal structure to another, whereas for these two, there is an empirical structure at one end, and hence the need for this distinction.

N3 If $U$ is any subset of $T$ such that (i) $t_0$ is in $U$ and (ii) whenever $t$ is in $U$ then so is $St$, then $U = T$.

N1 refers to the fact that, following the "flow of time", we cannot revert to the origin. N2 asserts that there is no branching in time, i.e., for any time instant there is only one instant immediately after it. Finally, N3 corresponds to the fact that our temporal experience necessarily involves all points in time and that, in following them one after another, using N1 and N2, none are left out.

In abstract terms, N1–N3 constitute a version of the Peano postulates, N3 being the axiom of mathematical induction. Now, all structures that satisfy the Peano postulates are isomorphic. Furthermore, for any specific such structure, there is a unique binary operation possessing the natural attributes of addition.[9] *We can thus say for the index set $T$ of time instants, which satisfies the Peano postulates, that it can be replaced for all practical purposes by the system of nonnegative integers under addition.* The idea of treating $\mathbb{N}$ as a model of the index set $T$ of time instants then stands justified.

Our arguments for the index set of discrete time instants hold equally for 1-D space of our spatial perception. There is, however, one basic difference between time and space. In our temporal experience, the notion of order amongst time instants, based upon the distinction between past, present and future, is an empirical necessity. In our spatial experience, however, such ordering as results from assuming a particular direction is an optional attribute. So in modeling 1-D space by $\mathbb{N}$ we are imposing upon it restrictions that are not empirically essential.[10]

We are so used to the idea of treating time and space in terms of numbers that we do not normally think that this calls for a justification. In modeling signals in terms of vector spaces, we are likewise blinkered by familiarity from looking at alternative algebraic models.[11] As already pointed out in Chapter 1 in the context of LTI systems, signal spaces can also be modeled in terms of algebraic structures known as rings, integral domains, and algebras. For a detailed discussion of this point of view and its unifying role and relevance, the reader is referred to the works of Püschel and others [11, 12].

[9] For a detailed discussion of these points, see Henkin [4].

[10] There are several basic issues related to modeling of time and space that we leave out of our discussions here. A glimpse of the complexities involved at the deeper end can be had from van Benthem [2]. A historically significant line of thinking about the modeling of continuous time is also to be found in the works of Hamilton, as documented in Neils and Otte [9]. As they point out, Hamilton took the view that *algebra is to time what geometry is to space*. Pursuing this viewpoint, Hamilton came up with a mathematical formulation (a formalization) of the notion of time that is apparently analogous to what Dedekind later obtained for the real number system.

[11] I am tempted at this point to borrow, rather cheekily, half a metaphor from Rawls [13, p. 118]. As he says, while deciding upon procedures for identifying just principles, we ought to operate from behind a *veil of ignorance*. We could in a similar spirit say that many deep questions about commonplace concepts are usually hidden from us by what we might call a *veil of familiarity*.

### 3.2.4 Models in General

The foregoing discussions about time and space connect strongly with the idea of modeling in general, as interpreted in Rosen [15, 16].

Rosen's work rests on the premise that the world of our perception separates into two domains, one of *natural systems*, and another of *formal systems*. These two domains are connected by what he chooses to call *encoding* and *decoding* procedures. His natural systems, conceived as a perceptual 'snapshot' of the real world, subsume what we have called here empirical structures, and formal systems correspond to formal structures. His treatment of the subject of modeling makes profound reading, addressed as it is to some very deep questions about the role and limits of models. Can living organisms be modeled as machines? What exactly is a machine? And above all, what is life? The tenor of his thoughts should be of direct interest to those concerned with the contemporary activities in the areas of signal and data (audio and video) representation and compression.[12]

Our discussions of modeling and measurement will not be really complete without taking note of the important role of metaphors in the development, conceptualization, and communication of mathematical ideas. It might seem rather far-fetched to suggest that metaphors have such a role. After all, to go by the common usage, a metaphor is a figure of speech, one that is used for literary adornment. What place could it have in mathematics, or in science, or in modeling and measurement, where precision and logic rule the roost? That it indeed has a place, and a very crucial one at that, is what I wish to very briefly highlight in the concluding section of this chapter.

## 3.3 Metaphors

> *What are we,*
> *But blobs on the canvas of Time!*

Metaphors pervade our language and thought, even in a technical discourse.[13] It is not uncommon, for instance, for text books on circuit analysis to say that a voltage source connected to a 2–port *sees* a driving point impedance of a certain value, say $50\,\Omega$. Here is a clear case of a metaphor being used in a routine manner, supposedly to promote understanding. Unlike a technical term explicitly defined in advance, here is a term—*seeing*, that is used without first defining it and certainly way outside the scope of its literal meaning, assuming that the underlying analogy between a human and a voltage source, and likewise between a tree out there and

[12] The book by Barrow [1] is also worth looking at in this connection. Yaglom [21] gives yet another noteworthy account of the multi-threaded connections between mathematical structures and the physical world.

[13] For a comprehensive account of the current thinking on the subject of metaphors in general, the collection of essays in Ortony [10] is a rich source.

the impedance of a 2–port, is self-evident. What is more, it is assumed that this helps comprehension as a figure of speech. Well, does it? To most, it is just another one of those common words that one glosses over. To them, it carries no special significance in the general flow of the main idea; they do not even notice it. But for the few who do notice it, it brings up a point to ponder on the side, and to realize that the use is not in a literal sense of the word, but rather as a metaphorical shortcut for an otherwise lengthy and 'dry' description of the relationship between a source and a load.

Mathematical, scientific, and technical literature is strewn with such metaphorical uses. Their role is primarily taken to be that of enlivening the process of communication and learning, without any formal implications for the subject matter under discussion.

But then there is more to metaphors than just that. Indeed, as cognitive scientists now tell us, metaphors are the linkages through which we develop our understanding of the real world, and of mathematics. A metaphor, or rather *a conceptual metaphor* as they choose to call it, is in this respect an attribute transferring mapping from one conceptual domain, *the source domain*, to another, *the target domain*.[14]

As very eloquently argued by Lakoff and Núñez [6], abstract mathematical ideas are also conceptualized and comprehended through chains and layers of such conceptual metaphors, with the initial source domains consisting of parts of our concrete bodily experiences. To support their arguments, they take up a wide range of mathematical concepts and bring to the fore their metaphorical moorings. From arithmetic to algebra, from basic set theory to the theory of hypersets, from the notions of limits, continuity and convergence to infinitesimals (nonstandard analysis), they all come under their scanner. While experts may debate the validity of their findings, Lakoff and Núñez undoubtedly open up radically new avenues of understanding for those concerned with learning and applying advanced mathematical concepts.

[14] Notice the structural similarity with the representational theory of measurement and modeling.

# References

1. John D. Barrow. *Impossibility: The Limits of Science and the Science of Limits*. Oxford University Press, New York, 1998.
2. J.F.A.K Van Benthem. *The Logic of Time*. D. Reidel, London, 1983.
3. Zoltan Domotor and Vadim Batitsky. The analytic versus representational theory of measurement: a philosophy of science perspective. *Measurement Science Review*, 8(6):129–146, 2008.
4. L. Henkin. On mathematical induction. *Amer. Math. Month.*, 67(4):823–843, 1960.
5. David H. Krantz, R. Duncan Luce, Patrick Suppes, and Amos Tversky. *Foundations of Measurement*. Academic, London, 1971.
6. George Lakoff and Rafael E. Núñez. *Where Mathematics Comes From*. Basic Books, New York, 2000.
7. Louis Narens. *Abstract Measurement Theory*. MIT Press, Cambridge, MA, 1985.
8. Louis Narens. *Introduction to the Theories of Measurement and Meaningfulness and the Use of Symmetry in Science*. Lawrence Erlbaum, London, 2007.
9. H. Neils and M. Otte. Origins of the program of arithmetization of mathematics. In H. Mehrtens, editor, *Social History of Nineteenth Century Mathematics*, pages 30–32. Birkhäuser, Boston, 1981.
10. Andrew Ortony. *Metaphor and Thought*. Cambridge University Press, Cambridge, 1993.
11. Markus Püschel and José M.F. Moura. The algebraic structure in signal processing: Time and space. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, 2006.
12. Markus Püschel and José M.F. Moura. Algebraic signal processing theory: foundations and 1-d time. *IEEE Transactions on Signal Processing*, 56(8):3572–3585, 2008.

13. John Rawls. *A Theory of Justice*. Oxford University Press, New York, 1999.

14. Fred S. Roberts. *Measurement Theory with Applications to Decision Making, Utility and the Social Sciences*. Addison Wesley, Reading, MA, 1979.

15. Robert Rosen. *Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations*. Pergamon, London, 1985.

16. Robert Rosen. *Life Itself*. Columbia University Press, New York, 1991.

17. D. Scott and P. Suppes. Foundational aspects of theories of measurement. *Jr. Sym. Logic*, pages 113–128, June 1958.

18. S.S. Stevens. On the theory of scales of measurement. *Science, New Series*, 103(2684):677–680, 1946.

19. P. Suppes and J.L. Zinnes. Basic measurement theory. In R.D. Luce, R.R. Bush, and E. Galanter, editors, *Handbook of Mathematical Psychology, vol 1*, pages 1–76. Wiley, New York, 1963.

20. Patrick Suppes. A comparison of the meaning and uses of models in mathematics and the empirical sciences. *Synthèse*, 12:287–301, 1960.

21. I.M. Yaglom. *Mathematical Structures and Mathematical Modelling*. Gordon & Breach, New York, 1986.

# Chapter 4

# Symmetries, Automorphisms and Groups

## 4.1 Introduction

*What beauty is to a young*
*Algebraist beholder*
*Automorphisms and symmetries are to*
*him as he grows older*
*But to an electrical engineer? Oh well!*
*Both are just like a boulder*
*Suffice for him a chip, a bus, and a man-*
*ual in a folder.*

*Or so it is, till he chances to catch amidst*
*his bustle and toil*
*Those whispered melodies of Hilbert and*
*Weyl*
*That gently float from just over his*
*shoulder.*

In this chapter we see how group theory enters in an explicit manner in the study of symmetry, and in what way the connections between group theory and symmetry are relevant to the study of signals and systems. The ideas discussed here will serve as a background for the study of modern harmonic analysis on groups, and of its applications in signal representation and processing.

V.P. Sinha, *Symmetries and Groups in Signal Processing*, Signals and Communication Technology,
DOI 10.1007/978-90-481-9434-6_4, © Springer Science+Business Media B.V. 2010

By symmetry, one commonly understands physical or geometric symmetry, the kind in which physical objects, geometrical figures or patterns, when subjected to certain reorientations, are found to be indistinguishable from what they were originally. A cube is symmetric in this sense, and so is an equilateral triangle. So also is a resistive 2–port whose ports can be interchanged without affecting the currents and voltages of the rest of the network to which the ports are connected. There are then similar symmetries of pictures, images, and multi–dimensional signals, both in the sample and spectral domains.[1]

Significantly, it is as much meaningful, and in essentially the same sense, to talk of symmetries in temporal events. The basic idea remains the same; only the context and interpretation changes. Thus, a periodic signal has the symmetry that it is indistinguishable from its translates resulting from shifting it in time by integral multiples of its time period. A more subtle instance of such symmetry is the one mentioned in Section 1.6—that of time– or shift–invariance of systems. In this case, it is the input–output relations that remain unchanged under time translations or shifts.

In the study of physical problems, presence of such symmetries invariably leads to a great deal of simplification, both in their visualization as well as in their mathematical encoding and analysis. In this respect, a unified mathematical framework for dealing with symmetries and their implications is provided by the theory of groups and their representations.

In the sciences, group theoretic techniques of exploiting symmetries have had a place of central importance since the early twenties.[2] In electrical engineering, however, a commensurate recognition for them has been rather slow in coming. Symmetries have no doubt long been exploited to simplify analysis in many results in circuit analysis.[3] But the links with group theory did not begin to attract general attention until the sixties, even though there were important pointers to their

[1] In the theory of 2–D and multidimensional digital filters, these symmetries are put to use in simplifying design methods. See Antoniou [12].

[2] These techniques, in the form that they are in use today, are largely the outcome of intense and closely knit activities in mathematics and mathematical physics begun in the twenties. See the collection of papers in MacKey [13] for a very illuminating historical account of their development and role. On the general subject of symmetry, see the classic by Weyl [19].

[3] The method of symmetrical components of power systems, the Bartlett's bisection theorem for 2–port networks, and the method of characterizing differential amplifiers in terms of common and differential mode gains are some of the well known examples of this. The notion of characteristic or image impedance of a 2–port, which is central to the classical theory of image parameter filters, is also another instance of the use of symmetry.

significance in the early literature, some of them in network theory,[4] and others in the area of waveguide design.[5]

The picture has changed radically over the last thirty years, and these links now play a vital role on many fronts.[6] Well, what really *are* these links? Let us see.

## 4.2 Symmetries and Automorphisms

Observe, to begin with, that the notion of symmetry of an object, be it a physical one, a pattern or an image, or a temporal event, rests essentially on our perception of it as consisting of two things: one, a set of constituent members (the points that form the surface of a cube, the vertices of a triangle, the voltages and currents at the ports of a network), and two, a relationship specified for them (lengths of the sides and the angles between them for the cube or the triangle, the voltage current relationships at the ports for the 2–port). A symmetry of the object then consists of the fact that, if some or all of the constituents are interchanged in a one-to-one fashion, each one acting for the one it replaces in the specified relationship, then the relationship remains unaltered.

More abstractly, an object is for us in this context a relational structure. It is good enough for the present purposes to assume that there is just one relation involved, so that we have as our starting point, a structure $(S, \mathcal{R})$, where $S$ is a set, representing the constituent members, and $\mathcal{R}$ is a relation, relating $S$ either just to itself or to some other given sets in addition.

Take the case of an equilateral triangle for instance. Let its vertices be labelled $a$, $b$ and $c$, and let its sides be of length $\lambda$, as shown in Figure 4.1a.

[4] A significant early contribution in this class, and perhaps amongst the first ones, to point out the role of group theory in the study of invariance and symmetry in circuit problems, is that of Howitt [9]. Although this paper drew attention to several significant avenues of work, including one that points to the state space approach, it does not appear to have received the kind of attention it deserved at the time. Note, in particular, the following comments in his concluding remarks: "What in electric circuit theory corresponds to the principal or normal coordinates in dynamic theory? . . . in the study of an electrical network . . . , one continually encounters many seemingly unrelated branches of mathematics, such as (1) continued fractions, . . . , (5) group theory, (6) Fourier series and transforms . . . etc.. It seems almost as if something were there, inarticulately trying to make itself understood. But perhaps it must await a modern Euler."

Another interesting early paper to independently make a strong case for the use of group theory in the analysis and classification of networks was that of Gaertner [6]. Through his treatment of 2–port networks and their cascade connections, he put forth the point that the group theoretic approach "justifies itself not only by the results of its derivations but also by adding to the understanding and insight and thus providing ideas and suggestions on how to handle certain problems." Also see in this connection the classical work by Brillouin [1, Chapter 9].

[5] Appearing in the fifties, these introduced group representation theory in the design of symmetrical waveguides. See Kerns [10] and Pannenborg [14].

[6] As a sample of the kinds of activities that have been going on in recent years, see Lenz [11] and Foote and Mirchandani [4, 5].

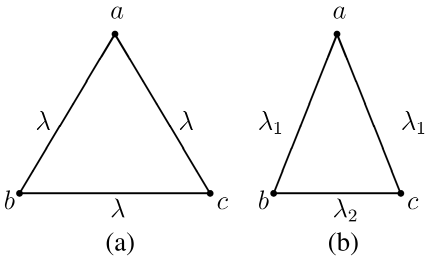

Figure 4.1 (**a**) An equilateral triangle; (**b**) an isosceles triangle

Looking upon it as a figure defined by the three points in the plane serving as vertices, together with the lines joining them, its symmetries consist of the attribute that in whatever manner we interchange the labels of the vertices, for any two vertices with the same labels before and after the interchange, the distance between them remains the same.[7]

To interpret this algebraically, let us concentrate on the set $S = \{a, b, c\}$ of the vertices of the triangle. A relabelling of the vertices is characterized by a one-to-one onto, i.e., invertible, mapping from $S$ to $S$. There are six such mappings, also called permutations, on a set of three elements. For the sake of convenience, let us call these mappings, $f_1, f_2, \ldots, f_6$, as listed below in Table 4.1.

| $f_i$ | $f_i(a)$ | $f_i(b)$ | $f_i(c)$ |
|---|---|---|---|
| $f_1$ | $a$ | $b$ | $c$ |
| $f_2$ | $b$ | $c$ | $a$ |
| $f_3$ | $c$ | $a$ | $b$ |
| $f_4$ | $a$ | $c$ | $b$ |
| $f_5$ | $c$ | $b$ | $a$ |
| $f_6$ | $b$ | $a$ | $c$ |

Table 4.1 One–one mapping from the set $S = \{a, b, c\}$ to itself

In addition to the vertices, we need also to specify the lengths of lines between them to completely specify the triangle. For an equilateral triangle, these lengths are the same, in our case $\lambda$. We may then look upon the triangle as a ternary relation $\mathcal{R}$, a subset of $S \times S \times \mathbb{R}^+$:

$$\mathcal{R} = \{(x, y, \lambda) | x, y \in S \quad \text{and} \quad x \neq y\}$$

[7] This is equivalent to saying that if we rotate it around its centre by multiples of $2\pi/3$ radians, or reflect it along the perpendicular from a vertex to the opposite side, it coincides with itself.

The relational structure $\mathbb{S} = (S, \mathcal{R})$, consisting of the set $S$ together with the relation $\mathcal{R}$, then allows us to algebraically describe the symmetries of the equilateral triangle.[8]

Note that any of the maps $f_i$ is such that the interchanges introduced by it amongst the members of $S$ do not affect the relation $\mathcal{R}$. Treating $\mathcal{R}$ as a set, this is equivalent to saying that the set of triples resulting from the interchanges is the same as the set $\mathcal{R}$:

$$\{(f_i(x), f_i(y), \lambda) | (x, y, \lambda) \in \mathcal{R}\} = \mathcal{R}\,. \tag{4.1}$$

In order to bring it in line with the sort of algebra we want to bring in here, let us rephrase this condition into an equivalent form, consisting of two parts,

$$\{(f_i(x), f_i(y), \lambda) | (x, y, \lambda) \in \mathcal{R}\} \subseteq \mathcal{R}\,\text{, and,} \tag{4.2a}$$

$$\mathcal{R} \subseteq \{(f_i(x), f_i(y), \lambda) | (x, y, \lambda) \in \mathcal{R}\}\,. \tag{4.2b}$$

Going a step further, (4.2a) and (4.2b) may be rewritten respectively as

$$(x, y, \lambda) \in \mathcal{R} \quad \text{implies that} \quad (f_i(x), f_i(y), \lambda) \in \mathcal{R}\,\text{, and,} \tag{4.3a}$$

$$(x, y, \lambda) \in \mathcal{R} \quad \text{implies that} \quad (f_i^{-1}(x), f_i^{-1}(y), \lambda) \in \mathcal{R}\,. \tag{4.3b}$$

We shall say that the one-to-one onto function $f_i$ is a *structure-preserving* map for the structure $\mathbb{S}$ if it satisfies the condition (4.3a). By the same token, $f_i^{-1}$ is also structure-preserving for $\mathbb{S}$ if it satisfies the condition (4.3b). Furthermore, we shall say that $f_i$ is an *automorphism* of $\mathbb{S}$ if $f_i$ as well as $f_i^{-1}$ are both structure-preserving, i.e., if conditions (4.3a) and (4.3b) are both satisfied.

As we shall see in the next section, for the kinds of structures in which we are interested here in connection with symmetry, it is enough to stipulate that $f_i$ be a structure-preserving map; this implies that $f_i^{-1}$ is also one.

A symmetry of the equilateral triangle may then be said to consist of the fact that there is a one-to-one onto map from the ground set $S$ to $S$ such that it is an automorphism of the structure $\mathcal{S} = (S, \mathcal{R})$. Table 4.2 shows the correspondence implicit in this point of view between the primitives in terms of which we describe symmetries of the triangle at the intuitive level, and their algebraic counterparts.

Consider as another geometrical example, the symmetries of the isosceles triangle shown in Figure 4.1b. Proceeding as in the case of the equilateral triangle, the pertinent relation here is

$$\mathcal{L} = \{(a, b, \lambda_1), (a, c, \lambda_1), (b, a, \lambda_1), (c, a, \lambda_1), (b, c, \lambda_2), (c, b, \lambda_2)\}.$$

[8] Recall that for a relational structure, the most commonly encountered situation is of the type in which the pertinent relations are given on the ground set or powers of it. In more general cases, the relations may be over not just over the ground set but on several other external sets in addition to it. For us here, $\mathbb{R}^+$ is such an external set for the relation $\mathcal{R}$. Admittedly, structures with external sets can not be clubbed with those without external sets. For the present purposes, however, we shall overlook this fact.

| Intuitive notions | Algebraic counterparts |
|---|---|
| Labelled vertices | A Set $S$ |
| Distances between the vertices | A relation $\mathcal{R}$ over $S$ |
| Equilateral triangle | Relational structure $\mathbb{S} = (S, \mathcal{R})$ |
| Symmetry operations on the triangle | Automorphisms of $\mathbb{S}$ |

Table 4.2 Equilateral triangle; symmetry operations and automorphisms

Characterizing the symmetry operations in this case, there are only two automorphisms of the structure $\mathbb{S} = (S, \mathcal{L})$. These are $f_1$ and $f_2$ as listed in Table 4.1.

The general idea that these two geometrical examples point to is that *the intuitive notion of symmetry can be given an algebraic formulation in terms of automorphisms of relational structures.* We can in fact treat the notion of symmetry as synonymous with that of automorphisms of relational structures.

Let us now see what shape this idea takes in the case of networks and systems. Consider first a 2–port passive resistive network shown in Figure 4.2 under *dc* excitations, and suppose that it has internal physical symmetry about the axis $XY$. This symmetry implies that with external connections at the ports remaining the same, the port voltages and currents do not change if the 2–port is flipped around the $XY$ axis. We now want to restate this fact in relational terms.

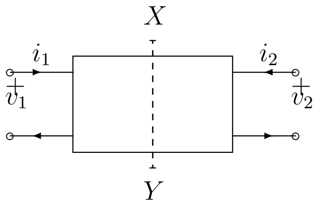

Figure 4.2 A resistive 2–port network

Let $\mathbf{x}$ denote the vector $[x_1 \;\; x_2]'$ of port variables, which may either be voltages or currents; since we are considering the *dc* case, it is a $2 \times 1$ vector of reals (a member of $\mathbb{R}^2$). A 2–port is typically characterized by a function giving two of the port variable in terms of the remaining two (e.g., $\mathbf{v} = f(\mathbf{i})$). It may alternatively be characterized as a binary relation $\mathcal{N}$ consisting of all admissible ordered pairs $(\mathbf{x}, \mathbf{y})$. We assume, to be specific, that the first member $\mathbf{x}$ is the port current vector and $\mathbf{y}$ is the corresponding voltage vector.

