AN INTRODUCTION TO SEMI-TENSOR PRODUCT OF MATRICES AND ITS APPLICATIONS

Daizhan Cheng • Hongsheng Qi • Yin Zhao

World Scientific

AN INTRODUCTION TO SEMI-TENSOR PRODUCT OF MATRICES AND ITS APPLICATIONS

This page intentionally left blank

AN INTRODUCTION TO SEMI-TENSOR PRODUCT OF MATRICES AND ITS APPLICATIONS

Daizhan Cheng
Hongsheng Qi
Yin Zhao

Chinese Academy of Sciences, China

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

AN INTRODUCTION TO SEMI-TENSOR PRODUCT OF MATRICES AND ITS APPLICATIONS

ISBN-13 978-981-4374-68-2
ISBN-10 981-4374-68-7

Printed in Singapore.

Preface

Matrix theory has long been a fundamental tool in mathematical disciplines as well as many other scientific fields. Unlike calculus, which was mainly created by two geniuses: Isaac Newton and Gottfried Leibniz, it is hard to tell who is the principal inventor of matrix theory (or its brother — linear algebra). Though in the 19th century some mathematicians such as Carl Friedrich Gauss, Arthur Cayley, and James Joseph Sylvester *et al* have made significant contributions to it, which became the main body of matrix theory, the matrix may have appeared long before that. There is an ancient Chinese book called "The Nine Chapters on the Mathematical Art" ("Jiu Zhang Suan Shu"), that appeared in 152 BC. In this book a sequence of rectangles of data (matrices) have been used to solve linear systems. Each rectangle contains the coefficients and constants of linear systems, and the sequence of rectangles are used to describe the Gauss substitution process. In fact, a rectangle is exactly an augmented matrix. In Chinese "algebraic equation" is called "Fang Cheng", which means "rectangle process" or "rectangle transformation".

In the epoch of computer, matrix theory becomes more and more important, because it is the fundamental tool in numerical computations. Nowadays the numerical computation is not only a tool for scientific calculations, but also one of the ways to discover truths.

But the classical matrix theory also has some disadvantages.

(1) Matrix is a proper tool to deal with linear and bilinear functions. To deal with multilinear functions or even nonlinear ones matrix theory seems incapable.

(2) Comparing with scalar product, matrix product is less convenient because (i) it has dimensional restriction, and (ii) it is not commutative.

To deal with trilinear functions (or, generally, three-dimensional data), a natural way to generalize the concept of matrix is to arrange the data into a cube. Cubic matrix was firstly proposed by Bates and Watts (Bates and Watts, 1980, 1981). Then it was systemized in Tsai (1983); Wei (1986). A brief survey can be found in Wang (2002). Cubic matrix has achieved some successful applications in statistics. But since it requires several new product rules, it is not very convenient to use. As for the fourth or even higher-dimensional data, cubic matrix theory is also not applicable. To deal with general multilinear mapping, Dr. Zhang Yingshan proposed a multi-edge matrix theory (Zhang, 1993), which is a creative work. Unfortunately, because of the complexity it is not commonly accepted.

Semi-tensor product (STP) is a generalization of the conventional matrix product. It is designed to deal with higher-dimensional data as well as multilinear mappings. The basic idea of STP is motivated by computer science. In a computer the higher-dimensional data can easily be treated without arranging them into a cube or even higher-dimensional cuboid. The data are simply arranged into a long line, and the hierarchies of data are indicated by some marks. For instance, in C-language the pointer, the pointer to pointer, and the pointer to pointer to pointer, *etc.* are used to indicate the hierarchies and then to manipulate the data. The STP of matrices is designed in such a way that the product rule can automatically search the proper position for each factor of multiplier (or subset of data).

Meanwhile, unlike conventional matrix product, the STP does not require the dimension match condition. It can be used for any two matrices. (We refer to Chapter 4 for detailed discussions.) In addition, the STP has certain commutative properties, called the pseudo-commutative property, by enlarging the sizes of factor matrices and/or using an auxiliary matrix, called the swap matrix. Hence, the STP can overcome the second inadequacy of the conventional matrix product in a certain sense. Because of these advantages, the STP becomes a powerful tool in dealing with multilinear and nonlinear calculations in computers. This fact can be seen throughout this book.

Roughly speaking, this book consists of two parts. The first part introduces the concepts and properties of this generalized matrix product, and the second part is various applications, including applications to Boolean functions, cryptograph, universal algebra, physics, and algebra, *etc.* Particularly, the applications to control problems, including fuzzy control systems, control of Boolean networks, and analysis and control of nonlinear systems.

It was pointed out in Emmott (2006) that “Concepts, Theorems and Tools developed within computer science are now being developed into new conceptual tools and technological tools of potentially profound importance, with wide-ranging applications outside the subject in which they originated, especially in sciences investigating complex systems, most notably in biology and chemistry." "The invention of key new conceptual tools (*e.g.* calculus) or technological tools (*e.g.* the telescope, electron microscope) typically form the building blocks of scientific revolutions that have, historically, changed the course of history and society. Such conceptual and technological tools are now emerging at the intersection of computer science, mathematics, biology, chemistry and engineering."

We are confident that STP of matrices is one of such concepts and tools. It is motivated from computer science. Some concepts are adopted from computer science and software programming. It is applied to various problems in cooperation with numerical calculation via computer. Since the concept of STP is so natural, in applications it is so powerful, and mathematically it is so simple, we were frequently asked such questions: "Is the STP a new concept? Why it is not discovered early?" Following sceptics' clues, we have tried to find the "original version" of semi-tensor product. But we failed to find even a similar thing. It seems to us that the STP of matrices can appear only in the epoch of computer. Without computer, you can hardly find even a meaningful example for the application of STP.

Apart from other applications, the application of STP to dynamic systems is significant. It consists of two classes: (i) application to continuous dynamic systems, and (ii) application to logical dynamic systems. We refer to two books for examples of these two kinds of applications: (i) The book of Mei *et al.* (2010) shows the applications of semi-tensor product to power systems, which are typical nonlinear dynamic systems. (ii) The book of Cheng *et al.* (2011b) demonstrates the applications of semi-tensor product to the analysis and control of Boolean networks, which are typical logical dynamic systems.

A brief version of this book was prepared for a graduate course in Shandong University. Then it was expanded into a comprehensive introduction to the theory of STP and its currently known main applications. The book consists of the following contents.

Chapter 1 considers multi-dimensional data, their arrangements, and some of their operations, *etc.* First, their matrix-type arrangements are discussed. Then some nonconventional matrix products are introduced, which are Kronecker product, Hadamard product, and Khatri-Rao product. Finally, some concepts about multi-dimensional data and their properties are investigated. They are (i) tensor form; (ii) Nash equilibrium; (iii) symmetric group; (iv) swap matrix. They are fundamental tools or objects used in the sequel.

Starting from multilinear mappings, Chapter 2 proposes the left STP as a new matrix product. Its basic properties are then studied. Particularly, one sees that the STP is a generalization of the conventional matrix product and it keeps all major properties of the conventional matrix product unchanged. Then some swapping properties of STP are presented, which show that this generalization has certain pseudo-commutative properties. Finally, as a bilinear mapping, some additional properties of the STP are revealed.

Chapter 3 considers the general linear mappings between vector spaces. The STP is used as a basic tool to reconsider the cross product on $\mathbb{R}^3$, structure of general linear algebra, and the mappings over matrices. Then the conversion of different matrix expressions is considered. Finally, two applications, namely, the Lie algebras of Lie groups, and the solvability of Sylvester equations, are discussed.

In Chapter 4 we first consider an alternative STP, namely, the right STP of matrices. Its basic properties and a comparison between left and right STPs are presented. Then both left and right STPs are extended to two matrices of arbitrary dimensions.

Chapter 5 considers some further properties of the STP, which consists of rank, pseudo-inverse, and positivity of the semi-tenor product of two matrices. This chapter is based on the works of a research group in Liaocheng University.

The matrix expression of logical functions is investigated in Chapter 6, where a logical expression is converted into its matrix form (also called its algebraic form). Using its algebraic form, certain fundamental properties of logical functions are revealed. This algebraic form is then used to solve logical equations and deal with logical inferences. Finally, multi-valued logic is introduced.

Chapter 7 proposes a new type of logic, called the mix-valued logic. Its normal form is determined. Then the general logical functions and their algebraic forms are considered. Certain formulas are obtained to convert one form to the other. Finally, some applications of the mix-valued logic are briefly introduced, which include the control of fuzzy systems and the strategy description of dynamic games.

In Chapter 8 fuzzy set and fuzzy logic are investigated. First, the matrices with entries of logical variables are considered, and the operations on them are constructed. Then a finite set, its power sets, and its fuzzy sets are expressed in a uniformed vector form. Finally, the mappings over finite sets, their power sets, and their fuzzy sets are expressed into a uniformed matrix form.

Chapter 9 considers the solvability of fuzzy relational equations. Through analyzing the structure of solutions of f, a new method for obtaining parameter set solutions via STP is proposed. Then, the set of all solutions are constructed by using parameter set solutions. Numerical examples are presented to describe the method.

Chapter 10 considers the fuzzy control problem. First, the multiple fuzzy relations are introduced, the products and compounds of multiple fuzzy relations are proposed. Their matrix expressions and computations are investigated. Then they are used to multiple fuzzy inference. A new concept, called the dual fuzzy relation, is proposed to convert an infinite universe of discourse into its dual one, which is finite. Using it and the matrix expression of multiple fuzzy relations, the design of fuzzification and defuzzification for fuzzy control systems with coupled fuzzy relations is studied.

Boolean functions are particularly useful in cryptography. They are discussed in Chapter 11, in which the polynomial expression of a Boolean function is firstly considered. Based on it, the Walsh transformation and the nonlinearity of Boolean functions are investigated. The conversion back and forth between vector form and the polynomial form of Boolean functions is then investigated and formulas are obtained for numerical calculation. Finally, the results are used to investigate the symmetry of Boolean functions.

Chapter 12 considers the bi-decomposition of Boolean functions, which is particularly important in circuit design. Both disjoint and non-disjoint cases are discussed and the necessary and sufficient conditions for each cases are presented. Then the results are extended to multi-valued and mix-valued cases and the corresponding necessary and sufficient conditions are also obtained.

The Boolean calculus is discussed in Chapter 13. First, the derivative of Boolean functions is defined. Using semi-tensor product, the formulas for calculating derivatives are obtained. Then the indefinite and definite integrals of Boolean functions are proposed, and the calculating formulas are also presented. Finally, some applications, including circuit fault detection *etc.*, are discussed.

Chapter 14 introduces the applications of STP to the lattice, graph and universal algebra, and explores their relationships. First, we consider the matrix expression of lattice. Certain structure properties of lattices are investigated. Then, the graph and its adjacent matrix are analyzed. Planar graph and coloring problem are discussed. Hypergraph is also introduced. Then, the finite universal algebra is discussed. The isomorphism and homomorphism of universal algebras are investigated via structure matrices of their operators. Finally, lattice-based logics, including quantum logic, are investigated.

Chapter 15 considers the application of semi-tensor product to the analysis of Boolean networks. The algebraic form of the dynamics of Boolean networks is proposed. Using it, the topological structure of Boolean networks is investigated. "Rolling gears" structure of large Boolean networks is proposed. Finally, the normal form of dynamic-static Boolean networks is discussed by using the technique developed in Chapter 12.

Chapter 16 considers the control of Boolean networks. A framework, including state space and various subspaces, is constructed. Based on this framework, the synthesis and control of Boolean networks are studied. The basic control problems concerned in this chapter include the controllability, observability, disturbance decoupling, identification, and optimal control, *etc.*

The application of STP to game theory is discussed in Chapter 17. We consider the game with finite players and each player has only finite strategies. Assume the strategies depend on the past finite historical strategies, the strategies can be expressed as a mix-valued logical dynamic system. Then the distance in strategy space can be established. And then the Nash and sub-Nash equilibriums can be obtained by using certain strategy optimization techniques.

Chapter 18 considers the matrix expressions of multi-variable polynomials. Two basic forms are proposed, and their conversions are established. Then the differential of multi-variable smooth functions is considered. Their Taylor expressions are expressed via STP, which are exactly the same form as the one of single variable functions. The basic differential formula is obtained and then it is used to calculate Lie derivatives etc.

Chapter 19 contains some applications of STP to mathematical problems in differential geometry and algebra. The connection in a differential manifold is considered and the Christofel matrix is investigated, and its expressions under different coordinates are obtained. Then, the contraction of tensor fields is finally investigated and the formula is proved. The second part of this chapter considers the application of STP to investigating the structure and properties of finite-dimensional algebras. The structure matrix of an algebra is introduced and investigated. Then the classification and properties of two- and three-dimensional algebras are revealed. Finally, the general and product algebras are investigated. All the discussions are based on the structure matrices of algebras.

Chapter 20 considers the Morgan's problem. That is, the input-output decoupling problem for linear control systems. The problem is a long-standing open problem. A simplified equivalent formula is obtained first. Using the simplified form and the semi-tensor product, the algorithm for a numerical solution to the problem is provided.

Chapter 21 considers some linearization problems of nonlinear dynamic (control) systems. First, the Carleman's linearization is considered. Then the non-regular state feedback linearization of nonlinear control systems is investigated in detail. The problem is reduced to a single-input linearization problem. As the major result, a numerical algorithm for the non-regular state feedback linearization, is presented.

Chapter 22 considers the stabilization of nonlinear control systems. Based on the center manifold theory, the stabilization of non-minimum phase nonlinear systems is solved by designing the center manifold. Derivative homogeneous Lyapunov function is proposed and the design technique, using STP, is obtained.

Some materials of this book come from Cheng and Qi (2007) with the permission of Science Press. So, this book might be considered as a second version of Cheng and Qi (2007), though in English and completely rewritten and much enlarged and updated.

Roughly speaking, Chapters 1-4 and 18 form the fundamental theory of STP, Chapter 5 contains some additional theoretical results, which are rarely used in other chapters. And all other chapters are applications. But (1) Chapter 18 is used only for the problems with some continuous dynamics; (2) some applications are closely related. For your convenience, the relation among chapters is depicted in Fig. 0.1.

Appendix A is used to explain relevant numerical calculations. A toolbox for the algorithms can be downloaded from `http://lsc.amss.ac.cn/~dcheng/`.

The authors are indebted to Prof. L. Guo, Prof. H.-F. Chen of Chinese Academy of Sciences, and to Prof. Q. Lu of Tsinghua University for their warmhearted support. The authors would like to express their sincere thanks to their colleagues, who have made significant contributions to this new field. Some of this long list are Prof. Y. Hong of Chinese Academy of Sciences, Prof. S. Mei, Prof. F. Liu of Tsinghua University, Prof. Y. Wang,

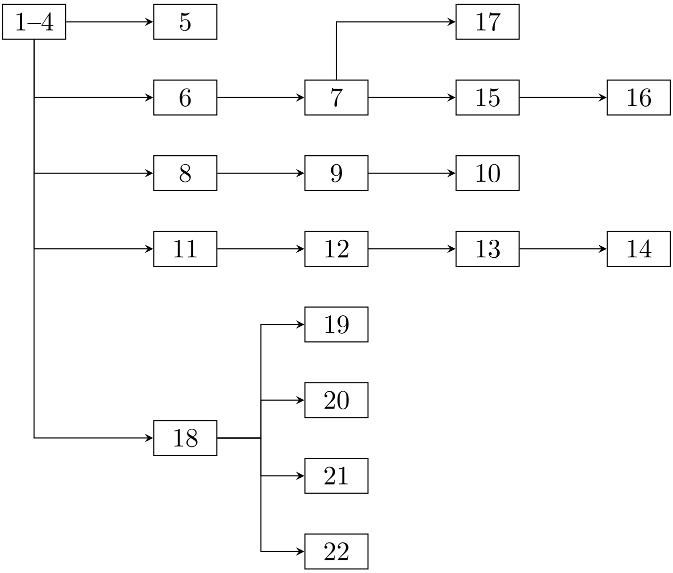


Fig. 0.1 Relation among chapters

Prof. J. Feng of Shandong University, Prof. Z. Li of Henan University of Economics and Law, Prof. J. Sun of Tongji University, Prof. J. Zhao of Liaocheng University, Prof. J. Liu of Bradley University, USA, Prof. M. Margaliot and Dr. D. Laschov of Tel Aviv University, Israel. Thanks also go to some colleagues, who have helped in preparing this manuscript. They are Dr. X. Yin of Washington University, USA, Dr. J. Kim of Glasgow University, *et al.* The works in this book are partly supported by National Natural Science Foundations of China under Grants 61074114, 60736022, 60821091 and 61104065 *etc.*

Finally, we thank World Scientific Publishing for their kind assistance in publishing this book.

Daizhan Cheng, Hongsheng Qi, and Yin Zhao
Chinese Academy of Sciences
November, 2011

Notations

$\mathbb{C}$	set of complex numbers
$\mathbb{R}$	set of real numbers
$\mathbb{Q}$	set of rational numbers
$\mathbb{Z}$	set of integers
$\mathbb{Z}_+$	set of nonnegative integers
$\mathbb{N}$	set of natural numbers
$\mathbb{Z}_n$	finite group $\{1, \cdots, n\}$ with $+(\text{mod } n)$
$:=$	"defined as $\cdots$"
$\mathcal{M}_{m \times n}$	set of $m \times n$ real matrices
$\mathcal{M}_n$	set of $n \times n$ real matrices
$\mathbb{C}_{m \times n}$	set of $m \times n$ complex matrices
$\text{id}(i_1, \cdots, i_k; n_1, \cdots, n_k)$	ordered multi-index
$A \succ_t B$	column number of A is t times of the row number of B
$A \prec_t B$	row number of B is t times of the column number of A
$\mathcal{I}m$	set of images
δ_n^k	kth column of I_n
$\text{lcm}\{\cdot, \cdot\}$	least common multiple
$\text{gcd}\{\cdot, \cdot\}$	greatest common divisor
$\mathcal{D}$	set $\{T, F\}$ or $\{1, 0\}$
$\mathcal{D}_k$	set $\{0, \frac{1}{k-1}, \cdots, \frac{k-2}{k-1}, 1\}$
$\mathcal{D}_\infty$	set $\{r \in \mathbb{R} \| 0 \leq r \leq 1\}$
$\mathcal{D}_k^{m \times n}$	set $m \times n$ matrices with entries in $\mathcal{D}_k$
Δ	set $\{\delta_2^1, \delta_2^2\}$
Δ_k	set of δ_k^i, $1 \leq i \leq k$
$\mathcal{L}(U; V)$	set of linear mapping from U to V

$\mathcal{L}(U_1, \cdots, U_k; V)$	set of multilinear mapping from $U_1 \times \cdots \times U_k$ to V
$\mathcal{T}_t^s$	set of tensors with covariant order s and contra-variant order t
$\mathrm{Col}(A)$	set of columns of matrix A
$\mathrm{Col}_i(A)$	ith column of matrix A
$\mathrm{Row}(A)$	set of rows of matrix A
$\mathrm{Row}_i(A)$	ith row of matrix A
$\mathrm{diag}(A_1, \cdots, A_k)$	block diagonal matrix whose diagonal blocks are $A_i, i = 1, \cdots, k$
A^{-T}	$A^{-T} := (A^T)^{-1}$
A^*	$A^* := (\bar{A})^T$
$\otimes$	tensor (or Kronecker) product
$A^{\otimes k}$	$\underbrace{A \otimes \cdots \otimes A}_{k}$
$\ltimes$	left semi-tensor product
$\rtimes$	right semi-tensor product
$V_c(A)$	column stacking form of matrix A
$V_r(A)$	row stacking form of matrix A
$\mathrm{tr}(A)$	trace of A
$\mathbf{S}_k$	symmetric group of k elements
$W_{[m,n]}$	swap matrix with index (m, n)
$W_{[n]}$	$W_{[n]} = W_{[m,n]}$
$\mathbf{1}_k$	$[\underbrace{1, 1, \cdots, 1}_{k}]^T$
$m \mid n$	m is a divisor of n
$\mathcal{L}_{m \times n}$	set of $m \times n$-dimensional logical matrices
$\delta_k[i_1, \cdots, i_s]$	logical matrix with $\delta_k^{i_j}$ as its jth column
$\delta_k\{i_1, \cdots, i_s\}$	$\left\{\delta_k^{i_1}, \cdots, \delta_k^{i_s}\right\} \subset \Delta_k$
$\mathcal{B}_{m \times n}$	set of $m \times n$-dimensional Boolean matrices
$\mathrm{Span} \cdots$	vector space spanned by $\cdots$
$\bowtie$	cross product on $\mathbb{R}^3$
$H < G$	H is a subgroup of G
$GL(n, \mathbb{R})$	nth order general linear group
$gl(n, \mathbb{R})$	nth order general linear algebra
$\neg$	negation
$\vee$	disjunction
$\wedge$	conjunction

$\rightarrow$	conditional
$\leftrightarrow$	biconditional
$\bar{\vee}$	exclusive or (EOR)
$\uparrow$	not and (NAND)
$\downarrow$	not or (NOR)
$\oslash$	rotator
$\triangledown$	confirmor
$\Rightarrow$	implication
$\Leftrightarrow$	equivalence
T_f	truth vector of f
$R(x_0)$	reachable set from x_0
$R_s(x_0)$	reachable set from x_0 at s steps
$\mathrm{Blk}_i(A)$	ith block of matrix A
$\mathcal{P}(E)$	set of subsets of E
$\mathcal{F}(E)$	set of fuzzy subsets of E
$\mathcal{P}(k)$	set of proper factors of $k \in \mathbb{Z}_+$
$\sqcup$	joint
$\sqcap$	meet
T_t	transient period
Ω	limit set
$(+)$	$\vee$-addition of k-valued matrices
$(\times)$	$\vee$-product of k-valued matrices
$A^{(k)}$	$\vee$-power of k-valued matrix A
$\langle + \rangle$	mod 2 addition of Boolean matrices
$\langle \times \rangle$	mod 2 product of Boolean matrices
$A^{\langle k \rangle}$	mod 2 power of Boolean matrix A
$D_v(A, B)$	vector distance of $A, B \in \mathcal{B}_{m\times n}$
$\mathcal{X}$	logical state space
$\mathcal{F}_\ell\{\cdots\}$	logical subspace generated by $\cdots$

This page intentionally left blank

Contents

Chapter 1

Multi-Dimensional Data

Roughly speaking, the classical matrix theory can mainly deal with one- or two-dimensional data. The main purpose of semi-tensor product of matrices is to use matrix tools to deal with higher-dimensional data. Hence the multi-dimensional data become the main objective of this book. This chapter considers how to arrange a set of higher-dimensional data into a vector or a matrix. First, the ordered multi-index is introduced to arrange a set of data into a properly ordered form. Then we briefly introduce some other matrix products, including the Kronecker Product (also called the tensor product) of matrices, which is a fundamental tool in this book; the Hadamard product and Khatri-Rao product, which are also used in the sequel. Tensor and Nash equilibrium are two useful examples for multi-dimensional data. They are introduced in this chapter and will be used and discussed again later. Symmetric group is another useful tool and it is also introduced here. Finally, we propose a special matrix, called the swap matrix, and its certain properties are discussed. It will be used largely to overcome the non-commutativity of matrix product.