Now, the operation of flipping the 2–port around the axis $XY$ is the same as interchanging the external connections at the two ports. Let $\phi_1 : \mathbb{R}^2 \to \mathbb{R}^2$ be the one-to-one onto map giving $\phi_1([x_1\ x_2]') = [x_2\ x_1]'$, and $\phi_0$ be the identity map giving $\phi_0([x_1\ x_2]') = [x_1\ x_2]'$. *The symmetry condition stipulated for the 2–port then amounts to saying that $\phi_0$ and $\phi_1$ are automorphisms of the relational structure $(\mathbb{R}^2, \mathcal{N})$ i.e., for $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^2$,*

$$\mathbf{x}\,\mathcal{N}\mathbf{y} \quad \text{if and only if} \quad \phi_i(\mathbf{x})\,\mathcal{N}\phi_i(\mathbf{y}) \quad i = 0, 1. \tag{4.4}$$

If we revert to the familiar functional description of the 2–port, $\mathbf{v} = f(\mathbf{i})$, then the symmetry condition becomes

$$\phi_i(\mathbf{v}) = \phi_i(f(\mathbf{i})) = f(\phi_i(\mathbf{i})). \tag{4.5}$$

For a linear 2–port, the function $f$ takes a matrix form in terms of its open–circuit parameters $r_{11}$, $r_{12}$, $r_{21}$ and $r_{22}$:

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} \tag{4.6}$$

or, in short,

$$\mathbf{v} = \mathbf{R}\mathbf{i}.$$

Furthermore, the functions $\phi_0$ and $\phi_1$ are better treated in this context as permutation matrices (to be called $\mathbf{P}_0$ and $\mathbf{P}_1$ respectively), i.e.,

$$\phi_0(\mathbf{x}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

and,

$$\phi_1(\mathbf{x}) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix}$$

The symmetry condition (4.5) then becomes that, in addition to (4.6), the matrix $\mathbf{R}$ should satisfy the condition

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} \tag{4.7}$$

Premultiplying (4.7) by $\mathbf{P}_1^{-1}$ and noting that (4.6) and (4.7) are to hold for all choices of $\mathbf{i}$, we conclude that the matrix $\mathbf{R}$ commutes with the two permutation matrices:

$$\mathbf{P}_i\mathbf{R} = \mathbf{R}\mathbf{P}_i, i = 0, 1 \tag{4.8}$$

or, in detail,

$$\begin{bmatrix} r_{21} & r_{22} \\ r_{11} & r_{12} \end{bmatrix} = \begin{bmatrix} r_{12} & r_{11} \\ r_{22} & r_{21} \end{bmatrix}$$

i.e., $r_{11} = r_{22}$ and $r_{12} = r_{21}$.[9] This is what we usually understand by symmetry of a linear 2–port.

Clearly, we can treat symmetries in this way for any number of ports, except that there will be an increasing number of possible symmetries and corresponding automorphisms that we will need to deal with. For a grounded linear 3–port (Figure 4.3) for example, there are six possible symmetries in all.

If all the six symmetries are present in the 3–port then (4.8) holds for the six $3 \times 3$ permutation matrices, $\mathbf{P_0} \cdots \mathbf{P_5}$ listed in Table 4.3, corresponding to the six possible permutations of the three ports.

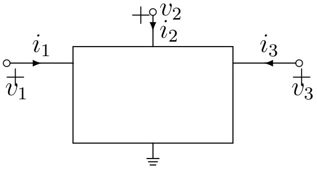

Figure 4.3 A resistive grounded 3–port network

$$\begin{array}{cccccc}
\mathbf{P_0} & \mathbf{P_1} & \mathbf{P_2} & \mathbf{P_3} & \mathbf{P_4} & \mathbf{P_5} \\
\begin{bmatrix} 1\,0\,0 \\ 0\,1\,0 \\ 0\,0\,1 \end{bmatrix} &
\begin{bmatrix} 0\,1\,0 \\ 0\,0\,1 \\ 0\,0\,1 \end{bmatrix} &
\begin{bmatrix} 0\,0\,1 \\ 1\,0\,0 \\ 0\,1\,0 \end{bmatrix} &
\begin{bmatrix} 1\,0\,0 \\ 0\,0\,1 \\ 0\,1\,0 \end{bmatrix} &
\begin{bmatrix} 0\,0\,1 \\ 0\,1\,0 \\ 1\,0\,0 \end{bmatrix} &
\begin{bmatrix} 0\,1\,0 \\ 1\,0\,0 \\ 0\,0\,1 \end{bmatrix}
\end{array}$$

Table 4.3 The matrices of automorphisms characterizing all possible symmetries of a grounded 3–port

Linked with six possible permutations of three objects as they are, these symmetries are essentially the same as those that we discussed for the equilateral triangle. There is, however, a difference between the two cases. While for an equilateral triangle all these symmetries are present by definition, for a 3–port there are various partial options too. These options can be identified in terms of certain subsets of the six matrices listed in Table 4.3. To start with, there is the subset containing only $\mathbf{P}_0$, which implies no symmetries except the identity. On the other extreme, we have the entire set $\{\mathbf{P}_0, \ldots, \mathbf{P}_5\}$ for the case of the 3–port having all possible symmetries. Then there is the subset $\{\mathbf{P}_0, \mathbf{P}_3\}$, which corresponds to interchanging ports 2 and 3, and likewise subset $\{\mathbf{P}_0, \mathbf{P}_4\}$, and $\{\mathbf{P}_0, \mathbf{P}_5\}$. There is also the subset $\{\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2\}$, which corresponds to cyclic interchanges of the ports. Why

[9] Incidentally, the second condition, $r_{12} = r_{21}$, is that of reciprocity, which means that symmetry in the linear 2–port implies reciprocity.

only these subsets and not any other, like the subset $\{\mathbf{P}_3, \mathbf{P}_4\}$ for instance? Before we turn to this question, which has to do with the structure of symmetry operations under composition (successive application), let us consider as one more example of symmetry, the notion of time–invariance of systems.

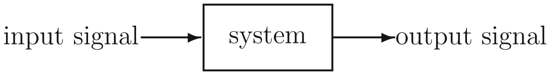

Figure 4.4 A time–invariant system $H$

Let $X$ denote the set of all continuous–time signals (real–valued functions on $-\infty \leq t \leq \infty$), and let $D_\tau : X \to X$ denote the operator that produces a delay of time $\tau$, i.e., for $x \in X$,

$$D_\tau x(t) = x(t - \tau).$$

Further, let $H$ be a time–invariant system that acts on an input $x$ to produce output $Hx$ (Figure 4.4). Then, as we have discussed earlier, time–invariance of $H$ means that

$$D_\tau H(x) = HD_\tau(x) \quad \text{for every} \quad x \in X \quad \text{and every} \quad \tau,$$

i.e., to put it in words, $H$ and $D_\tau$ commute.

Now, instead of looking at the system as a map on $X$, we may regard it as a binary relation $\mathcal{H}$ on $X$, consisting of the set of admissible input–output ordered pairs:

$$\mathcal{H} = \{(x, Hx) | x \in X\},$$

so that saying $x\mathcal{H}y$ is saying that $y$ is the output for input x. Time–invariance of $H$ then means that

$$\text{if} \quad x\mathcal{H}y \quad \text{then} \quad (D_\tau x)\mathcal{H}(D_\tau y) \quad \text{for every } \tau.$$

In other words, since the operators $D_\tau$ are one-to-one and onto maps on $X$, they are automorphisms of the relational structure $\mathbb{H} = (X, \mathcal{H})$.

In summary, the following three statements are equivalent.

(i) $H$ is time–invariant.

(ii) $H$ commutes with every $D_\tau$, i.e., $HD_\tau = D_\tau H$.

(iii) Every $D_\tau$ is an automorphism of the relational structure $\mathbb{H}$.

What we have seen for continuous–time systems is equally meaningful for other classes of systems, and we can say in general that *time–invariance or shift–invariance is in essence a symmetry property of systems that can be studied in terms of automorphisms of relational structures.*

## 4.3 Groups of Automorphisms

The way is now open for us to algebraically examine questions about symmetries and symmetry operations.

Clearly, when we apply one symmetry operation and then follow it up by another, the two together constitute, in terms of their net effect, another symmetry operation. This corresponds to the algebraic fact that two automorphisms of a structure produce under composition another automorphism of that structure.

Let us now examine the situation in detail for a relational structure $\mathbb{S} = (S, \mathcal{R})$, where $\mathcal{R}$ is a binary relation on the ground set $S$. Let $\mathbf{A}$ be the set of all one-to-one onto maps on $S$. Clearly, $\mathbf{A}$ is a group under composition. Consider now the subset $\mathbf{G}$ of $\mathbf{A}$, consisting of all automorphisms of $\mathbb{S}$.

To start with, if $f$ and $g$ are two of its automorphisms then $x\mathcal{R}y$ implies that $g(x)\mathcal{R}g(y)$, and this in turn implies that $f(g(x))\mathcal{R}f(g(y))$, i.e., $fg$ is structure-preserving. It also at the same time follows that $(fg)^{-1}$ too is structure-preserving. Moreover, since $f$ and $g$ are one-to-one onto on $S$, $h = fg$ is also one-to-one onto. Thus $h$ is an automorphism of $\mathbb{S}$. In other words, $\mathbf{G}$*, the set of all automorphisms of the structure* $\mathbb{S}$*, is closed under composition.*

Further, the identity map $e : S \to S$ ($e(x) = x$ for any $x \in S$) is evidently an automorphism of $\mathbb{S}$, and for any other automorphism $f$, $fe = ef = f$. Thus, $\mathbf{G}$ *contains an identity element for the operation of composition.*

What more can we say about $\mathbf{G}$? Well, there is the point about inverses that was mentioned on page 85: for certain kinds of relational structures, if $f$ is structure-preserving then so is $f^{-1}$. If the ground set is finite then the structure $\mathbb{S}$ is of this kind. An interesting way to see this is as follows.

Let $f$ be a one-to-one (and consequently, since $S$ is finite, also onto) map on $S$ that preserves the relation $\mathcal{R}$ i.e., $(x, y) \in \mathcal{R} \Rightarrow (f(x), f(y)) \in \mathcal{R}$. That is, the set $\mathcal{R}' = \{(f(x), f(y)) | \text{ for every}(x, y) \in \mathcal{R}\}$ is contained in the set $\mathcal{R}$. Let $f'$ be the map from $\mathcal{R}$ to $\mathcal{R}$ with values $f'((x, y)) = (f(x), f(y))$. Now, since $\mathcal{R}$ is finite and $f'$ is one-to-one, it is also onto. Then $\mathcal{R}'$, the image of $f'$, is the whole of $\mathcal{R}$. It then follows that

$$(f(x), f(y)) \in \mathcal{R} \Rightarrow (x, y) \in \mathcal{R},$$

or equivalently,

$$(u, v) \in \mathcal{R} \Rightarrow (f^{-1}(u), f^{-1}(v)) \in \mathcal{R}.$$

Thus, if $f$ is structure-preserving for a relational structure $\mathbb{S} = (S, \mathcal{R})$, where the ground set $S$ is finite, then $f^{-1}$ is also structure-preserving for $\mathbb{S}$.

Now suppose that the ground set is not finite but the structure-preserving maps of the structure $\mathbb{S}$ are finite in number. In this case too, if $f$ is structure-preserving then so is $f^{-1}$. To show this, we use a more crisp argument this time, one that holds for the previous case as well. We have already seen that $\mathbf{G}$, the set of all structure-preserving maps of $\mathbb{S}$ is closed under composition. Further, it is a subset of the group

$\mathbf{A}$ of all one-to-one onto maps on the ground set of $\mathbb{S}$. Being finite by hypothesis, $\mathbf{G}$ is then a subgroup of $\mathbf{A}$.[10] Thus if $f$ is in $\mathbf{G}$ then so is $f^{-1}$.

Such is not the case in general if the ground set is not finite. This is best explained with the help of a simple counter-example. Consider the structure $\mathbb{S} = (\mathbb{Z}, \mathcal{R})$, where $\mathcal{R} = \mathbb{Z}^+ \times \mathbb{Z}^+$, $\mathbb{Z}$ is the set of all integers, and $\mathbb{Z}^+$ is the set of nonnegative integers. Let $f : \mathbb{Z} \to \mathbb{Z}$ be the one-to-one onto function with values $f(n) = n + 1$ for $n \in \mathbb{Z}$. Then, since $(m, n) \in \mathcal{R}$ implies that $(f(m), f(n)) \in \mathcal{R}$, $f$ is an automorphism of $\mathbb{S}$. But for $(0, 0) \in \mathcal{R}$, $(f^{-1}(0), f^{-1}(0)) = (-1, -1)$ is not in $\mathcal{R}$. Thus $f^{-1}$ is not an automorphism of $\mathbb{S}$ in this case.

The situation is, however, not so disheartening on the whole. If the given relation is in fact a function or a binary operation, then there is no such problem.

In the case of a structure $\mathbb{S} = (S, g)$, where $g$ is a function (a unary operation) on $S$, the structure-preserving condition for a one-to-one and onto function $f$ on $S$ is that for any $x \in S$, $f(g(x)) = g(f(x))$, i.e., $f$ and $g$ commute. Then, for any $x \in S$,

$$g(x) = g(ff^{-1}(x)) = (gf)f^{-1}(x) = (fg)f^{-1}(x)$$

or,

$$f^{-1}g(x) = gf^{-1}(x).$$

Thus $f^{-1}$ is also structure-preserving.

Let us check this for a structure $\mathbb{S} = (S, \circ)$, where $\circ$ is a binary operation on $S$. For $a, b \in S$, if $f$ is structure-preserving one-to-one onto function on $S$, i.e., $f(x \circ y) = f(x) \circ f(y)$ for any $x, y \in S$, then

$$\begin{aligned} a \circ b &= (ff^{-1}a) \circ (ff^{-1}b) \\ &= f(f^{-1}(a) \circ f^{-1}(b)), \end{aligned}$$

i.e.,

$$f^{-1}(a \circ b) = f^{-1}(a) \circ f^{-1}(b).$$

That is, $f^{-1}$ is also structure-preserving for $\mathbb{S}$.

A doubt that may arise at this point is the following. An algebraic structure is also a relational structure in the sense that an $n$-ary operation is an $(n + 1)$-ary relation. What is special about it that forces $f^{-1}$ also to be structure-preserving along with $f$? Well, it is the condition that is imposed on a relation to become an

[10] We use here the standard result that if $\mathbf{A}$ is a group and $\mathbf{G}$ a finite nonempty subset of it that is closed under the group operation then $\mathbf{G}$ is a subgroup of $\mathbf{A}$. This follows easily from considering any element $a \in \mathbf{G}$ and its powers, and showing that $\mathbf{G}$ contains $a^{-1}$ as well as $e$, the identity element of $\mathbf{A}$. Consider the powers of $a$ as a sequence of successively generated elements $a^m = a^{m-1}a$ for $m = 2, 3, \ldots$. Since $\mathbf{G}$ is closed, all these elements are in $\mathbf{G}$. But, since $\mathbf{G}$ is finite, the elements must begin to repeat after a finite number of steps (at most equal to $n$, the cardinality of $\mathbf{G}$). Let us say $a^j = a^k$ for $j > k > 0$. Then by the cancellation law of $\mathbf{A}$, $a^{j-k} = e$. Thus the identity $e$, being a positive power of $a$, is in $\mathbf{G}$. Moreover, $a^{j-k} = aa^{j-k-1} = e$ and $j - k - 1 \geq 0$, i.e., $a^{j-k-1} = a^{-1} \in \mathbf{G}$.

operation. In the arguments given above, it is hidden in the definition of a function or a binary operation. Let us bring it out in the open by re-examining the situation in purely relational terms.

Consider again the same structure, with its binary operation treated as a ternary relation $\mathcal{R}$ ($\subseteq S \times S \times S$), and a one-to-one onto map $f : S \to S$, satisfying the conditions,

1. For $x, y \in S$, there is a $z \in S$ such that $(x, y, z) \in \mathcal{R}$,
2. If $(x, y, z) \in \mathcal{R}$ and $(x, y, w) \in \mathcal{R}$ then $z = w$,
3. The function $f : S \to S$ is one-to-one onto, and,
4. If $(x, y, z) \in \mathcal{R}$ then $(f(x), f(y), f(z)) \in \mathcal{R}$.

Conditions 1 and 2 are the additional conditions on $\mathcal{R}$ to make it a binary operation, and conditions 3 and 4 constitute the structure-preserving conditions for $f$. Let us now proceed in steps, explicitly indicating the particular given condition used in each step.

1. Suppose that $(x, y, z) \in \mathcal{R}$. Then, since $x \in S$, there is a $u \in S$ such that $f^{-1}(x) = u$, as $f$ is one-to-one and onto. *condition 3*
2. Next, for any $v \in S$, there is a $w \in S$ such that $(u, v, w) \in \mathcal{R}$. *condition 1*
3. It then follows that $(f(u), f(v), f(w)) \in \mathcal{R}$, so that by substituting $f^{-1}(x)$ for $u$, $(x, f(v), f(w)) \in \mathcal{R}$. *condition 4*
4. Now, if $f(v) = y$ then, since $(x, y, z) \in \mathcal{R}$ by hypothesis, $f(w) = z$. *condition 2*
5. Then, substituting for $u$, $v$ and $w$ as in steps 1 and 4, and in view of step 2, $(f^{-1}(x), f^{-1}(y), f^{-1}(z)) \in \mathcal{R}$.

Thus, if $(x, y, z) \in \mathcal{R}$ then $(f^{-1}(x), f^{-1}(y), f^{-1}(z)) \in \mathcal{R}$, i.e., $f^{-1}$ is also structure-preserving. Observe that all the given conditions, and none other, have been used in the arguments.

**Remark 5** The foregoing discussions should bring home the point that, compared to algebraic structures, relational structures are more intricate to handle when it comes to their automorphisms. In conceptualizing symmetry in a relational setting as I have, it is important to be aware of this. For a more formal treatment of the issues involved, see Dixon and Mortimer [2, Section 9.5, pp. 290–291], and Schmidt and Ströhlein [15, Chapter 7, pp. 142–144].

♠

With these clarifications, the basic facts that we need here about automorphisms are in place. Coming back to symmetries, we shall assume from now onwards that in the situations in which we are interested, there are only a finite number of symmetries. In that case, as we have already seen, the set of all automorphism of the corresponding structure always satisfies these conditions under composition: (a) closure and associativity, (b) existence of identity, and (c) existence of inverses. We thus find that *in the context of symmetries of interest to us here, the set of all automorphisms of a relational structure constitute a group under composition.*

The question (p. 89) about matrices $\mathbf{P}_i$ that together qualify as sets of symmetry operations for a 3-port network may now be answered. What is being asked is essentially this. For a structure $\mathbb{S} = (S, \mathcal{R})$, the set $\mathbf{G}$ of all one-to-one onto maps on $S$ is a group under composition. How is the set $\mathbf{H}$ of all automorphisms of $\mathbb{S}$ related to $\mathbf{G}$? Well, $\mathbf{H}$ is a group in its own right, and is at the same time a subset of $\mathbf{G}$. That is, the set $\mathbf{H}$ of all automorphisms of $\mathbb{S}$ is a subgroup of the group $\mathbf{G}$ of all one-to-one onto maps on the ground set $S$. Thus, besides $\mathbf{G}$ itself, the only other subsets admissible on this count are its subgroups $\{\mathbf{P}_0\}$, $\{\mathbf{P}_0, \mathbf{P}_3\}$, $\{\mathbf{P}_0, \mathbf{P}_4\}$, $\{\mathbf{P}_0, \mathbf{P}_5\}$, and $\{\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2\}$.

## 4.4 Symmetries of Linear Transformations

From symmetries in general, let us now turn to the situation in which the ground set has the structure of a vector space and the symmetries of interest to us are those of linear transformations on this space. A very wide variety of physical problems in which symmetries are exploited to simplify their study, the situation is essentially of this kind. The case of time– or shift–invariant linear systems and multi–port networks that we have discussed earlier serve as typical examples.

### 4.4.1 Symmetries and Symmetry Operations

So let us now consider the case in which the ground set is in fact a finite dimensional vector space $V$ over a field $\mathbb{F}$ (either $\mathbb{R}$ or $\mathbb{C}$). From what we have seen earlier about symmetries of structures in general, symmetry operations are those one-to-one onto maps on the ground set that preserve the defining relations of the structure.[11] Recall that a map $S$ on a set $X$ is said to preserve a (binary) relation $H$ on $X$ if, for $x, y \in X$, $xHy$ implies $S(x)HS(y)$. If $H$ is in fact a function or a map on $X$ then by $S$ preserving $H$ we equivalently mean that $H$ and $S$ commute under composition, i.e., $HS = SH$.

For the symmetries of a linear transformation $H$ on $V$, we thus have to look for one-to-one onto maps on $V$ that commute with $H$. These maps characterize the

[11] A function or a transformation is also a binary relation. So functions and transformations are equally well covered by this interpretation of symmetry.

symmetry operations. For our present purposes, we will confine ourselves to only those of such maps that meet the additional requirement of linearity.[12]

So, in view of our discussions so far, *to say that a linear transformation $H$ on a vector space $V$ has symmetries is to say that there is a group $\mathbf{P}$ of invertible linear transformations on $V$ such that $H$ commutes with every member of* $\mathbf{P}$:[13]

$$SH = HS \text{ for every } S \in \mathbf{P}. \tag{4.9}$$

Mathematically, $\mathbf{P}$ can be any arbitrary group of linear transformations on the given vector space, and it can be said to define symmetries of the transformation $H$ in a general sense.[14] From the point of view applications, however, it is generally the other way around. We start with an intuitive understanding of symmetries inherent in a problem, and then formally characterize them by a group of transformations. Such is the case for time–invariance of systems: we first identify operators that produce for signals what we intuitively consider as translations in time, and then view time–invariance of systems as symmetry in terms of these operators. Let us look at this special case a little more closely.

### 4.4.2 Translation Operators

Typically, in the study of signals and systems, we start with an index set $I$ (which may for instance represent points in time or space), and we consider a signal $f$ as a function on this set, $f : I \to \mathbb{F}$, where $\mathbb{F}$ is a field (generally, either $\mathbb{R}$ or $\mathbb{C}$). $V$ is correspondingly a vector space of such functions over $\mathbb{F}$. For members of $\mathbf{P}$, we have linear operators on $V$ defined through domain permutations of signals, or rather through a group, $\mathbf{G}$, of permutations acting on the index set $I$. Thus for a function $f$, let ${}^{\pi}\!f$, a *translate* of $f$, be the function with values ${}^{\pi}\!f(x) = f(\pi^{-1}(x))$ for $x \in I$, where $\pi$ is a permutation in $\mathbf{G}$.[15] Correspondingly we define an operator, a *translation operator*,[16] $D_\pi : V \to V$, $D_\pi f = {}^{\pi}\!f$ for $f \in V$, i.e.,

$$(D_\pi f)(x) = {}^{\pi}\!f(x) = f(\pi^{-1}(x)) \quad \text{for} \quad f \in V \quad \text{and} \quad x \in I.$$

[12] Besides being mathematically more tractable, the linear ones cover most of our needs of representing symmetry operations in physical problems that are of interest to us here. There is yet another angle from which the constraint of linearity is well justified. When the ground set is simply a set, every member stands on its own—$x$'s membership implies nothing about $y$'s membership of the set. On the other hand, if the ground set has additional structure, such as that of a vector space, membership of $x$ and $y$ implies membership of many other 'relatives' of theirs by definition ($\alpha x$ and $x + y$, for instance, for a vector space). For a map $\phi$ on the set, it is then pertinent to ask whether it is 'nice' enough to preserve this implied membership of the relatives. Specifically for a vector space, does it take $\alpha x$ into $\alpha\phi(x)$ and $x + y$ into $\phi(x) + \phi(y)$? In other words, is it a linear map? So long as our practical needs are met by them, it is reasonable to confine our attention to the linear ones.

[13] Note that $\mathbf{P}$ is a subgroup of the group of all invertible linear transformations on $V$ under composition.

[14] This would be in line with the ideas of Klein's Erlanger Programme.

[15] By a *permutation* I mean a one-to-one onto function from a set to itself. For a permutation $\pi$, $\pi^{-1}$ denotes the inverse permutation.

[16] For a purely mathematical justification for introducing such operators here, and for their role, see Edwards [3, vol. 1, pp. 16–17, pp. 57–59].

Note that the operator so defined on $V$ is linear. For, for two functions $f$ and $g$, we see that it is additive:

$$\begin{array}{rcl} D_\pi(f+g)(x) & = & (f+g)(\pi^{-1}x) \\ & = & f(\pi^{-1}x) + g(\pi^{-1}x) \\ & = & D_\pi f(x) + D_\pi g(x) \\ & = & (D_\pi f + D_\pi g)(x)\,, \end{array}$$

and it is also homogeneous, i.e., for any scalar $\alpha$ in the field of $V$,

$$\begin{array}{rcl} D_\pi(\alpha f)(x) & = & (\alpha f)(\pi^{-1}x) \\ & = & \alpha(f(\pi^{-1}x) \\ & = & \alpha D_\pi f(x). \end{array}$$

Further, the set $\{D_\pi | \pi \in \mathbf{G}\}$ of translation operators is a group isomorphic to $\mathbf{G}$. To check this, let $\phi : \mathbf{G} \rightarrow \mathbf{P}$ be the function with values $\phi(\pi) = D_\pi$. Clearly, $\phi$ is one-to-one and onto. Moreover, $D_\sigma D_\pi f(t) = f(\pi^{-1}\sigma^{-1}t) = f((\sigma\pi)^{-1}t) = D_{\sigma\pi} f(t)$ for any $t \in I$ and any $f \in V$. That is, $D_{\sigma\pi} = D_\sigma D_\pi$, or equivalently, $\phi(\sigma\pi) = \phi(\sigma)\phi(\pi)$. $\mathbf{P}$ is thus a group isomorphic to $\mathbf{G}$.

In the special case in which the group $\mathbf{G}$ is abelian, and its cardinality (i.e., the number of elements of the group) is the same as that of the index set $I$, there is an additional bonus. The action of $\mathbf{G}$ on $I$ can equivalently be characterized in terms of a binary operation that makes $I$ itself an (abelian) group isomorphic to $\mathbf{G}$.[17]

**Aside 1** A point about the notation for domain transformations and translation operators deserves attention here. As a concrete case, consider first a space of functions $f : \mathbb{R} \rightarrow \mathbb{R}$ (real valued functions of time say). A delay operator $D_\tau$ on this space is defined as one that produces from a function $f$ a function $D_\tau f$ with values $D_\tau f(t) = f(t - \tau)$. Why this '$-$' sign here? Why not work in terms of $f(t + \tau)$? One usually offers a physical explanation for this, saying that we are primarily interested in delays. This is fair enough, particularly because along with the group operation of addition, the real line also has a linear ordering of its points defined in terms of addition. But for other index sets, specially finite, such is not the case. Yet, in treating time shifts more generally in terms of one-to-one onto maps, or permutations, on the domain of the functions, we have analogously defined the translation operator as $(D_\pi f)(x) = f(\pi^{-1}x)$ for every $x \in I$. We have here the inverse of the permutation $\sigma$ coming into play. It is natural to ask why.

Well, this has to do with certain compulsions of notation that show up in the composition of operators. Let us see. Instead of defining the operator $D_\pi$ the way we have, let us define it instead as $(D_\pi f)(x) = f(\pi x)$. It is easily checked in that case that $D_{\pi\sigma} = D_\sigma D_\pi$. Observe the reversal in the order in which the subscripts appear on the two sides, making it somewhat awkward to keep a proper accounting of the orders. This does not happen with the notation we have adopted.

[17] If $I$ is finite then it has the structure of what is known as a *mixed radix number system*, of which modulo arithmetic is a special case. The theory of transforms such as the DFT and its other variants and generalizations are very closely linked to this case. For details see Siddiqi and Sinha [16, 17].

It must be added, however, that both notations are equally valid, and the results we obtain are in essence indifferent to the choice. It may be helpful in this context to take note of the following elementary result, which is left as an exercise.

**Exercise 4.4.1** *Let $G$ and $H$ be two finite groups of the same order, and let $h : G \to H$ be a one-to-one map satisfying the condition*

$$h(xy) = h(y)h(x)$$

*for any $x, y \in G$. Then the map $g : G \to H$ defined as $g(x) = h(x^{-1})$ for $x \in G$ is also one–one, and $g(xy) = g(x)g(y)$, i.e., the groups $G$ and $H$ are isomorphic.*

♡

## 4.5 Symmetry Based Decompositions

Let us now go back to the study of transformations with symmetries characterized by the constraint (4.9). Clearly, the class—let us call it $\mathcal{H}$—of all linear transformations on $V$ that satisfy this constraint with respect to a particular group $\mathbf{P}$ is closed under addition, scaling, and composition, i.e., if $H$ and $F$ satisfy this constraint, then so do $H + F$, $\alpha H$ for any $\alpha \in \mathbb{F}$, and $HF$. The class $\mathcal{H}$ is thus, under addition and scalar multiplication, a subspace of the vector space of all linear transformations on $V$.[18]

What special structural properties does $\mathcal{H}$ have on account of the fact that the set $\mathbf{P}$ is a group? And in what way, if any, these properties help us in exploiting symmetries in the study of signals and systems?

Considering that symmetries in general imply a reduction in complexity, it is reasonable to expect that transformations that satisfy (4.9) are in a significant way simpler in structure as compared to arbitrary linear transformations on $V$. As we shall presently see, they indeed are.

**Aside 2** Suppose you are to move from room A to room B and you want the arrangements of the articles in room B after moving to be exactly the same as they were in room A—articles on your main desk in A to move on to a corresponding desk in room B, those on your book rack on to a corresponding book rack in B, and so on. You could do this by first moving all the articles to room B at random, just finding a place for them without giving any thought to where they are to go finally. Having moved them all, you could then rearrange them to put them in the right places as they were in room A. But then the task would be much less laborious if you organize your moving from the start, transferring articles from desk in A to desk in B, from rack in A to rack in B, and so on. Although the final result is the same, this way you are better off in practice in terms of the effort you put in for the job. ♡

Very broadly, this is a result of the fact that the group property of $\mathbf{P}$ ensures for $V$ a special kind of direct sum decomposition into proper subspaces each of

[18] Under addition and composition, the set of all linear transformations forms a ring and with scaling also included, it forms an algebra. The class $\mathcal{H}$ is a subring and a subalgebra in the respective settings. Note that all this is true whether the set $\mathbf{P}$ is a group or not. The fact that $\mathbf{P}$ is a group makes $\mathcal{H}$ a very special kind of algebra with structural properties that are of central significance to us here.

which is invariant under every member of $\mathbf{P}$.[19] Such a decomposition has a special significance in problems of signals representation, because the subspaces can be associated with certain specific features of signals.

Associated with this decomposition, there is another one similarly disposed towards every $H \in \mathcal{H}$. The task of determining the action of any such $H$ on $V$ is as a result replaced by one of determining its actions separately on subspaces of smaller dimensions. Moreover, in signal processing applications, this is central to the idea of designing processors that act as filters. If the subspaces can be associated with different features of signals, the processors can be designed to act on the subspaces differently.

Interpreting the idea of decomposition in another way, constraint (4.9) ensures that there is a basis for $V$ with respect to which matrices of all transformations in $\mathcal{H}$ take a block–diagonal form, all of them having the same block structure. The important point about these basis vectors is that there is a general procedure to compute them, starting with the group of transformations $\mathbf{P}$, the abstract group that it is isomorphic to, and the so called irreducible representations of this abstract group. We will learn more about this procedure when we come to representation theory of groups. Let us at this point get a general idea of what this means in practice.

It is perhaps best to begin by first matrices, and relating it to linear transformations.

### 4.5.1 Block–Diagonalizability and Invariant Subspaces

Recall that a (nonsingular) matrix $A$ is said to diagonalize another matrix $B$ if the similarity transformation $A^{-1}BA$ produces a diagonal matrix; it is said to simultaneously diagonalize two or more matrices if it diagonalizes every one of them. More generally, it is said to block–diagonalize the given set of matrices if the same similarity transformation on all of them produces block–diagonal matrices with the same block structure.[20]

For a single (square) matrix of size $n$, with $n$ linearly independent eigenvectors, there is the very familiar result that it is diagonalized by the matrix made up of the eigenvectors as its columns. Moreover, if we have a set of two or more matrices of size $n$, and if they share a common set of $n$ linearly independent eigenvectors, then they are simultaneously diagonalized by the matrix made up of these common eigenvectors. But in that case these matrices necessarily commute amongst themselves.[21]

For a set of matrices that do not commute, simultaneous diagonalization is clearly ruled out (why?). There is, however, the option, under special circumstances,

[19] The term “invariant”, in the sense used here, is defined later in Section 5.5.

[20] A matrix is block–diagonal if it consists of square submatrices on the principal diagonal, and has zeros every where else. A diagonal matrix is also block–diagonal, with every diagonal block of size one.

[21] For extensive discussions on commutative matrices and their properties, see Suprunenko [18].

of simultaneously block–diagonalizing them. One such special circumstance is when the matrices of the given set form a group under multiplication.

Now consider matrices not just by themselves, but as matrices of linear transformations on a vector space with respect to a basis. Let us see what block–diagonalization means in this context.

Let us say for the sake of illustration that a vector space $V$ of dimension $n$ is a direct sum of two of its proper subspaces $V_1$ and $V_2$, $V = V_1 \oplus V_2$. Further, let $T$ be a linear transformation on $V$ such that $V_1$ and $V_2$ are both invariant under $T$, i.e., if $x \in V_i$ then $Tx \in V_i$.

Clearly, if the dimension of $V_1$ is $m$ (so that the dimension of $V_2$ is $(n - m)$) then $V$ has an ordered basis $(\phi_1, \phi_2, \ldots, \phi_m, \phi_{m+1}, \ldots, \phi_n)$, where $(\phi_1, \phi_2, \ldots, \phi_m)$ is an ordered basis of $V_1$ and $(\phi_{m+1}, \ldots, \phi_n)$ that of $V_2$.[22] With respect to this basis of $V$, the matrix of $T$ has a block–diagonal form:

$$\begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix},$$

where $T_1$ and $T_2$ are blocks of size $m$ and $(n - m)$ respectively.

The idea straightforwardly extends to the case involving more than two proper subspaces. Thus, suppose that $V = V_1 \oplus V_2 \oplus \cdots \oplus V_k$ for some $k$, where each $V_i$, $i = 1, 2, \ldots, k$, is invariant under the linear transformation $T$. Then there is a basis for $V$ with respect to which the matrix of $T$ has a block diagonal form consisting of $k$ blocks. We say that the transformation $T$ is *block–diagonalizable*.[23] If the subspaces are invariant under every member of a given set of transformations, then we say that the set is *simultaneously* block–diagonalizable. We are interested here in the case in which the given transformations form a group. In (4.9), this is the group $\mathbf{P}$. Before we consider the class $\mathcal{H}$, we look at this group itself.

## 4.5.2 Transformation Groups and Their Invariant Subspaces

Consider first the special case in which the group $\mathbf{P}$ is abelian. The invariant subspaces in this case turn out to be of dimension one, each spanned by a single basis vector. In other words, for the members of $\mathbf{P}$, there is in this case a common set of eigenvectors that together constitute a basis for $V$. With respect to this basis, the matrix of any member of $\mathbf{P}$ takes a diagonal form.[24] We have already encountered this situation in Section 1.6, where we discussed shift–invariance for systems on

[22] Although intuitively clear, you need to formally check that this is true. See Hoffman and Kunze [8, Lemma, Section 6.6, p. 209].

[23] With respect to different bases, a transformation has different matrices. So one can not talk of a transformation being in a block–diagonal form. Rather, one says that the transformation is block–diagonalizable, i.e., there is a basis with respect to which its matrix is block–diagonal.

[24] In purely matrix terms, this means that we have here a procedure for simultaneously diagonalizing all members of a given abelian group of matrices.

finite index sets.[25] Let us go back to the results of Examples (1.6.1) and (1.6.2), and see them from the viewpoint of vector spaces.

**Example 4.5.1** We started in Example 1.6.1 with 4–tuples of reals. To have greater flexibility in their representation, let us now treat them as 4–tuples of complex numbers. Consider the vector space $\mathbb{C}^4$ of all such 4–tuples under addition and scaling by complex numbers. For a 4–tuple $x = (x_0, x_1, \ldots, x_3)$, its coordinate (column) matrix with respect to the standard basis is then $[x_0\ x_1\ x_2\ x_3]'$.

Now, the four shift operators $D_0 \ldots D_3$ of Example 1.6.1 are linear transformations on $\mathbb{C}^4$, and the four matrices $D_0 \ldots D_3$ of Table 1.2, reproduced in Table 4.4, are the matrices of these transformations with respect to the standard basis on $\mathbb{C}^4$.

$$D_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad D_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$D_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \qquad D_3 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Table 4.4 Matrices of the shift operators on $\mathbb{C}^4$ (of Example 4.5.1) with respect to the standard basis

As pointed out in Example 1.6.2, these operators have a common set of four linearly independent eigenvectors $\phi_0,\ \phi_1,\ \phi_2,\ \phi_3$ whose coordinate vectors with respect to the standard basis are as given in Table 4.5.

$$\phi_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad \phi_1 = \begin{bmatrix} 1 \\ -j \\ -1 \\ j \end{bmatrix} \qquad \phi_2 = \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} \qquad \phi_3 = \begin{bmatrix} 1 \\ j \\ -1 \\ -j \end{bmatrix}$$

Table 4.5 A special basis for $\mathbb{C}^4$

.

□

[25] Equation (1.14) there is a matrix version of (4.9), in which the matrices $D_i$ form a cyclic group of order 4. We found in this case that there is a matrix $\mathbf{W}$ that simultaneously diagonalizes the matrices $D_i$.

One way to look at the four eigenvectors in this example is to consider each one of them individually as the basis of a one–dimensional subspace of $\mathbb{C}^4$. In other words, if $V_i$ denotes the subspace spanned by $\phi_i$, then $\mathbb{C}^4 = V_0 \oplus V_1 \oplus V_2 \oplus V_3$, and every $V_i$ is invariant under the operators $D_k$, $k = 0, 1, 2, 3$. Moreover, since the subspaces are each of dimension 1, we find that this is a decomposition in which each of the subspaces is clearly the smallest possible.

Generalizing the basic idea illustrated by this example, we are led to a general strategy for decomposing a vector space. It may be summed up as follows. *Given a vector space $V$ and a group $\mathbf{P}$ of linear transformations on $V$, identify a decomposition $V = V_1 \oplus V_2 \oplus \cdots \oplus V_k$ such that the subspaces $V_i$, $i = 1, \cdots, k$, are as small as possible, and are each invariant under every member of* $\mathbf{P}$.[26]

If the group $\mathbf{P}$ is abelian then it turns out that the subspaces of the decomposition are all of dimension 1. On the other hand, if the group is nonabelian then the subspaces of the decomposition could have higher dimensions. These dimensions are, however, decided by the structure of $\mathbf{P}$, independently of the dimension of the vector space $V$, and are comparatively much smaller than it. (Thus the dimension of $V$ may be 50 or 100 but irrespective of that, the subspaces will typically be of dimensions 1 or 2.)

About procedures for such a decomposition, we shall learn in the next chapter. As already pointed out, such decompositions are of central importance in the representation of signals.

### 4.5.3 Transformations with Symmetries

Accompanying the decomposition produced by the group of transformations $\mathbf{P}$, for $\mathcal{H}$ too, there is a decomposition $V = W_1 \oplus W_2 \oplus \cdots \oplus W_k$ for some $k$, such that each $W_i$ is invariant under every member of $\mathcal{H}$. In many situations, this decomposition coincides with the decomposition in terms of the $V_i$'s. But even if it does not, the two are very closely linked.

In more concrete matrix terms, this decomposition means the following. Suppose the matrices of the transformation $H$ and of the members of the group $\mathbf{P}$ are given to us with respect to some basis. Then using the structural properties of $\mathbf{P}$ and the fact that $H$ commutes with it, we can find a similarity transformation that places the given matrices in block diagonal forms. Here is an illustrative example.

**Example 4.5.2** Let us consider $\mathbb{R}^4$ as the given vector space. Let the group $\mathbf{P}$ consist of two transformations $P_1$ (the identity element) and $P_2$ whose matrices with respect to the standard basis on $\mathbb{R}^4$ are the following.

$$P_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } P_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

[26] I am essentially stating here a variation of what Gross [7] enunciates as the fundamental problem of harmonic analysis.

Further, let $H$ be a transformation whose matrix with respect to the standard basis is of the form

$$H = \begin{bmatrix} a & b & 0 & -1 \\ b & a & 1 & 0 \\ 0 & -1 & c & d \\ 1 & 0 & d & c \end{bmatrix},$$

where $a, b, c, d$ are arbitrary reals.

Using the procedure explained in the next chapter, we find the following matrix $\alpha$.

$$\alpha = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & -1 & 0 & 1 \end{bmatrix}.$$

Similarity transformation by $\alpha$ puts both $P_2$ and $H$ in the block diagonal form:[27]

$$\alpha^{-1} P_2 \alpha = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \text{ and,}$$

$$\alpha^{-1} H \alpha = \begin{bmatrix} a+b & 1 & 0 & 0 \\ 1 & c-d & 0 & 0 \\ 0 & 0 & a-b & -1 \\ 0 & 0 & 1 & c+d \end{bmatrix}$$

□

Notice that no matter what the constants $a, b, c, d$ in this example are, the same similarity transformation does the block diagonalization, producing the same block structure. This illustrates what is true in general.

[27] We need not check for $P_1$ as it is the identity matrix for this case.

# References

1. L. Brillouin. *Wave Propagation in Periodic Structures*. McGraw-Hill, New York, 1946.

2. John D. Dixon and Brian Mortimer. *Permutation Groups*. Springer, New York, 1996.

3. R.E. Edwards. *Fourier Series: A Modern Introduction*, volume 1. Springer-Verlag, New York, 1979.

4. R. Foote, G. Mirchandani, et al. A wreath product group approach to signal and image processing: part i–multiresolution analysis. *IEEE. Trans. on Signal Processing*, 48(1):102–132, 2000.

5. R. Foote, G. Mirchandani, et al. A wreath product group approach to signal and image processing: part ii–convolution, correlation, and applications. *IEEE. Trans. on Signal Processing*, 48(3):749–767, 2000.

6. Wolfgang Gaertner. The group theoretic aspect of four–pole theory. *IRE National Convention, Circuit Theory*, Part 2:36–43, 1954.

7. Kenneth I. Gross. On the evolution of noncommutative harmonic analysis. *Amer. Math. Month.*, 85:525–548, 1978.

8. Kenneth Hoffman and Ray Kunze. *Linear Algebra*. Prentice–Hall (India), New Delhi, 1972.

9. Nathan Howitt. Group theory and the electric circuit. *Physical Review*, 37:1583–1595, 1931.

10. David M. Kerns. Analysis of symmetrical waveguide junctions. *Jr. Res. Nat. Bur. Stand.*, 46:515–540, 1949.

11. Reiner Lenz. *Group Theoretical Methods in Image Processing*. Springer-Verlag, New York, 1990.

12. Wu-Sheng Lu and Andreas Antoniou. *Two–Dimensional Digital Filters*. Marcel Dekker, New York, 1992.

13. George W. Mackey. *The Scope and History of Commutative and Noncommutative Harmonic Analysis*, volume 5 of *History of Mathematics*. American Mathematical Society, 1992.

14. A.E. Pannenborg. On the scattering matrix of symmetrical waveguide junctions. *Phillips Res. Rep.*, 7:168–188, 1952.

15. Gunther Schmidt and Thomas Ströhlein. *Relations and Graphs*. Springer-Verlag, New York, 1993.

16. M.U. Siddiqi and V.P. Sinha. Generalized walsh functions and permutation–invariant systems. *IEEE EMC. Symp., Montreal*, pages 43–50, 1977.

17. M.U. Siddiqi and V.P. Sinha. Permutation–invariant systems and their application in the filtering of finite discrete data. *Proc. IEEE Int. Conf. on ASSP*, pages 352–355, 1977.

18. D.A. Suprunenko and R.I. Tyshkevich. *Commutative Matrices*. Academic, London, 1968.

19. Hermann Weyl. *Symmetry*. Oxford University Press, New York, 1952.

# Chapter 5

# Representations of Finite Groups

## 5.1 The Notion of Representation

Given a structure $S$, any other structure homomorphic to $S$ provides, in principle, a representation of $S$. Thus when you draw a Venn diagram to depict sets and their union and intersection, you resort to a representation. The given structure in this case is a set of sets under union and intersection, and its Venn diagram a geometrical representation of it. Likewise, when you draw a graph for a real-valued function of a real variable, you provide a representation for a given set of ordered pairs of reals, consisting of a collection of points on a plane referred to a particular pair of coordinates.

Not every representation is, however, of the same practical interest. First, it should relate the elements of the given structure to mathematical objects that are very familiar to us, e.g., numbers, matrices, or even drawings or geometrical constructions on a plane. Secondly, it should be such that representations similar to it can be provided for other structures that are of the same species as the given structure. For finite groups, both these requirements are well met by matrices under multiplication, producing what are called matrix representations. While the theory of such representations can be dealt with purely in terms of matrix algebra, a pedagogically more powerful approach is to present it within the framework of linear algebra. This is what I intend to do in this chapter. But before I take up that, let us informally examine through illustrations, the idea of a matrix representation of a group purely in terms of matrix algebra.

## 5.2 Matrix Representations of Groups

Consider an abstract group G of just two elements $a$ and $b$, with the multiplication table:

V.P. Sinha, *Symmetries and Groups in Signal Processing*, Signals and Communication Technology,
DOI 10.1007/978-90-481-9434-6_5, © Springer Science+Business Media B.V. 2010

| $*$ | $a$ | $b$ |
|---|---|---|
| $a$ | $a$ | $b$ |
| $b$ | $b$ | $a$ |

Consider in addition, the following two 2 × 2 matrices under matrix multiplication.

$$\mathrm{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \mathrm{B} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{5.1}$$

They too constitute a group, let us say M, its multiplication table being,

|  | A | B |
|---|---|---|
| A | A | B |
| B | B | A |

Between G and $\mathrm{M}$, if we now consider the map $\phi$, $\phi(a) = \mathrm{A}$ and $\phi(b) = \mathrm{B}$, then checking from the tables, for any $x$ and $y$ in G, $\phi(x * y) = \phi(x)\phi(y)$. We thus see that between the structures G and $\mathrm{M}$, the map $\phi$ induces an isomorphism. The matrix group $\mathrm{M}$ is then in this context an instance of a *matrix representation* of the group G. (Instead of the matrices, one could think of the function $\phi$ as the representation, as we shall do later when we come to vector spaces; the difference is simply in the way you visualize a function.)

Now consider the group $\mathrm{U}$ consisting of just one matrix, the matrix $\mathrm{A}$ of (5.1), and let $\psi$ be the map from G to U, $\psi(a) = \psi(b) = \mathrm{A}$. In this case too, for any $x$ and $y$ in G, $\psi(x * y) = \psi(x)\psi(y)$, so that $\psi$ is a homomorphism. U is then again a matrix representation of G.

Finally, the group $\mathrm{Q}$ consisting of the $2 \times 2$ matrices

$$\mathrm{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \mathrm{D} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{5.2}$$

is also a matrix representation of G. There are, apparently, endless possibilities. But are these representations *really* different? In particular, are the representations $\mathrm{M}$ and $\mathrm{Q}$ different? The answer is a qualified 'no'. They are different, and yet they are the same, somewhat in the sense that a picture is the same whether we look at it from this side or the other. To be more specific, consider the matrix $\mathrm{P}$,

$$\mathrm{P} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{5.3}$$

You may check that

$$\mathrm{C} = \mathrm{P}^{-1}\mathrm{AP}, \quad \text{and} \quad \mathrm{D} = \mathrm{P}^{-1}\mathrm{BP}. \tag{5.4}$$

That is, to use the standard language of matrix algebra, there is a single similarity transformation that converts all matrices of the representation $\mathrm{M}$ into those of $\mathrm{Q}$. It is easily checked that, *given a matrix representation of a group, any similarity transformation simultaneously applied to all its matrices produces yet another matrix representation of the same group.*

Then there is another point worth noting about the representations $\mathrm{M}$ and $\mathrm{Q}$. Hidden within these are two smaller representations of the same group $\mathrm{G}$. Look at the matrices of (5.2). We have

$$\begin{aligned} \mathrm{C} &= [1] \dot{+} [1] \\ \mathrm{D} &= [1] \dot{+} [-1] \end{aligned} \tag{5.5}$$

where '$\mathrm{A} \dot{+} \mathrm{B}$' denotes the operation of creating a new matrix by placing $\mathrm{A}$ and $\mathrm{B}$ as the first and second diagonal blocks, and zeros every where else. Note that the group consisting of the single $1 \times 1$ matrix $[1]$ also constitutes a representation of G, with both elements $a$ and $b$ mapping into [1]. Likewise, [1] and $[-1]$ also constitute a representation of G, with $a$ mapping into [1] and $b$ mapping into $[-1]$.