1.1 Multi-Dimensional Data

In scientific researches we have to deal with various kinds of data. First, we would like to clarify what do we mean the dimension of data. A set of data may depend on k factors. Assume each factor can have n_j levels, $j = 1, \cdots, k$, then to label this set of data, we may need k indices $i_1, i_2, \cdots, i_k$, and allow i_j runs from 1 to n_j. We, therefore, have a finite set of data as

$$D := \{d_{i_1,\cdots,i_k} \mid 1 \le i_j \le n_j,\, j = 1, \cdots, k\}. \tag{1.1}$$

Under this circumstances we say that the dimension of D is k, denoted by $\dim(D) = k$. It is not clear how to give a rigorous definition for the dimension of a set of data. Roughly speaking, the dimension of a set of data is the number of indices of the data. It is enough for our application. (But since indices are changeable, so this is not a rigorous definition.)

In this chapter only finite sets of data are considered, unless elsewhere stated.

Example 1.1.

(1) Consider a vector $X \in \mathbb{R}^n$. It can be expressed as $X = (x_1, x_2, \cdots, x_n)^T$. Hence a vector can be considered as a set of one-dimensional data.

(2) Consider a matrix $A \in \mathcal{M}_{m\times n}$. It can be expressed as

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

Hence, in a natural way, a matrix can be considered as a set of two-dimensional data.

(3) Let $y = f(x_1, x_2, \cdots, x_n)$ be a function of n variables. To get numerical expression of f, we let x_i take values $x_i^1, x_i^2, \cdots, x_i^{n_i}$. Then we have a set of n-dimensional data as

$$Y := \left\{ y_{i_1,\cdots,i_n} = f(x_1^{i_1}, \cdots, x_n^{i_n}) \,\middle|\, 1 \le i_j \le n_j,\ j = 1, \cdots, n \right\}.$$

To arrange a multi-dimensional data, or equivalently, to decide the order of a data, we introduce an ordered multi-index, briefly, multi-index.

Definition 1.1. A set of k-dimensional data (1.1) is said to be arranged by the (ordered) multi-index $\mathrm{id}(i_1, \cdots, i_k; n_1, \cdots, n_k)$, if the data are labeled by indices $i_1, \cdots, i_k$, and arranged in the order that $d_{p_1,\cdots,p_k} \prec d_{q_1,\cdots,q_k}$, if and only if there exists a $1 \le r \le k$, such that $p_i = q_i$, $i < r$ and $p_r < q_r$.

Example 1.2. Let

$$D = \{x_{i,j,k} \,|\, 1 \le i \le 2;\ 1 \le j \le 3;\ 1 \le k \le 4\}.$$

(i) If we arrange it by the multi-index $\mathrm{id}(i, j, k; 2, 3, 4)$, we have

$$[x_{111}, x_{112}, x_{113}, x_{114}, x_{121}, x_{122}, x_{123}, x_{124}, x_{131}, x_{132}, x_{133}, x_{134},$$
$$x_{211}, x_{212}, x_{213}, x_{214}, x_{221}, x_{222}, x_{223}, x_{224}, x_{231}, x_{232}, x_{233}, x_{234}].$$

(ii) If we arrange it by the multi-index $\mathrm{id}(j, i, k; 3, 2, 4)$, we have

$$[x_{111}, x_{112}, x_{113}, x_{114}, x_{211}, x_{212}, x_{213}, x_{214}, x_{121}, x_{122}, x_{123}, x_{124},\\ x_{221}, x_{222}, x_{223}, x_{224}, x_{131}, x_{132}, x_{133}, x_{134}, x_{231}, x_{232}, x_{233}, x_{234}].$$

(iii) If we arrange it by the multi-index $\mathrm{id}(k, j, i; 4, 3, 2)$, we have

$$[x_{111}, x_{211}, x_{121}, x_{221}, x_{131}, x_{231}, x_{112}, x_{212}, x_{122}, x_{222}, x_{132}, x_{232},\\ x_{113}, x_{213}, x_{123}, x_{223}, x_{133}, x_{233}, x_{114}, x_{214}, x_{124}, x_{224}, x_{134}, x_{234}].$$

Assume we have a set of n data, where $n = \prod_{i=1}^{k} n_i$. Then we can either use a single index to label the data as

$$D = [x_1, x_2, \cdots, x_n], \tag{1.2}$$

or use multi-index $\mathrm{id}(i_1, \cdots, i_k; n_i, \cdots, n_k)$ to arrange the data as

$$D = [x_{11\cdots 1}, x_{11\cdots 2}, \cdots, x_{11\cdots n_k}, \cdots, x_{n_1 n_2 \cdots n_k}]. \tag{1.3}$$

Then we need to find formulas to convert the single index to multi-index and vice versa. In the following we deduce the formulas. For notational ease, we introduce some notations as follows.

- Let $a \in \mathbb{Z}$, $0 < b \in \mathbb{Z}_+$. As in C-language, the $a\%b$ is used for the remaining of a/b.
- Denote by $[t]$ the largest integer, which is less than or equal to t.

For instance,

$$100\%3 = 1, \quad 100\%7 = 2, \quad (-7)\%3 = 2;$$
$$\left[\frac{7}{3}\right] = 2, \quad [-1.25] = -2.$$

It is easy to see that

$$a = \left[\frac{a}{b}\right] b + a\%b. \tag{1.4}$$

Next, we give the converting formulas between single index and multi-index of a set of data. We leave the proves to the reader.

Proposition 1.1. *Let D be a set of $n = \prod_{i=1}^{k} n_i$ data. It has been labeled by single index as in (1.2) and by multi-index as in (1.3). An element $x \in D$ is labeled by single index p and multi-index $\mu_1 \cdots \mu_k$. That is, $x \in D$ is expressed as*

$$x = x_p = x_{\mu_1 \cdots \mu_k}.$$

Then we have the following converting formulas:

(1) Set $p_k := p-1$, then $(\mu_1,\cdots,\mu_k)$ can be calculated iteratively by

$$\begin{cases}\mu_k = p_k\%n_k + 1,\\ p_j = \left[\frac{p_{j+1}}{n_{j+1}}\right], \quad \mu_j = p_j\%n_j + 1, \quad j = k-1,\cdots,1.\end{cases} \tag{1.5}$$

(2) Conversely, from single index to multi-index we have:

$$p = \sum_{j=1}^{k-1}(\mu_j - 1)n_{j+1}n_{j+2}\cdots n_k + \mu_k. \tag{1.6}$$

The following example shows the conversions.

Example 1.3. Consider a set of data

$$D = \{d_1, d_2, \cdots, d_{100}\}.$$

(1) Given a number $p = 35$, what is the multi-label of d_p under multi-index $\mathrm{id}(i_1, i_2, i_3; 4, 5, 5)$?
Using (1.5), we have

$$\begin{aligned} p_3 &= p - 1 = 34,\\ \mu_3 &= p_3\%n_3 + 1 = 34\%5 + 1 = 4 + 1 = 5,\\ p_2 &= [p_3/n_3] = [34/5] = 6,\\ \mu_2 &= p_2\%n_2 + 1 = 6\%5 + 1 = 1 + 1 = 2,\\ p_1 &= [p_2/n_2] = [6/4] = 1,\\ \mu_1 &= p_1\%n_1 + 1 = 1\%4 + 1 = 1 + 1 = 2. \end{aligned}$$

Thus the multi-index of p is $(2, 2, 5)$.

(2) Assume the multi-label of $d_q \in D$ under multi-index $\mathrm{id}(i_1, i_2, i_3; 5, 2, 10)$ is $(3, 2, 8)$. What is the single index q?
Using (1.6), we have

$$q = (\mu_1 - 1)n_2n_3 + (\mu_2 - 1)n_3 + \mu_3 = 2\cdot 2\cdot 10 + 10 + 8 = 58.$$

1.2 Arrangement of Data

Let D be a k-dimensional data as in (1.1) with $n = \prod_{i=1}^{k} n_i$. When $k = 1$, $n = n_1$, the data can be arranged into a vector as

$$V_D = (d_1, d_2, \cdots, d_n)^T. \tag{1.7}$$

When $k = 2$, the data can naturally be arranged into a matrix as

$$M_D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n_2} \\ d_{21} & d_{22} & \cdots & d_{2n_2} \\ \vdots & & & \\ d_{n_1 1} & d_{n_1 2} & \cdots & d_{n_1 n_2} \end{bmatrix}. \tag{1.8}$$

Vector and matrix are two major objects for matrix theory or linear algebra. Roughly speaking, matrix theory is a theory for one- or two-dimensional data. Now when $k \geq 3$, then what can we do? Let us first consider the case of $k = 3$. Say, we have a set of data D as

$$D = \{d_{ijk} \,|\, i = 1, \cdots, p; j = 1, \cdots, m; k = 1, \cdots, n\}. \tag{1.9}$$

Then how can we arrange the data? It was proposed by some researchers that the data are arranged into a cube, which consists of p layers and each layer is an $m \times n$ matrix. Such a compounded matrix is called a cubic matrix (Fig. 1.1). To manipulate such matrices several new product rules have to be proposed for the product of cubic matrix with cubic matrix, or with normal matrix, or with vector. Many new properties about the products need also to be developed (Bates and Watts, 1980; Tsai, 1983). Furthermore, a natural question is: when $k > 3$ what can we do? It seems that this is not a proper way to deal with the higher-dimensional data.

It is well known that a set of higher-dimensional data can easily be stored in a computer memory, where they are not arranged into a cubic or even higher-dimensional cuboid. In fact, the data are arranged into a line regardless the dimension of the data. Then how a computer to find the hierarchy structure of the data? they use some marks. Say, in C-language the pointer, pointer to pointer, pointer to pointer to pointer etc. are used to indicate the data structure.

Motivated by the computer technology, we propose to arrange a set of data into either a vector or a matrix. One may ask why not just use vector only. Could it be more convenient? In fact, to use tools developed in matrix theory, both vector and matrix forms are necessary.

To arrange a set of multi-dimensional data into a vector is rather easy. Formula (1.6) provides the single index label for each data. We are also interested in arranging the entries of a matrix into a vector.

Definition 1.2. Consider the matrix M_D in (1.8).

(1) Its row stacking form, denoted by $V_r(M_D)$, is defined as

$$\begin{aligned} V_r(M_D) = (&d_{11}, d_{12}, \cdots, d_{1n_2}, d_{21}, d_{22}, \cdots, d_{2n_2}, \\ &\cdots, d_{n_1 1}, d_{n_1 2}, \cdots, d_{n_1 n_2})^T. \end{aligned} \tag{1.10}$$

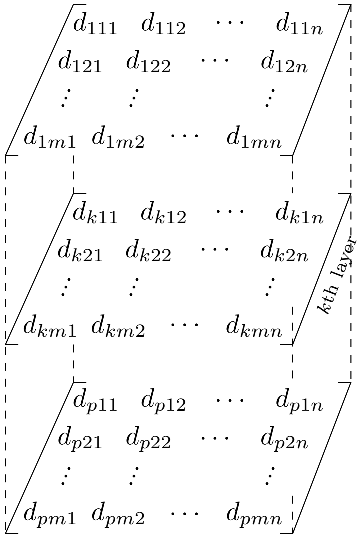

d_{111} d_{112} $\cdots$ d_{11n}
d_{121} d_{122} $\cdots$ d_{12n}
d_{1m1} d_{1m2} $\cdots$ d_{1mn}
d_{k11} d_{k12} $\cdots$ d_{k1n}
d_{k21} d_{k22} $\cdots$ d_{k2n}
d_{km1} d_{km2} $\cdots$ d_{kmn}
kth layer
d_{p11} d_{p12} $\cdots$ d_{p1n}
d_{p21} d_{p22} $\cdots$ d_{p2n}
d_{pm1} d_{pm2} $\cdots$ d_{pmn}

Fig. 1.1 A cubic matrix

(2) Its column stacking form, denoted by $V_c(M_D)$, is defined as

$$V_c(M_D) = (d_{11}, d_{21}, \cdots, d_{n_1 1}, d_{12}, d_{22}, \cdots, d_{n_1 2}, \\ \cdots, d_{1n_2}, d_{2n_2}, \cdots, d_{n_1 n_2})^T. \tag{1.11}$$

By definition, it is obvious that for any matrix A,

$$V_r(A) = V_c(A^T), \quad V_c(A) = V_r(A^T). \tag{1.12}$$

Next, we consider how to arrange a set of multi-dimensional data into a matrix form. Let $\mathrm{id}(i_1, \cdots, i_k; n_1, \cdots, n_k)$ be a multi-index, and

$$\{i_{j_1}, \cdots, i_{j_p}\} \subset \{i_1, \cdots, i_k\}.$$

Then $\mathrm{id}\left(i_{j_1}, \cdots, i_{j_p}; n_{j_1}, \cdots, n_{j_p}\right)$ is called a sub-index of $\mathrm{id}(i_1, \cdots, i_k; n_1, \cdots, n_k)$.

Definition 1.3. Let D be a k-dimensional data as in (1.1). Assume there are two disjoint sub-indices which form a partition of the index set of D as

$$\left\{i_{\alpha_1}, i_{\alpha_2}, \cdots, i_{\alpha_p}\right\} \cup \left\{i_{\beta_1}, i_{\beta_2}, \cdots, i_{\beta_q}\right\} = \left\{i_i, i_2, \cdots, i_k\right\}.$$

D is said to be arranged into a matrix (M_D) in the order of

$$\mathrm{id}\left(i_{\alpha_1}, \cdots, i_{\alpha_p}; n_{\alpha_1}, \cdots, n_{\alpha_p}\right) \times \mathrm{id}\left(i_{\beta_1}, \cdots, i_{\beta_q}; n_{\beta_1}, \cdots, n_{\beta_q}\right),$$

if the multi-indexed matrix M_D is defined as follows:

(i) $M_D \in \mathcal{M}_{n_r \times n_c}$, where $n_r = \prod_{j=1}^{p} n_{\alpha_j}$ and $n_c = \prod_{j=1}^{q} n_{\beta_j}$.

(ii) The rows of M_D is labeled by multi-index id $\left(i_{\alpha_1}, i_{\alpha_2}, \cdots, i_{\alpha_p}; n_{\alpha_1}, n_{\alpha_2}, \cdots, n_{\alpha_p}\right)$ and its columns is labeled by multi-index id $\left(i_{\beta_1}, i_{\beta_2}, \cdots, i_{\beta_q}; n_{\beta_1}, n_{\beta_2}, \cdots, n_{\beta_q}\right)$.

(iii) the $((\alpha_1, \cdots, \alpha_p), (\beta_1, \cdots, \beta_q))$th element of M_D has its label $\{\alpha_1, \cdots, \alpha_p\} \cup \{\beta_1, \cdots, \beta_q\}$ (in the corresponding order).

Example 1.4. Given a 4-dimensional data

$$D = \{d_{i,j,k,r} | i = 1, 2; j = 1, 2, 3; k = 1, 2, 3, 4; r = 1, 2\}.$$

(1) Arranging it into a matrix in the order of id$(i, j; 2, 3) \times$ id$(k, r; 4, 2)$, we have

$$M_D = \begin{bmatrix} d_{1111} & d_{1112} & d_{1121} & d_{1122} & d_{1131} & d_{1132} & d_{1141} & d_{1142} \\ d_{1211} & d_{1212} & d_{1221} & d_{1222} & d_{1231} & d_{1232} & d_{1241} & d_{1242} \\ d_{1311} & d_{1312} & d_{1321} & d_{1322} & d_{1331} & d_{1332} & d_{1341} & d_{1342} \\ d_{2111} & d_{2112} & d_{2121} & d_{2122} & d_{2131} & d_{2132} & d_{2141} & d_{2142} \\ d_{2211} & d_{2212} & d_{2221} & d_{2222} & d_{2231} & d_{2232} & d_{2241} & d_{2242} \\ d_{2311} & d_{2312} & d_{2321} & d_{2322} & d_{2331} & d_{2332} & d_{2341} & d_{2342} \end{bmatrix} \in \mathcal{M}_{6 \times 8}.$$

(2) Arranging it into a matrix in the order of id$(i, k; 2, 4) \times$ id$(j, r; 3, 2)$, we have

$$M_D = \begin{bmatrix} d_{1111} & d_{1112} & d_{1211} & d_{1212} & d_{1311} & d_{1312} \\ d_{1121} & d_{1122} & d_{1221} & d_{1222} & d_{1321} & d_{1322} \\ d_{1131} & d_{1132} & d_{1231} & d_{1232} & d_{1331} & d_{1332} \\ d_{1141} & d_{1142} & d_{1241} & d_{1242} & d_{1341} & d_{1342} \\ d_{2111} & d_{2112} & d_{2211} & d_{2212} & d_{2311} & d_{2312} \\ d_{2121} & d_{2122} & d_{2221} & d_{2222} & d_{2321} & d_{2322} \\ d_{2131} & d_{2132} & d_{2231} & d_{2232} & d_{2331} & d_{2332} \\ d_{2141} & d_{2142} & d_{2241} & d_{2242} & d_{2341} & d_{2342} \end{bmatrix} \in \mathcal{M}_{8 \times 6}.$$

1.3 Matrix Products

In addition to conventional matrix product, there are some other matrix products, which will be used throughout this book. This section gives a brief survey on their definitions and basic properties without proves. In the following we refer to some standard references of matrix theory for details.

1.3.1 *Kronecker Product of Matrices*

The Kronecker product of matrices is also called the tensor product of matrices. This product is applicable to any two matrices. It will be used from time to time throughout this book. We refer to Horn and Johnson (1991) for a complete discussion.

Definition 1.4. Let $A = (a_{ij}) \in \mathcal{M}_{m\times n}$ and $B = (b_{ij}) \in \mathcal{M}_{p\times q}$. The Kronecker product of A and B is defined as

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{11}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & & & \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix} \in \mathcal{M}_{mp\times nq}. \tag{1.13}$$

Next, we introduce some basic properties of the Kronecker product:

Proposition 1.2.

(1) (Associative Law)

$$A \otimes (B \otimes C) = (A \otimes B) \otimes C. \tag{1.14}$$

(2) (Distributive Law)

$$(\alpha A + \beta B) \otimes C = \alpha(A \otimes C) + \beta(B \otimes C), \tag{1.15}$$

$$A \otimes (\alpha B + \beta C) = \alpha(A \otimes B) + \beta(A \otimes C), \quad \alpha, \beta \in \mathbb{R}. \tag{1.16}$$

Proposition 1.3.

(1)

$$(A \otimes B)^T = A^T \otimes B^T. \tag{1.17}$$

(2) Assume A and B are invertible. Then

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}. \tag{1.18}$$

(3)

$$\operatorname{rank}(A \otimes B) = \operatorname{rank}(A)\operatorname{rank}(B). \tag{1.19}$$

(4) Let $A \in \mathcal{M}_{m\times m}$ and $B \in \mathcal{M}_{n\times n}$. Then

$$\det(A \otimes B) = (\det(A))^n(\det(B))^m. \tag{1.20}$$

$$\operatorname{tr}(A \otimes B) = \operatorname{tr}(A)\operatorname{tr}(B). \tag{1.21}$$

The next proposition is very useful.

Proposition 1.4. *Let* $A \in \mathcal{M}_{m\times n}$*,* $B \in \mathcal{M}_{p\times q}$*,* $C \in \mathcal{M}_{n\times r}$*, and* $D \in \mathcal{M}_{q\times s}$*. Then*

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD). \tag{1.22}$$

Particularly, we have

$$A \otimes B = (A \otimes I_p)(I_n \otimes B). \tag{1.23}$$

The next proposition is about the vector form of matrices.

Proposition 1.5.

(1) Let $X \in \mathbb{R}^n$ *and* $Y \in \mathbb{R}^n$ *be two column vectors. Then*

$$V_c(XY^T) = Y \otimes X. \tag{1.24}$$

(2) Let $A \in \mathcal{M}_{m\times p}$*,* $B \in \mathcal{M}_{p\times q}$*, and* $C \in \mathcal{M}_{q\times n}$*. Then*

$$V_c(ABC) = (C^T \otimes A)V_c(B). \tag{1.25}$$

1.3.2 *Hadamard Product*

Hadamard product of matrices is another useful product in certain problems. It will also be used in the sequel. The reader is referred to Horn and Johnson (1991); Zhang (2004) for more details about it.

Definition 1.5. Let $A = (a_{i,j}), B = (b_{i,j}) \in \mathcal{M}_{m\times n}$. The Hadamard product of A and B is defined as

$$A \odot B = (a_{i,j}b_{i,j}) \in \mathcal{M}_{m\times n}. \tag{1.26}$$

Hadamard product has some important properties as

Proposition 1.6.