As to M, we have,

$$\begin{aligned} \mathrm{A} &= \mathrm{P}([1] \dot{+} [1])\mathrm{P}^{-1} \\ \mathrm{B} &= \mathrm{P}([1] \dot{+} [-1])\mathrm{P}^{-1}. \end{aligned} \tag{5.6}$$

Relationships (5.2)–(5.6) can be looked upon in two ways: one, that a given matrix representation may be reduced through similarity transformations to a form in which it is a direct sum of smaller representations, or two, that from a given set of representations, other bigger ones may be constructed from their direct sums and similarity transformations over them.

Thus we find that by a combination of similarity transformations and direct summations of a given set of representations, we can produce any number of new representations. It does not mean, however, that any two representations are related to each other in this manner. How many really different representations are there for a finite group? How are the others related to these? Are there matrix representations for any finite group whatever? In order to answer such questions and many others of this kind, it is better at this point to recast the representation problem in the language of linear algebra, associating matrices with linear transformations on vector spaces.

**Exercise 5.2.1** For the abstract group of two elements, examine whether there are $2 \times 2$ matrix representations that cannot be reduced by similarity transformation to the representation given in (5.2).

**Exercise 5.2.2** Show that isomorphisms between groups induce an equivalence relation over the set of all groups.

**Exercise 5.2.3** Given a group of matrices, show that if one of them is nonsingular, all of them must be nonsingular. Create an example using $2 \times 2$ matrices to show that singular matrices may also constitute a group.

**Exercise 5.2.4** Show that a set of permutation matrices of the same size that commute pair wise can be simultaneously diagonalized, i.e, there is a single similarity transformation that puts all of them in the diagonal form.

## 5.3 Automorphisms of a Vector Space

The structure that we now look at for the representation of groups is that of automorphisms of a vector space under composition. Given a vector space $V$ over a field $\mathbb{F}$, a map $F : V \rightarrow V$ is an *automorphism* of $V$ if it is an invertible linear transformation from $V$ to $V$, i.e., if (i) $F$ is one-to-one and onto, and (ii) $F(\alpha\ x + \beta\ y) = \alpha\ F(x) + \beta\ F(y)$, for any $x, y \in V$ and any $\alpha$, $\beta \in \mathbb{F}$. You may easily verify that the set of all automorphisms of $V$ constitutes a group under composition. This group is commonly referred to as *the general linear group*, $\mathbf{GL}(V)$.

For invertible linear transformations on a vector space, their matrices are also invertible, irrespective of the basis. Taking this fact into account, we conclude that for a vector space of dimension $n$, the group $\mathbf{GL}(V)$ may be looked upon concretely in terms of the group of all $n \times n$ invertible matrices. It is this connection between vector space automorphisms and matrices that we exploit in systematically studying matrix representations of finite groups. Moreover, we concentrate on a special $n$-dimensional vector space, $\mathbb{F}^n$, the vector space of $n$-tuples of scalars from a field $\mathbb{F}$, with componentwise addition and scalar multiplication. This is because any $n$-dimensional vector space $V$ over $\mathbb{F}$ is isomorphic to it, and by confining to it we lose little but gain a good deal of clarity in computational work. For $\mathbb{F}^n$, the general linear group is commonly written as $\mathbf{GL}(n, \mathbb{F})$. Our interest here lies in the fields $\mathbb{R}$ and $\mathbb{C}$, and correspondingly in the general linear groups $\mathbf{GL}(n, \mathbb{R})$ and $\mathbf{GL}(n, \mathbb{C})$.

## 5.4 Group Representations in GL(V)

Using the language of vector spaces, we now say that a *representation* of a finite group $G$ in a vector space $V$ is a homomorphism $\rho$ from $G$ into $\mathbf{GL}(V)$, and we call $V$ the *representation space* of $\rho$. If $V$ is a finite dimensional vector space, which is what we are exclusively concerned with here, we use the notation $\rho : G \rightarrow \mathbf{GL}(n, \mathbb{F})$, where $\mathbb{F}$ is the field of $V$ and $n$ its dimension. Recall that homomorphisms over groups preserve the group structure. Thus when we look for a group representation, we are essentially looking for a subgroup of $\mathbf{GL}(n, \mathbb{F})$ onto which the given group homomorphically maps. If the order of this subgroup is the same as that of the original group G, we say the representation is a *faithful* one.

With respect to a certain basis in $V$, the matrices of the transformations $\rho(s)$, $s \in G$, are the *matrices of the representation* $\rho$; for different bases, the matrices of a representation are different. By looking at representations as groups of linear transformations on a vector space rather than purely as matrices, we make them coordinate free as it were. This is very helpful in studying their general properties, as we shall presently see. It is, however, inevitable that when we want to use them in practice, we end up using them in one matrix form or another. We work with their matrices with respect to a specific basis in a specific vector space. The following two examples illustrate the idea.

**Example 5.4.1** Consider the group G of two elements that we examined at the beginning of Section 5.2. Let us look at its representations in $\mathbb{R}$ treated as a one-dimensional vector space of real numbers under addition and scalar multiplication by reals. Any real number serves as its single basis vector. As its standard basis we choose $\delta_1 = 1$. $\mathbf{GL}(1,\mathbb{R})$ is simply the set of nonzero real numbers under multiplication. As our first choice for a representation $\rho^1 : G \to \mathbf{GL}(1,\mathbb{R})$, we can take $\rho^1(a) = \rho^1(b) = 1$. As our second choice, let $\rho^2(a) = +1$ and $\rho^2(b) = -1$. That there can be no other choice can be seen from the fact that there is no other way in which the equalities $x^2 = y^2 = x$, $xy = y$ can be satisfied for real $x$ and $y$. Note that the sets $\{1\}$ and $\{1,-1\}$ are subgroups of $\mathbf{GL}(1,\mathbb{R})$ under multiplication. We thus get two representations $\rho^1$ and $\rho^2$ in this case:

| $x$ | $a$ | $b$ |
|---|---|---|
| $\rho^1(x)$ | 1 | 1 |
| $\rho^2(x)$ | 1 | $-1$ |

**Example 5.4.2** Consider now the group $H$ to be the cyclic group of order 3, with the following multiplication table.

| $*$ | $a$ | $b$ | $c$ |
|---|---|---|---|
| $a$ | $a$ | $b$ | $c$ |
| $b$ | $b$ | $c$ | $a$ |
| $c$ | $c$ | $a$ | $b$ |

This time we choose for our representation space, the vector space of the set of complex numbers under addition, and scalar multiplication by complex numbers. We have to look for homomorphisms from $H$ to $\mathbf{GL}(1,\mathbb{C})$ in this case. Three such homomorphisms, $\rho^1$, $\rho^2$ and $\rho^3$, we easily identify to be the following.

| $x$ | $a$ | $b$ | $c$ |
|---|---|---|---|
| $\rho^1(x)$ | 1 | 1 | 1 |
| $\rho^2(x)$ | 1 | $r$ | $r^2$ |
| $\rho^3(x)$ | 1 | $r^2$ | $r$ |

$; r = \exp\jmath(2\pi/3)$

Amongst representations, there is one type that is identified very easily: just put $\rho(s) = 1$ for every $s \in H$. More precisely, choose $\rho : H \to \mathbf{GL}(1,\mathbb{R})$, such that with respect to the standard basis on $\mathbb{R}$, $\rho(s) = 1$ for all $s \in H$. This is the so called *trivial* representation, the kind of which is available for any (finite) group whatever. The first ones in the two examples just given are trivial representations.

Then there is another representation for any finite group, the *regular* representation, which, as we shall see later, carries within it basic information about all possible representations of the group. A mechanical way of constructing it directly in a matrix form is the following.

1. Given a group G, start with its Cayley table.
2. Rearrange its columns, replacing the column corresponding to a group element $s$ by the column corresponding to the element $s^{-1}$.
3. For any element $s$, put 1 in the rearranged table wherever $s$ appears, and zero every where else; treating the resulting table of 0s and 1s as a matrix, we get the matrix $\rho(s)$.
4. For a group of order $n$, the $n$ matrices each of size $n \times n$ that we get this way constitute under multiplication the regular (matrix) representation of the group.

We may alternatively look at the regular representation of a group G of order $n$ in terms of appropriately chosen linear transformations on a vector space V of dimension $n$. Let $\{\epsilon_s | s \in \text{G}\}$ be a basis of V indexed by the elements of G (i.e., the elements of G are assigned an arbitrary order and the basis vectors are arranged in that order, $\epsilon_s$ corresponding to $s$). Further, for every $s \in$ G let $\rho(s)$ be a linear transformation from $V$ to $V$ such that $\rho(s)\epsilon_u = \epsilon_{su}$ for every $u \in$ G. It follows that $\rho(vs)\epsilon_u = \epsilon_{vsu} = \rho(v)\epsilon_{su} = \rho(v)\rho(s)\epsilon_u$ for every $u, v, s \in$ G. Thus for any vector $x \in$ V, $\rho(vs)x = \rho(v)\rho(s)x$, i.e., $\rho(vs) = \rho(v)\rho(s)$ for all $v, s \in$ G. In $\rho$ :G $\rightarrow$ **GL**(V) therefore we have a representation of G. This is in fact the same regular representation that we earlier outlined, as you will see if you examine its matrices over the basis $\{\epsilon_s\}$. Observe that a regular representation is a faithful representation, whereas a trivial one is evidently not.

**Example 5.4.3** For the groups G and H of the two previous examples, let us now consider their regular representations on $\mathbb{R}^n$, where $n$ is the order of the group. Taking up G first, since the element $b$ is its own inverse, no column interchange in the Cayley table is needed. The matrices $\rho(a)$ and $\rho(b)$, with respect to the standard basis on $\mathbb{R}^2$, which constitute its regular representation, are then the following.

$$\rho_{ij}(x) = \begin{cases} 1, & \text{if the Cayley table entry in row } i \text{ and column } j \text{ is } x; \\ 0, & otherwise. \end{cases}$$

These matrices are

$$\rho(a) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \rho(b) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Now for the regular representation of the group H, we take $\mathbb{R}^3$ as the representation space and reindex its standard bases $\delta_1$, $\delta_2$, $\delta_3$ as $\delta_a$, $\delta_b$, $\delta_c$ respectively. Since elements $b$ and $c$ are each other's inverses we interchange their corresponding rows in the Cayley table and get

| $*$ | $a$ | $b^{-1}$ | $c^{-1}$ |
|---|---|---|---|
| $a$ | $a$ | $c$ | $b$ |
| $b$ | $b$ | $a$ | $c$ |
| $c$ | $c$ | $b$ | $a$ |

To get $\rho(s)$ from this table, we put $1$ wherever there is $s$, and $0$ elsewhere. We thus get the matrices of the representation $\rho$ :H $\rightarrow$ $\mathbf{GL}(3, \mathbb{R})$, with respect to the standard basis on $\mathbb{R}^3$:

$$\rho(a) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \rho(b) = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \rho(c) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

To get these matrices in a more formal way, we start with the condition $\rho(s)\delta_u = \delta_{su}$. Then for any vector $x = [x_a \;\; x_b \;\; x_c]$, $\rho(s)x = x_a\delta_{sa} + x_b\delta_{sb} + x_c\delta_{sc}$. For instance, $\rho(b)x = x_a\delta_{ba} + x_b\delta_{bb} + x_c\delta_{bc} = x_a\delta_b + x_b\delta_c + x_c\delta_a$, so that the matrix of $\rho(b) = [\delta_b \;\; \delta_c \;\; \delta_a]$. This checks with what we obtained earlier.

What other representations are possible besides the trivial and regular representations? As we observed in Section 5.2, once we have identified one matrix representation, we can produce from it through similarity transformations an apparently limitless number of other representations. A salient point about infinite, or even very large sets, that one needs to remember in this context is that to enumerate their members is at best to classify them according to some crucial invariant features such that they are separated into a 'coarser' and more manageable number of classes. On this count, representations related through similarity transformations are not to be treated as different from one another. Let us make this idea more precise.

Consider two representations $\rho^1 : \mathrm{G} \rightarrow \mathbf{GL}(V_1)$ and $\rho^2 : \mathrm{G} \rightarrow \mathbf{GL}(V_2)$ of a group G. We say that $\rho^1$ is *equivalent* to $\rho^2$, or simply that the two representations are equivalent, if there exists an invertible linear transformation, $h : \mathrm{V}_1 \rightarrow \mathrm{V}_2$ (i.e., $h$ is one-to-one onto, so that $V_1$ and $V_2$ are isomorphic, and are of the same dimension) such that

$$\rho^1(s) = h^{-1}\rho^2(s)h \quad \text{for all } s \in \mathrm{G}. \tag{5.7}$$

If there is no such map then we say that the representations are *inequivalent*. We call them so because relationships of the form (5.7) induce an equivalence relation over the set of all representations, and a corresponding partitioning of it into equivalence classes.

If we give (5.7) a matrix interpretation then it means that matrices of the two representations with respect to specific bases are related through a similarity transformation under the matrix of $h$ for the same bases. The two representations (5.1) and (5.2) were of this kind.

Related to the idea of equivalent representations, an important matter of detail about their matrices is the following: *for any representation of a finite group, one can always choose a basis for the representation space such that the matrices of the representation with respect to it are all unitary ($i.e.$, the conjugate transpose equals the inverse).* More concretely, given a matrix representation of a group, we can always turn it through a similarity transformation into another representation in which all the matrices are unitary. It is enough for us here to know that this is possible. See Jansen and Boon [4, p. 83], Tinkham [9, p. 20] for a detailed proof.

For most purposes, equivalent representations are indistinguishable from one another and they may be treated as one and the same, to within isomorphic changes over representation spaces. Seen from this angle, the problem of enumerating all possible representations of a group is essentially one of identifying its inequivalent representations.

Even inequivalent representations are, however, far too numerous to merit separate attention as they are. The saving grace is that, carrying overlapping information as they do, they can be further slashed down in number to only a few ‘atomic’ ones, that need be considered as really different from one another, and from which all the others are built by simple linear algebraic operations, or rather, to which all others can be reduced.

## 5.5 Reducible and Irreducible Representations

In Section 5.2 we saw that a matrix representation may be such that it can be put in a block diagonal form in which it is the direct sum of several smaller representations. To see what this means in the language of linear algebra, let us begin by examining whether it is possible to ‘shrink’ a given representation $\rho : \mathrm{G} \to \mathbf{GL}(V)$, and obtain from it a smaller one. One such important possibility arises if $V$ has a proper subspace, say $U$, such that $\rho(s)$ maps $U$ into $U$, for every $s \in \mathrm{G}$.

For a subspace $U$ of a vector space $V$, if there is a linear transformation, $f$, on $V$ such that $f$ maps $U$ into $U$, then we say that $U$ is *invariant* under $f$; if there is a set of such transformations under each of which $U$ is invariant then we say that $U$ is invariant under this set. In the same spirit, if for a representation $\rho : \mathrm{G} \to \mathbf{GL}(V)$, a subspace $U$ is invariant under $\rho(s)$ for every $s \in \mathrm{G}$, then we say, in short, that $U$ is invariant under G.[1]

[1] Some authors use the term “stable” instead of “invariant”, as for instance Serre [8].

For a subspace $U$ invariant under G, let $\rho^{\mathrm{U}}(s)$ denote the *restriction* of $\rho(s)$ to U, i.e., the transformation from $U$ to $U$ with values

$$\rho^{\mathrm{U}}(s)x = \rho(s)x \text{ for } x \in \mathrm{U}$$

Then for $x \in U$ and $s, v \in$ G,

$$\rho^{\mathrm{U}}(sv)x = \rho(sv)x = \rho(s)\rho(v)x$$

$$= \rho(s)\rho^{\mathrm{U}}(v)x = \rho^{\mathrm{U}}(s)\rho^{\mathrm{U}}(v)x,$$

since $U$ is invariant under G. That is, $\rho^{\mathrm{U}}$ is a homomorphism from G to **GL**($U$). Thus, to sum up: *Given a representation* $\rho$ : G $\rightarrow$ **GL**($V$), *if* $U$ *is a subspace of* $V$ *invariant under* G, *then* $\rho^{\mathrm{U}}$ : G $\rightarrow$ **GL**($U$), *where* $\rho^{\mathrm{U}}(s)$ *is the restriction of* $\rho(s)$ *to* U, *is also a representation of* G.

Of the representation space $V$, if $U$ is a proper subspace invariant under G then on $U$ we get a representation of a smaller size than that on $V$. In view of this, we say that the original representation on $V$ is *reducible*. It may well be that $U$ is likewise reducible. If it is, then we are led to yet another representation space that is of a still smaller dimension. The process may admit of being continued and we may ultimately reach a one-dimensional representation space, a stage beyond which no further reduction is possible (leaving out the null space). On the other hand, the process may terminate earlier, ending on a representation space whose dimension is greater than one. The representation that is reached at this stage in either case is not further reducible, or is what is called an *irreducible representation*, one for which the representation space does not have a proper subspace invariant under the group being represented.

**Example 5.5.1** Let us consider as G the dyadic group of order 4, whose elements $g_1$, $g_2$, $g_3$ and $g_4$ satisfy the following Cayley table.

| $*$ | $g_1$ | $g_2$ | $g_3$ | $g_4$ |
|---|---|---|---|---|
| $g_1$ | $g_1$ | $g_2$ | $g_3$ | $g_4$ |
| $g_2$ | $g_2$ | $g_1$ | $g_4$ | $g_3$ |
| $g_3$ | $g_3$ | $g_4$ | $g_1$ | $g_2$ |
| $g_4$ | $g_4$ | $g_3$ | $g_2$ | $g_1$ |

Consider as the given representation, $\rho$ :G $\rightarrow$ **GL**($\mathbb{R}^4$), whose values are the following matrices obtained by using the procedure of Example 5.4.3.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

We treat these as matrices of $\rho(g_i)$ with respect to the usual standard basis in $\mathbb{R}^4$. Let $U$ be the subspace spanned by the basis vectors $\psi_1 = [1 \;\; 1 \;\; 1 \;\; 1]'$ and $\psi_2 = [1 \;\; 1 \;\; -1 \;\; -1]'$. We shall presently see that $U$ is invariant under G, i.e., for any $x \in U$, $\rho(g_i)x \in U$. To begin with, $\rho(g_1)x \in U$. Further,

$$\begin{aligned} \rho(g_2)x &= \rho(g_2)[\alpha_1\, \psi_1 + \alpha_2\, \psi_2] \\ &= \alpha_1 \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} \\ &= \alpha_1\psi_1 + \alpha_2\psi_2 \quad \in \text{U}. \end{aligned}$$

Likewise we find that

$$\begin{aligned} \rho(g_3)x &= [\alpha_1\, \psi_1 - \alpha_2\, \psi_2] \quad \in U \\ &\text{and} \\ \rho(g_4)x &= [\alpha_1\, \psi_1 - \alpha_2\, \psi_2] \quad \in U. \end{aligned}$$

Thus $U$ is invariant under G. The representation $\rho^U$ of G is then, with respect to the basis vectors $\psi_1$, $\psi_2$ on $U$, given by

$$\rho^U(g_1) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \rho^U(g_2)$$

$$\rho^U(g_3) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = \rho^U(g_4).$$

Observe that these two matrices are in diagonal form. This suggests that we may carry out further reduction if we choose within $U$ the subspace $W$ that is spanned by the single basis vector $\psi_1$, or alternatively the subspace $W'$ spanned by the vector $\psi_2$. Proceeding as in moving from $V$ to $U$, we get on $W$ and $W'$ the following one-dimensional representations.

| | $g_1$ | $g_2$ | $g_3$ | $g_4$ |
|---|---|---|---|---|
| $\rho^W(g_i)$ | $[+1]$ | $[+1]$ | $[+1]$ | $[+1]$ |
| $\rho^{W'}(g_i)$ | $[+1]$ | $[+1]$ | $[-1]$ | $[-1]$. |

We have thus arrived at two irreducible representations of the dyadic group of order 4.

About reducing representations the way we have just seen, one significant fact is that for a reduced representation that we get this way, we automatically get another one, as though the original one were split into two. This is so because for a given subspace $U$ of $V$ invariant under G, there is a complement of $U$ that is also invariant under G. Let us examine this fact in some detail.

Going back to Example 5.5.1, let us consider in addition to $\psi_1$ and $\psi_2$, two other vectors $\psi_3 = [1 \ \ -1 \ \ 0 \ \ 0]'$ and $\psi_4 = [0 \ \ 0 \ \ 1 \ \ -1]'$. Being linearly independent, $\psi_1$, $\psi_2$, $\psi_3$, $\psi_4$ constitute a basis of $V$. Let $U_0$ be the subspace spanned by $\psi_3$, $\psi_4$. Then any vector $x \in V$ has a unique decomposition $x = x_1 + x_2$, where $x_1 \in U$ and $x_2 \in U_0$. The subspace $U_0$ is, in other words, a complement of $U$, and $V$ is the *direct sum* of $U$ and $U_0$ ($V = U \oplus U_0$, to use the common notation). Further, since,

$$\begin{array}{ll}
\rho(g_1)\psi_3 = +\psi_3 & \rho(g_1)\psi_4 = +\psi_4 \\[1em]
\rho(g_2)\psi_3 = +\psi_3 & \rho(g_2)\psi_4 = +\psi_4 \\
\rho(g_3)\psi_3 = +\psi_3 & \rho(g_3)\psi_4 = +\psi_4 \\
\rho(g_4)\psi_3 = +\psi_3 & \rho(g_4)\psi_4 = +\psi_4
\end{array}$$

for any $x \in U_0$, $\rho(g_i)x \in U_0$. Thus $U_0$ is invariant under G, and on $U_0$ we get a reduced representation $\rho^{U_0}$ with the following values with respect to the basis $\psi_3$, $\psi_4$.

$$\begin{array}{ccccc}
 & g_1 & g_2 & g_3 & g_4 \\
\rho^{U_0}(g_i) & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} & \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} & \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}
\end{array}$$

**Exercise 5.5.1** Carry out further reduction of $\rho^{U_0}$, obtaining two one-dimensional representations of the group G of Example 5.5.1.

## 5.6 Reducibility of Representations

Let us now convince ourselves that the decomposition we obtained for the representation of Example 5.5.1 is not just a coincidence but an instance of a general possibility. Recall that a subspace $U$ of a vector space $V$ will in general have many complements, one corresponding to every way of completing the basis of $U$ to a basis of $V$. Thus if $V$ is the Euclidean plane, and $U$ is a line through the origin (i.e., $V$ is $\mathbb{R}^2$ and $U$ is a 1-dimensional subspace of $\mathbb{R}^2$), then any other line through the origin is a complement of $U$. For a complement $U'$ of $U$, a vector $x \in V$ has a unique decomposition, $x = x_1 + x_2$, such that $x_1 \in U$ and $x_2 \in U'$. We may look upon this decomposition as affected through a projection.

Recall further that a linear transformation $E$ on $V$ is a *projection* if $E^2 = E$. Given a direct-sum decomposition $V = U \oplus U'$, we can associate with it a unique transformation, $E$, defined by the condition $Ex = x_1$, where $x = x_1 + x_2$, with $x_1 \in U$ and $x_2 \in U'$. First, it is linear, and secondly, $E^2x = Ex_1 = x_1 = Ex$, and $Eu = 0$ for any $u \in U'$. Thus $E$ is a projection, whose range space is $U$ and null space is $U'$; we say it is a projection of *$V$ on $U$ along $U'$*.

Conversely, given a projection $E$, its range and null spaces $U$ and $U'$ provide a unique direct sum decomposition $V = U \oplus U'$.

**Exercise 5.6.1** Show that a linear transformation $E$ on $V$, with range space $U$ and null space $U'$, is a projection (on $U$ along $U'$) if and only if for any $x \in U$, $Ex = x$.

Now, if $\rho$ is a representation of a group G on $V$, and $U$ a subspace of $V$ invariant under G, then our decomposition problem is to establish that $U$ has a complement that is also invariant under G. We establish this by constructing one, starting with an arbitrary complement $U'$. Let $E$ be the projection of $V$ on $U$ along $U'$. Then construct the average $E_0$,

$$E_0 = \frac{1}{|G|} \sum_{s \in \mathrm{G}} \rho(s) E \, \rho(s)^{-1}. \tag{5.8}$$

$E_0$ has the important property that

(a) it is a projection of $V$ on $U$. To check that it is a projection is to check that for any $x \in U$, $E_0 x = x$. Now for any $x \in U$, since $U$ is the range of the projection $E$, and it is invariant under G, $E\rho(s)^{-1}x = \rho(s)^{-1}x$. That is, $\rho(s)E\rho(s)^{-1}x = x$, for any $x \in U$ and for every $s \in G$. Then for any $x \in U$, $E_0 x = x$.

(b) its null space $U_0$ is invariant under G. Note first that $\rho(s)E_0\rho(s)^{-1} = E_0$, i.e., $\rho(s)E_0 = E_0\rho(s)$ for every $s \in G$. Then for $x \in U_0$, the null space of $E_0$, $E_0\rho(s)x = \rho(s)E_0 x = 0$. Thus, if $x \in U_0$ then so is $\rho(s)x$ for every $s \in G$ i.e., $U_0$ is invariant under $G$.

It then follows that $V = U \oplus U_0$ is the desired decomposition for which both $U$ and $U_0$ are invariant under $G$. We have thus established the following.[2]

**Proposition 5.6.1** *Given a representation $\rho$ :* G $\to$ **GL**$(V)$*, and a subspace $U$ of $V$ invariant under* G*, there is for $U$ a complement $U_0$ that is also invariant under* G.

It follows that if we choose for $V$ a basis consisting of a basis of $U$ completed by a basis of $U_0$, then with respect to these bases for $V$, $U$ and $U_0$, the matrices of the representation $\rho$ take the block diagonal form

$$\rho(s) \;=\; \begin{bmatrix} \rho^{\mathrm{U}}(s) & 0 \\ 0 & \rho^{\mathrm{U_0}}(s) \end{bmatrix}, s \in \mathrm{G} \tag{5.9}$$

where $\rho^{\mathrm{U}}(s)$ and $\rho^{\mathrm{U_0}}(s)$ here denote the matrices of the restrictions of $\rho(s)$ to $U$ and $U_0$ respectively. Keeping this in mind, $\rho$ is called the *direct sum* of the representations $\rho^{\mathrm{U}}$ and $\rho^{\mathrm{U_0}}$.

Now, what we did to the representation $\rho$, we may in turn do to $\rho^{\mathrm{U}}$ and $\rho^{\mathrm{U_0}}$, and unless they are irreducible, further split each into two. Repeating this exercise, we reach a stage when no further splitting of this kind is possible. For the representation space, $V$, we thus finally get a direct sum decomposition $V = U_1 \oplus U_2 \oplus \cdots \oplus U_m$, in which each $U_i$, called an *irreducible subspace*, is invariant under G, serving as the representation space of an irreducible representation. Organizing this line of argument into a proof by induction, we are led to the following result.

[2] I have followed here the approach given in Serre [8, Section 1.3, pp. 5–7].

**Proposition 5.6.2** *Every representation is a direct sum of irreducible representations.*

When decomposed into a direct sum of irreducible representations, the block diagonalized matrices of a representation are said to be in the *completely reduced form.*

**Example 5.6.1** To highlight the finer points of this result, let us now consider as an example the group G of order 6, with the following multiplication table.

| $*$ | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ | $g_6$ |
|---|---|---|---|---|---|---|
| $g_1$ | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ | $g_6$ |
| $g_2$ | $g_2$ | $g_3$ | $g_1$ | $g_5$ | $g_6$ | $g_4$ |
| $g_3$ | $g_3$ | $g_1$ | $g_2$ | $g_6$ | $g_4$ | $g_5$ |
| $g_4$ | $g_4$ | $g_6$ | $g_5$ | $g_1$ | $g_3$ | $g_2$ |
| $g_5$ | $g_5$ | $g_4$ | $g_6$ | $g_2$ | $g_1$ | $g_3$ |
| $g_6$ | $g_6$ | $g_5$ | $g_4$ | $g_3$ | $g_2$ | $g_1$ |

Consider as the given a representation, $\rho : \mathrm{G} \longrightarrow \mathbf{GL}(\mathbb{R}^3)$, whose values $\rho(g_i)$, with respect to the standard basis on $\mathbb{R}^3$, are the following matrices $\mathrm{P}_i$,

$$P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad P_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad P_3 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$P_4 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad P_5 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad P_6 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Let $U$ denote the subspace spanned by the vectors

$$\psi_1 = [0 \;\; 1/\sqrt{2} \;\; -1/\sqrt{2}]' \text{ and } \psi_2 = [\sqrt{2/3} \;\; -1/\sqrt{6} \;\; -1/\sqrt{6}]'.$$

You may check that $U$ is invariant under G. We are to obtain a complement of $U$ that is invariant under G.

To start with, we take an arbitrary complement $U'$, spanned by the vector $\phi_3 = [1 \;\; 2 \;\; 3]'$, which is independent of $\psi_1$ and $\psi_2$. $U'$ is not invariant under G, as can be verified by examining $P_i\phi_3$. With respect to the basis $\{\psi_1, \psi_2, \phi_3\}$, the projection $\bar{E}$ of $\mathbb{R}^3$ on $U$ along $U'$ has the matrix,

$$\bar{E} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and with respect to the standard basis, it is

$$E = \begin{bmatrix} 5/6 & -1/6 & -1/6 \\ -1/3 & 2/3 & -1/3 \\ -1/2 & -1/2 & 1/2 \end{bmatrix},$$

where $E = P\bar{E}P^{-1}$, and $P$ is the matrix consisting of $\psi_1$, $\psi_2$, and $\phi_3$ as columns (in that order).

Then the mean $E_0$ defined in (5.8) is in this case

$$E_0 = \begin{bmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{bmatrix}.$$

By inspection, we find that its null space $U_0$ is the subspace spanned by the (normalized) vector $\psi_3 = (1/\sqrt{3})[1 \;\; 1 \;\; 1]'$. For $U$, $U_0$ is the desired complement that is invariant under G. Completing the basis of $U$ by the basis of $U_0$, we get for $\mathbb{R}^3$ the basis $\{\psi_1, \psi_2, \psi_3\}$. With respect to this basis the matrices of the representation $\rho$ become:

$$\bar{P}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \bar{P}_2 = \begin{bmatrix} -1/2 & -\sqrt{3}/2 & 0 \\ \sqrt{3}/2 & -1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \bar{P}_3 = \begin{bmatrix} -1/2 & \sqrt{3}/2 & 0 \\ -\sqrt{3}/2 & -1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\bar{P}_4 = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \bar{P}_5 = \begin{bmatrix} 1/2 & -\sqrt{3}/2 & 0 \\ -\sqrt{3}/2 & -1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \bar{P}_6 = \begin{bmatrix} 1/2 & \sqrt{3}/2 & 0 \\ \sqrt{3}/2 & -1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where $\bar{\mathrm{P}}_i = \Psi^{-1}P_i\Psi$, and $\Psi$ is the matrix with $\psi_1$, $\psi_2$ and $\psi_3$ as its columns from the left.

Note that these matrices are in a reduced block diagonalized form, their corresponding blocks being of the same size. Thus each set of blocks has the same multiplication table as the original given matrices, and is a matrix representation of the group G. As we shall see later, the upper blocks constitute an irreducible representation of G, and therefore no further reduction is possible for them.

**Exercise 5.6.2** In the example just given, choose some other basis vector $\phi_3$ and verify that it leads to the same reduction.

In purely matrix terms, Proposition 5.6.2 means that matrices representing a group can be simultaneously reduced by a similarity transformation to a block diagonal form such that no further reduction of the blocks is possible in this manner. These blocks are matrices of irreducible representations. It is these irreducible representations that are the 'atomic' ones I mentioned earlier. What are they, and how many of them (nonisomorphic ones) are there in all for a particular group?

Questions like these, and many others about representations in general, are best answered by first examining some basic facts about relationships between any two irreducible representations of a group. The main tool for this is an important result of linear algebra, known as Schur's lemma, that deals with two (finite dimensional) vector spaces and sets of operators on these spaces.

## 5.7 Schur's Lemma and the Orthogonality Theorem

Consider two irreducible representations $\rho^1 : \mathrm{G} \to \mathbf{GL}(V_1)$ and $\rho^2 : \mathrm{G} \to \mathbf{GL}(V_2)$ of a finite group G. If they are inequivalent, there is, by definition, no one-to-one linear map, $h$, from $V_1$ onto $V_2$, satisfying (5.7); in fact $V_1$ and $V_2$ may not even be isomorphic i.e.there is no one-to-one onto map from $\mathrm{V}_1$ to $V_2$ to start with. There may nevertheless be a one-to-one linear map, $h$, from $V_1$ into $V_2$ that satisfies an altered form of identity (5.7):

$$h\rho^1(s) \quad = \quad \rho^2(s)h \qquad \text{for all } s \in \mathrm{G} \tag{5.10}$$

which is admissible irrespective of whether the two representations are equivalent or inequivalent. We may depict this identity by the following commutative diagram.

$$\begin{tikzcd}
\mathrm{V}_1 \arrow[r, "h"] \arrow[d, "\rho^1(s)"'] & \mathrm{V}_2 \arrow[d, "\rho^2(s)"] \\
\mathrm{V}_1 \arrow[r, "h"'] & \mathrm{V}_2
\end{tikzcd}$$

The map $h$ may be looked upon as inducing a homomorphism from the structure $(V_1, \rho^1(s))$ to $(V_2, \rho^2(s))$, preserving the unary operations $\rho^i(s)$ for all $s \in G$.

Consider now any (nonzero) linear map $g$ from $V_1$ to $V_2$, and construct from it the map

$$g_0 \quad = \quad \frac{1}{|G|}\sum_{s\in \mathrm{G}} \left(\rho^2(s)\right)^{-1} g\, \rho^1(s). \tag{5.11}$$

Then we find that $g_0$ satisfies (5.10):

$$\begin{aligned}
\left(\rho^2(r)\right)^{-1} g_0\, \rho^1(r) \quad &= \quad \frac{1}{|G|}\sum_{s\in \mathrm{G}} \left(\rho^2(r)\right)^{-1}\left(\rho^2(s)\right)^{-1} g\, \rho^1(s)\rho^1(r) \\
&= \quad \frac{1}{|G|}\sum_{s\in \mathrm{G}} \left(\rho^2(sr)\right)^{-1} g\, \rho^1(sr) \\
&= \quad g_0,
\end{aligned}$$

i.e., $g_0\, \rho^1(s) = \rho^2(s)\, g_0$ for all $s \in \mathrm{G}$.

As we shall presently see, (5.10) places on $h$ a strong constraint which, when considered for $g_0$, leads us to important results about how $\rho^1$ and $\rho^2$ are related. Let us first see what the constraint is.

**Proposition 5.7.1 (Schur's lemma)** *Let $V_1$ and $V_2$ be finite-dimensional vector spaces, and let $\{A_1, A_2, \ldots, A_n\}$ and $\{B_1, B_2, \ldots, B_n\}$ be two sets of linear operators on $V_1$ and $V_2$ respectively, such that $V_1$ has no proper subspaces invariant under $A_i$ and likewise $V_2$ has no proper subspaces invariant under $B_i$ for all $i = 1, 2, \ldots, n$. If $h$ is a linear map from $V_1$ to $V_2$ satisfying the condition $hA_i = B_i h$ for every $A_i$ and $B_i$, then either $h = 0$ or, $h$ is invertible $(i.e.$, one-to-one and onto).*[3]

The proof is as follows. Given that $h\ A_i = B_i\ h$, $h = 0$ is of course one possibility, but what if $h \neq 0$?

Let $W_1$ be the set of all $x \in V_1$ such that $hx = 0$. (Note that $W_1$, which is called the *kernel* of $h$ and commonly written $\ker(h)$, is a subspace of $V_1$. Then,

$$h\ A_i x = B_i\ hx = 0 \text{ for } i = 1, 2, \cdots, n.$$

Thus if $x$ is in $W_1$ then $A_i x$ is also in $W_1$ i.e., $W_1$ is a subspace of $V_1$ invariant under all $A_i$. But by hypothesis, $V_1$ has no proper subspaces invariant under $A_i$. So, either $W_1 = 0$ or $W_1 = V_1$. The latter is ruled out since $h \neq 0$. It follows that $W_1 = 0$. Now the kernel of a linear map is 0 iff the map is one-to-one. So the map $h$ is one-to-one.

Next, let us look at $W_2$, the image of $h$, which is a subspace of $V_2$. For any $y \in W_2$, since $h$ is one-to-one, there is a unique $x \in V_1$ for which $hx = y$. Then,

$$B_i y = B_i\ hx = h\ A_i x \in W_2.$$

So $W_2$ is invariant under all $B_i$. But again, since there are no such proper subspaces of $V_2$, either $W_2 = 0$ or $W_2 = V_2$. The former is not true because we have assumed that $h \neq 0$. That means $W_2 = V_2$, i.e., the linear map $h$ is onto. Thus, on the whole, if $h \neq 0$ then it is invertible. This completes the proof of the proposition.

**Corollary** *If in the above proposition, $V_1 = V_2$ and $A_i = B_i$ then $h$ is a multiple of the identity operator $I$ on $V_1$ (or $V_2$).*

To see this, note that in this case $(h - cI)A_i = A_i\ (h - cI)$ for any complex number $c$. Then by Schur's lemma, either $(h - cI) = 0$ or $(h - cI)$ is invertible. But $h$ has at least one complex eigenvalue, say $\lambda$, so that $(h - \lambda I)$ does not have an inverse. Thus $(h - \lambda I) = 0$, or equivalently, $h = \lambda I$, confirming the corollary.

Coming back to condition (5.10), it immediately follows in the light of Schur's lemma that *either* $\rho^1$ and $\rho^2$ are equivalent representations and $h$ is invertible, *or* if

[3] This version of the lemma is taken from Greub [1, p. 54]. It points to the fact that for the result to hold the two sets of operators need not be groups.

they are inequivalent then $h = 0$. Thus we find that for inequivalent representations $\rho^1$ and $\rho^2$,

$$g_0 = \frac{1}{|G|} \sum_{s \in \mathrm{G}} \left(\rho^2(s)\right)^{-1} g\, \rho^1(s) = 0, \tag{5.12}$$

where $g$ is any linear map from $V_1$ to $V_2$.

On the other hand, if $\rho^1$ and $\rho^2$ are the same representations, i.e., $V_1 = V_2$ and $\rho^1(s) = \rho^2(s)$ then using the corollary, we find that

$$g_0 = \frac{1}{|G|} \sum_{s \in \mathrm{G}} \left(\rho^1(s)\right)^{-1} g\, \rho^1(s) = \lambda I, \tag{5.13}$$

where $g$ is any linear map on $V_1$, $I$ the identity operator on it, and $\lambda$ is a real or complex number that depends on $h$.

Let us now examine the implications of conditions (5.12) and (5.13) for the matrices of the representations $\rho^1$ and $\rho^2$, for an arbitrary choice of bases on $V_1$ and $V_2$. Considering the matrix form of condition (5.12), and noting that $\left(\rho^2(s)\right)^{-1} = \rho^2(s^{-1})$, we get for the elements of the matrix of $g_0$ the following relation:

$$(g_0)_{il} = \frac{1}{|\mathrm{G}|} \sum_{p,q} \left[ \sum_{s \in \mathrm{G}} {\rho^2}_{ip}(s^{-1})\, {\rho^1}_{ql}(s) \right] g_{pq} = 0.$$

This is true for any arbitrary $g$, and therefore the coefficients of $g_{pq}$ are identically zero. In other words,

$$\frac{1}{|\mathrm{G}|} \sum_{s \in \mathrm{G}} {\rho^2}_{ip}(s^{-1})\, {\rho^1}_{ql}(s) = 0, \tag{5.14}$$

for every admissible $i$, $p$, $q$ and $l$.

As for condition (5.13) in its matrix form, we first observe that[4]

$$\lambda = \frac{1}{n}\mathrm{Tr}(g_0) = \frac{1}{n}\mathrm{Tr}(g) = \frac{1}{n}\sum_{p,q} \delta_{pq}\, g_{pq}$$

where $n$ is the dimension of $V_1$, and $\delta_{pq}$ is the Kronecker delta ($= 1$ if $p = q$ and zero otherwise). This follows if we look at the trace of $g_0$, and use the fact that the trace is distributive over addition and invariant under similarity transformation. Then for the elements of $g_0$ in condition (5.13), we get the identity,

$$(g_0)_{il} = \frac{1}{|\mathrm{G}|} \sum_{p,q} \left[ \sum_{s \in \mathrm{G}} {\rho^1}_{ip}(s^{-1})\, {\rho^1}_{ql}(s) \right] g_{pq} = \frac{1}{n} \sum_{p,q} \delta_{il}\, \delta_{pq}\, g_{pq}.$$

[4] $\mathrm{Tr}(A)$ denotes the trace of the matrix $A$.

Again, since $g$ is arbitrary, equating coefficients of $g_{pq}$, we finally obtain the relation,

$$\frac{1}{|\mathrm{G}|}\sum_{s\in \mathrm{G}} \rho^1{}_{ip}(s^{-1})\, \rho^1{}_{ql}(s) = \frac{1}{n}\delta_{il}\,\delta_{pq}. \tag{5.15}$$

If we add the condition that the representation matrices are unitary, then identities (5.14) and (5.15) may be combined in the form of the following result.

**Proposition 5.7.2 (Orthogonality Theorem)** *For any two irreducible representations $\rho^j$ and $\rho^k$ of a finite group* G *of order $m$, the elements of their matrices in unitary form satisfy the following orthogonality relations:*

$$\sum_{s\in \mathrm{G}} \rho^j{}_{pi}(s)^*\, \rho^k{}_{ql}(s) \;=\; \frac{m}{\sqrt{n_j\, n_k}}\,\delta_{jk}\,\delta_{il}\,\delta_{pq}, \tag{5.16}$$

*where $n_j$ and $n_k$ are the dimensions of the representation spaces of $\rho^j$ and $\rho^k$ respectively.*

**Remark**: In the identity (5.16), the constant $\sqrt{n_j\, n_k}$ is simply a "dressing" that makes the formula look symmetrical; keeping the $\delta's$ in mind, it has a role only when the two representations are the same, i.e., $j = k$ and $n_j = n_k$.

One helpful way to look at this result is as follows. Consider one of the representations and think of the elements in the same position in each of its $m$ (unitary) matrices as constituting an $m$-dimensional vector. Do the same for the other representation. Then the $n_j^2 + n_k^2$ vectors that we get this way are mutually orthogonal.

As we shall soon see, this orthogonality property leads us to several important conclusions about irreducible representations. But before we turn to these, there is yet another important notion concerning representations that we need to bring in here—that of characters.

## 5.8 Characters and Their Properties

As noted earlier, the trace of a matrix is invariant under similarity transformations, i.e., $\mathrm{Tr}(A) = \mathrm{Tr}(B^{-1}AB)$. For a linear transformation $t$ on a vector space V, this allows us, although perhaps in a 'backhanded manner', to talk of its trace, $\mathrm{Tr}(t)$, as the sum of the diagonal elements of its matrix with respect to some basis in V; it is a function of $t$, independent of the basis we choose. Extending this idea to group representations, we say that for a representation $\rho$ of a group G, its *character* is a complex valued function, $\chi$, over G, with values $\chi(s) = \mathrm{Tr}(\rho(s)), s \in$ G. When we have more than one representation to deal with at the same time, we index the characters the same way as the representations. Thus for $\rho^1$ and $\rho^2$, we write their characters as $\chi_1$ and $\chi_2$ respectively. The characters of irreducible representations we call *irreducible characters*.

**Example 5.8.1** Refer to the group G of order 6 and its representation $\rho$ given in example 6.1. Its character $\chi$ has the following values.

| $g_i$ | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ | $g_6$ |
|---|---|---|---|---|---|---|
| $\chi(g_i)$ | 3 | 0 | 0 | 1 | 1 | 1 |

**Aside 3** The term "character", incidentally, has some interesting history. Introduced by Gauss around 1801 in the course of his studies on quadratic forms, it started its journey towards its modern meaning and definition through the works of Dirichilet, Dedekind and Weber on abelian groups; the chi ($\chi$) notation is due to Dirichilet. It was in the theory of group representations initiated in 1896 by Frobenius that its scope was extended to nonabelian groups. For fuller details, see Hawkins [3] and Williams [10]. ♡

The remarkable thing about characters, specifically about those of irreducible representations, is that we can very often compute them easily even without knowing concretely what the representations are; for more on this, see Hall [2, Chapters 3 and 5]. Furthermore, they provide us a great deal of advance information about these representations. Consequently, they serve as very useful tools in constructing irreducible matrix representations. Let us now look at some of the properties of characters that are important from this angle.

To begin with, we have these three basic facts about the values of characters.

**Proposition 5.8.1** *For any representation of a group G. its character has the value $\chi(e) = n$, where $e$ is the identity element of the group G, and $n$ is the dimension of the representation space.*

**Proposition 5.8.2** *For any $r, s \in$ G, $\chi(r^{-1}sr) = \chi(s)$. Equivalently, $\chi(sr) = \chi(rs)$. In words, character values for elements of the same class of the group are the same.*

A complex valued function $f$ on a group G is called a $class\ function$ if it satisfies the condition $f(rs) = f(sr)$ for any $r, s \in$G. Class functions on a group constitute a vector space over the complex number field. Characters are distinguished members of this vector space. More about this later.

**Proposition 5.8.3** *For any $s \in$ G, $\chi(s^{-1}) = \chi(s)^*$.*

To check the first and second propositions, just focus on some matrix form of a representation and use the invariance properties of traces. As for the third, it follows from the fact that for any representation $\rho$, its matrices have a unitary form in which the elements satisfy the condition $\rho_{ij}(s^{-1}) = \rho_{ji}(s)^*$.

Amongst representations, the two key relationships are those of equivalence and direct sum. These translate into simple relationships amongst their characters.

**Proposition 5.8.4** *Two representations of a group are equivalent if and only their characters are the same.*

That is, the representations are uniquely determined, to within equivalences, by the characters.

**Proposition 5.8.5** *If a representation $\rho$ is the direct sum of representations $\rho^1$ and $\rho^2$, with characters $\chi$, $\chi_1$ and $\chi_2$ respectively, then $\chi = \chi_1 + \chi_2$.*

Now, every representation is a direct sum of irreducible representations, with some of them possibly repeated and others altogether absent. If $\chi$ is its character, and $\chi_1$, $\chi_2$, ... are the characters of the irreducible representations, then Proposition 5.8.5 implies that

$$\chi = n_1\chi_1 + n_2\chi_2 + \cdots \ ,$$

where $n_i$ is the number of times the $i$th irreducible representation is repeated in the direct sum, treating its absence as repeating zero times.