(1) (Commutativity) For any two matrices $A, B \in \mathcal{M}_{m\times n}$

$$A \odot B = B \odot A. \tag{1.27}$$

(2) (Associative Law) Let $A, B, C \in \mathcal{M}_{m\times n}$*. Then*

$$(A \odot B) \odot C = A \odot (B \odot C). \tag{1.28}$$

(3) (Distributive Law) Let $A, B, C \in \mathcal{M}_{m\times n}$*. Then*

$$(\alpha A + \beta B) \odot C = \alpha(A \odot C) + \beta(B \odot C), \quad \alpha, \beta \in \mathbb{R}. \tag{1.29}$$

Proposition 1.7.

(1)

$$(A \odot B)^T = A^T \odot B^T. \tag{1.30}$$

(2) Let $A \in \mathcal{M}_n$ and $E = \mathbf{1}_n$. Then

$$A \odot (EE^T) = A = (EE^T) \odot A. \tag{1.31}$$

(3) Let $X, Y \in \mathbb{R}^n$ be two column vectors. Then

$$(XX^T) \odot (YY^T) = (X \odot Y)(X \odot Y)^T. \tag{1.32}$$

Define

$$H_n = \operatorname{diag}(\delta_n^1, \cdots, \delta_n^n).$$

Then we have

Proposition 1.8. *Let $A, B \in \mathcal{M}_{m\times n}$. Then*

$$A \odot B = H_m^T (A \otimes B) H_n. \tag{1.33}$$

Proposition 1.9 (Schur's Theorem). *Let $A, B \in \mathcal{M}_n$ be symmetric.*

(i) If $A \geq 0$ and $B \geq 0$, then $A \circ B \geq 0$;
(ii) If $A > 0$ and $B > 0$, then $A \circ B > 0$.

Proposition 1.10 (Oppenbeim's Theorem). *Let $A, B \in \mathcal{M}_n$ be symmetric. If $A \geq 0$ and $B \geq 0$, then*

$$\det(A \odot B) \geq \det(A)\det(B). \tag{1.34}$$

1.3.3 *Khatri-Rao Product*

We refer to Ljung and Söderström (1982) or Zhang (2004) for details of Khatri-Rao product of matrices.

Definition 1.6. Let $A \in \mathcal{M}_{m\times r}$ and $B \in \mathcal{M}_{n\times r}$. The Khatri-Rao product of A and B is defined as

$$A * B = \left[\operatorname{Col}_1(A) \otimes \operatorname{Col}_1(B), \operatorname{Col}_2(A) \otimes \operatorname{Col}_2(B), \cdots, \operatorname{Col}_r(A) \otimes \operatorname{Col}_r(B)\right]. \tag{1.35}$$

Proposition 1.11.

(1) (Associative Law) Let $A \in \mathcal{M}_{m\times r}$, $B \in \mathcal{M}_{n\times r}$, and $C \in \mathcal{M}_{p\times r}$. Then

$$(A * B) * C = A * (B * C). \tag{1.36}$$

(2) (Distributive Law) Let $A, B \in \mathcal{M}_{m\times r}$ *and* $C \in \mathcal{M}_{n\times r}$*. Then*

$$(aA + bB) * C = a(A * C) + b(B * C), \quad a, b \in \mathbb{R}. \tag{1.37}$$

$$C * (aA + bB) = a(C * A) + b(C * B), \quad a, b \in \mathbb{R}. \tag{1.38}$$

The following example is useful in the sequel.

Example 1.5. A matrix $A \in \mathcal{M}_{m\times r}$ is called a logical matrix if all of its columns are of the form δ_m^i, $1 \le i \le m$. The set of $m \times r$ logical matrices is denoted by $\mathcal{L}_{m\times r}$.

Assume $A \in \mathcal{L}_{m\times r}$ and $B \in \mathcal{L}_{n\times r}$. Then

$$A * B \in \mathcal{L}_{mn\times r}.$$

Remark 1.1. In addition to conventional matrix product, we have introduced Kronecker product, Hadamard product, and Khatri-Rao product of matrices. One sees easily that the associativity and distributivity are two common properties. These two properties may be considered as two fundamental requirements for any matrix products.

1.4 Tensor

Tensor is a typical multilinear mapping. This section is a brief introduction. We refer to Boothby (1986) for details.

Let V be an n-dimensional vector space with a basis $\{d_1, \cdots, d_n\}$. Denote by V^* the dual space of V, that is V^* is the set of linear functions on V. Let $\{e_1, \cdots, e_n\} \subset V^*$ be a basis of V^*, which is dual to $\{d_1, \cdots, d_n\}$. That is,

$$e_i(d_j) = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

Then $X = \sum\limits_{i=1}^{n} x_i d_i \in V$ can be expressed as a column vector $X = (x_1, \cdots, x_n)^T$, and $\omega = \sum\limits_{i=1}^{n} \omega_i e_i \in V^*$ as a row vector $\omega = (\omega_1, \cdots, \omega_n)$.

Definition 1.7.

(1) Let $f : V^s \to \mathbb{R}$ be an s-linear mapping, and

$$f(d_{i_1}, \cdots, d_{i_s}) = \mu_{i_1 i_2 \cdots i_s}, \quad 1 \le i_p \le n,\ p = 1, \cdots, s.$$

Arrange $\{\mu_{i_1 i_2 \cdots i_s} | 1 \le i_j \le n,\ j = 1, \cdots, s\}$ into a row by using multi-index $\mathrm{id}(i_1, \cdots, i_s; n, \cdots, n)$ as

$$M_f = [\mu_{11\cdots 1}\ \cdots\ \mu_{11\cdots n}\ \cdots\ \mu_{nn\cdots n}]. \tag{1.39}$$

M_f is called the structure matrix of f. By the linearity, it is easy to check that for $X_1, \cdots, X_s \in V$

$$f(X_1, \cdots, X_s) = M_f(X_1 \otimes \cdots \otimes X_s). \tag{1.40}$$

f is called a tensor of covariant order s. The set of tensors on V of covariant order s is denoted by $\mathcal{T}^s$.

(2) Let $f : (V^*)^t \to \mathbb{R}$ be a t-linear mapping, and

$$f(e_{j_1}, \cdots, e_{j_t}) = \mu^{j_1 j_2 \cdots j_t}, \quad 1 \le j_q \le n,\ q = 1, \cdots, t.$$

Arrange $\{\mu^{j_1 j_2 \cdots j_t} | 1 \le j_q \le n,\ q = 1, \cdots, t\}$ into a column by using multi-index $\mathrm{id}(j_1 j_2 \cdots j_t; n, \cdots, n)$ as

$$M_f = [\mu^{11\cdots 1}\ \cdots\ \mu^{11\cdots n}\ \cdots\ \mu^{nn\cdots n}]^T. \tag{1.41}$$

M_f is called the structure matrix of f. By the linearity, it is easy to check that for $\omega_1, \cdots, \omega_t \in V^*$

$$f(\omega_1, \cdots, \omega_t) = (\omega_1 \otimes \cdots \otimes \omega_t) M_f. \tag{1.42}$$

f is called a tensor of contra-variant order t. The set of tensors on V of contra-variant order t is denoted by $\mathcal{T}_t$.

(3) Let $f : V^s \times (V^*)^t \to \mathbb{R}$ be an $(s+t)$-linear mapping, and

$$\begin{aligned} &f(d_{i_1}, \cdots, d_{i_s}, e_{j_1}, \cdots, e_{j_t}) = \mu^{j_1 j_2 \cdots j_t}_{i_1 i_2 \cdots i_s}, \\ &\quad 1 \le i_p \le n,\ p = 1, \cdots, s;\ 1 \le j_q \le n,\ q = 1, \cdots, t. \end{aligned}$$

Arrange $\left\{\mu^{j_1 j_2 \cdots j_t}_{i_1 i_2 \cdots i_s} | 1 \le i_p \le n,\ p = 1, \cdots, s;\ 1 \le j_q \le n,\ q = 1, \cdots, t\right\}$ into a matrix in the order

$$\mathrm{id}(j_1, \cdots, j_t; n, \cdots, n) \times \mathrm{id}(i_1, \cdots, i_s; n, \cdots, n).$$

(Precisely, its columns are labeled by multi-index $\mathrm{id}(i_1, \cdots, i_s; n, \cdots, n)$ and rows are labeled by multi-index $\mathrm{id}(j_1, \cdots, j_t; n, \cdots, n)$.) Then we have

$$M_f = \begin{bmatrix} \mu^{11\cdots 1}_{11\cdots 1} & \mu^{11\cdots 1}_{11\cdots 2} & \cdots & \mu^{11\cdots 1}_{11\cdots n} & \cdots & \mu^{11\cdots 1}_{nn\cdots n} \\ \mu^{11\cdots 2}_{11\cdots 1} & \mu^{11\cdots 2}_{11\cdots 2} & \cdots & \mu^{11\cdots 2}_{11\cdots n} & \cdots & \mu^{11\cdots 2}_{nn\cdots n} \\ \vdots & & & & & \\ \mu^{11\cdots n}_{11\cdots 1} & \mu^{11\cdots n}_{11\cdots 2} & \cdots & \mu^{11\cdots n}_{11\cdots n} & \cdots & \mu^{11\cdots n}_{nn\cdots n} \\ \vdots & & & & & \\ \mu^{nn\cdots n}_{11\cdots 1} & \mu^{nn\cdots n}_{11\cdots 2} & \cdots & \mu^{nn\cdots n}_{11\cdots n} & \cdots & \mu^{nn\cdots n}_{nn\cdots n} \end{bmatrix}. \tag{1.43}$$

By the linearity, it is also easy to check that for $X_1, \cdots, X_s \in V$ and $\omega, \cdots, \omega_t \in V^*$

$$f(X_1, \cdots, X_s, \omega_1, \cdots, \omega_t) = (\omega_1 \otimes \cdots \otimes \omega_t) M_f (X_1 \otimes \cdots \otimes X_s). \tag{1.44}$$

f is called a tensor of covariant order s and contra-variant order t. The set of tensors on V of covariant order s and contra-variant t is denoted by $\mathcal{T}_t^s$.

In the following we assume contra-variant order $t = 0$.

Definition 1.8. A tensor $f \in \mathcal{T}^s$ is symmetric if for any $i \neq j$

$$\begin{aligned} f(X_1, \cdots, X_i, \cdots, X_j \cdots, X_s) = f(X_1, \cdots, X_j, \cdots, X_i \cdots, X_s),& \\ X_1, \cdots, X_s \in V.& \end{aligned} \tag{1.45}$$

A tensor $f \in \mathcal{T}^s$ is skew-symmetric if for any $i \neq j$

$$\begin{aligned} f(X_1, \cdots, X_i, \cdots, X_j \cdots, X_s) = -f(X_1, \cdots, X_j, \cdots, X_i \cdots, X_s),& \\ X_1, \cdots, X_s \in V.& \end{aligned} \tag{1.46}$$

From high school algebra we know that in $\mathbb{R}^3$ two products were defined: (i) inner product; (ii) cross product. Fix a basis $\{\vec{i}, \vec{j}, \vec{k}\}$ as $\vec{i} = (1,0,0)^T$, $\vec{j} = (0,1,0)^T$, and $\vec{k} = (0,0,1)^T$. Let $X = (x_1, x_2, x_3)^T$ and $Y = (y_1, y_2, y_3)^T$. The inner product is defined as

$$\langle X, Y \rangle := x_1 y_1 + x_2 y_2 + x_3 y_3. \tag{1.47}$$

The cross product, denoted by $\bowtie$, is defined as

$$X \bowtie Y = \det \begin{bmatrix} \vec{i} & \vec{j} & \vec{k} \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix}.$$

Example 1.6.

(1) The inner product on $\mathbb{R}^3$ is a tensor of covariant order $r = 2$. It is symmetric. (This result is also true for $\mathbb{R}^n$.)

(2) Let $X, Y, Z \in \mathbb{R}^3$.

$$v := \langle X, Y \bowtie Z \rangle.$$

We leave to the reader to check that v is a tensor of covariant order $r = 3$. It is skew-symmetric. In fact, v is the volume of the parallelepiped (as X, Y, Z satisfy the right hand rule, otherwise, it is the negative volume).

1.5 Nash Equilibrium

This section gives a brief introduction to some basic concepts of game theory, including strategic form and Nash equilibrium. Chapter 17 will continue the discussion. We also refer to Fudenberg and Tirole (1991) for more details.

Assume a game has n players, denoted by $P_1, \cdots, P_n$, and each player P_i can have $n_i > 0$ possible actions, called his strategies, denoted by $S_i = \{s^i_1, s^i_2, \cdots, s^i_{n_i}\}$, $i = 1, \cdots, n$. Let

$$f^i(s^1_{k_1}, s^2_{k_2}, \cdots, s^n_{k_n}) : \prod_{i=1}^{n} S_i \to \mathbb{R}, \quad i = 1, \cdots, n \tag{1.48}$$

be the payoff of the ith player, which means what the player i obtains from the game when P_j takes his strategy $s^j_{k_j}$, $j = 1, \cdots, n$. For compactness, denote by

$$\begin{aligned} \mu^i_{k_1,k_2,\cdots,k_n} := f^i(s^1_{k_1}, s^2_{k_2}, \cdots, s^n_{k_n}),\ i = 1, \cdots, n,\\ k_j = 1, \cdots, n_j,\ j = 1, \cdots, n. \end{aligned} \tag{1.49}$$

It is reasonable to assume that each player is pursuing his maximum payoff.

Definition 1.9. A set of strategies $\{s^1_*, s^2_*, \cdots, s^n_*\}$ is called a Nash equilibrium, if

$$f^i(s^1_*, s^2_*, \cdots, s^i_*, \cdots, s^n_*) \geq f^i(s^1_*, s^2_*, \cdots, s^i, \cdots, s^n_*), \quad \forall\, s^i,\ i = 1, \cdots, n. \tag{1.50}$$

Nash equilibrium is extremely important because once it is reached, each player intends to stick on this strategy forever.

Now let us see how to find the Nash equilibrium. The following procedure comes from its definition directly. For each i we can put the data

$$D_i = \{\mu^i_{k_1,k_2,\cdots,k_n} \,|\, k_j = 1, \cdots, n_j,\ j = 1, \cdots, n\}$$

into a matrix M_i, which is ordered by

$$\mathrm{id}(k_i; n_i) \times \mathrm{id}(k_1, \cdots, k_{i-1}, k_{i+1}, \cdots, k_n; n_1, \cdots, n_{i-1}, n_{i+1}, \cdots, n_n).$$

Then for each column of M_i we can find at least one n-index $(k^*_1, \cdots, k^*_n)$, which corresponds to the largest value of μ^i. Denote by K_i the set of n-indices found from each columns of M_i. Then

$$E_N := \bigcap_{i=1}^{n} K_i$$

is the set of Nash equilibriums.

We give some examples to depict this.

Example 1.7 (Prisoner's Dilemma). Two suspects are arrested by policemen. The policemen have insufficient evidence for a conviction, and, having separated the prisoners, visit each of them to offer the same deal. If one testifies for the prosecution against the other (defects) and the other remains silent (cooperates), the defector goes free and the silent accomplice receives the full 10-year sentence. If both remain silent, both prisoners are sentenced to only 1 year in jail for a minor charge. If each betrays the other, each receives a 5-year sentence. Each prisoner must choose to betray the other or to remain silent. Each one is assured that the other would not know about the betrayal before the end of the investigation. How should the prisoners act?

The payoff bi-matrix is given in Table 1.1.

Table 1.1 Payoff of Prisoner's Dilemma

$P_1\backslash P_2$	C	D
C	$-1 \quad -1$	$-10 \quad 0$
D	$0 \quad -10$	$-5 \quad -5$

Now we have M_1 and M_2 as

$$M_1 = \begin{bmatrix} -1 & -10 \\ \underline{0} & \underline{-5} \end{bmatrix}, \quad M_2 = \begin{bmatrix} -1 & -10 \\ \underline{0} & \underline{-5} \end{bmatrix},$$

where the underline elements are the column maximal elements.

Note that in M_1 the row index is k_1 and the column index is k_2 while in M_2 the row index is k_2 and the column index is k_1, hence we have the K_i set as

$$K_1 = \{(2,1),(2,2)\}; \quad K_2 = \{(1,2),(2,2)\}.$$

It follows that the set of Nash equilibrium(s) is

$$E_N = K_1 \cap K_2 = \{(2,2)\},$$

which means (D, D) is the only Nash equilibrium.

Example 1.8. Assume a game has three players. Their strategies are: $S_1 = \{s_1^1, s_2^1, s_3^1\}$, $S_2 = \{s_1^2, s_2^2\}$, $S_3 = \{s_1^3, s_2^3\}$. And the payoffs are shown in Table 1.2.

Then we have $M_i, i = 1, 2, 3$ as

$$M_1 = \begin{bmatrix} \underline{3} & -1 & \underline{2} & \underline{2} \\ 2 & \underline{1} & 1 & 0 \\ 1 & -2 & 0 & \underline{2} \end{bmatrix}, M_2 = \begin{bmatrix} \underline{1} & \underline{2} & \underline{2} & -1 & \underline{3} & 1 \\ -1 & 0 & 1 & \underline{1} & 2 & \underline{4} \end{bmatrix}, M_3 = \begin{bmatrix} \underline{2} & \underline{3} & 0 & \underline{2} & -1 & -1 \\ 1 & -1 & \underline{2} & \underline{2} & \underline{1} & \underline{3} \end{bmatrix}.$$

Table 1.2 Payoffs

P_2	s^2_1						s^2_2					
$P_1\backslash P_3$	s^3_1			s^3_2			s^3_1			s^3_2		
s^1_1	3	1	2	−1	2	1	2	−1	3	2	0	−1
s^1_2	2	2	0	1	−1	2	1	1	2	0	1	2
s^1_3	1	3	−1	−2	1	−1	0	2	−1	2	4	3

Note that in M_1 the row index is k_1 and the column index is $\mathrm{id}(k_2, k_3; 2, 2)$, it follows that

$$K_1 = \{(1,1,1), (2,1,2), (1,2,1), (1,2,2), (3,2,2)\}.$$

In M_2 the row index is k_2 and the column index is $\mathrm{id}(k_1, k_3; 3, 2)$, it follows that

$$K_2 = \{(1,1,1), (1,1,2), (2,1,1), (2,2,2), (3,1,1), (3,2,2)\}.$$

In M_3 the row index is k_3 and the column index is $\mathrm{id}(k_1, k_2; 3, 2)$, hence

$$K_3 = \{(1,1,1), (1,2,1), (2,1,2), (2,2,1), (2,2,2), (3,1,2), (3,2,2)\}.$$

Therefore,

$$E_N = K_1 \cap K_2 \cap K_3 = \{(1,1,1), (3,2,2)\}.$$

That is, (s^1_1, s^2_1, s^3_1) and (s^1_3, s^2_2, s^3_2) are two Nash equilibriums.

1.6 Symmetric Group

Let $S = \{1, 2, \cdots, k\}$. A permutation σ on S is a one-to-one mapping from S onto S. All the permutations on S with the product as the combination of two permutations form a group, called the symmetric group on k letters, or kth order symmetric group, denoted by $\mathbf{S}_k$ (Hungerford, 1974).

We use some numerical examples to depict it. For instance, let $k = 5$. A $\sigma \in \mathbf{S}_5$ may be expressed as

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 5 & 1 & 4 & 3 \end{pmatrix}.$$

Let another permutation $\tau \in \mathbf{S}_5$ be expressed as

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 5 & 3 & 4 & 1 & 2 \end{pmatrix}.$$

The product on $\mathbf{S}_5$ is defined as

$$\tau\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 5 & 1 & 4 & 3 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 3 & 2 & 5 & 1 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 3 & 2 & 5 & 1 & 4 \end{pmatrix}.$$

An alternative expression of an element in $\mathbf{S}_k$ is expressing it as a product of cycles. For instance, we can express $\sigma = (1\ 2\ 5\ 3)$, $\tau = (1\ 5\ 2\ 3\ 4)$, and $\tau\sigma = (1\ 3\ 5\ 4)$. Consider another example, let

$$\mu = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 1 & 4 & 5 & 6 & 3 \end{pmatrix} \in \mathbf{S}_6.$$

Then it can be expressed as $\mu = (1\ 2)(3\ 4\ 5\ 6)$.

It is easy to check that the cardinal number $|\mathbf{S}_k| = k!$.

A cycle of two elements, such as $(a, b) \in \mathbf{S}_k$, is called a transposition.

Proposition 1.12. *Every permutation can be expressed as a product of transpositions.*

Proof. We have only to prove that each cycle can be expressed as a product of transpositions. Assume the length of a cycle is 1: we have $(r_1) = (r_1\ r_2)(r_2\ r_1)$. Assume the length of a cycle is greater than 1, then we have $(r_1\ r_2\ \cdots\ r_k) = (r_1\ r_k)(r_1\ r_{k-1})\cdots(r_1\ r_2)$. □

Note that a permutation σ can have different products of transpositions, but the number of transpositions can either be even or odd, but not both (Hungerford, 1974). When the number of the transpositions is even we say $\operatorname{sgn}(\sigma) = 1$, otherwise, $\operatorname{sgn}(\sigma) = -1$.

For a $\sigma \in \mathbf{S}_k$ define a matrix M_σ as

$$M_\sigma = \delta_k[\sigma(1)\ \sigma(2)\ \cdots\ \sigma(k)].$$

Then M_σ can realize the permutation as

$$(\sigma(1)\ \sigma(2)\ \cdots\ \sigma(k)) = (1\ 2\ \cdots\ k)M_\sigma.$$

Moreover,

$$\operatorname{sgn}(\sigma) = \det(M_\sigma). \tag{1.51}$$

Note that sometimes to label a set of data the index order and the index arrange order may not coincide. Say the data are labeled by multi-index $(i_1, \cdots, i_k; n_1, \cdots, n_k)$ and the multi-index may be ordered by id $\left(i_{\sigma(1)}, \cdots, i_{\sigma(k)}; n_{\sigma(1)}, \cdots, n_{\sigma(k)}\right)$. See the following example.