Characters exhibit important orthogonality properties with respect to the *scalar product*, $(\phi, \psi)$, of complex valued functions $\phi$ and $\psi$ over a group G, defined as follows.

$$(\phi, \psi) = \frac{1}{|\mathrm{G}|}\sum_{g\in \mathrm{G}} \phi(g)\psi(g)^*.$$

**Proposition 5.8.6** *If $\chi_1$ and $\chi_2$ are two irreducible characters of a group G, then they satisfy the orthogonality relation,*

$$(\chi_1, \chi_2) = \begin{cases} 1, & \textit{if the representations are equivalent;} \\ 0, & \textit{if they are inequivalent.} \end{cases}$$

**Proposition 5.8.7** *If $\chi$ is the character of a representation of a group G then $(\chi, \chi)$ is a positive integer; it equals 1 if and only if the representation is irreducible.*

**Proposition 5.8.8** *The vector space H of class functions on a group G is of dimension equal to the number of classes of G. Furthermore, the characters of the irreducible representations of G form an orthonormal basis of H.*

Combining the two parts of this proposition, we come to this conclusion about irreducible representations.

**Proposition 5.8.9** *The number of inequivalent irreducible representations of a group is equal to the number of its classes.*

Proposition 5.8.8 tells us that a class function,$\psi$, has a 'Fourier' expansion in terms of the irreducible characters:

$$\psi = \sum_{i=1}^{k} a_i \chi_i ; \quad a_i = (\,\psi, \chi_i)$$

where the group has $k$ irreducible representations and the $\chi_i$'s are their characters.

In particular, the character of any representation has such an expansion. In view of Proposition 5.8.5, the coefficient $a_i$ is simply the number of times the i-th irreducible representation is repeated in its completely reduced form. Thus from the character, we are able to know the structure of the completely reduced form of a representation. In order to fully determine this form, we need to do some more work that will be explained later in Section 5.10.

About the dimensions of irreducible representations, we can say the following.

**Proposition 5.8.10** *If a group G has $k$ inequivalent irreducible representations of dimensions $l_1, \ldots, l_k$, then,*

$$\sum_{i=1}^{k} {l_i}^2 = |G|.$$

For an abelian group, each member forms a class by itself, making the number of classes equal to its order. Thus Propositions 5.8.9 and 5.8.10 imply that, for an abelian group, the number of its inequivalent irreducible representations equals its order, and each of these representations is of dimension 1. Indeed, the characters themselves are in this case synonymous with the irreducible matrix representations.

Whether an irreducible representation has a real matrix form or not is easily determined using the following result.

**Proposition 5.8.11** *Jensen and Boon* [4, p. 130] *An irreducible representation of a group G has a real matrix form if and only if its character $\chi$ satisfies the condition,*

$$\sum_{g \in \mathrm{G}} \chi(s^2) = |G|.$$

Lastly, take note of this important feature of regular representations.

**Proposition 5.8.12** *The regular representation of a group contains in its completely reduced form every irreducible representation of the group, each repeated a number of times equal to its dimension.*

All these results have their roots in the ideas leading to the orthogonality theorem. Simple and powerful as they are, they tend to evoke a sense of wonder and disbelief when you first encounter them. The groundwork done in the previous section should help soften the disbelief part. In any case, if doubts hold you back from going ahead, see Jensen and Boon [4, Chapter 2], Serre [8, Chapter 2], Lomont [6, Chapter 2] for detailed proofs. For elaborate illustrations, see Tinkham [9, Chapter 3].

A fine point about the order in which these results are presented needs to be mentioned. One usually presents theorems and propositions in such an order that their proofs draw upon the ones that have already appeared but not on those that are to come later. Thus Proposition 5.8.12 should on this count appear before Proposition 5.8.9, as you will see if you check the proofs. With the proofs left out, they are for the present purposes better arranged as they are.

As I mentioned towards the end of Section 5.2, the purpose of moving over from matrices to **GL**(V) was to avail of the advantage that linear algebra offers in studying the properties of group representations. Now that we have a good idea of the various important properties of representations, we revert to their matrix form which is what we ultimately use in practical applications.

## 5.9 Constructing Irreducible Representations

Let us now put to use the understanding gained so far to see how irreducible matrix representations (*irreps*, for short) can be determined. For all groups of practical importance the irreps are already available in the literature. So from the point of view of applications, we do not need to construct them afresh. It is, however, instructive to see what steps are involved in their construction. This is what we do now. We follow an informal approach, half guessing and half checking for the conditions that the irreps and their characters uniquely satisfy.

The basic steps that we go through are the following.

1. Obtain the Cayley table; for the symmetries of a geometrical object, do this by actually performing the symmetry operations on the object and noting down how the labels defining the object (the labels of the vertices of an equilateral triangle or a square, for example) permute.

2. Working with the Cayley table, determine the classes of the group. The number of classes is also the number of irreps of the group (See Proposition 5.8.9).

3. Determine the dimensions of the irreps using Proposition 5.8.10.

4. Determine the characters for the irreps. Use their orthogonality properties and the fact that their values for elements of the same class are the same.

5. The irreps may either be real or complex. Check whether you are in luck, and all the irreps are real (Proposition 5.8.11).

6. For the one-dimensional irreps, stop here; the character values, read off as $1 \times 1$ matrices are the desired one-dimensional irreps.

7. For the remaining irreps, do some more intelligent guess work. Go back to the Cayley table, and start with the entries for one of the irrep matrices whose square is the identity. Without any loss of generality, impose the condition that the irrep matrices are unitary (normal, if real). This should enable you to make a first guess at the irrep matrix. Try guesses like this for the others, exploiting every time the constraints that the Cayley table puts on their products. Remember that there is a great deal of redundancy in the constraints that the Cayley table imposes, in the sense that you need not check for all of them. As soon as the point is reached when you have used a certain set of constraints to determine all the matrices of the irrep, the other constraints will automatically be satisfied. All the same, once you have all the matrices, do finally check whether they do.

To try out these steps, let us now consider as an example the group of symmetry operations on a square.

**Example 5.9.1** Consider the group of symmetry operations of the square $abcd$ shown in Figure 5.1. Let us construct the irreducible representations of this group.

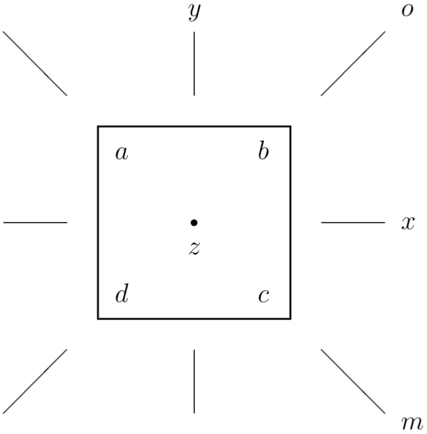

Figure 5.1 Square $abcd$ and its axes of symmetry

As labels for the symmetry operations, let us use $\mathrm{Q}_e$ to denote the identity operation, $\mathrm{Q}_1$, $\mathrm{Q}_2$, $\mathrm{Q}_3$ for clockwise rotations in steps of 90°, and $\mathrm{Q}_x$, $\mathrm{Q}_y$, $\mathrm{Q}_o$ and $\mathrm{Q}_m$ for the reflections along the $x, y, o$ and $m$ axes respectively. Let $\mathrm{G}_r$ be the abstract group isomorphic to the group of these symmetry operations, with elements $g_1, g_2, \ldots, g_8$ corresponding to $\mathrm{Q}_e$, $\mathrm{Q}_x$, $\mathrm{Q}_y$, $\mathrm{Q}_1$, $\mathrm{Q}_2$, $\mathrm{Q}_3$, $\mathrm{Q}_o$, $\mathrm{Q}_m$ respectively.

*Cayley Table*: First, generate the Cayley table for the symmetry operations. A simple way to do this is to cut out a cardboard square, label it as shown in Figure 5.1, and then to determine the table entries by physically checking how the symmetry operations combine when applied one after another.

The table that is obtained by carrying out these operations on the card board square is shown in Table 5.1. Note that the entry in the row $i$ and column $j$ is the operation $\mathrm{Q}_i\mathrm{Q}_j$, which is the result of first applying $\mathrm{Q}_j$ and then $\mathrm{Q}_i$ to the square. For the abstract group $\mathrm{G}_r$, the table is obtained by simply replacing the $\mathrm{Q}_i$'s by the respective $g_i$'s.

| | $\mathrm{Q}_e$ | $\mathrm{Q}_x$ | $\mathrm{Q}_y$ | $\mathrm{Q}_1$ | $\mathrm{Q}_2$ | $\mathrm{Q}_3$ | $\mathrm{Q}_o$ | $\mathrm{Q}_m$ |
|---|---|---|---|---|---|---|---|---|
| $\mathrm{Q}_e$ | $\mathrm{Q}_e$ | $\mathrm{Q}_x$ | $\mathrm{Q}_y$ | $\mathrm{Q}_1$ | $\mathrm{Q}_2$ | $\mathrm{Q}_3$ | $\mathrm{Q}_o$ | $\mathrm{Q}_m$ |
| $\mathrm{Q}_\mathrm{x}$ | $\mathrm{Q}_x$ | $\mathrm{Q}_e$ | $\mathrm{Q}_2$ | $\mathrm{Q}_o$ | $\mathrm{Q}_y$ | $\mathrm{Q}_m$ | $\mathrm{Q}_1$ | $\mathrm{Q}_3$ |
| $\mathrm{Q}_\mathrm{y}$ | $\mathrm{Q}_y$ | $\mathrm{Q}_2$ | $\mathrm{Q}_e$ | $\mathrm{Q}_m$ | $\mathrm{Q}_x$ | $\mathrm{Q}_o$ | $\mathrm{Q}_3$ | $\mathrm{Q}_1$ |
| $\mathrm{Q}_1$ | $\mathrm{Q}_1$ | $\mathrm{Q}_m$ | $\mathrm{Q}_o$ | $\mathrm{Q}_2$ | $\mathrm{Q}_3$ | $\mathrm{Q}_e$ | $\mathrm{Q}_x$ | $\mathrm{Q}_y$ |
| $\mathrm{Q}_2$ | $\mathrm{Q}_2$ | $\mathrm{Q}_y$ | $\mathrm{Q}_x$ | $\mathrm{Q}_3$ | $\mathrm{Q}_e$ | $\mathrm{Q}_1$ | $\mathrm{Q}_m$ | $\mathrm{Q}_o$ |
| $\mathrm{Q}_3$ | $\mathrm{Q}_3$ | $\mathrm{Q}_o$ | $\mathrm{Q}_m$ | $\mathrm{Q}_e$ | $\mathrm{Q}_1$ | $\mathrm{Q}_2$ | $\mathrm{Q}_y$ | $\mathrm{Q}_x$ |
| $\mathrm{Q}_o$ | $\mathrm{Q}_o$ | $\mathrm{Q}_3$ | $\mathrm{Q}_1$ | $\mathrm{Q}_y$ | $\mathrm{Q}_m$ | $\mathrm{Q}_x$ | $\mathrm{Q}_e$ | $\mathrm{Q}_2$ |
| $\mathrm{Q}_m$ | $\mathrm{Q}_m$ | $\mathrm{Q}_1$ | $\mathrm{Q}_3$ | $\mathrm{Q}_x$ | $\mathrm{Q}_o$ | $\mathrm{Q}_y$ | $\mathrm{Q}_2$ | $\mathrm{Q}_e$ |

Table 5.1 Cayley Table for the symmetry operations

*The Number of Irreps:* Determine now the classes of this group. Recall that the class of an element $y$ of a group G is the set $\{\, x \mid x = u^{-1}yu, u \in \mathrm{G}\,\}$. There are five classes in this case. Taking up the elements one by one, we find that the classes are, $\{\,\mathrm{Q}_e\,\}$, $\{\,\mathrm{Q}_2\,\}$, $\{\,\mathrm{Q}_x, \mathrm{Q}_y\,\}$, $\{\,\mathrm{Q}_1, \mathrm{Q}_3\,\}$ and $\{\,\mathrm{Q}_o, \mathrm{Q}_m\,\}$. The number of irreps for the group G is therefore 5.

*The Dimensions of Irreps:* Having determined the Cayley table and the classes with the help of the group of symmetry operations of the square, we now turn to the abstract group $\mathrm{G}_r$. Call its irreps $\rho^1, \rho^2, \ldots, \rho^5$, their dimensions $l_1, l_2, \ldots, l_5$, and their characters $\chi_1, \chi_2, \ldots, \chi_5$ respectively. Then, since the order of the group is 8, we have,

$$l_1{}^2 + l_2{}^2 + l_3{}^2 + l_4{}^2 + l_5{}^2 = 8.$$

This equality is satisfied uniquely by the $l_i$'s, to within reordering of their indices. We may treat the solution to be $l_1 = l_2 = l_3 = l_4 = 1$, and $l_5 = 2$.

*Characters:* We know from Proposition 5.8.7 that the sum of the squares of the moduli of character values of an irrep equals the order of the group. So the values of the characters of each of the four one-dimensional irreps are either $+1$ or $-1$, and likewise the character values for the one two-dimensional irrep are 2,$-2$ or 0. Further, character values for members of the same class are the same. Trials with these constraints in mind lead us to the unique solutions for the characters as shown in Table 5.2.

| $\chi_i$ | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ | $g_6$ | $g_7$ | $g_8$ |
|---|---|---|---|---|---|---|---|---|
| $\chi_1$ | $+1$ | $+1$ | $+1$ | $+1$ | $+1$ | $+1$ | $+1$ | $+1$ |
| $\chi_2$ | $+1$ | $+1$ | $+1$ | $-1$ | $+1$ | $-1$ | $-1$ | $-1$ |
| $\chi_3$ | $+1$ | $-1$ | $-1$ | $-1$ | $+1$ | $-1$ | $+1$ | $+1$ |
| $\chi_4$ | $+1$ | $-1$ | $-1$ | $+1$ | $+1$ | $+1$ | $-1$ | $-1$ |
| $\chi_5$ | $+2$ | $0$ | $0$ | $0$ | $-2$ | $0$ | $0$ | $0$ |

Table 5.2 Characters of irreps of the group $\mathrm{G}_r$

*The Irreps:* The four one-dimensional irreps are given by the characters. For the two-dimensional irrep, let us first apply the realness test (Proposition 5.8.11). Looking up the Cayley table, we get the squares of the group elements and the corresponding character values as shown in Table 5.3.

| $g_i$ | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ | $g_6$ | $g_7$ | $g_8$ |
|---|---|---|---|---|---|---|---|---|
| ${g_i}^2$ | $g_1$ | $g_1$ | $g_1$ | $g_5$ | $g_1$ | $g_5$ | $g_1$ | $g_1$ |
| $\chi_5({g_i}^2)$ | $+2$ | $+2$ | $+2$ | $-2$ | $+2$ | $-2$ | $+2$ | $+2$ |

Table 5.3 Test for realness of the two-dimensional irrep.

The last row in this table adds to 8, the order of the group. The two-dimensional representation is therefore real. Moreover, the general unitary requirement for irreps makes the irrep matrices normal in this case, i.e., the transpose of $\rho^5(g_i)$ equals its inverse. One could get all these matrices by alternately guessing and checking with the Cayley table, but we have an easier way out through the rotational and reflection symmetry operations of the square treated as those of the plane. More specifically, we have a representation of this group on $\mathbb{R}^2$, which has the following matrices with respect to the usual standard basis on $\mathbb{R}^2$.

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$$

All these matrices are normal. Furthermore, the representation that they constitute satisfies the criterion for irreducibility given in Proposition 5.8.8. But we know that there is just one two-dimensional irrep in this case. So these matrices are indeed the matrices of the two-dimensional representation $\rho^5$. Table 5.4 shows all the irreps together.

| | $g_1$ | $g_2$ | $g_3$ | $g_4$ | $g_5$ | $g_6$ | $g_7$ | $g_8$ |
|---|---|---|---|---|---|---|---|---|
| $\rho^1$ | $[+1]$ | $[+1]$ | $[+1]$ | $[+1]$ | $[+1]$ | $[+1]$ | $[+1]$ | $[+1]$ |
| $\rho^2$ | $[+1]$ | $[+1]$ | $[+1]$ | $[-1]$ | $[+1]$ | $[-1]$ | $[-1]$ | $[-1]$ |
| $\rho^3$ | $[+1]$ | $[-1]$ | $[-1]$ | $[-1]$ | $[+1]$ | $[-1]$ | $[+1]$ | $[+1]$ |
| $\rho^4$ | $[+1]$ | $[+1]$ | $[-1]$ | $[-1]$ | $[+1]$ | $[+1]$ | $[-1]$ | $[-1]$ |
| $\rho^5$ | $\begin{bmatrix}+1 & 0\\ 0 & +1\end{bmatrix}$ | $\begin{bmatrix}+1 & 0\\ 0 & -1\end{bmatrix}$ | $\begin{bmatrix}-1 & 0\\ 0 & +1\end{bmatrix}$ | $\begin{bmatrix}0 & +1\\ -1 & 0\end{bmatrix}$ | $\begin{bmatrix}-1 & 0\\ 0 & -1\end{bmatrix}$ | $\begin{bmatrix}0 & -1\\ +1 & 0\end{bmatrix}$ | $\begin{bmatrix}0 & +1\\ +1 & 0\end{bmatrix}$ | $\begin{bmatrix}0 & -1\\ -1 & 0\end{bmatrix}$ |

Table 5.4 Irreps for the group of symmetry operations of a square

**Exercise 5.9.1** For the symmetries of an equilateral triangle, and a regular hexagon, carry out a similar exercise.

## 5.10 Complete Reduction of Representations

We now come back to representations in general and their direct sum decomposition into irreducible representations. For our purposes, it is good enough to consider representations on $\mathbb{F}^n$, where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$.

Let us say that we have a representation $\rho : \mathrm{G} \rightarrow \mathbf{GL}(V)$ of a group G. We already know from the discussions following Proposition 5.8.9 what the structure of its completely reduced form is—which irreducible representations appear (to within isomorphisms) in it and how many times. Given the matrices of the representation with respect to some basis in V, we are interested in determining a similarity transformation that brings them into the completely reduced block diagonal form. We are, in other words, interested in identifying for the representation space an appropriate basis that is made of bases of its irreducible subspaces, so that with respect to it the matrices of the representation take the desired completely reduced form.

Let $\rho^1, \ldots, \rho^h$ be the irreducible representations of G, with dimensions $n_1, \ldots, n_h$, and characters $\chi_1, \ldots, \chi_h$ respectively. Consider now a complete reduction of $\rho$ in which the equivalent irreducible representations are collected together. Let us say that the blocks equivalent to the irreducible representation $\rho^i$ are $a_i$ in number, with the corresponding irreducible subspaces labelled as $V^i_1, \cdots, V^i_{a_i}$. Further, let $V_i$ denote the direct sum of those of these subspaces over which the subrepresentations are equivalent to $\rho^i$, $i = 1, \ldots, h$. We are interested in the decomposition

$$\begin{aligned} V &= \sum_{i=1}^{h} \sum_{j=1}^{a_i} \oplus V^i_j \\ &= \sum_{i=1}^{h} \oplus V_i \end{aligned}$$

where each $V^i_j$, as also $V_i$, is invariant under G.

A basis that we seek for $V$ is then a basis obtained by putting together the bases of the irreducible subspaces $V_j^i$. In some situations we may even be content with the coarser decomposition in terms of the $V_i$'s, in which case we will get a partial reduction of $\rho$, its $i$th block further reducible to $a_i$ blocks, each equivalent to the irreducible representation $\rho^i$. Moreover, if in the completely reduced form of the given representation the irreducible representations appear not more than once, the $V_i$'s themselves are the irreducible subspaces. Keeping this in mind, and the fact that the procedure for obtaining bases for $V_i$ is relatively simpler, we consider this first.

So our first task is the following. Given a basis for $V$, and the matrices of $\rho(s)$ with respect to this basis, determine a basis for each $V_i$.

A very effective method of doing this is to exploit the connections between projections and direct sum decompositions. Recall (Section 5.6) that for a vector space $V = V_1 \oplus V_2$, the projection $p_1$ of $V$ on $V_1$ along $V_2$ has a "filter like" action on any $x \in$ V. Thus if $x = x_1 + x_2$, with (unique) components $x_1$ and $x_2$ in $V_1$ and $V_2$ respectively, then $p_1 x = x_1$. Likewise, for the projection $p_2$ of $V$ on $V_2$ along $V_1$, $p_2 x = x_2$. Now suppose that we have a basis $\alpha_1, \alpha_2, \ldots, \alpha_n$ for the space $V$, and that we know the matrices of projections $p_1$ and $p_2$ with respect to this basis. Then the vectors $p_1\alpha_k$, $k = 1, 2, \ldots, n$ have precisely as many linearly independent ones amongst them that constitute a basis of $V_1$. In particular, if the original representation space is $\mathbb{F}^n$, and if we know the matrix of the projection $p_1$ with respect to the standard basis, then the set of all linearly independent columns of this matrix form for $V_1$ a basis of the kind we seek. Likewise, we get a basis for $V_2$.

Observe that the points just made about a decomposition into two subspaces are equally valid for a decomposition into any finite number of subspaces, and that the procedure of using projections to determine their bases is valid in general.

Coming back to our reduction problem, the relevant projections are given by the following formula.

$$p_i = \frac{n_i}{|\mathrm{G}|} \sum_{s \in \mathrm{G}} \chi_i(s)^* \rho(s) \tag{5.17}$$

Using arguments based on the Schur's lemma, or alternatively on the orthogonality theorem, as we shall see more clearly a little later, it can be shown that $p_i$ is indeed a projection of $V$ onto $V_i$. The projections $p_i$ are easily constructed, and so are the bases for $V_i$. Here is an example.

**Example 5.10.1** Consider for the dyadic group of four elements (Example 5.5.1), the representation $\rho :\rightarrow \mathbf{GL}(\mathbb{R}^4)$ whose matrices with respect to the standard basis are

$$\overset{\rho(g_1)}{\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}}\quad\overset{\rho(g_2)}{\begin{bmatrix}0&1&0&0\\1&0&0&0\\0&0&0&1\\0&0&1&0\end{bmatrix}}\quad\overset{\rho(g_3)}{\begin{bmatrix}0&0&1&0\\0&0&0&1\\1&0&0&0\\0&1&0&0\end{bmatrix}}\quad\overset{\rho(g_4)}{\begin{bmatrix}0&0&0&1\\0&0&1&0\\0&1&0&0\\1&0&0&0\end{bmatrix}}$$

There are four irreducible representations of this group, each of dimension one, and their characters (or, to put it loosely, the representations themselves) are as shown in Table 5.5.

| $\chi_i$ | $g_1$ | $g_2$ | $g_3$ | $g_4$ |
|---|---|---|---|---|
| $\chi_1$ | $+1$ | $+1$ | $+1$ | $+1$ |
| $\chi_2$ | $+1$ | $+1$ | $-1$ | $-1$ |
| $\chi_3$ | $+1$ | $-1$ | $+1$ | $-1$ |
| $\chi_4$ | $+1$ | $-1$ | $-1$ | $+1$ |

Table 5.5 Irreducible characters of the dyadic group of four elements

The given representation is in fact the regular representation of the group, in whose completely reduced form each irreducible representation appears once and only once, its dimension being one (Proposition 5.8.12). Thus the decomposition characterized by the projections of the formula (5.17) is the final decomposition we seek. With respect to the standard basis on $\mathbb{R}^4$, the matrices for the projections $p_i, i = 1, 2, 3, 4$ are:

$$\begin{aligned}
p_1 &= [\rho(g_1) + \rho(g_2) + \rho(g_3) + \rho(g_4)]\\
&= \frac{1}{4}\begin{bmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{bmatrix}\\
p_2 &= [\rho(g_1) + \rho(g_2) - \rho(g_3) - \rho(g_4)]\\
&= \frac{1}{4}\begin{bmatrix}1&1&-1&-1\\1&1&-1&-1\\-1&-1&1&1\\-1&-1&1&1\end{bmatrix}\\
p_3 &= \frac{1}{4}\begin{bmatrix}1&-1&1&-1\\-1&1&-1&1\\1&-1&1&-1\\-1&1&-1&1\end{bmatrix}
\end{aligned}$$

$$p_4 = \frac{1}{4}\begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

Scanning the columns of these matrices by turn, we find that the irreducible spaces in this case are the four one-dimensional spaces $V_1$, $V_2$, $V_3$ and $V_4$ spanned by the vectors $[1\ \ 1\ \ 1\ \ 1]'$, $[1\ \ 1\ \ -1\ \ -1]'$, $[1\ \ -1\ \ 1\ \ -1]'$ and $[1\ \ -1\ \ -1\ \ 1]'$ respectively. The (normalized) matrix of the similarity transformation that completely reduces the matrices of the representation $\rho$ is then

$$\alpha = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

The completely reduced representation matrices $\alpha^{-1}\rho(g_i)\alpha$ are then the following.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Observe, as a final check, that entries in the first, second, third and fourth diagonal positions of these matrices together constitute the respective irreducible representations.