Example 1.9. Consider a set of data

$$D = \{d_{i_1,i_2,i_3} \,|\, i_1 = 1, 2; i_2 = 1, 2, 3; i_3 = 1, 2, 3\}.$$

Let $\sigma = (1, 3, 2) \in \mathbf{S}_3$. Assume D is required to be arranged in the order of $\mathrm{id}(i_{\sigma(1)}, i_{\sigma(2)}, i_{\sigma(3)}; n_{\sigma(1)}, n_{\sigma(2)}, n_{\sigma(3)})$. Since $\sigma(1) = 3$, $\sigma(2) = 1$, and $\sigma(3) = 2$, the data are arranged in the order of $\mathrm{id}(i_3, i_1, i_2; 3, 2, 3)$. Hence we have

$$\begin{array}{l} d_{111}\ d_{121}\ d_{131}\ d_{211}\ d_{221}\ d_{231}\ d_{112}\ d_{122}\ d_{132} \\ d_{212}\ d_{222}\ d_{232}\ d_{113}\ d_{123}\ d_{133}\ d_{213}\ d_{223}\ d_{233}. \end{array}$$

1.7 Swap Matrix

In this section we define a special matrix, called the swap matrix. It is very useful in order to overcome the non-commutativity of the matrix product. Swap matrix was firstly introduced in Horn and Johnson (1991), where it is called commutation matrix.

Definition 1.10. A swap matrix $W_{[m,n]} \in \mathcal{M}_{mn \times mn}$ is constructed in the following way:

Step 1. Label its columns by index (i, j) in the order of $\mathrm{id}(i, j; m, n)$ and its rows by index (I, J) in the order of $\mathrm{id}(J, I; n, m)$.

Step 2. The entry at row (I, J) and column (i, j), denoted by $w_{(I.J),(i,j)}$, is assigned as

$$w_{(I.J),(i,j)} = \begin{cases} 1, & I = i, \text{ and } J = j \\ 0, & \text{otherwise.} \end{cases} \tag{1.52}$$

We give some examples to depict the swap matrices.

Example 1.10.

(1) Consider $W_{[2,3]}$. Labeling its columns by (i, j) in the order of $\mathrm{id}(i, j; 2, 3)$ and its rows by (I, J) in the order of $\mathrm{id}(J, I; 3, 2)$, then

the swap matrix can be constructed as

$$W_{[2,3]} = \begin{array}{c} \begin{array}{cccccc} (11) & (12) & (13) & (21) & (22) & (23) \end{array} \\ \left[\begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \end{array} \begin{array}{c} \\ (11) \\ (21) \\ (12) \\ (22) \\ (13) \\ (23) \end{array} .$$

(2) Consider $W_{[3,2]}$. Labeling its columns by (i,j) in the order of $\mathrm{id}(i,j;3,2)$ and its rows by (I,J) in the order of $\mathrm{id}(J,I;2,3)$, then the swap matrix can be constructed as

$$W_{[3,2]} = \begin{array}{c} \begin{array}{cccccc} (11) & (12) & (21) & (22) & (31) & (32) \end{array} \\ \left[\begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \end{array} \begin{array}{c} \\ (11) \\ (21) \\ (31) \\ (12) \\ (22) \\ (32) \end{array} .$$

According to the construction of swap matrix, the following two propositions are immediate consequences.

Proposition 1.13. *Let $A \in \mathcal{M}_{m\times n}$. Then*

$$W_{[m,n]}V_r(A) = V_c(A); \quad W_{[n,m]}V_c(A) = V_r(A). \tag{1.53}$$

Proposition 1.14.

(1) Let $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^n$ be two column vectors. Then

$$W_{[m,n]}(X \otimes Y) = Y \otimes X. \tag{1.54}$$

(2) Let $\omega \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^n$ be two row vectors. Then

$$(\omega \otimes \mu)W_{[n,m]} = \mu \otimes \omega. \tag{1.55}$$

This proposition has an equivalent statement.

Corollary 1.1. *Let $D = \{x_{ij} \mid i = 1, \cdots, m, j = 1, \cdots, n\}$ be a set of data. X is a column vector of the elements of D, labeled by (i,j) and arranged in the order of $\mathrm{id}(i,j;m,n)$, and Y is a column vector of the elements of D, labeled by (i,j) and arranged in the order of $\mathrm{id}(j,i;n,m)$. Then*

$$W_{[m,n]}X = Y; \quad W_{[n,m]}Y = X. \tag{1.56}$$

Swap matrix has some special properties, which follow from its definition immediately.

Proposition 1.15.

(1) A swap matrix is an orthogonal matrix. It satisfies

$$W_{[m,n]}^T = W_{[m,n]}^{-1} = W_{[n,m]}. \tag{1.57}$$

(2) When $m = n$, (1.57) becomes

$$W_{[n,n]} = W_{[n,n]}^T = W_{[n,n]}^{-1}. \tag{1.58}$$

(3)

$$W_{[1,n]} = W_{[n,1]} = I_n. \tag{1.59}$$

$m = n$ is particularly useful in the sequel. To simplify the notation, we denote

$$W_{[n]} := W_{[n,n]}.$$

Exercises

1.1 Prove the formulas (1.5) and (1.6).

1.2 Express data D in Example 1.2 into a matrix in the order of

(i) $\mathrm{id}(i;2) \times \mathrm{id}(j,k;3,4)$;

(ii) $\mathrm{id}(i,k;2,4) \times \mathrm{id}(j;3)$.

1.3 Given $D = \{d_1, d_2, \cdots, d_{120}\}$.

(i) What is the multi-label of $p = 72$ under the order of $\mathrm{id}(i,j,k;6,4,5)$;

(ii) The multi-label of d_p in the order of $\mathrm{id}(i,j,k;4,5,6)$ is $(2,4,3)$. Find p.

1.4 A k-dimensional data D is as in (1.1). Assume there is a partition of index set as

$$\left\{i_{\alpha_1}, \cdots, i_{\alpha_p}\right\} \cup \left\{i_{\beta_1}, \cdots, i_{\beta_q}\right\} = \{i_1, \cdots, i_k\}.$$

D is arranged into a matrix M_D in the order of

$$\mathrm{id}\left(i_{\alpha_1}, \cdots, i_{\alpha_p}; n_{\alpha_1}, \cdots, n_{\alpha_p}\right) \times \mathrm{id}\left(i_{\beta_1}, \cdots, i_{\beta_q}; n_{\beta_1}, \cdots, n_{\beta_q}\right).$$

Find the positions of the rth element of $V_r(M_D)$ and the sth element of $V_c(M_D)$ in M_D.

1.5 Prove the following products satisfy the associative law and the distributive law. (i) Kronecker product; (ii) Hadamard product; (iii) Khatri-Rao product.

1.6 Let $Z \in \mathcal{L}_{mn\times k}$. Show that there exist unique $X \in \mathcal{L}_{m\times k}$ and $Y \in \mathcal{L}_{n\times k}$ such that Z is the Khatri-Rao product of X and Y. That is,

$$Z = X * Y.$$

1.7 Let ξ be an eigenvector of A with respect to the eigenvalue $\lambda \in \sigma(A)$ and η be an eigenvector of B with respect to the eigenvalue $\mu \in \sigma(B)$. Prove that $\xi \otimes \eta$ is an eigenvector of AB with respect to the eigenvalue $\lambda\mu \in \sigma(AB)$.

1.8 Check that the v defined in Example 1.6 is a skew-symmetric tensor of covariant order $r = 3$.

1.9 Let $V = \mathbb{R}^4$ with canonical basis. For $X_1, X_2 \in V$ and $\sigma_1, \sigma_2 \in V^*$ we define $\pi : V \times V \times V^* \times V^* \to \mathbb{R}$ by

$$\pi(\sigma_1, \sigma_2, X_1, X_2) = \sigma_1(X_1) \times \sigma_2(X_2).$$

(Where $\sigma \in V^*$ is expressed as a row and $X \in V$ is expressed as a column, and $\sigma(X) = \sigma X$.)

(i) Show that $\pi \in \mathcal{T}_2^2(V)$.

(ii) Calculate the structure matrix M_π.

1.10 Let $V = \mathbb{R}^3$ with canonical basis. For $X_1, X_2, X_3 \in V$ and $\sigma_1, \sigma_2 \in V^*$ we define $\pi : V \times V \times V \times V^* \times V^* \to \mathbb{R}$ by

$$\pi(\sigma_1, \sigma_2, X_1, X_2, X_3) = \sigma_1(X_1) \times \sigma_2(X_2 \bowtie X_3).$$

(i) Show that $\pi \in \mathcal{T}_2^3(V)$.

(ii) Calculate the structure matrix M_π.

1.11 Consider the game illustrated in Table 1.3.

Table 1.3 Payoffs

P_3	A						B					
$P_1\backslash P_2$	L			R			L			R		
U	0	0	10	−5	−5	0	−2	−2	0	−5	−5	0
D	−5	−5	0	1	1	−5	−5	−5	0	−1	−1	5

(i) Find Nash equilibrium of this game.

(ii) Assume Player 1 and Player 2 form a coalition, the coalition's payoff is the sum of their payoffs. Construct the payoff bi-matrix for $P_1 - P_2$ as one side and P_3 as the other side.

(iii) Does the Nash equilibrium found in (i) remains an equilibrium for (ii)?

1.12 (i) Give an example, where the game has no Nash equilibrium;

(ii) Give an example, where the game has more than one equilibriums.

1.13 Assume $\sigma, \tau, \pi \in \mathbf{S}_6$ are given as $\sigma = (1, 2, 3)(4, 5)$, $\tau = (2, 3)(4, 5, 6)$, and $\pi = (1, 6)(3, 5)$.

(i) Calculate $\sigma\tau$, $\tau\pi$, and $\pi\sigma$;

(ii) Calculate $(\sigma\tau)\pi$ and $\sigma(\tau\pi)$ to verify the associativity.

1.14 (i) Prove $\mathbf{S}_k$ is a group;

(ii) Let $\sigma = (1,3), \tau = (4,5,3)$. Find the subgroup $H < \mathbf{S}_5$, generated by $\{\sigma, \tau\}$.

1.15 A set of data D is arranged by $\mathrm{id}(i_1, \cdots, i_k; n_1, \cdots, n_k)$, and under this order a data $d_p \in D$ is the pth element. $\sigma \in \mathbf{S}_k$ is a known permutation. Find the multi-index of d_p in the order of $\mathrm{id}\left(i_{\sigma(1)}, \cdots, i_{\sigma(k)}; n_{\sigma(1)}, \cdots, n_{\sigma(k)}\right)$.

1.16 Let V be an n-dimensional vector space with a basis $\{d_1, \cdots, d_n\}$, which has dual basis $\{e_1, \cdots, e_n\}$. Consider a tensor $\omega \in T^r_s(V)$. The structure matrix of ω under these bases is M_ω. Let $\{\tilde{d}_1, \cdots, \tilde{d}_n\}$ be another basis of V, with dual basis $\{\tilde{e}_1, \cdots, \tilde{e}_n\}$. Moreover,

$$\begin{bmatrix} \tilde{d}_1 \\ \vdots \\ \tilde{d}_n \end{bmatrix} = A \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix}.$$

Find the structure matrix of ω under new basis.

1.17 (i) Calculate $W_{[2,4]}$ and $W_{[4,2]}$;

(ii) Verify that $W^T_{[2,4]} = W^{-1}_{[2,4]} = W_{[4,2]}$.

1.18 Prove the following two alternative expressions of the swap matrix:

$$W_{[m,n]} = \begin{bmatrix} I_m \otimes (\delta^1_n)^T \\ I_m \otimes (\delta^2_n)^T \\ \vdots \\ I_m \otimes (\delta^n_n)^T \end{bmatrix}, \tag{1.60}$$

$$W_{[m,n]} = \left[I_n \otimes \delta^1_m \ I_n \otimes \delta^2_m \ \cdots \ I_n \otimes \delta^m_m \right]. \tag{1.61}$$

1.19 Let $X \in \mathbb{R}^m$, $Y \in \mathbb{R}^n$, $Z \in \mathbb{R}^p$ be columns. Prove the following equations:

$$Y \otimes X \otimes Z = \left(W_{[m,n]} \otimes I_p\right) X \otimes Y \otimes Z, \tag{1.62}$$

$$X \otimes Z \otimes Y = \left(I_m \otimes W_{[n,p]}\right) X \otimes Y \otimes Z, \tag{1.63}$$

$$Z \otimes Y \otimes X = \left(W_{[n,p]} \otimes I_m\right) \left(I_n \otimes W_{[m,p]}\right) \left(W_{[m,n]} \otimes I_p\right) X \otimes Y \otimes Z. \tag{1.64}$$

1.20 Let $A \in \mathcal{M}_{m\times n}$. Prove the following equations:

$$\begin{aligned} V_c(A^T) &= W_{[n,m]} V_c(A), \\ V_r(A^T) &= W_{[m,mn]} V_r(A). \end{aligned} \tag{1.65}$$

Chapter 2

Semi-Tensor Product of Matrices

Starting from bilinear functions we show an alternative way of calculating bilinear mapping instead of using matrix form, and then extend this new product method to multilinear case. This new method leads to the definition of general left semi-tensor product (STP) of matrices, which is a generalization of conventional matrix product. Then certain basic properties of STP are revealed. Roughly speaking, all the major properties of the conventional matrix product remain true for this generalized product. In the light of swap matrix, certain pseudo-commutative properties of STP are obtained, which show one of the advantages of STP over the conventional matrix product. Finally, the bilinearity property of the STP of two vectors is investigated.

2.1 Multilinear Function

Definition 2.1. Let V_i, $i = 0, 1, \cdots, k$ be real vector spaces. A mapping $f : V_1 \times V_2 \times \cdots \times V_k \to V_0$ is called a k-linear mapping, if

$$\begin{aligned} & f(X_1, \cdots, \alpha X_i + \beta Y_i, \cdots, X_k) \\ & = af(X_1, \cdots, X_i, \cdots, X_k) + bf(X_1, \cdots, Y_i, \cdots, X_k), \\ & \qquad X_j \in V_j,\ j = 1, \cdots, k;\ Y_i \in V_i,\ \alpha, \beta \in \mathbb{R}. \end{aligned} \tag{2.1}$$

When $V_0 = \mathbb{R}$, it is called a k-linear function.

Let $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ be a bilinear function. Assume

$$f(\delta_m^i, \delta_n^j) = \mu_{i,j}, \quad i = 1, \cdots, m;\ j = 1, \cdots, n.$$

Then we can arrange the data $D = \{\mu_{i,j} \mid i = 1, \cdots, m;\ j = 1, \cdots, n\}$ into a matrix as

$$M_f = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2n} \\ \vdots & & & \\ \mu_{m1} & \mu_{m2} & \cdots & \mu_{mn} \end{bmatrix},$$

which is called the structure matrix of f. Now for any $X = (x_1, \cdots, x_m)^T$ and $Y = (y_1, \cdots, y_n)^T$, (precisely, $X = \sum_{i=1}^m x_i \delta_m^i$, etc.) we have

$$f(X,Y) = \sum_{i=1}^m \sum_{j=1}^n \mu_{i,j} x_i y_j. \tag{2.2}$$

Using the structure matrix, we have a matrix expression of (2.2) as

$$f(X,Y) = X^T M_f Y. \tag{2.3}$$

It was mentioned in Chapter 1 that to deal with the multilinear function, say 3-linear function, the cubic matrix has been proposed by Bates and Watts (1980); Tsai (1983). Certain product rules between cubic matrix and conventional matrix etc. have also been developed. It has some successful applications (Wang, 2002). But they are rather complicated. Moreover, this approach can hardly be extended to higher-dimensional data. An attempt of using matrix to deal with higher-dimensional data is so-called multi-edge matrix, proposed by Zhang (1993). Because of the complexity, it can also hardly be used.

We intended to develop a universal and easy way to deal with the multilinear mappings. Recall the bilinear case, alternatively, we may arrange D into a row vector in the order of $\mathrm{id}(i, j; m, n)$ as

$$V_f = (\mu_{11}\ \mu_{12}\ \cdots\ \mu_{1n}\ \cdots \mu_{m1}\ \cdots\ \mu_{mn}) = V_r(M_f).$$

Note that if we split V_f into m equal-size blocks as

$$V_f = [V_1\ V_2\ \cdots\ V_m],$$

then the data in V_1 has the first index $i = 1$, in V_2 has the first index $i = 2$ and so on. Then it is clear that

$$f(X,Y) = \left(\sum_{i=1}^m x_i V_i \right) Y. \tag{2.4}$$

Here we may consider $\sum_{i=1}^m x_i V_i$ as the "product" of V_f with X.

The advantage of this new "product" lies on that it can be easily extended to higher-order case. For instance, consider $f : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^t \to \mathbb{R}$. Let

$$\mu_{i,j,k} := f(\delta_m^i, \delta_n^j, \delta_t^k), \quad i = 1, \cdots, m,\ j = 1, \cdots, n,\ k = 1, \cdots, t.$$

Then we arrange $\{\mu_{i,j,k}\}$ into a row vector V_f in the order of $\mathrm{id}(i, j, k; m, n, t)$. Let $X \in \mathbb{R}^m$, $Y \in \mathbb{R}^n$, and $Z \in \mathbb{R}^t$. We still split V_f into m equal-size blocks as

$$V_f = [V_1\ V_2\ \cdots\ V_m],$$

then

$$V_{i_0} = [\mu_{i_{011}}\ \mu_{i_{012}}\ \cdots\ \mu_{i_{01t}}\ \mu_{i_{0n1}}\ \cdots\ \mu_{i_{0nt}}],$$

which is the data of the first index $i = i_0$. Using the new product, we have

$$V_f X = \sum_{i=1}^{m} V_i x_i.$$

You now can see that x_i has been multiplied to the proper segment of data. Continuing this procedure, we split $V_f X$ into n blocks as

$$V_f X = [U_1, U_2, \cdots, U_n].$$

Then you can deal with the second factor, using this new product, as

$$V_f XY = \sum_{j=1}^{n} U_j y_j.$$

Now y_j has been multiplied to the proper segment of data too. Finally, we split $V_f XY$ as

$$V_f XY = [W_1, W_2, \cdots, W_t].$$

Using the new product again, which is now the same as the conventional inner product, we have

$$V_f XYZ = \sum_{k=1}^{t} W_k z_k.$$

Again, z_k is multiplied to the proper segment of data, and finally we have

$$V_f XYZ = f(X, Y, Z). \tag{2.5}$$

Based on this observation, we give the following rigorous definition for this new "product":

Definition 2.2.

(1) Let $X \in \mathbb{R}^{mn}$ be a row and $Y \in \mathbb{R}^m$ be a column. Then we split X into m equal-size blocks as $(X^1\ X^2\ \cdots\ X^m)$, such that $X^i \in \mathbb{R}^n$, $i = 1, \cdots, m$, and define the left STP of X and Y, denoted by $X \ltimes Y$, as

$$X \ltimes Y := \sum_{i=1}^{m} X^i y_i \in \mathbb{R}^n. \tag{2.6}$$

(2) Let $X \in \mathbb{R}^m$ be a row and $Y \in \mathbb{R}^{mn}$ be a column. Then we define the left STP of X and Y as

$$X \ltimes Y := \left(Y^T \ltimes X^T\right)^T \in \mathbb{R}^n. \tag{2.7}$$

Using matrix product to express a bilinear function as in (2.3) is very convenient. Unfortunately, it can hardly be used for multilinear functions. The advantage of (2.4) is that each set of data is arranged as a vector, and the product can be realized by a sequence of products between two vectors. Then it can easily be extended to 3-linear case as in (2.5), as well as to general multilinear functions.

Consider a function $f : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_k} \to \mathbb{R}$. Denote

$$\mu_{i_1,i_2,\cdots,i_k} = f(\delta_{n_1}^{i_1}, \delta_{n_2}^{i_2}, \cdots, \delta_{n_k}^{i_k}), \quad i_j = 1, \cdots, n_j,\ j = 1, \cdots, k.$$

Denote by $X_j = (x_1^j, \cdots, x_{n_j}^j)^T \in \mathbb{R}^{n_j}$, $j = 1, 2, \cdots, k$. Then

$$f(X_1, \cdots, X_k) = \sum_{i_1=1}^{n_1} \cdots \sum_{i_k=1}^{n_k} \mu_{i_1,i_2,\cdots,i_k} x_{i_1}^1 x_{i_2}^2 \cdots x_{i_k}^k.$$

Arrange

$$D = \left\{ \mu_{i_1,i_2,\cdots,i_k} = f(\delta_{n_1}^{i_1}, \cdots, \delta_{n_k}^{i_k}) \,\middle|\, i_j = 1, \cdots, n_i,\ i = 1, \cdots, k \right\}$$

into a row vector V_f in the order of $\mathrm{id}(i_1, i_2, \cdots, i_k; n_1, n_2, \cdots, n_k)$. Then it is easy to see that

$$f(X_1, \cdots, X_k) = ((\cdots(V_f \ltimes X_1) \ltimes X_2) \ltimes \cdots \ltimes X_k) \cdots). \tag{2.8}$$

Remark 2.1.

(1) From (2.8) one sees that the left STP can search for each factor vector its corresponding index automatically. Hence it is very convenient in dealing with multi-dimensional data.

(2) It is clear from the discussion of multilinear functions that to perform this product the dimension of one factor vector should be a multiple of that of the other factor vector. This "multiplier dimension" requirement for two factor matrices is particularly important. Through this book we mainly focus on this particular case, rather than considering the product of two arbitrary matrices.

Example 2.1.

(1) Let $X = [1\ 3\ 2\ 4]$ and $Y = [2\ \ -1]^T$. Then

$$X \ltimes Y = [1\ 3] \times 2 + [2\ 4] \times (-1) = [0\ 2].$$

(2) Let $X = [1\ 2\ \ -1]$ and $Y = [2\ 1\ \ -1\ 0\ \ -2\ 1]^T$. Then

$$X \ltimes Y = 1 \times \begin{bmatrix} 2 \\ 1 \end{bmatrix} + 2 \times \begin{bmatrix} -1 \\ 0 \end{bmatrix} + (-1) \times \begin{bmatrix} -2 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

(3) Recall Example 1.6. The volume tensor is defined as

$$v := \langle X, Y \times Z \rangle, \quad X, Y, Z \in \mathbb{R}^3.$$

Then

$$\mu_{111} = \left\langle \delta_3^1, \delta_3^1 \times \delta_3^1 \right\rangle = 0;\ \mu_{112} = \left\langle \delta_3^1, \delta_3^1 \times \delta_3^2 \right\rangle = 0;\ \cdots .$$

Finally, we have the vector form of $\{\mu_{ijk} \mid i, j, k = 1, 2, 3\}$ as

$$V_v = [0\ 0\ 0\ 0\ 0\ 1\ 0\ \ -1\ 0\ 0\ 0\ \ -1\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ \ -1\ 0\ 0\ 0\ 0\ 0].$$

Now assume $X = [1, 1, -1]^T$, $Y = [2, 1, -2]^T$, $Z = [-1, 0, -2]^T$. Then we have

$$\begin{aligned} v(X, Y, Y) &= (((V_v \ltimes X) \ltimes Y) \ltimes Z) \\ &= (([0\ \ -1\ \ -1\ 1\ 0\ 1\ 1\ \ -1\ 0] \ltimes Y) \ltimes Z) \\ &= [-1\ 0\ \ -1] \ltimes Z \\ &= 3. \end{aligned}$$

2.2 Left Semi-Tensor Product of Matrices

Let $A \in \mathcal{M}_{m\times n}$ and $B \in \mathcal{M}_{p\times q}$.

(i) If $n = p$, A and B are said to be of "equal dimension".

(ii) If $n = tp$ or $nt = p$ (where $t \in \mathbb{Z}_+$, then A and B are said to be of "multiplier dimension". If $n = tp$, we denote it by $A \succ_t B$, and if $nt = p$ we denote it by $A \prec_t B$.

(iii) Otherwise, we say A and B are of arbitrary dimension.

We use $\mathrm{Row}(A)$ ($\mathrm{Col}(A)$) for the set of rows (columns) of A, and $\mathrm{Row}_i(A)$ ($\mathrm{Col}_i(A)$) the ith row (column) of A.

Definition 2.3. Let $A \in \mathcal{M}_{m\times n}$ and $B \in \mathcal{M}_{p\times q}$, and A and B are of multiplier dimension. Then the left STP of A and B is defined as

$$A \ltimes B = \begin{bmatrix} \mathrm{Row}_1(A) \ltimes \mathrm{Col}_1(B) & \mathrm{Row}_1(A) \ltimes \mathrm{Col}_2(B) & \cdots & \mathrm{Row}_1(A) \ltimes \mathrm{Col}_q(B) \\ \mathrm{Row}_2(A) \ltimes \mathrm{Col}_1(B) & \mathrm{Row}_2(A) \ltimes \mathrm{Col}_2(B) & \cdots & \mathrm{Row}_2(A) \ltimes \mathrm{Col}_q(B) \\ \vdots & & & \\ \mathrm{Row}_m(A) \ltimes \mathrm{Col}_1(B) & \mathrm{Row}_m(A) \ltimes \mathrm{Col}_2(B) & \cdots & \mathrm{Row}_m(A) \ltimes \mathrm{Col}_q(B) \end{bmatrix}. \tag{2.9}$$

Example 2.2. Let

$$X = \begin{bmatrix} 1 & 2 & -1 & 2 \\ 0 & 1 & 2 & 3 \\ 3 & 3 & 1 & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} 1 & 2 \\ -1 & 3 \end{bmatrix}.$$

Then

$$\begin{aligned} X \ltimes Y &= \begin{bmatrix} (1\ 2) - (-1\ 2) & 2(1\ 2) + 3(-1\ 2) \\ (0\ 1) - (2\ 3) & 2(0\ 1) + 3(2\ 3) \\ (3\ 3) - (1\ 1) & 2(3\ 3) + 3(1\ 1) \end{bmatrix} \\ &= \begin{bmatrix} 2 & 0 & -1 & 10 \\ -2 & -2 & 6 & 11 \\ 2 & 2 & 9 & 9 \end{bmatrix}. \end{aligned}$$

Remark 2.2. Let $A \in \mathcal{M}_{m\times n}$ and $B \in \mathcal{M}_{p\times q}$. It follows from the definition that when $n = p$ we have

$$A \ltimes B = AB.$$

That is, when the conventional matrix product of A and B is defined the left STP of A and B coincides with their conventional product. In fact, the product rule defined in (2.9) is exactly the same as conventional matrix product, only each "inner product" now could be between two vectors of different sizes. This fact shows that the left STP is a generalization of the conventional matrix product. Because of this fact, the symbol "$\ltimes$" may be omitted. We can always consider AB as $A \ltimes B$, when A and B meet the equal-dimension requirement, the product becomes conventional matrix product automatically. In the sequel, unless we want to emphasize the product is left STP, the symbol "$\ltimes$" is mostly omitted.

In Chapter 1, in addition to conventional matrix product, some other matrix products have been introduced, which are Kronecker product, Hadamard product, and Khatri-Rao product. One sees from there that all the different matrix products satisfy two fundamental properties: associative law and distributive law. These two properties may be considered as fundamental requirements for a matrix product. Without them, the fundamental matrix algebraic structure will be destroyed. So we have to show that the left STP also satisfies these two laws. First, we give a lemma.

Lemma 2.1. *Given $A \in \mathcal{M}_{m\times n}$ and $B \in \mathcal{M}_{p\times q}$, where $n|p$ or $p|n$. Then*

$$\begin{aligned} A \ltimes B &= \begin{bmatrix} \mathrm{Row}_1(A) \ltimes B \\ \vdots \\ \mathrm{Row}_m(A) \ltimes B \end{bmatrix} \\ &= \left[A \ltimes \mathrm{Col}_1(B) \ \cdots \ A \ltimes \mathrm{Col}_q(B) \right]. \end{aligned} \tag{2.10}$$

Proof. According to the definition, a straightforward computation, starting from the first (second) form of (2.10) by expanding B column by column (expanding A row by row), then both two forms degenerate to (2.9) immediately. □

Theorem 2.1. *Assume the dimensions of the matrices involved in the following equations (2.11) and (2.12) meet the dimension requirement such that the $\ltimes$ is well defined. Then we have*

(1) (Distributive Law)

$$\begin{cases} F \ltimes (aG \pm bH) = aF \ltimes G \pm bF \ltimes H, \\ (aF \pm bG) \ltimes H = aF \ltimes H \pm bG \ltimes H, \quad a, b \in \mathbb{R}. \end{cases} \tag{2.11}$$

(2) (Associative Law)

$$(F \ltimes G) \ltimes H = F \ltimes (G \ltimes H). \tag{2.12}$$

Proof. Equation (2.11) can be proved by a straightforward computation. We leave to the reader to check it. In the following we prove (2.12).

First of all, we have to show that if F, G and H have feasible dimensions for $(F \ltimes G) \ltimes H$ the dimensions are also feasible for $F \ltimes (G \ltimes H)$.

Case 1: $F \succ G$ and $G \succ H$. So the dimensions of F, G and H can be assumed as $m \times np$, $p \times qr$ and $r \times s$ respectively.

Now the dimension of $F \ltimes G$ is $m \times nqr$. It is good for $(F \ltimes G) \ltimes H$. On the other hand, the dimension of $G \ltimes H$ is $p \times qs$. It is good for $F \ltimes (G \ltimes H)$.

Case 2: $F \prec G$ and $G \prec H$. So the dimensions of F, G and H can be assumed as $m \times n$, $np \times q$ and $rq \times s$ respectively.

Now the dimension of $F \ltimes G$ is $mp \times q$. It is good for $(F \ltimes G) \ltimes H$. On the other hand, the dimension of $G \ltimes H$ is $npr \times s$. It is good for $F \ltimes (G \ltimes H)$.

Case 3: $F \prec G$ and $G \succ H$. So the dimensions of F, G and H can be assumed as $m \times n$, $np \times qr$ and $r \times s$ respectively.

Now the dimension of $F \ltimes G$ is $mp \times qr$. It is good for $(F \ltimes G) \ltimes H$. On the other hand, the dimension of $G \ltimes H$ is $np \times qs$. It is good for $F \ltimes (G \ltimes H)$.

Case 4: $F \succ G$ and $G \prec H$. So the dimensions of F, G and H can be assumed as $m \times np$, $p \times q$ and $rq \times s$ respectively.

Now the dimension of $F \ltimes G$ is $m \times nq$. To make it feasible for $(F \ltimes G) \ltimes H$, we need

Case 4.1: $(F \ltimes G) \succ H$, i.e., $n = n'r$. It is good for $F \ltimes (G \ltimes H)$;

Case 4.2: $(F \ltimes G) \prec H$, i.e., $r = nr'$. It is good for $F \ltimes (G \ltimes H)$;

The dimension of $G \ltimes H$ is $pr \times s$. To make it feasible for $(F \ltimes G) \ltimes H$, we need

Case 4.3: $F \succ (G \ltimes H)$, i.e., $n = n'r$. It is good for $(F \ltimes G) \ltimes H$;

Case 4.4: $F \prec (G \ltimes H)$, i.e., $r = nr'$. It is good for $(F \ltimes G) \ltimes H$;

Next, we prove the associativity. We have to prove it case by case. But Cases 1–3 are similar, we prove only Case 1, i.e. $F \succ G$ and $G \succ H$.

Let $F_{m\times np}$, $G_{p\times qr}$ and $H_{r\times s}$ be given. Based on Lemma 2.1 we can, without loss of generality, assume $m = 1$ and $s = 1$. Split F as

$$F = [F_1, \cdots, F_p],$$

where F_i, $i = 1, \cdots, p$ are $1 \times n$ blocks. Then

$$\begin{aligned} F \ltimes G &= \begin{bmatrix} F_1, \cdots, F_p \end{bmatrix} \ltimes \begin{bmatrix} g^1_{11} & \cdots & g^1_{1q} & \cdots & g^1_{r1} & \cdots & g^1_{rq} \\ \vdots & & & & & & \\ g^p_{11} & \cdots & g^p_{1q} & \cdots & g^p_{r1} & \cdots & g^p_{rq} \end{bmatrix} \\ &= \left[\sum_{i=1}^{p} F_i g^i_{11}, \cdots, \sum_{i=1}^{p} F_i g^i_{1q}, \cdots, \sum_{i=1}^{p} F_i g^i_{r1}, \cdots, \sum_{i=1}^{p} F_i g^i_{rq} \right]. \end{aligned}$$

Then we have

$$\begin{aligned} (F \ltimes G) \ltimes H &= (F \ltimes G) \ltimes \begin{bmatrix} h_1 \\ \vdots \\ h_r \end{bmatrix} \\ &= \left[\sum_{j=1}^{r} \sum_{i=1}^{p} F_i g^i_{j1} h_j, \cdots, \sum_{j=1}^{r} \sum_{i=1}^{p} F_i g^i_{jq} h_j \right]. \end{aligned} \tag{2.13}$$

On the other hand,

$$\begin{aligned}
&\begin{bmatrix} g_{11}^1 \cdots g_{1q}^1 & \cdots & g_{r1}^1 \cdots g_{rq}^1 \\ \vdots & & \\ g_{11}^p \cdots g_{1q}^p & \cdots & g_{r1}^p \cdots g_{rq}^p \end{bmatrix} \ltimes \begin{bmatrix} h_1 \\ \vdots \\ h_r \end{bmatrix} \\
&= \begin{bmatrix} \sum\limits_{j=1}^r g_{j1}^1 h_j & \cdots & \sum\limits_{j=1}^r g_{jq}^1 h_j \\ \vdots & & \\ \sum\limits_{j=1}^r g_{j1}^p h_j & \cdots & \sum\limits_{j=1}^r g_{jq}^p h_j \end{bmatrix}.
\end{aligned}$$

Then

$$\begin{aligned}
F \ltimes (G \ltimes H) &= [F_1, \cdots, F_p] \ltimes (G \ltimes H) \\
&= \left[\sum_{j=1}^r \sum_{i=1}^p F_i g_{j1}^i h_j, \cdots, \sum_{j=1}^r \sum_{i=1}^p F_i g_{jq}^i h_j \right],
\end{aligned}$$

which is the same as equation (2.13).

Since Cases 4.1–4.4 are similar, we prove Case 4.1 only.

Let $F_{m\times npr}$, $G_{p\times q}$ and $H_{rq\times s}$ be given. We also assume $m = 1$ and $s = 1$. Split F as

$$F = [F_{11}, \cdots, F_{1r}, \cdots, F_{p1}, \cdots, F_{pr}],$$

where each F_{ij}, $i = 1, \cdots, p$, $j = 1, \cdots, r$ are $1 \times n$ blocks.

$$G = \begin{bmatrix} g_{11} & \cdots & g_{1q} \\ \vdots & & \\ g_{p1} & \cdots & g_{pq} \end{bmatrix}, \quad H = [h_{11}, \cdots, h_{1r}, \cdots, h_{q1}, \cdots, h_{qr}]^T.$$

A careful computation shows that

$$(F \ltimes G) \ltimes H = F \ltimes (G \ltimes H) = \sum_{i=1}^p \sum_{j=1}^r \sum_{k=1}^q F_{ij} g_{ik} h_{kj}.$$

□

In the above proof, we leave some cases for the reader to verify.

Before exploring more properties of left STP, we consider an interesting example, which may provide a convincing reason for this generalization of matrix product.

Example 2.3. Let $X, Y, Z, W \in \mathbb{R}^n$. Then

$$A := (XY^T)(ZW^T) \in \mathcal{M}_{n\times n}$$

is a well defined matrix. Denote the entries of A as $A=(a_{ij})$. Then

$$a_{ij}=\sum_{k=1}^{n} x_i y_k z_k w_j, \quad i,j=1,\cdots,n.$$

Noticing that the matrix product is associative and $Y^T Z$ is a scalar, we can do the following calculation:

$$A = XY^T ZW^T = X(Y^T Z)W^T = (Y^T Z)XW^T = Y^T(ZX)W^T. \tag{2.14}$$

Now we meet a puzzle: What is the expression ZX in (2.14)? The puzzle shows that the conventional matrix product has a "bug". Because an illegal item can be produced through legal algebraic transformations.

Since the conventional matrix product can be considered as a special case of the left STP of matrices, we may ignore the conventional product and consider all the products involved in this example are the left STP. Then we have

$$\begin{aligned} Y^T \ltimes (Z \ltimes X) \ltimes W^T &= y\left[z_1x_1 \ \cdots \ z_1x_n \ \cdots \ z_nx_1 \ \cdots \ z_nx_n\right]^T w \\ &= \left[\sum_{k=1}^{n} y_k z_k x_1 \ \cdots \ \sum_{k=1}^{n} y_k z_k x_n\right]^T w \\ &= \begin{bmatrix} \sum\limits_{k=1}^{n} y_k z_k x_1 w_1 & \cdots & \sum\limits_{k=1}^{n} y_k z_k x_1 w_n \\ & \vdots & \\ \sum\limits_{k=1}^{n} y_k z_k x_n w_1 & \cdots & \sum\limits_{k=1}^{n} y_k z_k x_n w_n \end{bmatrix} \\ &= A. \end{aligned} \tag{2.15}$$

2.3 Fundamental Properties

This section provides some other fundamental properties of the left STP. For statement ease, we simply call it the STP. Throughout this book the default STP is the left STP.

Proposition 2.1.

(1) Let $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^n$ be two column vectors. Then $X \ltimes Y$ is well defined. Moreover,

$$X \ltimes Y = X \otimes Y. \tag{2.16}$$

(2) Let $\omega \in \mathbb{R}^m$ and $\sigma \in \mathbb{R}^n$ be two row vectors. Then $\omega \ltimes \sigma$ is well defined. Moreover,

$$\omega \ltimes \sigma = \sigma \otimes \omega. \tag{2.17}$$

Proof. Both equalities can be verified directly by definition and straightforward computations. □

This proposition is simple but useful. It converts Kronecker product of vectors into STP. We give some examples to show how to use it.

Example 2.4.

(1) If X is a row vector or a column vector, then

$$X^k := \underbrace{X \ltimes X \ltimes \cdots \ltimes X}_{k}$$

is always well defined.

(2) Recall the tensor formula (1.44) in Chapter 1. Assume $f \in \mathcal{T}_t^s$. Using (2.16) and (2.17), we have

$$f(X_1, \cdots, X_s, \omega_1, \cdots, \omega_t) = \omega_t \ltimes \cdots \ltimes \omega_1 M_f X_1 \ltimes \cdots \ltimes X_s. \tag{2.18}$$

The advantage of (2.18) over (1.44) is, (2.18) involves only one matrix product, and since the STP has associative property, no parentheses are necessary any more. Later on, you can see that in the form of (2.18), a tensor will be manipulated much easier.

(3) Let $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^q$ be two columns and $A \in \mathcal{M}_{m\times n}$, $B \in \mathcal{M}_{p\times q}$ be two given matrices. Then

$$(AX) \ltimes (BY) = (A \otimes B)(X \ltimes Y). \tag{2.19}$$

In general let $X_i \in \mathbb{R}^{n_i}$ and $A_i \in \mathcal{M}_{m_i\times n_i}$, $i = 1, \cdots, k$. Then

$$\ltimes_{i=1}^k (A_i X_i) = \left(\otimes_{i=1}^k A_i\right)\left(\ltimes_{i=1}^k X_i\right). \tag{2.20}$$

Particularly,

$$(AX)^k = \left(\underbrace{A \otimes \cdots \otimes A}_{k}\right) X^k := A^{\otimes k} X^k. \tag{2.21}$$

(4) Let $\omega \in \mathbb{R}^m$, $\sigma \in \mathbb{R}^p$ be two rows and $A \in \mathcal{M}_{m\times n}$, $B \in \mathcal{M}_{p\times q}$ be two given matrices. Then

$$(\omega A) \ltimes (\sigma B) = (\omega \ltimes \sigma)(B \otimes A). \tag{2.22}$$

In general let $\omega_i \in \mathbb{R}^{m_i}$ be rows and $A_i \in \mathcal{M}_{m_i\times n_i}$, $i = 1, \cdots, k$. Then

$$\ltimes_{i=1}^k (\omega_i A_i) = \left(\ltimes_{i=1}^k \omega_i\right)\left(\otimes_{i=1}^k A_{k+1-i}\right). \tag{2.23}$$

Particularly,

$$(\omega A)^k = \omega^k A^{\otimes k}. \tag{2.24}$$

The verification of equations (2.19)–(2.24) is left to the reader.

The following proposition is about the block multiplication rule.

Proposition 2.2. *Assume $A \succ_t B$ (or $A \prec_t B$). Decompose A and B into blocks as*

$$A = \begin{bmatrix} A^{11} & \cdots & A^{1s} \\ \vdots & & \vdots \\ A^{r1} & \cdots & A^{rs} \end{bmatrix}, \quad B = \begin{bmatrix} B^{11} & \cdots & B^{1t} \\ \vdots & & \vdots \\ B^{s1} & \cdots & B^{st} \end{bmatrix}.$$

If $A^{ik} \succ_t B^{kj}$, $\forall\, i, j, k$ (correspondingly, $A^{ik} \prec_t B^{kj}$, $\forall\, i, j, k$), then

$$A \ltimes B = \begin{bmatrix} C^{11} & \cdots & C^{1t} \\ \vdots & & \vdots \\ C^{r1} & \cdots & C^{rt} \end{bmatrix}, \tag{2.25}$$

where

$$C^{ij} = \sum_{k=1}^{s} A^{ik} \ltimes B^{kj}.$$

Proof. Using the definition, a careful calculation by multiplying C^{ij} out and collecting terms leads to the conclusion. □

In fact, Definition 2.3 can be considered as a particular case of the above proposition.

Next, we consider the power of a matrix A.

Definition 2.4. Assume $A \in \mathcal{M}_{m\times n}$, where either $m|n$ or $n|m$. Then A^n is recursively defined as

$$\begin{cases} A^1 = A \\ A^{k+1} = A^k \ltimes A, \quad k = 1, 2, \cdots . \end{cases} \tag{2.26}$$

It is easy to see that A^k is well defined. Moreover, if $m = nt$, then $A^{k+1} \in \mathcal{M}_{t^k n\times n}$, and if $mt = n$, then $A^{k+1} \in \mathcal{M}_{m\times t^k m}$.

In the following remark we briefly discuss the dimension of the STP of matrices. We leave the verification of the claims in the following remark to the reader.

Remark 2.3.

(1) The dimension of the STP of matrices can be easily obtained by the common factor elimination of the second index of the leading matrix with the first index of the following matrix. For instance

$$A_{p\times qr} \ltimes B_{r\times s} \ltimes C_{qst\times l} = (A \ltimes B)_{p\times qs} \ltimes C_{qst\times l} = (A \ltimes B \ltimes C)_{pt\times l}.$$

To get the first equality r is canceled and to get the second equality qs is canceled. This way is obviously the generalization of the conventional matrix product. For instance, $A_{p\times s}B_{s\times q} = (AB)_{p\times q}$. It can be considered as s has been canceled.

(2) Unlike conventional multiplication, even if both $A \ltimes B$ and $B \ltimes C$ are well defined, $A \ltimes B \ltimes C = (A \ltimes B) \ltimes C$ may not be defined. A counter example is: $A \in M_{3\times 4}$, $B \in M_{2\times 3}$ and $C \in M_{9\times 1}$. (A general version of STP will be introduced in Chapter 4, where the STP is defined for arbitrary factor matrices.)

(3) If $A \succ_s B$ ($A \prec_s B$) and $B \succ_t C$ ($B \prec_t C$), then $A \ltimes B \succ_{st} C$ (Correspondingly, $A \ltimes B \prec_{st} C$). Hence if $A_1 \prec A_2 \prec \cdots \prec A_k$ or $A_1 \succ A_2 \succ \cdots \succ A_k$, then $\ltimes_{i=1}^{k} A_i$ is well defined.