Now suppose that for a particular representation, we seek a complete reduction, and that in its completely reduced form the irreducible representations appear more than once. In that case the subspaces $V_i$ must be further decomposed into $V^i_j$. It is easier for this purpose to skip the step of explicitly identifying the bases of $V_i$ and, starting from the original space $V$ itself, directly determine bases for $V^i_j$. This is what we discuss now.

Let us label the desired basis vectors of $V^i_j$ as $e^i_{jk}$, $1 \leq k \leq n_i$. We impose on these vectors one condition: they are so chosen that in the completely reduced matrices of the given representation, the $a_i$ blocks corresponding to the irreducible representation $\rho^i$ are all identical and unitary; the entries of this common matrix we shall write as $\rho^i_{pq}$; $1 \leq p, q \leq n_i$. Such a choice is indeed possible because these $a_i$ blocks are matrices of subrepresentations equivalent to the irreducible representation $\rho^i$.

With this understanding, let us now see how to go about determining basis vectors of the desired kind. First, examine the tabular arrangement shown in Figure 5.2 for them for one particular $V_i$.

Observe that the vectors shown in this table, if grouped row wise as shown by the solid rectangles, form bases of the irreducible subspaces $V^i_j$, $1 \leq j \leq a_i$.

$$\begin{array}{cccccc}
V_i = & W_1^i & \oplus & W_2^i & \oplus\cdots\oplus & W_{n_i}^i \\
= & & & & & \\
V_1^i & e_{11}^i & & e_{12}^i & \cdots & e_{1n_i}^i \\
\oplus & & & & & \\
V_2^i & e_{21}^i & & e_{22}^i & \cdots & e_{2n_i}^i \\
\oplus & & & & & \\
\vdots & \vdots & & \vdots & & \vdots \\
\oplus & & & & & \\
V_{a_i}^i & e_{a_i1}^i & & e_{a_i2}^i & \cdots & e_{a_in_i}^i
\end{array}$$

Figure 5.2 Two decompositions of $V_i$

But if we group them column wise as shown by the dashed rectangles, they give another decomposition of $V_i$ in terms of the subspaces $W_k^i$, $1 \leq k \leq n_i$. The trick for determining these vectors lies in exploiting the following linear transformations that cross-connect these two decompositions,

$$p_{jk}^{(i)} = \frac{n_i}{|\mathrm{G}|}\sum_{s\in \mathrm{G}} \rho_{jk}^i(s)^*\rho(s) \quad \text{for } j \text{ and } k \text{ from } 1 \text{ to } n_i, \tag{5.18}$$

where $\rho^i$ is the $i$-th irreducible representation in (unitary) matrix form.

As we shall presently see, we start with the given basis of $V$ and then, using one of the transformations, $p_{11}^{(i)}$, determine basis vectors of $W_1^i$. Then on these we apply the transformations $p_{k1}^{(i)}$ to obtain the basis vectors of $W_k^i$, $2 \leq k \leq a_i$. Having obtained all of these in this manner, we arrange them in lexical order, proceeding from left to right in the first row of the tabular form, then from left to right of the next row, and so on. So ordered, they form a basis for $V_i$ of the desired kind. (In view of the way they are obtained, for each row, the vectors of the second column onwards are commonly referred to as *partners* of the first one of that row.) Finally, bases of $V_1$ to $V_h$ together give us a basis of $V$ of the desired kind. Vectors forming such a basis are, in the parlance of physics, commonly called *symmetry-adapted basis vectors*. This nomenclature is meant to highlight the fact that in dealing with physical problems, the relevant groups are groups of symmetries.

Now the details. We need first to understand what the transformations (5.18) actually do. Let us take a 'close-up' view of a specific hypothetical example. Let us say that for a group G of order 6 we have a representation $\rho$ on $\mathbb{R}^4$ and its matrices with respect to the standard basis on $\mathbb{R}^4$ are given to us. Suppose that in the completely reduced form of this representation, just one irreducible representation $\rho^1$ of dimension 2 appears twice. Then with respect to the desired basis, $(e^1_{11}, e^1_{12}, e^1_{21}, e^1_{22})$, its matrices are

$$\begin{bmatrix} \rho^1_{11}(s) & \rho^1_{12}(s) & 0 & 0 \\ \rho^1_{21}(s) & \rho^1_{22}(s) & 0 & 0 \\ 0 & 0 & \rho^1_{11}(s) & \rho^1_{12}(s) \\ 0 & 0 & \rho^1_{21}(s) & \rho^1_{22}(s) \end{bmatrix}$$

The transformations (5.18) are in this case $p^{(1)}_{11}, p^{(1)}_{12}, p^{(1)}_{21}$ and $p^{(1)}_{22}$. Applying the orthogonality theorem to the entries of their matrices, we find that they are the following matrices of $1s$ and $0s$.

$$p^{(1)}_{11} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} p^{(1)}_{12} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$p^{(1)}_{21} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} p^{(1)}_{22} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

You may then check that the basis vectors $e^1_{11}, e^1_{12}, e^1_{21}$ and $e^1_{22}$ transform under $p^{(1)}_{11}$ as follows.

$$\begin{array}{ll} p^{(1)}_{11} e^1_{11} = e^1_{11} & p^{(1)}_{11} e^1_{21} = e^1_{21} \\ p^{(1)}_{11} e^1_{12} = 0 & p^{(1)}_{11} e^1_{22} = 0 \end{array}$$

That is, $p^{(1)}_{11}$ leaves $e^1_{11}$ and $e^1_{21}$, the basis vectors of $W^1_1$, unchanged and annihilates the others. It follows that it is a projection of $V$ onto $W^1_1$. Likewise, $p^{(1)}_{22}$ is a projection onto $W^1_2$. So, using $p^{(1)}_{11}$, we can determine a basis for $W^1_1$. (These facts implicitly clarify the behaviour of transformations (5.17). Observe that $p_1 = p^{(1)}_{11} + p^{(1)}_{22}$, an identity matrix of size four. Thus $p_1$ is a projection onto $V_1$.)

What about $p^{(1)}_{21}$? Well, the basis vectors transform under it as follows.

$$\begin{array}{ll} p^{(1)}_{21} e^1_{11} = e^1_{12} & p^{(1)}_{21} e^1_{21} = e^1_{22} \\ p^{(1)}_{21} e^1_{12} = 0 & p^{(1)}_{21} e^1_{22} = 0 \end{array}$$

Clearly, $p^{(1)}_{21}$ is not a projection. It is, however, a one-to-one onto map from $W^1_1$ to $W^1_2$, so that a basis of the former gets transformed by it into a basis of the latter.

Thus, on the whole, $p^{(1)}_{11}$ leads us to a basis of $W^1_1$, and then $p^{(1)}_{21}$ leads us to a basis of $W^1_2$. Having obtained these, it is only a matter of arranging them in the right order, as outlined earlier, to get a basis of the desired kind for the representation subspace $V_1$.

Based on this 'close-up' view, it is reasonable to come to this conclusion about the transformations (5.18) in general: $p_{11}^{(i)}$ is projection of $V_i$ onto $W_1^i$, and $p_{k1}^{(i)}$ is an isomorphism from $W_1^i$ to $W_k^i$, $2 \leq k \leq n_i$. For a formal and detailed proof that indeed it is so, and for many other properties of these transformations, see texts such as Tinkham [9, Chapter 3, pp. 39–41], Serre [8, Section 2.7], McWeeny [7, Chapter 5, pp. 126–133] and Miller [5, Chapter 3, pp. 92–96].

With these clarifications, we are now ready to deal with the general complete reduction problem: Given the matrices of a representation $\rho : \mathrm{G} \to \mathbf{GL}(V)$ with respect to some basis, determine a basis for $V$, and correspondingly the transformation matrix $\alpha$, that puts the matrices of $\rho$ in the completely reduced form.

We consider $V$ to be $\mathbb{R}^n$ or $\mathbb{C}^n$, and the basis given on it to be the standard basis. The following steps then lead us to a basis for it of the desired kind, the basis vectors in this case being column vectors of size $n \times 1$.

**STEP 1** For the given group G, ascertain the irreducible representations $\rho^i$, $1 \leq i \leq h$, where $h$ is the number of its irreducible representations .

**STEP 2** For each $\rho^i$, compute the matrix $p_{11}^{(i)}$. It has $a_i$ linearly independent columns. Identify any one set of these and after normalizing them, label them $e_{11}^i, e_{21}^i, \ldots, e_{a_i 1}^i$.

**STEP 3** Compute the matrices $p_{21}^{(i)}, p_{31}^{(i)}, \ldots, p_{n_i 1}^{(i)}$.

**STEP 4** Compute the partners of the vectors obtained in Step 2. For $e_{l1}^i$, $1 \leq l \leq a_i$, the partners are the vectors $p_{k1}^{(i)} e_{l1}^i$, $2 \leq k \leq n_i$ after they have been normalized.

**STEP 5** Collect the normalized vectors obtained in Steps 2 and 4 and arrange them in the order: $e_{11}^i$ followed by its partners, then $e_{21}^i$ followed by its partners, and so forth. (To put it another way, arrange the vectors $e_{jk}^i$ in the lexical order of the index set $ijk$.)

**STEP 6** Having done this for every $i = 1, \ldots, h$, arrange as columns of a matrix, the ordered vectors obtained in Step 5 for $i = 1$ followed by those for $i = 2$, and so forth. The resulting matrix, which will be of size $n \times n$, is the desired matrix $\alpha$.

It should be noted at this point that the matrix $\alpha$ that we get this way, or rather the basis of $V$ identified in this manner, is not unique. It depends on the choice we make in Step 2 for the basis of the range of $p_{11}^{(i)}$. In contrast, the coarse decomposition obtained through (22) does not entail such ambiguity.

**Example 5.10.2** Consider this time the group of order 6 of Example 5.6.1. Its regular representation consists of the following matrices $\mathrm{P}_1$ to $\mathrm{P}_6$, each of size 6.

$$\begin{bmatrix} 1&0&0&0&0&0\\ 0&1&0&0&0&0\\ 0&0&1&0&0&0\\ 0&0&0&1&0&0\\ 0&0&0&0&1&0\\ 0&0&0&0&0&1 \end{bmatrix} \begin{bmatrix} 0&0&1&0&0&0\\ 1&0&0&0&0&0\\ 0&1&0&0&0&0\\ 0&0&0&0&0&1\\ 0&0&0&1&0&0\\ 0&0&0&0&1&0 \end{bmatrix} \begin{bmatrix} 0&1&0&0&0&0\\ 0&0&1&0&0&0\\ 1&0&0&0&0&0\\ 0&0&0&0&1&0\\ 0&0&0&0&0&1\\ 0&0&0&1&0&0 \end{bmatrix}$$

$$\begin{bmatrix} 0&0&0&1&0&0\\ 0&0&0&0&0&1\\ 0&0&0&0&1&0\\ 1&0&0&0&0&0\\ 0&0&1&0&0&0\\ 0&1&0&0&0&0 \end{bmatrix} \begin{bmatrix} 0&0&0&0&1&0\\ 0&0&0&1&0&0\\ 0&0&0&0&0&1\\ 0&1&0&0&0&0\\ 1&0&0&0&0&0\\ 0&0&1&0&0&0 \end{bmatrix} \begin{bmatrix} 0&0&0&0&0&1\\ 0&0&0&0&1&0\\ 0&0&0&1&0&0\\ 0&0&1&0&0&0\\ 0&1&0&0&0&0\\ 1&0&0&0&0&0 \end{bmatrix}$$

This group has one irreducible representation of dimension 2, and two irreducible representations of dimension 1. We need them in their unitary (i.e., normal in the real case) forms. These are

$$\rho^1 : \begin{bmatrix} 1 & 0\\ 0 & 1 \end{bmatrix} \begin{bmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2}\\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2}\\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} -1 & 0\\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2}\\ \frac{-\sqrt{3}}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2}\\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{bmatrix}$$

$$\rho^2 : \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix}$$

$$\rho^3 : \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} -1 \end{bmatrix} \begin{bmatrix} -1 \end{bmatrix} \begin{bmatrix} -1 \end{bmatrix}$$

In the completely reduced form of $\rho$, $\rho^1$ appears twice, and $\rho^2$ and $\rho^3$ each appear once. Thus there are no partners to be computed for $\rho^2$ and $\rho^3$. The relevant transformations (5.18) are then in this case $p_{11}^{(1)}$, $p_{21}^{(1)}$, $p_{11}^{(2)}$ and $p_{11}^{(3)}$.

Taking up $\rho^1$ first, we get

$$p_{11}^{(1)} = \begin{bmatrix} \frac{1}{3} & -\frac{1}{6} & -\frac{1}{6} & -\frac{1}{3} & \frac{1}{6} & \frac{1}{6}\\ -\frac{1}{6} & \frac{1}{3} & -\frac{1}{6} & \frac{1}{6} & \frac{1}{6} & -\frac{1}{3}\\ -\frac{1}{6} & -\frac{1}{6} & \frac{1}{3} & \frac{1}{6} & -\frac{1}{3} & \frac{1}{6}\\ -\frac{1}{3} & \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & -\frac{1}{6} & -\frac{1}{6}\\ \frac{1}{6} & \frac{1}{6} & -\frac{1}{3} & -\frac{1}{6} & \frac{1}{3} & -\frac{1}{6}\\ \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} & \frac{1}{3} \end{bmatrix}$$

It has two linearly independent columns, for which let us pick the first and the second. After normalizing them, we get

$$\begin{aligned} e_{11}^1 &= \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{2\sqrt{3}} & -\frac{1}{2\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} \end{bmatrix}'\\ e_{21}^1 &= \begin{bmatrix} -\frac{1}{2\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} & -\frac{1}{\sqrt{3}} \end{bmatrix}' \end{aligned}$$

For computing their partners, we first compute $p^{(1)}_{21}$:

$$\begin{aligned}
p^{(1)}_{21} &= \sum_{s\in \mathrm{G}} \rho^1_{21}(s)^* \rho(s) \\
&= \frac{\sqrt{3}}{2}[\mathrm{P}_2 - \mathrm{P}_3 - \mathrm{P}_5 + \mathrm{P}_6] \\
&= \frac{\sqrt{3}}{2}\begin{bmatrix} 0 & -1 & 1 & 0 & -1 & 1 \\ 1 & 0 & -1 & -1 & 1 & 0 \\ -1 & 1 & 0 & 1 & 0 & -1 \\ 0 & -1 & 1 & 0 & -1 & 1 \\ -1 & 1 & 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & -1 & 1 & 0 \end{bmatrix}
\end{aligned}$$

The vectors $e^1_{12} = p^{(1)}_{21} e^1_{11}$ and $e^1_{22} = p^{(1)}_{21} e^1_{21}$, after normalization, are the partners of $e^1_{11}$ and $e^1_{21}$ respectively. These come out to be:

$$\begin{aligned}
e^1_{12} &= \begin{bmatrix} 0 & \tfrac{1}{2} & -\tfrac{1}{2} & 0 & -\tfrac{1}{2} & \tfrac{1}{2} \end{bmatrix}' \\
e^1_{22} &= \begin{bmatrix} -\tfrac{1}{2} & 0 & \tfrac{1}{2} & -\tfrac{1}{2} & \tfrac{1}{2} & 0 \end{bmatrix}'
\end{aligned}$$

This completes the computations for $\rho^1$. Following the same steps for $\rho^2$ and $\rho^3$, we use $p^{(2)}_{11}$ and $p^{(3)}_{11}$ to get

$$\begin{aligned}
e^2_{11} &= \frac{1}{\sqrt{6}}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}' \\
e^3_{11} &= \frac{1}{\sqrt{6}}\begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix}'
\end{aligned}$$

There are no partner vectors to be computed for these. So, arranging the vectors in the order $e^1_{11}, e^1_{12}, e^1_{21}, e^1_{22}, e^2_{11}, e^3_{11}$, we finally get the matrix $\alpha$:

$$\alpha = \begin{bmatrix}
\frac{1}{\sqrt{3}} & 0 & -\frac{1}{2\sqrt{3}} & -\frac{1}{2} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\
-\frac{1}{2\sqrt{3}} & \frac{1}{2} & \frac{1}{\sqrt{3}} & 0 & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\
-\frac{1}{2\sqrt{3}} & -\frac{1}{2} & -\frac{1}{2\sqrt{3}} & \frac{1}{2} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\
-\frac{1}{\sqrt{3}} & 0 & \frac{1}{2\sqrt{3}} & -\frac{1}{2} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \\
\frac{1}{2\sqrt{3}} & -\frac{1}{2} & \frac{1}{2\sqrt{3}} & \frac{1}{2} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \\
\frac{1}{2\sqrt{3}} & \frac{1}{2} & -\frac{1}{\sqrt{3}} & 0 & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}}
\end{bmatrix}$$

You may check that under similarity transformation by $\alpha$, the matrices, $\mathrm{P}_1,\ 1 \leq k \leq 6$, of the given representation do indeed acquire the completely reduced block diagonal form,

$$\alpha^{-1}\mathrm{P}_k\alpha = \begin{bmatrix} \rho^1(g_k) & & & \\ & \rho^1(g_k) & & \\ & & \rho^2(g_k) & \\ & & & \rho^3(g_k) \end{bmatrix}$$

**Exercise 5.10.1** Use the procedure outlined in this section to obtain a completely reduced form of the matrices $\mathrm{P}_1$ to $\mathrm{P}_6$ of Example 5.6.1.

## 5.11 Further on Reduction

The reduction technique that we learnt in the last section has additional powers; it also block diagonalizes matrices that commute with the matrices of a group representation. Let us now examine this result.

First, a simple fact about commuting transformations on a vector space. On a vector space $V$, let $p$ and $h$ be two linear transformations that commute, i.e., $ph = hp$. It is easily seen that the range and null spaces of $p$ are invariant under $h$, and those of $h$ are invariant under $p$.

For any $u$ in the range $R_p$ of $p$, there is $x \in V$, such that $px = u$. So $hu = hpx = p(hx) \in R_p$. That is, $R_p$ is invariant under $h$. As to the null space $N_p$ of $p$, for any $x \in N_p$, $p(hx) = hpx = 0$, and therefore $hx \in N_p$. Thus the range and null space of $p$ are both invariant under $h$. From symmetry, the range and null space of $h$ are likewise invariant under $p$.

Consider now a representation $\rho : \mathrm{G} \to \mathbf{GL}(V)$, and a linear transformation $h$ on $V$ that commutes with $\rho(s)$ for every $s \in \mathrm{G}$. Suppose that we have already determined a set of symmetry-adapted basis vectors for $V$, which completely reduce the matrices of $\rho$. We shall presently see that the same vectors, suitably reordered, also block diagonalize the matrix of $h$.

Going back to (5.18), recall that $W_k^i$ is the range space of $p_{kk}^i$, $1 \leq k \leq n_i$. Since the given transformation $h$ commutes with every $\rho(s)$,

$$\begin{aligned} hp_{kk}^{(i)} &= \frac{n_i}{|\mathrm{G}|} \sum_{s \in \mathrm{G}} \rho_{kk}^i(s)^* (h\rho(s)) \\ &= \left[ \frac{n_i}{|\mathrm{G}|} \sum_{s \in \mathrm{G}} \rho_{kk}^i(s)^* \rho(s) \right] h \\ &= p_{kk}^{(i)} h, \end{aligned}$$

and therefore, $W_k^i$ is invariant under $h$ for every $k, 1 \leq k \leq n_i$. It then follows that the basis vectors $e_{jk}^i$, read off column wise in the tabular form given in Figure 5.2, i.e., in the lexical order of the index set $ikj$, form an ordered basis that block diagonalizes the matrix of $h$. More specifically, if our representation space is $\mathbb{R}$ ($\mathbb{C}$), then the matrix of $h$ with respect to the standard basis gets block diagonlized under similarity transformation by the matrix $\beta$ whose columns are the symmetry-adapted vectors taken in this order. For the example considered in the previous section, the vectors are then taken in the order $e^1_{11}, e^1_{21}, e^1_{12}, e^1_{22}, e^2_{11}, e^3_{11}$, so that the matrix $\beta$ is in this case,

$$\beta \;=\; \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{2\sqrt{3}} & 0 & -\frac{1}{2} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ -\frac{1}{2\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{2} & 0 & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ -\frac{1}{2\sqrt{3}} & -\frac{1}{2\sqrt{3}} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{3}} & \frac{1}{2\sqrt{3}} & 0 & -\frac{1}{2} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{2\sqrt{3}} & \frac{1}{2\sqrt{3}} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{2\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{2} & 0 & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \end{bmatrix}$$

# References

1. Werner H. Greub. *Linear Algebra*. Springer, New York, 1981.
2. G.G. Hall. *Applied Group Theory*. American Elsevier, New York, 1967.
3. T. Hawkins. The origins of the theory of group characters. *Arch. Hist. Exact Sci.*, 7:142–170, 1971.
4. L. Jansen and M. Boon. *Theory of Finite Groups: Applications in Physics*. North–Holland, Amsterdam, 1967.
5. W. Miller (Jr.). *Symmetry Groups and Their Applications*. Academic, London, 1972.
6. J.S. Lomont. *Applications of Finite Groups*. Academic, London, 1959.
7. R. McWeeny. *Symmetry: An Introduction to Group Theory and its Applications*. Pergamon, London, 1963.
8. J.-P. Serre. *Linear representation of Finite Groups*. Springer-Verlag, New York, 1977.
9. M. Tinkham. *Group Theory and Quantum Mechanics*. McGraw-Hill, New York, 1964.
10. F.L. Williams. History and variation on the theme of the frobenius reciprocity theorem. *Mathematical Intelligencer*, 13(3):68–71, 1991.

# Chapter 6

# Signal Processing and Representation Theory

The focus in the previous chapter was on the representation of finite groups by matrices and linear transformations (over the field of real or complex numbers). The related theory is very rich in content, and stands on its own as an area of study in mathematics. One could very justifiably ask at this point: What connections could group representation theory have with signal processing? The answer lies in relating the theory to what has been said in Chapters 1 and 4 about signals and systems.

Recall that, for an abelian group, all its irreducible representations on the complex number field are scalar functions, i.e., functions from the group to complex numbers under multiplication. For nonabelian groups, however, their irreps can not all be scalars, simply because numbers are necessarily commutative. The next possible option then is to consider matrices or, at a more abstract level, to consider linear transformations on a vector space, for their representation, and thus bypass the commutativity constraint. This is indeed what representation theory does. It deals with the nature and structure of such representations.

A high point of this theory is the result that every representation of a group is a direct sum of its irreducible representations. Equivalently, it means that the underlying representation space can be decomposed into a direct sum of subspaces of relatively small dimensions such that (a) each of these subspaces is an invariant subspace under each of the linear transformations representing the group and (b) each one of the subspaces is as small as possible. Representation theory supplies the algebraic tools for constructing such decompositions.