(4) Let $p \geq 2$ be an integer. Define a set of matrices as

$$\mathcal{M}^p := \cup_{i,j\in\mathbb{Z}_+}\mathcal{M}_{p^i\times p^j}.$$

Then it is easy to see that for any $A, B \in \mathcal{M}^p$ their semi-tensor product is always well defined. Moreover, it is also closed, that is, $\ltimes : \mathcal{M}^p \times \mathcal{M}^p \to \mathcal{M}^p$. When the p-valued logic is considered ($p = 2$ is the standard logic), the matrices used there are of this form.

Proposition 2.3. *Assume $A \ltimes B$ is well defined, then*

$$(A \ltimes B)^T = B^T \ltimes A^T. \tag{2.27}$$

Proof. Assume X is a row, Y is a column and $X \ltimes Y$ is well defined, then a straightforward computation shows that

$$X \ltimes Y = \left(Y^T \ltimes X^T\right)^T. \tag{2.28}$$

Now consider $A \ltimes B$. By definition, the (i, j)th block of $A \ltimes B$ is

$$\mathrm{Row}_i(A) \ltimes \mathrm{Col}_j(B).$$

Meanwhile, the (j, i)th block of $B^T \ltimes A^T$ is

$$\mathrm{Row}_j(B^T) \ltimes \mathrm{Col}_i(A^T) = [\mathrm{Col}_j(B)]^T \ltimes [\mathrm{Row}_i(A)]^T.$$

According to (2.28),

$$\left([\mathrm{Col}_j(B)]^T \ltimes [\mathrm{Row}_i(A)]^T\right)^T = \mathrm{Row}_i(A) \ltimes \mathrm{Col}_j(B).$$

That is, the transpose of (i, j)th block of $A \ltimes B$ is the (j, i)th block of $B^T \ltimes A^T$. □

The following proposition shows that the STP of two matrices can easily be realized by using conventional product plus Kronecker product.

Proposition 2.4.

(1) If $A \succ_t B$, then

$$A \ltimes B = A(B \otimes I_t). \tag{2.29}$$

(2) If $A \prec_t B$, then

$$A \ltimes B = (A \otimes I_t)B. \tag{2.30}$$

Proof. We prove (2.29) only. The proof of (2.30) is similar. Say, $B \in \mathcal{M}_{p\times q}$. Then

$$B \otimes I_t = [\mathrm{Col}_1(B) \otimes I_t \;\; \mathrm{Col}_2(B) \otimes I_t \;\; \cdots \;\; \mathrm{Col}_q(B) \otimes I_t\,].$$

Using this form and Proposition 2.2, we can, without loss of generality, assume A is a row and B is a column. Then a straightforward computation verifies the equality. □

Proposition 2.4 is of particular importance. Many properties can easily be obtained via (2.29) and (2.30). In fact, (2.29) and (2.30) can be considered as an alternative definition of the left STP of matrices.

Proposition 2.5. *Assume A and B are square matrices and both $A \ltimes B$ and $B \ltimes A$ are well defined, then*

(1) $A \ltimes B$ and $B \ltimes A$ have the same characteristic functions.

(2)

$$\mathrm{tr}(A \ltimes B) = \mathrm{tr}(B \ltimes A). \tag{2.31}$$

(3) If at least one of A and B is invertible, then

$$A \ltimes B \sim B \ltimes A, \tag{2.32}$$

where "$\sim$" stands for the similarity of two matrices.

(4) If both A and B are upper triangular (lower triangular, diagonal, orthogonal), then $A \ltimes B$ is also upper triangular (lower triangular, diagonal, orthogonal correspondingly).

(5) If both A and B are invertible, then

$$(A \ltimes B)^{-1} = B^{-1} \ltimes A^{-1}. \tag{2.33}$$

(6) If $A \prec_t B$, then

$$\det(A \ltimes B) = [\det(A)]^t \det(B). \tag{2.34}$$

If $A \succ_t B$, then

$$\det(A \ltimes B) = \det(A)[\det(B)]^t. \tag{2.35}$$

Proof. Using (2.29) and (2.30) to convert the STP into conventional product plus Kronecker product, then the above properties can be easily obtained via known properties of either conventional or Kronecker products. As an example, we prove (5): Assume $A \prec_t B$, then

$$\begin{aligned}(A \ltimes B)^{-1} &= (A(B \otimes I_t))^{-1} = (B \otimes I_t)^{-1} A^{-1}\\ &= (B^{-1} \otimes I_t) A^{-1} = B^{-1} \ltimes A^{-1}.\end{aligned}$$

□

The STP of a matrix with an identity matrix has some special properties. Roughly speaking, if the size of I_k is larger than the size of matrix M (comparing the column number of the first factor with the row number of the second factor), then it will enlarge M, otherwise, it keeps M unchanged.

Proposition 2.6.

(1) Let $M \in \mathcal{M}_{m \times pn}$. Then

$$M \ltimes I_n = M. \tag{2.36}$$

(2) Let $M \in \mathcal{M}_{m \times n}$. Then

$$M \ltimes I_{pn} = M \otimes I_p. \tag{2.37}$$

(3) Let $M \in \mathcal{M}_{pm \times n}$. Then

$$I_p \ltimes M = M. \tag{2.38}$$

(4) Let $M \in \mathcal{M}_{m \times n}$. Then

$$I_{pm} \ltimes M = I_p \otimes M. \tag{2.39}$$

Proof. All the equalities follow from Proposition 2.4 immediately. □

2.4 Pseudo-Commutativity via Swap Matrix

One major inferior of matrix product to scalar product is that it is not commutative. Using swap matrix etc., the STP can change the order of its factors in certain sense. We call these properties the pseudo-commutativity. It is very useful in applications. The following proposition is a re-statement of Proposition 1.14.

Proposition 2.7.

(1) Let $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^n$ be two column vectors. Then

$$W_{[m,n]} XY = YX. \tag{2.40}$$

(2) Let $\omega \in \mathbb{R}^m$ and $\sigma \in \mathbb{R}^n$ be two row vectors. Then

$$\omega\sigma W_{[m,n]} = \sigma\omega. \tag{2.41}$$

Equations (2.40) and (2.41) may tell you why $W_{[m,n]}$ is called "swap matrix". Later on, you will see that (2.40) and (2.41) are extremely useful, because they are used to overcome the non-commutative shortage of the matrix product.

The following is a generalization of Proposition 2.7, which shows that if $X_i \in \mathbb{R}^{n_i}$, $i = 1, \cdots, k$, are k column vectors. Then we can use swap matrix to swap the factors X_t with X_{t+1} in the product $\ltimes_{i=1}^k X_i$, and similar for row vectors.

Proposition 2.8.

(1) Let $X_i \in \mathbb{R}^{n_i}$, $i = 1, \cdots, k$, be k column vectors. Setting $\alpha = \prod_{j=1}^{t-1} n_j$, $\beta = \prod_{j=t+1}^{k} n_j$, we have

$$\begin{aligned} &\left[I_\alpha \otimes W_{[n_t,n_{t+1}]} \otimes I_\beta\right] X_1 X_2 \cdots X_k \\ &= X_1 \cdots X_{t-1} X_{t+1} X_t X_{t+2} \cdots X_k. \end{aligned} \tag{2.42}$$

(2) Similarly, let $\omega_i \in \mathbb{R}^{n_i}$, $i = 1, \cdots, k$ be k row vectors. Then we have

$$\begin{aligned} &\omega_1 \omega_2 \cdots \omega_k \left[I_\beta \otimes W_{[n_t,n_{t+1}]} \otimes I_\alpha\right] \\ &= \omega_1 \cdots \omega_{t-1} \omega_{t+1} \omega_t \omega_{t+2} \cdots \omega_k. \end{aligned} \tag{2.43}$$

Proof. We prove (2.42) only. The proof of (2.43) is similar.

$$\begin{aligned} LHS = &\left[I_\alpha \otimes W_{[n_t,n_{t+1}]} \otimes I_\beta\right] \\ &\times \left[(X_1 \cdots X_{t-1}) \otimes X_t X_{t+1} \otimes (X_{t+2} \cdots X_k)\right] \\ = &\left[I_\alpha \times (X_1 \cdots X_{t-1})\right] \otimes \left[W_{[n_t,n_{t+1}]} \times X_t X_{t+1}\right] \\ &\otimes \left[I_\beta \times (X_{t+2} \cdots X_k)\right] \\ = &(X_1 \cdots X_{t-1}) \otimes (X_{t+1} X_t) \otimes (X_{t+2} \cdots X_k) \\ = &(X_1 \cdots X_{t-1}) \ltimes (X_{t+1} X_t) \ltimes (X_{t+2} \cdots X_k) = RHS. \end{aligned}$$

□

In fact, the swap matrix can also be used to exchange the order of blocks in a matrix. The following is a further generalization of Proposition 2.7, or a generalization of Proposition 2.8.

Proposition 2.9.

(1) Assume a matrix A is split into a row of blocks as

$$A = [A_{11}, \cdots, A_{1n}, \cdots, A_{m1}, \cdots, A_{mn}],$$

where all the blocks have the same size. Moreover, the blocks are ordered by multi-index $\mathrm{id}(i, j; m, n)$. *Then*

$$AW_{[n,m]} = [A_{11}, \cdots, A_{m1}, \cdots, A_{1n}, \cdots, A_{mn}], \tag{2.44}$$

in which the blocks are ordered by multi-index $\mathrm{id}(j, i; n, m)$.

(2) Let

$$B = \left[B_{11}^T, \cdots, B_{1n}^T, \cdots, B_{m1}^T, \cdots, B_{mn}^T\right]^T$$

be a column of blocks, in which the equal-size blocks are ordered by the multi-index $\mathrm{id}(i, j; m, n)$. *Then*

$$W_{[m,n]}B = \left[B_{11}^T, \cdots, B_{m1}^T, \cdots, B_{1n}^T, \cdots, M_{mn}^T\right]^T, \tag{2.45}$$

in which the blocks are ordered by multi-index $\mathrm{id}(j, i; n, m)$.

We leave the proof to the reader.

In the following we consider the swap of matrices with vectors. we need some auxiliary properties.

Lemma 2.2.

(1) Let Z be a t-dimensional row vector and $A \in \mathcal{M}_{m\times n}$. Then

$$ZW_{[m,t]}A = AZW_{[n,t]} = A \otimes Z. \tag{2.46}$$

(2) Let Y be a t-dimensional column vector and $A \in \mathcal{M}_{m\times n}$. Then

$$AW_{[t,n]}Y = W_{[t,m]}YA = A \otimes Y. \tag{2.47}$$

Proof. (1) Using equation (1.60) in Exercise 2.1, a direct computation shows that

$$ZW_{[m,t]} = \sum_{j=1}^{t} z_j I_m \otimes (\delta_n^j)^T = I_m \otimes Z. \tag{2.48}$$

Using Proposition 2.4, we have

$$ZW_{[m,t]}A = (I_m \otimes Z)A = (I_m \otimes Z)(A \otimes I_t) = A \otimes Z.$$

Similarly, we have

$$AZW_{[n,t]} = A(I_n \otimes Z) = (A \otimes I_1)(I_n \otimes Z) = A \otimes Z.$$

(2) Starting from (2.46), we replace A by A^T and replace Z by Y^T, and then take transpose on both sides. Noting that $W_{[m,n]}^T = W_{[n,m]}$, (2.47) follows immediately.

□

Lemma 2.3. *Let $A \in \mathcal{M}_{m\times n}$ and $X \in \mathcal{M}_{n\times q}$. Then*

$$V_r(AX) = A \ltimes V_r(X); \tag{2.49}$$

and

$$V_c(AX) = (I_q \otimes A)\, V_c(X). \tag{2.50}$$

We leave the proves of (2.49) and (2.50) to the reader.

The following lemma is useful in the sequel.

Lemma 2.4. *Let $A \in \mathcal{M}_{m\times n}$. Then*

$$W_{[m,q]} \ltimes A \ltimes W_{[q,n]} = I_q \otimes A. \tag{2.51}$$

Equivalently, we have

$$W_{[q,m]}(I_q \otimes A) W_{[n,q]} = A. \tag{2.52}$$

Proof. Let $X \in M_{n\times q}$. According to (2.49) we have

$$V_r(AX) = A \ltimes V_r(X) = A \ltimes W_{[q,n]} V_c(X). \tag{2.53}$$

Multiplying both sides of (2.53) by $W_{[m,q]}$ yields

$$V_c(AX) = (W_{[m,q]} \ltimes A \ltimes W_{[q,n]}) V_c(X). \tag{2.54}$$

Comparing (2.50) with (2.54) and taking into consideration that the entries of X are arbitrary, it is clear that (2.51) is true.

Left multiplying $W_{[q,m]}$ and right multiplying $W_{[n,q]}$ convert (2.51) to (2.52). □

Now we are ready to present the following result, which may be considered as the pseudo-commutativity between matrices and vectors. It is of particular importance.

Proposition 2.10. *Given $A \in \mathcal{M}_{m\times n}$.*

(1) Let $\omega \in \mathbb{R}^t$ be a row vector. Then

$$A\omega = \omega W_{[m,t]} A W_{[t,n]} = \omega (I_t \otimes A). \tag{2.55}$$

(2) Let $Z \in \mathbb{R}^t$ be a column vector. Then

$$ZA = W_{[m,t]} A W_{[t,n]} Z = (I_t \otimes A) Z. \tag{2.56}$$

(3) Let $X \in \mathbb{R}^m$ be a column vector. Then

$$X^T A = [V_r(A)]^T X. \tag{2.57}$$

(4) Let $Y \in \mathbb{R}^n$ be a column vector. Then

$$AY = Y^T V_c(A). \tag{2.58}$$

(5) Let $X \in \mathbb{R}^m$ be a column vector and $\omega \in \mathbb{R}^n$ be a row vector. Then

$$X\omega = \omega W_{[m,n]} X. \tag{2.59}$$

Proof. Right multiplying both sides of the first equality of (2.46) by $W_{[t,n]}$, and using the fact that $W_{[n,t]}^{-1} = W_{[t,n]}$ yield the first equality of (2.55). S-tarting from the first equality and using (2.51), we have the second equality.

Similarly, left multiplying both sides of the first equality of (2.47) by $W_{[m,t]}$ yields the first equality of (2.56). Applying (2.51) to the first equality yields the second equality.

We leave the proves of (2.57), (2.58), and (2.59) to the reader.

Note that $W_{[1,t]} = W_{[t,1]} = I_t$. Then (2.59) is an immediate consequence of (2.55) or (2.56). □

The following result "swaps" two factors of a Kronecker product.

Proposition 2.11. *Let $A \in \mathcal{M}_{m\times n}$ and $B \in \mathcal{M}_{s\times t}$. Then*

$$A \otimes B = W_{[s,m]} \ltimes B \ltimes W_{[m,t]} \ltimes A = (I_m \otimes B) \ltimes A. \tag{2.60}$$

Proof. Denoting $A^i := \mathrm{Row}_i(A)$, $i = 1, \cdots, m$, and $B^j := \mathrm{Row}_j(B)$, $j = 1, \cdots, s$, a straightforward computation shows

$$A \otimes B = \begin{bmatrix} a_{11}B^1 & \cdots & a_{1n}B^1 \\ a_{11}B^2 & \cdots & a_{1n}B^2 \\ \vdots & & \\ a_{11}B^s & \cdots & a_{1n}B^s \\ \vdots & & \\ a_{m1}B^1 & \cdots & a_{mn}B^1 \\ a_{m1}B^2 & \cdots & a_{mn}B^2 \\ \vdots & & \\ a_{m1}B^s & \cdots & a_{mn}B^s \end{bmatrix} = \begin{bmatrix} B^1 \ltimes A^1 \\ B^2 \ltimes A^1 \\ \vdots \\ B^s \ltimes A^1 \\ \vdots \\ B^1 \ltimes A^m \\ B^2 \ltimes A^m \\ \vdots \\ B^s \ltimes A^m \end{bmatrix}. \tag{2.61}$$

Applying (2.46) to each row of B yields

$$B \ltimes W_{[m,t]} \ltimes A = \begin{bmatrix} A \otimes B^1 \\ A \otimes B^2 \\ \vdots \\ A \otimes B^s \end{bmatrix} = \begin{bmatrix} B^1 \ltimes A^1 \\ B^1 \ltimes A^2 \\ \vdots \\ B^1 \ltimes A^m \\ \vdots \\ B^s \ltimes A^1 \\ B^s \ltimes A^2 \\ \vdots \\ B^s \ltimes A^m \end{bmatrix}. \tag{2.62}$$

Comparing (2.61) with (2.62), one sees easily that the blocks of

$$\{B^i \ltimes A^j \,|\, i = 1, \cdots, s;\ j = 1, \cdots, m\}$$

are arranged in (2.62) by the order of $\mathrm{id}(i, j; s, m)$ and arranged in (2.61) by the order of $\mathrm{id}(j, i; m, s)$. Using Proposition 2.9, the first equality of (2.60) follows.

Applying (2.51) to the first equality of (2.60), its second equality follows. □

We give an example for this.

Example 2.5. Assume

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix},$$

then $m = n = 2$, $s = 3$, $t = 2$. Hence we have

$$W_{[32]} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad W_{[22]} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$
\begin{aligned}
&W_{[32]} \ltimes B \ltimes W_{[22]} \ltimes A\\
&= \begin{bmatrix} b_{11} & b_{12} & 0 & 0\\ b_{21} & b_{22} & 0 & 0\\ b_{31} & b_{32} & 0 & 0\\ 0 & 0 & b_{11} & b_{12}\\ 0 & 0 & b_{21} & b_{22}\\ 0 & 0 & b_{31} & b_{32} \end{bmatrix} \ltimes \begin{bmatrix} a_{11} & a_{12}\\ a_{21} & a_{22} \end{bmatrix}\\
&= \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12}\\ a_{11}b_{21} & a_{11}b_{22} & a_{12}b_{21} & a_{12}b_{22}\\ a_{11}b_{31} & a_{11}b_{32} & a_{12}b_{31} & a_{12}b_{32}\\ a_{21}b_{11} & a_{21}b_{12} & a_{22}b_{11} & a_{22}b_{12}\\ a_{21}b_{21} & a_{21}b_{22} & a_{22}b_{21} & a_{22}b_{22}\\ a_{21}b_{31} & a_{21}b_{32} & a_{22}b_{31} & a_{22}b_{32} \end{bmatrix} = A \otimes B.
\end{aligned}
$$

Finally, we show some factorization properties of the swap matrix. Later on, you will see that they are useful in manipulating data order.

Proposition 2.12. *Consider a swap matrix.*

(1) When $n = qr$ is a composite number, $W_{[p,n]}$ can be factorized as

$$W_{[p,qr]} = (I_q \otimes W_{[p,r]})(W_{[p,q]} \otimes I_r) = (I_r \otimes W_{[p,q]})(W_{[p,r]} \otimes I_q). \tag{2.63}$$

(2) When $m = pq$ is a composite number, $W_{[m,r]}$ can be factorized as

$$W_{[pq,r]} = (W_{[p,r]} \otimes I_q)(I_p \otimes W_{[q,r]}) = (W_{[q,r]} \otimes I_p)(I_q \otimes W_{[p,r]}). \tag{2.64}$$

Proof. We prove (2.63) only. Let $X_1 \in \mathbb{R}^p$, $X_2 \in \mathbb{R}^q$, and $X_3 \in \mathbb{R}^r$. Then

$$W_{[p,qr]}X_1X_2X_3 = X_2X_3X_1.$$

Meanwhile,

$$(W_{[p,q]} \otimes I_r)X_1X_2X_3 = X_2X_1X_3,$$

and

$$(I_q \otimes W_{[p,r]})(W_{[p,q]} \otimes I_r)X_1X_2X_3 = (I_q \otimes W_{[p,r]})X_2X_1X_3 = X_2X_3X_1.$$

That is,

$$W_{[p,qr]}X_1X_2X_3 = (I_q \otimes W_{[p,r]})(W_{[p,q]} \otimes I_r)X_1X_2X_3.$$

Since X_1, X_2 and X_3 are arbitrarily chosen, the above equation shows the first equality in (2.63). Exchanging q and r yields the second equality of (2.63).

The proof of (2.64) is similar. □

2.5 Semi-Tensor Product as Bilinear Mapping

When the STP is applied to two vectors, then it becomes a bilinear mapping. This mapping has particular importance in further applications. This section is devoted to explore this.

Definition 2.5. Let E, F, G be three vector spaces. A mapping $\phi : E \times F \to G$ is called a bilinear mapping, if

$$\begin{aligned} \phi(aX_1 + bX_2, Y) &= a\phi(X_1, Y) + b\phi(X_2, Y); \\ \phi(X, cY_1 + dY_2) &= c\phi(X, Y_1) + d\phi(X, Y_2), \end{aligned} \tag{2.65}$$

where $a, b, c, d \in \mathbb{R}$; $X, X_1, X_2 \in E$; $Y, Y_1, Y_2 \in F$.