As discussed in Chapter 4, in a wide variety of signal processing problems, our starting point consists of a vector space of signals, together with a class of systems (processors) whose symmetries are characterized by a group of linear transformations on this space. A common objective in these cases is to identify a suitable decomposition of the signal space into subspaces that can be separately handled by the processing systems. In more precise terms, this amounts to the objective set out in Section 4.5.2 on page 100, which I reproduce here for convenience:

DOI 10.1007/978-90-481-9434-6_6, © Springer Science+Business Media B.V. 2010

**Remark 6** *Given a vector space $V$ and a group $\mathbf{P}$ of linear transformations on $V$, identify a decomposition $V = V_1 \oplus V_2 \oplus \cdots \oplus V_k$ such that the subspaces $V_i$, $i = 1, \ldots, k$, are as small as possible, and are each invariant under every member of*$\mathbf{P}$. ♠

It is in this context that the reduction techniques of group representation theory become directly relevant in signal processing. We first consider the case in which signals are functions on a group.

## 6.1 Signals as Functions on Groups

Consider the real line, $\mathbb{R}$, as the domain of functions that represent continuous-time signals. A crucial operation on functions in this context is that of translation, or shift, in time, as discussed in Section 4.4.2. Thus, for a function $f : \mathbb{R} \to \mathbb{R}$, we talk of its translates, or shifted versions, ${}^{\tau}\!f$, ${}^{\tau}\!f(t) = f(t - \tau)$. Furthermore, scaling of signals in time is of no primary concern in the theory of LTI systems, i.e., for a function $f : \mathbb{R} \to \mathbb{R}$, we are not interested here in functions $g$, $g(t) := f(\tau t)$ for $\tau \in \mathbb{R}$.

On the whole, we can say that we are concerned here with $\mathbb{R}$ not merely as a set but as one with additional structure. Admittedly, we are not interested here in the full structure of the real line, which is that of a *field*, with addition and multiplication as its two binary operations. We do, however, take into account the fact that, with respect to the operation of addition, it has the structure of a group. In the specific case of continuous-time signals, it is the group $(\mathbb{R}, +)$. In the discrete-time case, it is the group $(\mathbb{Z}, +)$. In the discrete finite case, one of the choices is the group $\mathbb{Z}_n$ under addition modulo $n$. When we talk of a translate of a function, we implicitly assume such a group structure for the index set on which the function is defined. This is the motivation for thinking of signals very generally as functions on a group, and of systems for processing such signals as linear transformations with symmetries characterized by an isomorphic group of translation operators.[1]

Going back to Remark 6, let the signal space $V$ be a vector space of functions from a group $\mathbf{G}$ to $\mathbb{C}$ (or $\mathbb{R}$), and let $\mathbf{P}$ be the group of associated translation operators isomorphic to $\mathbf{G}$. Then the subspaces $V_i$ are the smallest possible subspaces of $V$, each of which is invariant under every member of $\mathbf{P}$. A signal $f$ in $V$ is then uniquely decomposable into components belonging to the subspaces $V_i$, and the action of every member of $\mathbf{P}$ on the components can be determined separately. These components are the so called *harmonics* (or *spectral components*) of the signal $f$.

**Remark 7** In the special case when the group $\mathbf{G}$, and consequently the associated isomorphic group $\mathbf{P}$ of translation operators, is abelian (commutative) then

[1] Translation operators, as explained in Section 4.4.2, may be visualized as generalizations of the familiar delay lines of circuit and system theory. You may recall the way they are used in a transversal filter for simulating an LTI system with a given impulse response.

the subspaces $V_i$ are all of dimension one and their basis vectors coincide with the common eigenvectors of the members of $\mathbf{P}$. Thus, if the group is $\mathbb{R}$ under addition (as in the case of continuous-time signals) then the complex exponentials emerge as the harmonic components of signals. For electrical engineers, the term “harmonics” typically means sinusoids or complex exponentials. Within the group theoretic framework, the term has acquired a much broader interpretation. The spirit is nonetheless the same throughout. ♠

Very broadly, the subject of harmonic analysis on groups (also known as Fourier analysis on groups) deals with such decompositions of functions on groups (commutative as well as noncommutative), and their applications in mathematics and engineering. For an introduction to the subject, the paper by Gross [7] may be a good starting point. Edwards [4] and Devito [3] should provide further insights. For engineering applications in digital system design, see Stanković [15] and Karpovsky [10]. Applications in the area of image understanding are discussed in Kanatani [9].

## 6.2 Symmetries of Linear Equations

Another area of applications of representation theory is in the solution of algebraic and differential equations with symmetries. The collection of papers in Allgower [1] should give an idea of the kind of work that has of late been going on in this direction. One particular line of such work has to do with symmetry based block-diagonalization of a system of linear equations. A very simple example related to filter design is what I give next to illustrate the basic idea.

**Example 6.2.1** Consider the $LC$ 2-port analog filter shown in the Figure 6.1. It is the circuit of an insertion-loss low-pass filter of order 5, working between an input source $V_s$ and an output load resistance $R_L$. A common property of such filters of odd orders is that their circuits have physical symmetry with respect to the input and output ports. In this particular example, this is reflected in the fact that the end inductors are both of value $L_1$ and the two capacitors are both of value $C$.

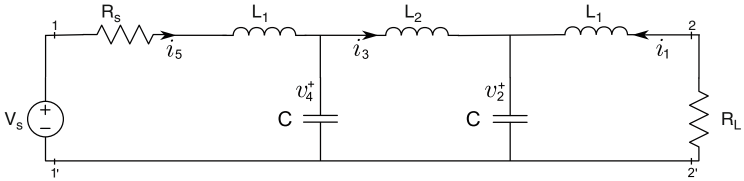

Figure 6.1 An insertion-loss low-pass filter

The state equations of the circuit, with $v_2$, $v_4$, $i_1$, $i_3$, and $i_5$ as the state variables, can be seen to be:

$$\frac{d}{dt}\begin{bmatrix} v_2 \\ v_4 \\ i_1 \\ i_3 \\ i_5 \end{bmatrix} = \begin{bmatrix} 0 & 0 & a & a & 0 \\ 0 & 0 & 0 & -a & a \\ -b & 0 & c & 0 & 0 \\ -d & d & 0 & 0 & 0 \\ 0 & -b & 0 & 0 & c \end{bmatrix}\begin{bmatrix} v_2 \\ v_4 \\ i_1 \\ i_3 \\ i_5 \end{bmatrix} + V_s\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ b \end{bmatrix} \tag{6.1}$$

where coefficients $a$, $b$, $c$, and $d$ are determined by the component values of the circuit. Let $A$ denote the coefficient matrix of Eq. (6.1):

$$A = \begin{bmatrix} 0 & 0 & a & a & 0 \\ 0 & 0 & 0 & -a & a \\ -b & 0 & c & 0 & 0 \\ -d & d & 0 & 0 & 0 \\ 0 & -b & 0 & 0 & c \end{bmatrix} \tag{6.2}$$

Based on symmetry arguments, we can see that the coefficient matrix $A$ commutes with the following matrix $P$:

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \tag{6.3}$$

Note that $P$ and $I$ (the identity matrix of size 5) together form a group of order 2. Using the reduction procedure discussed in Chapter 5, we are then in a position to block-diagonalize the matrix $A$. The procedure leads us to the matrix $\alpha$:

$$\alpha = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & -1 & 0 \end{bmatrix} \tag{6.4}$$

with the property that, under similarity transformation by it, the matrix $\hat{A} = \alpha^{-1}A\alpha$ is in a block-diagonal form with two blocks, one of size 2 and another of size 3:

$$\hat{A} = \begin{bmatrix} 0 & a & 0 & 0 & 0 \\ -b & c & 0 & 0 & 0 \\ 0 & 0 & 0 & a & 2a \\ 0 & 0 & -b & c & 0 \\ 0 & 0 & -d & 0 & 0 \end{bmatrix}$$

Now, define a new set of variables $\hat{v}_2$, $\hat{v}_4$, $\hat{i}_1$, $\hat{i}_3$, and $\hat{i}_5$ by the equality

$$\begin{bmatrix} \hat{v}_2 \\ \hat{v}_4 \\ \hat{i}_1 \\ \hat{i}_3 \\ \hat{i}_5 \end{bmatrix} = \alpha^{-1} \begin{bmatrix} v_2 \\ v_4 \\ i_1 \\ i_3 \\ i_5 \end{bmatrix} \tag{6.5}$$

Then the Eq. (6.1) takes the following equivalent form:

$$\frac{d}{dt} \begin{bmatrix} \hat{v}_2 \\ \hat{v}_4 \\ \hat{i}_1 \\ \hat{i}_3 \\ \hat{i}_5 \end{bmatrix} = \begin{bmatrix} 0 & a & 0 & 0 & 0 \\ -b & c & 0 & 0 & 0 \\ 0 & 0 & 0 & a & 2a \\ 0 & 0 & -b & c & 0 \\ 0 & 0 & -d & 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{v}_2 \\ \hat{v}_4 \\ \hat{i}_1 \\ \hat{i}_3 \\ \hat{i}_5 \end{bmatrix} + \alpha^{-1} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ bV_s \end{bmatrix} \tag{6.6}$$

Owing to the block-diagonal form of the coefficient matrix, Eq. (6.6) separates out into two sets of uncoupled differential equations. This helps simplify the task of realizing the original $LC$ filter as an equivalent active state-variable filter.[2] □

**Remark 8** As in the example given in Section 4.5.3, here too the matrix $\alpha$ does not depend on the specific values of the parameters $a$, $b$, $c$, and $d$. It is entirely decided by the symmetries of the coefficient matrix $A$. What holds in this respect in the specific case of a filter of order 5, it also holds for all such insertion-loss filters of odd orders in general. ♠

## 6.3 Fast Discrete Signal Transforms

Yet another major connection of representation theory with signal processing is in the design of what are called *fast algorithms* for a discrete finite transform, such as the discrete Fourier transform (DFT).

Admittedly, the DFT has not come for explicit mention in our discussions. It is, however, reasonable to assume that the reader is familiar with its wide ranging applications in signal processing. The reader must also be familiar with the class of algorithms, collectively known as fast Fourier transform (FFT) algorithms, that are used in efficiently carrying out DFT computations. For someone interested in revising the basics, presented in the setting of linear algebra, a good introduction is given in Frazier [6, Chapter 2]. Although Frazier does not bring in group theory in his treatment, he does relate the DFT to the eigenvectors of translation-invariant linear transformations on finite dimensional vector spaces. In that sense his approach does

[2] To see more clearly the potential relevance of this idea in this context, the reader may consult texts on active filter design. Two recent ones are Pactitis [13] and Irons [8].

connect with group representation theory. A tutorial overview of a group theoretic interpretation of the FFT can be found in Rockmore [14].

Egner and Püschel [5] may be consulted for more on the subject of symmetry based fast algorithms. They address the very general issue of deriving fast algorithms for a very wide class of what they have called "discrete signal transforms", of which the DFT and the discrete cosine transform (DCT) are two special kinds. A proper understanding of these algorithms and related ideas requires as a prerequisite a background in group theory. Hopefully the contents of this book will initiate the reader in building up such a background.

In closing, I should like to add that several new avenues of exploiting symmetry considerations in signal processing are just about beginning to open up. One such avenue is related to the presence of partial symmetries in signals and systems. As pointed out in Veblen and Whitehead [16, p. 32] and Lawson [11, p. 2], there are spaces whose groups of automorphisms reduce to the identity.[3] Group theory as a tool for studying symmetries does not take us very far in such cases. As Lawson [11] argues, the theory of inverse semigroups, which deals with partial symmetries, may provide a more meaningful framework. This calls for fresh and substantive new investigations.

One can also make a case for considering signals not as members of a vector space, but rather as subspaces of a vector space. (We do a similar thing when we talk of events as subsets in probability theory.) In that case the signal space has the structure of a non-distributive lattice. What about the role of symmetry in the context of signal spaces so conceived? Will all this be relevant in the modeling of some real-life situation? The classic paper by Birkhoff and von Neumann [2] seems to suggest that there may well be such relevance, even in the area of signal processing. ***But all that is part of another story!***

[3] The real number field $\mathbb{R}$ serves as an example. This point is also discussed in Marquis [12, p. 35].

# References

1. Eugene L. Allgower, Kurt Georg, and Rick Miranda, editors. *Exploiting Symmetry in Applied and Numerical Analysis*. American Mathematical Society, Providence, 1993.

2. G. Birkhoff and J. von Neumann. The logic of quantum mechanics. *Annals of Mathematics*, 37(2):823–843, 1936.

3. Carl L. DeVito. *Harmonic Analysis: A Gentle Introduction*. Jones & Bartlett, Boston, 2007.

4. R.E. Edwards. *Fourier Series: A Modern Introduction*, volume 1. Springer-Verlag, New York, 1979.

5. Sebastian Egner and Mark Püschel. Automatic generation of fast discrete signal transforms. *IEEE Trans. on Signal Processing*, 49(9):1992–2002, 2001.

6. Michael W. Frazier. *Introduction to Wavelets Through Linear Algebra*. Springer-Verlag, New York, 2000.

7. Kenneth I. Gross. On the evolution of noncommutative harmonic analysis. *Amer. Math. Month.*, 85:525–548, 1978.

8. Fred H. Irons. *Active Filters for Integrated-Circuit Applications*. Artech House, Boston, 2005.

9. Kenichi Kanatani. *Group-Theoretical Methods in Image Understanding*. Springer-Verlag, London, 1990.

10. Mark G. Karpovsky, Radomir S. Stanković, and Jaakko T. Astola. *Spectral Logic and Its Applications for the Design of Digital Devices*. Wiley, New Jersey, 2008.

11. M.V. Lawson. *Inverse Semigroups: The Theory of Partial Symmetries*. World Scientific, Singapore, 1998.

12. Jean-Pierre Marquis. *From a Geometrical Point of View: A Study of the History and Philosophy of Category Theory*. Springer, Dordrecht, 2009.

13. S.A. Pactitis. *Active Filters: Theory and Design*. CRC Press, Boca Raton, 2008.

14. Daniel N. Rockmore. The fft: An algorithm the whole family can use. *Computing in Science and Engineering*, 2(1):60–64, 2000.

15. Radomir S. Stanković, Claudio Moraga, and Jaakko T. Astola. *Fourier Analysis on Finite Groups with Applications in Signal Processing and System Design*. Interscience Publishers, New York, 2005.

16. Oswald Veblen and J.H.C. Whitehead. *The Foundations of Differential Geometry*. Cambridge University Press, Cambridge, 1932, 1953, 1960.

# Appendix A

# Parentheses, Their Proper Pairing, and Associativity

## A.1 Proper Pairing of Parentheses

We all know how to deal with parentheses in mathematical expressions; what is their role, how do they pair, what are the rules of pairing, and so on. Let us try to formalize all this.

To begin with, we have an alphabet of two symbols, the left parenthesis, written "(" and called here in short "$lp$", and the right parenthesis ")" called "$rp$". Then we have finite strings of such parentheses. Not all such strings are, however, of interest to us. We want to look at only those that consist of parentheses that are 'properly' paired, or matched. Intuitively, we know what proper matching means: for every $lp$ there is precisely one matching $rp$ and vice versa, every $lp$ is to the left of its matching $rp$, and no two matching pairs interlace.

How do we say all this in the language of algebra? Well, one way is to start with an alphabet $A = \{(,)\}$, consisting of the set of the two symbols denoting the two parentheses, and the set of all finite sequences (or strings) of these symbols.[1] Let $p_i$ denote the entry in the $i$th position of a string (counted from the left); it may be either an $lp$ or an $rp$. A string is of *length* $n$ if it has $n$ such entries $p_1, p_2, \ldots p_n$. It is convenient here to treat a string s as an indexed set $S$, with the entries $p_i$ as its members: $S = \{p_i\}$.

Let $f : S \rightarrow S$ be a partial function on $S$ that associates an $lp$ of $S$ with an $rp$ of $S$.[2] Then, to say that $f$ accomplishes the matching we have in mind is to say that it meets the following four conditions: (i) $f$ is one-to-one (ii) if $p_j$ in $S$ is an $rp$ then there is an $lp$, say $p_i$, such that $f(p_i) = p_j$ (iii) if $f(p_i) = p_j$ then $i < j$ (i.e., $p_i$ is

[1] Here, and in the subsequent discussions on parentheses, I follow Manaster [3, pp. 5–6].

[2] A partial function from a set $A$ to a set $B$ is a function from a subset of $A$ to a subset of $B$.We talk of a partial function here because it is defined over only some of the members of $B$, the right parentheses.

to the left of $p_j$ in s), and (iv) there is no interlacing of $lp$–$rp$ pairs, i.e., if $p_i$ and $p_j$ are $lp$s with $i < j$, and $f(p_i) = p_{i'}, f(p_j) = p_{j'}$, and if $j < i'$, then $j < j' < i'$.[3]

If such a function exists for a string then we say that the string has a *proper pairing*. It is such a proper pairing of parentheses that we rely on to specify the order in which operations have to be performed in an algebraic expression. Crucial in this respect is the fact that there can not be two alternative proper pairings for the same string of parentheses.[4]

**Theorem A.1.1** *A finite string of left and right parentheses admits of at most one proper pairing.*

PROOF: To check that it is indeed so, we argue by induction on lengths of strings. We do this by first identifying a pair formed by an $lp$ followed immediately by an $rp$, suppressing the pair, and then relating the resulting string of smaller length to the original one.

Let $p_i$ be the leftmost $rp$ in a string. Assuming that there is a proper pairing of the parentheses, there is then an $lp$, say $p_{i'}$, $i' < i$, for which $f(p_{i'}) = p_i$.[5] Then it must be that $i' = i - 1$. For, suppose there is $j'$, $i' < j' < i$, such that $p'_j$ is an $lp$, pairing with $p_j$, an $rp$ (necessarily) to the right of $p_i$ (i.e., $j > i$). This will amount to an interlacing, contradicting our assumption that we have a proper pairing here. Thus $i' = i - 1$. In all, if the original string has a proper pairing then there is in it an $lp$–$rp$ pair, $p_{i-1}p_i$.

Now the inductive part. Proper pairing requires an even number of parentheses in a string. So, we need to look at strings of lengths $2(n+1)$ for $n = 0, 1, 2 \ldots$. For $n = 0$, there is just one string with proper pairing—the string “()”, and this pairing is unique. Let us now assume that for any $n$, i.e., for any string of length $2(n+1)$, if a proper pairing exists then it is unique. Moving from $n$ to $n+1$, consider now a string $p_1p_2p_3 \ldots p_{2n+2}p_{2n+3}p_{2n+4}$ of length $2((n+1)+1) = 2n+4$ that has a proper pairing. As argued earlier, this string has an innermost pair consisting of $p_{i-1}$, an $lp$, paired with $p_i$, the left most $rp$ of the string. Suppressing this pair, we get a string for which there is at most one proper paring by hypothesis. On the whole, the original string then also has at most one proper pairing. It then follows by induction that for any finite string of parentheses, there is at most one proper pairing. □

This would be a good point to look at a text such as Manaster [3, Chapter 1], whose line of argument I have used here, for additional material related to logic and formal languages.

[3] For example, the pairings resulting in $\overset{(}{p_i}\overset{(}{p_j}\overset{)}{p_{i'}}\overset{)}{p_{j'}}$ are not allowed, whereas those resulting in $\overset{(}{p_i}\overset{(}{p_j}\overset{)}{p_{j'}}\overset{)}{p_{i'}}$ are allowed.

[4] In organizing a mathematical expression in parentheses, we have in mind a certain order in which the operations are to be performed, and the parentheses are meant to indicate that order. But if another proper pairing were possible, then some one evaluating that expression could go by this other pairing and get a result altogether different from the one intended.

[5] It can not be an $rp$ since we have already picked up the leftmost.

## A.2 Parentheses and the Associative Law

Consider a binary operation given on a set $S$. Adopting the multiplicative notation for the operation, the associative law says that for any three elements $x_1, x_2, x_3 \in S$, their products formed in the two possible ways for the given order are equal: $(x_1x_2)x_3 = x_1(x_2x_3)$. So, if the given operation is associative, we omit the parentheses and simply write $x_1x_2x_3$ for the unique product.

Now, if we have more than three elements, we can form their product in successive steps in various different ways. Thus for four elements, $x_1, x_2, x_3, x_4$, for instance, we get $(x_1x_2)(x_3x_4)$, $(x_1(x_2x_3))x_4$, $((x_1x_2)x_3)x_4$, amongst several others. Clearly, the second and the third ones are equal—applying the associative law to the first factor of the former, we get the latter. What about the first one? Well, again by the associative law, we have $(x_1x_2)(x_3x_4) = ((x_1x_2)x_3)x_4$.

We wish to show that the associative law implies equality of such products for any finite number of elements. To be more specific, for elements $x_1, x_2, \ldots, x_n \in S$, $n \geq 3$, consider their products formed in various different ways, each of these in several successive steps indicated by parentheses. We use induction to show the following.

**Proposition A.2.2** *For the given order in which the elements $x_1, x_2, \ldots, x_n$ appear, their product under an associative binary operation is in the end the same for any $n \geq 3$, irrespective of the differences in the successive steps. This unique product we simply write as $x_1x_2 \ldots x_n$, leaving out the parentheses.*

Let us say that the proposition is true for elements numbering less than $n$. Now, for $n$ elements, any particular way of forming their product will in the last step have two factors having less than $n$ elements in each. The product thus takes the form $(x_1x_2 \ldots x_k)(x_{k+1} \ldots x_n)$ for some $k < n$. Likewise, some other way of forming the product yields $(x_1x_2 \ldots x_l)(x_{l+1} \ldots x_n)$ for some $l < n$. Without any loss of generality, let us say that $l < k$. Then, by the associative law, we have

$$\begin{aligned}(x_1x_2 \ldots x_k)(x_{k+1} \ldots x_n) &= (x_1x_2 \ldots x_l)(x_{l+1} \ldots x_k)(x_{k+1} \ldots x_n)\\ &= (x_1x_2 \ldots x_l)(x_{l+1} \ldots x_n)\,.\end{aligned}$$

In other words, any two different ways of forming the product give the same result. Given that the proposition is true for $n = 3$ (the associative law), it follows that it is true for all $n \geq 3$.[1]

**Aside 4** There are some fine points that we need to take notice of here. In many situations, associativity is not directly stipulated for the given binary operation. It inherits it from that of another operation in terms of which it is defined. Consider for instance the case of union of sets. We say that for two sets $S_1$ and $S_2$, their

[1] For more on this and other related issues, see Kochendörffer [2, pp. 1–4] and Artin [1, pp. 40–41].

union $S_1 \cup S_2 = \{x|x \in S_1 \,\text{or}\; x \in S_2\}$. So defined in terms of the set membership relation '$\in$' and the logical connective 'or' (a binary operation on propositions), the union operation is associative because the connective 'or' is so by definition. We may thus invoke Proposition A.2.2 for the union operation applied to $n$ sets and write the union as $S_1 \cup S_2 \cup \cdots \cup S_n$.

We may alternatively define union of $n$ sets in general as $S_1 \cup S_2 \cup \cdots \cup S_n = \{x|x \in S_1 \,\text{or}\; x \in S_2 \,\text{or}\; \cdots \,\text{or}\; x \in S_n\}$. In this definition, Proposition A.2.2 is invoked at the inner logical level for the connective 'or', and $S_1 \cup S_2$ becomes a special case. Of course, whether we define union this way or we first define it as a binary operation and then through repeated applications for $n$ sets, effectively the same purpose is served.

Finally, there is also a point about notation. In $S_1 \cup S_2$, the symbol '$\cup$' denotes a binary operation whereas in $S_1 \cup S_2 \cup \cdots \cup S_n$, its repeated presence denotes an $n$–ary operation. Considering the close connection between the two and the resulting convenience, this ambiguity is harmless. ♡

# References

1. Michael Artin. *Algebra*. Prentice–Hall, New York, 1991.
2. Rudolf Kochendörffer. *Group Theory*. McGraw-Hill, New York, 1970.
3. Alfred B. Manaster. *Completeness, Compactness, and Undecidability*. Prentice–Hall (India), New Delhi, 1975.

# Index