Let $\{e_1, \cdots, e_m\}$ and $\{f_1, \cdots, f_n\}$ be bases of E and F respectively. Denote by

$$t_{i,j} = \phi(e_i, f_j), \quad 1 \leq i \leq m,\ 1 \leq j \leq n,$$

and let

$$T = \text{Span}\{t_{i,j}\} \subset G.$$

Then T is the smallest subspace containing $\mathcal{I}m(\phi)$. Assume

$$\{\, t_{i,j} \mid 1 \leq i \leq m, 1 \leq j \leq n\}$$

are linearly independent, then they form a basis of T. Arrange them into a matrix form by using multi-index $\text{id}(i, j; m, n)$ as

$$B_T = (t_{11}, t_{12}, \cdots, t_{1n}, \cdots, t_{m1}, \cdots, t_{mn}).$$

Let

$$X = \sum_{i=1}^{m} x_i e_i \in E; \quad Y = \sum_{j=1}^{n} y_j f_j \in F.$$

Then

$$\phi(X, Y) = \sum_{i=1}^{m} x_i \sum_{j=1}^{n} y_j \phi(e_i, f_j) = B_T \ltimes (x_i, \cdots, x_m)^T \ltimes (y_1, \cdots, y_n)^T.$$

Simply express a vector by its coefficients as $X \sim (x_i, \cdots, x_m)^T$ etc. Then we have $\phi : E \times F \to T$ is described as

$$\phi(X, Y) = B_T \ltimes X \ltimes Y. \tag{2.66}$$

Observe that $\mathcal{I}m(\phi) \neq T$. Particularly, we would like to emphasize that $\mathcal{I}m(\phi)$ is not a vector space. For instance, assume $E = F = \mathbb{R}^2$ with their canonical basis $\{\delta_2^1, \delta_2^2\}$. Then

$$t_{i,j} = \delta_2^i \ltimes \delta_2^j, \quad i, j = 1, 2.$$

Note that $\{t_{i,j} | i = 1, 2,\ j = 1, 2\}$ are linearly independent, and

$$T = \text{Span}\,\{t_{i,j} | i = 1, 2,\ j = 1, 2\} = \mathbb{R}^4.$$

But it is easy to verify that

$$\mathcal{I}m(\phi) = \left\{ z \in \mathbb{R}^4 \;\middle|\; z_1 z_4 = z_2 z_3 \right\}.$$

Particularly, we use the STP as a mapping $\ltimes : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^{mn}$, defined in a natural way as in (2.66). Denote by $Z = X \ltimes Y \in \mathbb{R}^{mn}$ and label the entries of Z by $\text{id}(i, j; m, n)$. Then the image is

$$\mathcal{I}m(\ltimes) = \left\{ Z = (z_{11}, \cdots, z_{1n}, \cdots, z_{mn})^T \;\middle|\; z_{i,p} z_{j,q} = z_{i,q} z_{j,p} \right\}.$$

That is, there are

$$\binom{m}{2}\binom{n}{2} = \frac{m(m-1)n(n-1)}{4}$$

constrained equations, which are

$$z_{i,p} z_{j,q} = z_{i,q} z_{j,p}, \quad 1 \leq i \neq j \leq m;\ 1 \leq p \neq q \leq n.$$

Definition 2.6. Let E and F be two finite dimensional vector spaces. A bilinear mapping $\otimes : E \times F \to T$, where $T \supset \mathcal{I}m(\otimes)$, is called a universal mapping, if for any bilinear mapping $\phi : E \times F \to H$, there exists a unique mapping $f : T \to H$, such that Fig. 2.1 is commutative. That is,

$$\phi = f \circ \otimes.$$

$$\begin{tikzcd}
E \times F \arrow[rr, "\phi"'] \arrow[d, "\otimes"] & & H \\
T \arrow[urr, "f"'] & &
\end{tikzcd}$$

Fig. 2.1 Universal bilinear mapping

Proposition 2.13. *The STP $\ltimes : E \times F \to T$ is a universal bilinear mapping, where T is the space generated by the image of $\ltimes$.*

Proof. Let $\phi : E \times F \to H$ be a bilinear mapping. Then

$$\phi(X, Y) = M_\phi XY, \quad X \in E,\ Y \in F,$$

where M_ϕ is the structure matrix of ϕ. Define $f : T \to H$ as a linear mapping $Z \mapsto M_\phi Z$, Then $\phi = f \circ \ltimes$. The uniqueness of f comes from the fact that the structure matrix of a bilinear mapping is unique. □

In the following we prove a property concerning about the STP as a bilinear mapping.

Proposition 2.14. *Let $X_1, \cdots, X_k \in E$, $Y_1, \cdots, Y_k \in F$, and*

$$\sum_{i=1}^{k} X_i \ltimes Y_i = 0. \tag{2.67}$$

(1) If $Y_1, \cdots, Y_k$ are linearly independent, then $X_1 = \cdots = X_k = 0$.
(2) If $X_1, \cdots, X_k$ are linearly independent, then $Y_1 = \cdots = Y_k = 0$.

Proof. (1) Denote

$$X_i = (x_i^1, \cdots, x_i^m)^T, \quad i = 1, \cdots, k.$$

Then

$$\sum_{i=1}^{k} X_i \ltimes Y_i = \begin{bmatrix} \sum\limits_{i=1}^{k} x_i^1 Y_i \\ \vdots \\ \sum\limits_{i=1}^{k} x_i^m Y_i \end{bmatrix} = 0.$$

The conclusion follows.

(2)

$$\sum_{i=1}^{k} Y_i \ltimes X_i = W_{[m,n]} \sum_{i=1}^{k} X_i \ltimes Y_i = 0.$$

The above conclusion implies this one.

□

We refer to Greub (1978) for a completed description of multilinear mappings.

Exercises

2.1 Let $f : V_1 \times V_2 \times \cdots \times V_k \to V_0$ be a multilinear mapping. Denote by $\{e^i_1, \cdots, e^i_{n_i}\}$ the basis of V_i, $i = 0, 1, \cdots, k$.

(i) Give the matrix expression of f. That is, find a matrix M_f, which is called the structure matrix of f, such that

$$X_0 = f(X_1, \cdots, X_k) = M_f \ltimes_{i=1}^{k} X_i, \quad X_i \in V_i,\ i = 0, 1, \cdots, k.$$

(ii) Let $(\tilde{e}^i_1, \cdots, \tilde{e}^i_{n_i})$ be another basis of V_i, satisfying

$$\begin{bmatrix} \tilde{e}^i_1 \\ \vdots \\ \tilde{e}^i_{n_i} \end{bmatrix} = A_i \begin{bmatrix} e^i_1 \\ \vdots \\ e^i_{n_i} \end{bmatrix}, \quad i = 0, 1, \cdots, k.$$

Find the structure matrix of f under this new basis.

2.2 Let $V = \mathcal{M}_2$ be the vector space of 2×2 matrices with a basis $\{e_1, e_2, e_3, e_4\}$ as

$$e_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad e_3 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad e_4 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then each $X \in V$ can be expressed in vector form as $X = (x_1\ x_2\ x_3\ x_4)^T$ (Precisely, $X = \sum\limits_{i=1}^{4} x_i e_i$.)

(i) Define a product on V as the conventional matrix product:

$$X * Y = XY.$$

Find the structure matrix of the product M_*, such that

$$X * Y = M_* XY.$$

(ii) Define a product on V as the lie bracket:

$$X * Y = XY - YX.$$

Find the structure matrix of the product M_*, such that

$$X * Y = M_* XY.$$

2.3 Prove the following alternative expression of swap matrix.

$$W_{[m,n]} = \left[\delta^1_n \ltimes \delta^1_m \ \cdots \ \delta^n_n \ltimes \delta^1_m \ \cdots \ \delta^1_n \ltimes \delta^m_m \ \cdots \ \delta^n_n \ltimes \delta^m_m\right]. \qquad (2.68)$$

2.4 Complete the proof of Theorem 2.1.

2.5 Prove equations (2.19)–(2.24).

2.6 (i) Let $X \in V$. Define a mapping, which reduce the covariant order of a tensor by one, denoted by $i_X : \mathcal{T}_s^r \to T_s^{r-1}$, and defined as (Boothby, 1986)

$$i_X(\omega) = \omega(X, \cdots ; \cdots), \quad \omega \in \mathcal{T}_s^r. \tag{2.69}$$

Prove that the structure matrix of $i_X(\omega)$ is

$$M_{i_X(\omega)} = M_\omega \ltimes X. \tag{2.70}$$

(ii) Let $\pi : \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^3 \to \mathbb{R}$ with its structure matrix as

$$M_\pi = \begin{bmatrix} 1 & -1 & 0 & 2 & 1 & -2 & 3 & -1 & 2 & 1 & 1 & -1 \end{bmatrix}.$$

Find the structure matrix of $i_x(\pi)$, where $X = (1 \ -1)^T$.

2.7 Let $\sigma \in V^*$. A mapping $i_\sigma : T_s^r \to T_{s-1}^r$ is defined as

$$i_\sigma(\omega) = \omega(\cdots ; \sigma, \cdots). \tag{2.71}$$

Prove that the structure matrix of $i_\sigma(\omega)$ is

$$M_{i_\sigma(\omega)} = \sigma \ltimes M_\omega. \tag{2.72}$$

2.8 Let η be a multilinear mapping:

$$\eta \in L\left(V_1 \times V_2 \times \cdots \times V_k; W\right),$$

where $\dim(V_i) = n_i$, $\dim(W) = n_0$. Given $X \in V_t$, we define a mapping

$$i_X^t(\eta) : L\left(V_1 \times \cdots \times V_k; W\right) \to L\left(V_1 \times \cdots V_{t-1} \times V_{t+1} \times \cdots \times V_k; W\right)$$

as

$$\begin{aligned} i_X^t(\eta)(Y_1, \cdots, Y_{k-1}) = \eta(Y_1, \cdots, Y_{t-1}, X, Y_t, \cdots, Y_{k-1}), \\ \forall\, Y_i \in V_i,\ i < t; Y_i \in V_{i+1},\ i \geq t. \end{aligned} \tag{2.73}$$

Denote by M_η and M_ζ the structure matrices of η and $\zeta = i_X^t(\eta)$ respectively. Prove that

$$M_\zeta = M_\eta \ltimes (I_{n_1 + \cdots + n_{t-1}} \otimes X). \tag{2.74}$$

2.9 Prove Proposition 2.5.

2.10 Prove equations (2.49) and (2.50).

2.11 Prove equations (2.57), (2.58), and (2.59).

2.12 Let $\sigma \in \mathbf{S}_k$ be a kth-order permutation. $n = n_1 \times \cdots \times n_k$, $n_i \geq 1$. Define an $n \times n$ matrix W_σ as follows: Label its columns by k indices

$i_1, \cdots, i_k$ in the order of $\mathrm{id}(i_1, \cdots, i_k; n_1, \cdots, n_k)$, and label its rows by k indices $j_1, \cdots, j_k$ in the order of $\mathrm{id}(j_{\sigma_1}, \cdots, j_{\sigma_k}; n_{\sigma_1}, \cdots, n_{\sigma_k})$. Then set its entry at $i_1 \cdots i_k$ column and $j_1 \cdots j_k$ row as

$$\omega_{j_1, \cdots, j_k}^{i_1, \cdots, i_k} = \begin{cases} 1, & i_1 = j_1, \cdots, i_k = j_k \\ 0, & \text{otherwise.} \end{cases}$$

Call the matrix W_σ the permutation matrix of σ.

(i) Given column vectors $X_i \in \mathbb{R}^{n_i}$, $i = 1, \cdots, k$, prove that

$$W_\sigma X_1 \cdots X_k = X_{\sigma_1} \cdots X_{\sigma_k}. \tag{2.75}$$

(ii) Given row vectors $\omega_i \in \mathbb{R}^{n_i}$, $i = 1, \cdots, k$, prove that

$$\omega_1 \cdots \omega_k W_\sigma = \omega_{\sigma_1} \cdots \omega_{\sigma_k}. \tag{2.76}$$

2.13 Let $\sigma = \sigma_k \sigma_{k-1} \cdots \sigma_1 \in \mathbf{S}_n$. Then

$$W_\sigma = W_{\sigma_k} W_{\sigma_{k-1}} \cdots W_1. \tag{2.77}$$

Prove it.

2.14 Assume $n_1 = n_2 = n_3 = 2$, $\sigma = (123) \in \mathbf{S}_3$.

(i) Construct the permutation matrix W_σ.

(ii) Let $\sigma_1 = (12)$, $\sigma_2 = (13)$. Then $\sigma = \sigma_2 \sigma_1$. Construct the swap matrices W_{σ_1} and W_{σ_2}. Then check that

$$W_\sigma = W_{\sigma_2} W_{\sigma_1}.$$

2.15 Consider sets of vectors as $X_i \in \mathbb{R}^u$, $Y_{ij} \in \mathbb{R}^v$, $Z_{ijk} \in \mathbb{R}^s$, $W_{ijk} \in \mathbb{R}^t$. Assume $\{X_i \,|\, 1 \le i \le \alpha\}$ are linearly independent, $\{Y_{ij} \,|\, 1 \le i \le \alpha; 1 \le j \le \beta\}$ are linearly independent, $\{Z_{ijk} \,|\, 1 \le i \le \alpha; 1 \le j \le \beta; 1 \le k \le \gamma\}$ are also linearly independent. Moreover,

$$\sum_{i=1}^{\alpha} \sum_{j=1}^{\beta} \sum_{k=1}^{\gamma} X_i Y_{ij} Z_{ijk} W_{ijk} = 0.$$

Then

$$W_{ijk} = 0, \quad 1 \le i \le \alpha; 1 \le j \le \beta; 1 \le k \le \gamma.$$

Prove it.

2.16 Let X be a square matrix and $p(x)$ a polynomial. Moreover, the polynomial $p(x)$ is expressed as $p(x) = q(x)x + p_0$.

(i) Prove that

$$V_r(p(X)) = q(X)V_r(X) + p_0V_r(I). \tag{2.78}$$

(ii) Give the column stacking form expression of (2.78).

2.17 Let $A \prec_t B$. Prove that

$$\text{rank}(AB) \leq \min(t \times \text{rank}(A), \text{rank}(B)).$$

Similarly, let $A \succ_t B$. Prove that

$$\text{rank}(AB) \leq \min(\text{rank}(A), t \times \text{rank}(B)).$$

2.18 (i) Use $W_{[2]}$ to express $W_{[2,8]}$;

(ii) Use $W_{[2,3]}$ to express $W_{[4,9]}$.

2.19 (i) A quadratic vector form of $x = (x_1\ x_2\ \cdots\ x_n)^T$ is expressed as

$$Y = \begin{bmatrix} X^TM_1X \\ X^TM_2X \\ \vdots \\ X^TM_mX \end{bmatrix}.$$

Find a matrix M, such that

$$Y = MX^2. \tag{2.79}$$

(ii) Given

$$Y = \begin{bmatrix} x_1^2 + 2x_2^2 + x_2x^3 \\ x_1x_2 - x_2x_3 \\ x_1x_3 + x_2^2 - x_3^2 \end{bmatrix},$$

find M such that Y can be expressed in the form of (2.79).

2.20 Let $A \in \mathcal{M}_{m\times n}$, $X \in \mathcal{M}_{n\times p}$, and $B \in \mathcal{M}_{p\times q}$. Prove that

$$V_r(AXB) = AW_{[q,n]}(I_n \times B^T)V_r(X).$$

(Hint: Use Lemma 2.3 and equation (1.65).)

Chapter 3

Multilinear Mappings among Vector Spaces

In this chapter we consider how to use STP to describe and analyze multilinear mappings among vector spaces. First, we consider the cross product on $\mathbb{R}^3$. Using this simple example, the structure matrix of multilinear mappings is introduced. Then, as two typical multilinear mappings, the Lie algebra and the linear mappings of matrices are investigated. Finally, two applications are considered: One is the general linear group and its algebra, and the other one is solving matrix equations, including the Hautus and Sylvester equations.

3.1 Cross Product on $\mathbb{R}^3$

In Chapter 2 we have investigated multilinear functions. By introducing the structure matrices, it was shown that the semi-tensor product is a proper tool to describe and analyze multilinear functions. In fact, a multilinear mapping can also be expressed via structure matrix and STP.

Definition 3.1. Let V_i, $i = 0, 1, \cdots, k$ be n_i-dimensional vector spaces with $\{e^i_1, \cdots, e^i_{n_i}\}$ as the fixed bases of V_i, and $\phi : V_1 \ltimes \cdots \ltimes V_k \to V_0$ be a multilinear mapping. Denote

$$\phi(e^1_{i_1}, \cdots, e^k_{i_k}) = \sum_{i_0=1}^{n_0} \mu^{i_0}_{i_1,\cdots,i_k} e^0_{i_0}, \quad i_j = 1, \cdots, n_j,\ j = 1, \cdots, k.$$

Then the matrix

$$M_\phi = \begin{bmatrix} \mu^1_{11\cdots1} & \cdots & \mu^1_{11\cdots n_k} & \cdots & \mu^1_{n_1n_2\cdots n_{k-1}1} & \cdots & \mu^1_{n_1n_2\cdots n_k} \\ \mu^2_{11\cdots1} & \cdots & \mu^2_{11\cdots n_k} & \cdots & \mu^2_{n_1n_2\cdots n_{k-1}1} & \cdots & \mu^2_{n_1n_2\cdots n_k} \\ \vdots & & & & & & \\ \mu^{n_0}_{11\cdots1} & \cdots & \mu^{n_0}_{11\cdots n_k} & \cdots & \mu^{n_0}_{n_1n_2\cdots n_{k-1}1} & \cdots & \mu^{n_0}_{n_1n_2\cdots n_k} \end{bmatrix} \tag{3.1}$$

is called the structure matrix of ϕ.

Let $X_i = [x_1^i, \cdots, x_{n_i}^i]^T \in V_i$, $i = 1, \cdots, k$ (precisely, $X_i = \sum_{j=1}^{n_i} x_j^i e_j^i$ and we use its coefficients as a column vector expression of X_i, when the basis of V_i is fixed). Then the following result is obvious.

Proposition 3.1. *Let $X_i \in V_i$, $i = 1, \cdots, k$. Then*

$$\phi(X_1, \cdots, X_k) = M_\phi \ltimes_{i=1}^k X_i. \tag{3.2}$$

This expression is very convenient in investigating multilinear mappings. In this section we mainly consider the cross product on $\mathbb{R}^3$.

Consider the cross product $\bowtie$ on $\mathbb{R}^3$ again. It is easy to verify that $\bowtie$: $\mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$ is a bilinear mapping. Fix the canonical basis of $\mathbb{R}^3$ as $\{e_1, e_2, e_3\}$, where $e_i = \delta_3^i$, $i = 1, 2, 3$. Denote

$$e_i \bowtie e_j = \mu_{ij}^1 e_1 + \mu_{ij}^2 e_2 + \mu_{ij}^3 e_3, \quad i, j = 1, 2, 3.$$

Then it is easy to calculate that

$$\begin{array}{llllll}
\mu_{11}^1 = 0, & \mu_{11}^2 = 0, & \mu_{11}^3 = 0, & \mu_{12}^1 = 0, & \mu_{12}^2 = 0, & \mu_{12}^3 = 1, \\
\mu_{13}^1 = 0, & \mu_{13}^2 = -1, & \mu_{13}^3 = 0, & \mu_{21}^1 = 0, & \mu_{21}^2 = 0, & \mu_{21}^3 = -1, \\
\mu_{22}^1 = 0, & \mu_{22}^2 = 0, & \mu_{22}^3 = 0, & \mu_{23}^1 = 1, & \mu_{23}^2 = 0, & \mu_{23}^3 = 0, \\
\mu_{31}^1 = 0, & \mu_{31}^2 = 1, & \mu_{31}^3 = 0, & \mu_{32}^1 = -1, & \mu_{32}^2 = 0, & \mu_{32}^3 = 0, \\
\mu_{33}^1 = 0, & \mu_{33}^2 = 0, & \mu_{33}^3 = 0. & & &
\end{array}$$

Hence, the structure matrix of $\bowtie$ is

$$M_c = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{3.3}$$

Assume $X = [x_1, x_2, x_3]^T$ and $Y = [y_1, y_2, y_3]^T$. Then

$$\begin{aligned}
X \bowtie Y &= M_c \ltimes X \ltimes Y \\
&= M_c \ltimes X \ltimes Y \\
&= M_c [x_1y_1, x_1y_2, x_1y_3, x_2y_1, x_2y_2, x_2y_3, x_3y_1, x_3y_2, x_3y_3]^T \\
&= [x_2y_3 - x_3y_2, -x_1y_3 + x_3y_1, x_1y_2 - x_2y_1]^T.
\end{aligned} \tag{3.4}$$

Using structure matrix to calculate the cross product of two vectors is not an efficient way. But the analytic expression of cross product, as

$$X \bowtie Y = M_c XY,$$

could be very convenient in theoretical analysis. A simple example is for multiple cross product. We have that

$$Y_1 \bowtie Y_2 \bowtie \cdots \bowtie Y_k = M_c^k \ltimes_{i=1}^k Y_i. \tag{3.5}$$

Particularly, we have

$$\underbrace{Y \ltimes Y \ltimes \cdots \ltimes Y}_{k} = M_c^k Y^k. \tag{3.6}$$

Following example may convince you on the convenience of this expression.

Example 3.1. In mechanics it is easy to see that the angular momentum of a rigid body about its mass center is

$$H = \int r \ltimes (\omega \ltimes r) \mathrm{d}m, \tag{3.7}$$

where $r = [x, y, z]$ is the position arrow, starting from the mass center; $\omega = [\omega_x, \omega_y, \omega_z]^T$ is the angular velocity vector. We want to prove the following equation for angular momentum of a rigid body (Sidi, 1997)

$$\begin{bmatrix} H_x \\ H_y \\ H_z \end{bmatrix} = \begin{bmatrix} I_x & -I_{xy} & -I_{zx} \\ -I_{xy} & I_y & -I_{yz} \\ I_{zx} & -I_{yz} & I_z \end{bmatrix} \begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix}, \tag{3.8}$$

where

$$\begin{array}{lll} I_x = \int (y^2 + z^2)\mathrm{d}m, & I_y = \int (z^2 + x^2)\mathrm{d}m, & I_z = \int (x^2 + y^2)\mathrm{d}m, \\ I_{xy} = \int xy\mathrm{d}m, & I_{yz} = \int yz\mathrm{d}m, & I_{zx} = \int zx\mathrm{d}m. \end{array}$$

Let M be the moment of the force, acting on the rigid body. We first prove that the dynamic equation of a rotating solid body is

$$\frac{\mathrm{d}H}{\mathrm{d}t} = M. \tag{3.9}$$

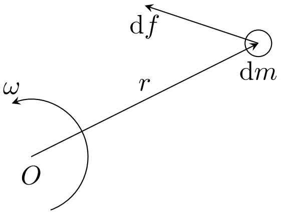


Fig. 3.1 Rotation

Consider a mass dm, with O as its rotating center, r as the position vector (from O to dm). df is the force acting on it, (refer to Fig. 3.1). From Newton's second law:

$$\mathrm{d}f = a\mathrm{d}m = \frac{\mathrm{d}v}{\mathrm{d}t}\mathrm{d}m = \frac{\mathrm{d}}{\mathrm{d}t}(\omega \ltimes r)\mathrm{d}m.$$

Now consider the moment of force on it, which is

$$\mathrm{d}M = r \bowtie \mathrm{d}f = r \bowtie \frac{\mathrm{d}}{\mathrm{d}t}(\omega \bowtie r)\mathrm{d}m.$$

Integrating it over the solid body, we have

$$M = \int r \bowtie \frac{\mathrm{d}}{\mathrm{d}t}(\omega \bowtie r)\mathrm{d}m. \tag{3.10}$$

We claim that

$$r \bowtie \frac{\mathrm{d}}{\mathrm{d}t}(\omega \bowtie r) = \frac{\mathrm{d}}{\mathrm{d}t}\left[r \bowtie (\omega \bowtie r)\right]. \tag{3.11}$$

$$\begin{aligned} RHS &= \tfrac{\mathrm{d}}{\mathrm{d}t}(r) \bowtie (\omega \bowtie r) + r \bowtie \tfrac{\mathrm{d}}{\mathrm{d}t}(\omega \bowtie r) \\ &= (\omega \bowtie r) \bowtie (\omega \bowtie r) + r \bowtie \tfrac{\mathrm{d}}{\mathrm{d}t}(\omega \bowtie r) \\ &= 0 + r \bowtie \tfrac{\mathrm{d}}{\mathrm{d}t}(\omega \bowtie r) = LHS. \end{aligned}$$

Applying this to (3.10), we have

$$\begin{aligned} M &= \textstyle\int \tfrac{\mathrm{d}}{\mathrm{d}t}[r \bowtie (\omega \bowtie r)]\mathrm{d}m \\ &= \tfrac{\mathrm{d}}{\mathrm{d}t} \textstyle\int r \bowtie (\omega \bowtie r)\mathrm{d}m \\ &= \tfrac{\mathrm{d}}{\mathrm{d}t} H. \end{aligned}$$

Next, we prove the angular momentum equation (3.8). According to (3.3), we have

$$\begin{aligned} H &= \textstyle\int r \bowtie (\omega \bowtie r)\mathrm{d}m \\ &= \textstyle\int M_c r M_c \omega r \mathrm{d}m \\ &= \textstyle\int M_c (I_3 \otimes M_c) r \omega r \mathrm{d}m \\ &= \textstyle\int M_c (I_3 \otimes M_c) W_{[3,9]} r^2 \omega \mathrm{d}m \\ &= \textstyle\int M_c (I_3 \otimes M_c) W_{[3,9]} r^2 \mathrm{d}m \omega. \\ &:= \textstyle\int \Psi r^2 \mathrm{d}m\omega, \end{aligned}$$

where

$$\begin{aligned} \Psi &= M_c (I_3 \otimes M_c) W_{[3,9]} \\ &= \begin{bmatrix} 0&0&0&0&-1&0&0&0&-1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0 \\ 0&1&0&0&0&0&0&0&0&-1&0&0&0&0&0&0&0&-1&0&0&0&0&0&0&0&1&0 \\ 0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&-1&0&0&0&-1&0&0&0&0 \end{bmatrix}. \end{aligned}$$

Then we have

$$\Psi r^2 = \begin{bmatrix} y^2+z^2 & xy & -xz \\ -xy & x^2+z^2 & -yz \\ -xz & -yz & x^2+y^2 \end{bmatrix}.$$

(3.8) follows immediately.

3.2 General Linear Algebra

This section considers the structure matrix of general linear algebra or its subalgebra. General linear algebra is an important Lie algebra. We first introduce Lie algebra.

Definition 3.2.

(1) Let V be a vector space over $\mathbb{R}$ and a mapping $* : V \times V \to V$ satisfies:

(i) (Bilinearity)

$$(\alpha X + \beta Y) * Z = \alpha(X * Z) + \beta(Y * Z), \quad \alpha, \beta \in \mathbb{R};\ X, Y, Z \in V; \tag{3.12}$$

(ii) (Skew-symmetry)

$$X * Y = -Y * X, \quad X, Y \in V; \tag{3.13}$$

(iii) (Jacobi Identity)

$$(X * Y) * Z + (Y * Z) * X + (Z * X) * Y = 0, \quad X, Y, Z \in V. \tag{3.14}$$

Then $(V, *)$ is called a Lie algebra.

(2) Let $W \subset V$ be a subspace. W is called a Lie subalgebra of $(V, *)$, if

$$X * Y \in W, \quad \forall\, X, Y \in W. \tag{3.15}$$

Consider $\mathcal{M}_n$. It is obviously a vector space over $\mathbb{R}$. We then can define a product over $\mathcal{M}_n$, called Lie bracket, as

$$[A, B] := AB - BA, \quad A, B \in \mathcal{M}_n. \tag{3.16}$$

We leave it to the reader to verify that $(\mathcal{M}_n, [\cdot, \cdot])$ is a Lie algebra. This algebra is called the nth order general linear algebra, and denoted by $gl(n, \mathbb{R})$. It is very important because any finite-dimensional Lie algebra is isomorphic to a subalgebra of the general linear algebra (Varadarajan, 1984).

Let $\{\Delta_{i,j} \mid i, j = 1, \cdots, n\}$ be a basis of $gl(n, \mathbb{R})$, where $\Delta_{ij} = (d^{i,j}_{p,q}) \in \mathcal{M}_n$ determined by

$$d^{i,j}_{p,q} = \begin{cases} 1, & p = i, \text{ and } q = j \\ 0, & \text{otherwise.} \end{cases}$$

Then it is clear that

$$[\Delta_{i,j}, \Delta_{\alpha,\beta}] = \begin{cases} \Delta_{i,\beta}, & j = \alpha, \text{ and } i \neq \beta \\ -\Delta_{\alpha,j}, & \beta = i, \text{ and } j \neq \alpha \\ \Delta_{i,\beta} - \Delta_{\alpha,j}, & j = \alpha \neq i = \beta \\ 0, & \text{otherwise.} \end{cases} \tag{3.17}$$

Now assume the basis is arranged into a row in the order of $\mathrm{id}(i,j;n,n)$, then the structure matrix, denoted by M_{L_n}, can easily be constructed using (3.17). Then we have

$$V_r([A,B]) = M_{L_n} V_r(A) V_r(B), \quad A, B \in gl(n,\mathbb{R}). \tag{3.18}$$

We give a numerical example for this.

Example 3.2. Consider $gl(2,\mathbb{R})$. The basis, defined by (3.17), is:

$$\Delta_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \Delta_{12} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \Delta_{21} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \Delta_{22} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Using (3.17), it is easy to calculate $M_{gl(2,\mathbb{R})}$ as

$$M_{gl(2,\mathbb{R})} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{3.19}$$

Proposition 3.2. *Consider $gl(n,\mathbb{R})$.*

(1) Define

$$sl(n,\mathbb{R}) := \{ A \in gl(n,\mathbb{R}) \mid \mathrm{tr}(A) = 0 \}. \tag{3.20}$$

$sl(n,\mathbb{R})$ is a subalgebra of $gl(n,\mathbb{R})$, which is called the special linear algebra.

(2) Define

$$o(n,\mathbb{R}) := \left\{ A \in gl(n,\mathbb{R}) \mid A^T = -A \right\}. \tag{3.21}$$

$o(n,\mathbb{R})$ is a subalgebra of $gl(n,\mathbb{R})$, which is called the orthogonal algebra.

(3) Define

$$sp(2n,\mathbb{R}) := \left\{ A \in gl(2n,\mathbb{R}) \mid A^T J + JA = 0 \right\}, \tag{3.22}$$

where

$$J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}.$$

$sp(n,\mathbb{R})$ is a subalgebra of $gl(2n,\mathbb{R})$, which is called the symplectic algebra.

Table 3.1 Lie bracket on $sl(2,\mathbb{R})$

	e_1	e_2	e_3
e_1	0	$-2e_1$	e_2
e_2	$2e_1$	0	$-2e_3$
e_3	$-e_2$	$2e_3$	0

We leave the proof of Proposition 3.2 to the reader.

Example 3.3. Consider $sl(2,\mathbb{R})$. Choose a basis as

$$e_1 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad e_3 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

From Table 3.1 the structure matrix of $sl(2,\mathbb{R})$ is obtained easily as

$$M_{sl(2,\mathbb{R})} = \begin{bmatrix} 0 & -2 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -2 & 0 & 2 & 0 \end{bmatrix}. \tag{3.23}$$

Definition 3.3. Let $(V,*)$ be a Lie algebra and $X \in V$. Then $\mathrm{ad}_X : V \to V$ defined by

$$\mathrm{ad}_X(Y) := X * Y, \quad Y \in V, \tag{3.24}$$

is called the adjoint representation of X.

Since $*$ is a bilinear mapping, ad_X is a linear mapping. Assume V is an n-dimensional vector space and fix a basis of V as $\{e_1, \cdots, e_n\}$. Then the corresponding structure matrix is uniquely expressed as M_V. Then we have

$$\mathrm{ad}_X(Y) = M_V XY, \quad \forall Y \in V.$$

Moreover, we have

Proposition 3.3. *Under a fixed basis the matrix expression of the adjoint representation is*

$$M_{\mathrm{ad}_X} = M_V X. \tag{3.25}$$

Example 3.4. Let

$$X = \begin{bmatrix} -2 & 0 \\ 1 & 2 \end{bmatrix} \in sl(2,\mathbb{R}).$$

Using the basis in Example 3.3, $X = -2e_2 + e_3 \sim (0 \ -2 \ 1)^T$. Then the matrix expression of ad_X is

$$M_{\mathrm{ad}_X} = M_{sl(2,\mathbb{R})} X = \begin{bmatrix} -4 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 2 & 4 \end{bmatrix}.$$

3.3 Mappings over Matrices

The row stacking and column stacking forms of a matrix have been investigated in Chapter 1. These expressions are sometimes very convenient in use. In this section we consider the matrix expression of mappings between matrices, which are expressed in their row or column stacking forms.

We denote by $L(V, W)$ the set of linear mappings from vector space V to vector space W. Particularly, $L(\mathcal{M}_{p\times q}, \mathcal{M}_{m\times n})$ is the set of linear mappings from $\mathcal{M}_{p\times q}$ to $\mathcal{M}_{m\times n}$. We start with some examples.

Example 3.5 (Lyapunov mapping). Given a square matrix $A \in \mathcal{M}_n$. Consider the following mapping $L_A : \mathcal{M}_n \to \mathcal{M}_n$, defined as

$$L_A(X) = AX + XA^T, \quad X \in \mathcal{M}_n. \tag{3.26}$$

A square matrix A is said to be Hurwitz, if all the eigenvalues of A have negative real parts. The following result is well known (Willems, 1970): A is a Hurwitz matrix, if and only if, for any negative definite matrix $Q < 0$, $L_A(X) = Q$ has unique solution, which is positive definite.

As a linear mapping on $\mathcal{M}_n$, L_A has a matrix expression as (Ooba and Funahashi, 1997)

$$M^c_{L_A} = A \otimes I + I \otimes A. \tag{3.27}$$

The precise meaning of this matrix expression is that as both the argument matrix and the image matrix are expressed into column stacking form, we have

$$V_c(L_A(X)) = M^c_{L_A} V_c(X). \tag{3.28}$$

Note that we use superscript c to indicate that the matrix expression is over column stacking forms. When the row stacking forms are considered, we should have

$$V_r(L_A(X)) = M^r_{L_A} V_r(X).$$

Since

$$L^T_A(X) = X^T A^T + AX^T = L_A(X^T),$$

we have

$$\begin{aligned} V_r(L_A(X)) &= V_c(L^T_A(X)) = V_c(L_A(X^T)) \\ &= M^c_{L_A} V_c(X^T) = M^c_{L_A} V_r(X). \end{aligned}$$

It follows that

$$M^r_{L_A} = M^c_{L_A}. \tag{3.29}$$

Example 3.6 (Symplectic mapping). Recall that a $2n\times 2n$ matrix $X \in sp(2n,\mathbb{R})$, if and only if, X satisfies (3.22). In general, we can replace J by an arbitrary matrix $N \in gl(n,\mathbb{R})$ and define a mapping (Cheng *et al.*, 1998)

$$L_N(X) = NX + X^T N, \quad X \in \mathcal{M}_n. \tag{3.30}$$

It was proved that (Cheng *et al.*, 1998)

$$\mathcal{G}_N = \{\, X \in gl(n,\mathbb{R}) \mid L_N(X) = 0\}$$

is a subalgebra of $gl(n,\mathbb{R})$.

It is not difficult to show that the matrix expression of L_N is

$$M^c_{L_N} = I_n \otimes N + (N^T \otimes I_n)W_{[n]}. \tag{3.31}$$

Denote by

$$GL(n,\mathbb{R}) := \{A \in \mathcal{M}_n \mid \det(A) \neq 0\}.$$

Then it is easy to check that $GL(n,\mathbb{R})$ with matrix product is a group. This group is called the general linear group, which is a Lie group (Boothby, 1986), with its corresponding Lie-algebra $gl(n,\mathbb{R})$. In fact, $\mathcal{G}_N$ is the Lie algebra of the Lie group

$$G_N = \{Z \in GL(n,\mathbb{R}) \,|\, Z^T N Z = N\},$$

which is a Lie subgroup of $GL(n,\mathbb{R})$ (Cheng *et al.*, 1998).

The following property of L_N is interesting.

Proposition 3.4 (Cheng *et al.*, 1998). *For any $N \in \mathcal{M}_n$, $L_N(X) = 0$ has at least a solution $X \neq 0$. In other words, 0 is an eigenvalue of L_N.*

Let $\rho : \mathcal{M}_{p\times q} \to \mathcal{M}_{m\times n}$. We always have two matrix expressions M^c_ρ and M^r_ρ corresponding to column stacking form and row stacking form respectively. The following proposition shows that these two expressions are easily convertible.

Proposition 3.5. *Let $\rho \in L(\mathcal{M}_{p\times q}, \mathcal{M}_{m\times n})$. Then*

$$\begin{cases} M^r_\rho = W_{[n,m]} M^c_\rho W_{[p,q]}, \\ M^c_\rho = W_{[m,n]} M^r_\rho W_{[q,p]}. \end{cases} \tag{3.32}$$

Particularly, if $\rho \in L(\mathcal{M}_n, \mathcal{M}_n)$, then (3.32) becomes

$$\begin{cases} M^r_\rho = W_{[n]} M^c_\rho W_{[n]}, \\ M^c_\rho = W_{[n]} M^r_\rho W_{[n]}. \end{cases} \tag{3.33}$$

Proof. Consider the first equality of (3.32). Using the formulas in Proposition 1.13, we have

$$\begin{aligned} V_r(\rho(X)) &= W_{[n,m]} V_c(\rho(X)) \\ &= W_{[n,m]} M_\rho^c V_c(X) \\ &= W_{[n,m]} M_\rho^c W_{[p,q]} V_r(X). \end{aligned}$$

Since $X \in \mathcal{M}_{p\times q}$ is arbitrary, the first equality of (3.32) follows.

Left multiplying both sides of the first equality of (3.32) by $W_{[m,n]}$, and right multiplying both sides by $W_{[q,p]}$, and using equality (1.53), we have the second equality of (3.32). □

For convenience, we take M_ρ^c as a default matrix expression of the linear mapping ρ on matrices.

Let $Z \in \mathcal{M}_{n\times p}$. We consider a general linear mapping, formed by matrix products. Precisely, let $\rho : \mathcal{M}_{n\times p} \to \mathcal{M}_{m\times q}$, be defined as

$$Z \mapsto AZB + CZ^T D, \tag{3.34}$$

where $A \in \mathcal{M}_{m\times n}$, $B \in \mathcal{M}_{p\times q}$, $C \in \mathcal{M}_{m\times p}$, $D \in \mathcal{M}_{n\times q}$.

It is not difficult to find that several useful linear matrix mappings, such as Lyapunov mapping, symplectic mapping etc., are particular forms of ρ, defined by (3.34). When we find the matrix expression of ρ, we have the matrix expressions of all such linear matrix mappings. In fact, we have the following:

Proposition 3.6. *The matrix expression of ρ defined by (3.34), is*

$$M^c = (B^T \otimes A) + (D^T \otimes C) W_{[p,n]}. \tag{3.35}$$

Proof. We first prove the matrix expressions of four fundamental linear matrix mappings in Table 3.2. In the following we assume the concerned matrices have their dimensions as in (3.34).

Table 3.2 Expression under column stacking forms

ρ	$\boldsymbol{M_\rho^c}$
$Z \mapsto AZ$	$I_p \otimes A$
$Z \mapsto ZB$	$B^T \otimes I_n$
$Z \mapsto CZ^T$	$(I_n \otimes C)W_{[p,n]}$
$Z \mapsto Z^T D$	$(D^T \otimes I_p)W_{[p,n]}$

We prove them one by one.

(i) Let $A \in \mathcal{M}_{m\times n}$ and $Z \in \mathcal{M}_{n\times p}$. Then

$$V_c(AZ) = (I_p \otimes A)V_c(Z). \tag{3.36}$$

A straightforward computation shows that

$$\begin{aligned} V_c(AZ) &= V_c\left[A\,\mathrm{Col}_1(Z)\ A\,\mathrm{Col}_2(Z)\ \cdots\ A\,\mathrm{Col}_p(Z)\ \right] \\ &= \begin{bmatrix} A\,\mathrm{Col}_1(Z) \\ A\,\mathrm{Col}_2(Z) \\ \vdots \\ A\,\mathrm{Col}_p(Z) \end{bmatrix} \\ &= \begin{bmatrix} A & 0 & \cdots & 0 \\ 0 & A & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & A \end{bmatrix} \begin{bmatrix} \mathrm{Col}_1(Z) \\ \mathrm{Col}_2(Z) \\ \vdots \\ \mathrm{Col}_p(Z) \end{bmatrix} \\ &= (I_p \otimes A)V_c(Z). \end{aligned}$$

(ii) Let $B \in \mathcal{M}_{p\times q}$ and $Z \in \mathcal{M}_{n\times p}$. Then

$$V_c(ZB) = (B^T \otimes I_n)V_c(Z). \tag{3.37}$$

We use (i) to prove it.

$$\begin{aligned} V_c(ZB) &= V_r(B^T Z^T) \\ &= W_{[n,q]}V_c(B^T Z^T) \\ &= W_{[n,q]}(I_n \otimes B^T)V_r(Z) \\ &= W_{[n,q]}(I_n \otimes B^T)W_{[n,p]}V_c(Z) \\ &= B^T V_c(Z). \end{aligned}$$

Note that the last equality comes from (2.52). Finally, according to Proposition 2.4, it follows that

$$B^T V_c(Z) = (B^T \otimes I_n)V_c(Z).$$

(iii) Let $C \in \mathcal{M}_{m\times p}$ and $Z \in \mathcal{M}_{n\times p}$. Then

$$V_c(CZ^T) = (I_n \otimes C)W_{[p,n]}V_c(Z). \tag{3.38}$$

To see this we have

$$\begin{aligned} V_c(CZ^T) &= (I_n \otimes C)V_c(Z^T) = (I_n \otimes C)V_r(Z) \\ &= (I_n \otimes C)W_{[p,n]}W_{[n,p]}V_r(Z) = (I_n \otimes C)W_{[p,n]}V_c(Z). \end{aligned}$$

(iv) Let $D \in \mathcal{M}_{n\times q}$ and $Z \in \mathcal{M}_{n\times p}$. Then

$$V_c(Z^T D) = (D^T \otimes I_p) W_{[p,n]} V_c(Z). \tag{3.39}$$

To get this formula, we have that

$$\begin{aligned} V_c(Z^T D) &= (D^T \otimes I_p) V_c(Z^T) = (D^T \otimes I_p) V_r(Z) \\ &= (D^T \otimes I_p) W_{[p,n]} W_{[n,p]} V_r(Z) = (D^T \otimes I_p) W_{[p,n]} V_c(Z). \end{aligned}$$

The mapping ρ of (3.34) can be considered as a linear combination of two compounded mappings. Using the matrix expressions of the four fundamental mappings we can get the matrix expressions of the two compounded mappings of ρ. Then adjusting the dimensions of the matrices in the mapping (3.34), we can have that

$$M^c = (B^T \otimes I_m)(I_p \otimes A) + (D^T \otimes I_m)(I_n \otimes C) W_{[p,n]}.$$

A simplification leads to (3.35). □

In the following we give some numerical examples to illustrate the formula (3.35).

Example 3.7. Assume $A, C \in \mathcal{M}_{3\times 2}$, $B, D \in \mathcal{M}_{2\times 4}$, and

$$A = \begin{bmatrix} 1 & -1 \\ 2 & 1 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ -1 & 1 \end{bmatrix},$$

$$B = \begin{bmatrix} 1 & -1 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 1 & 1 & -1 \\ 1 & 0 & 2 & 1 \end{bmatrix}.$$

(1) Assume $\rho : \mathcal{M}_2 \to \mathcal{M}_{3\times 2}$, defined as $Z \mapsto AZ$. Then

$$M_\rho = I_2 \otimes A = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Hence

$$V_c(AZ) = M_\rho V_c(Z) = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} z_{11} \\ z_{21} \\ z_{12} \\ z_{22} \end{bmatrix} = \begin{bmatrix} z_{11} - z_{21} \\ 2z_{11} + z_{21} \\ z_{21} \\ z_{12} - z_{22} \\ 2z_{12} + z_{22} \\ z_{22} \end{bmatrix}. \tag{3.40}$$

A direct computation shows that

$$AZ = \begin{bmatrix}1 & -1\\ 2 & 1\\ 0 & 1\end{bmatrix}\begin{bmatrix}z_{11} & z_{12}\\ z_{21} & z_{22}\end{bmatrix} = \begin{bmatrix}z_{11}-z_{21} & z_{12}-z_{22}\\ 2z_{11}+z_{21} & 2z_{12}+z_{22}\\ z_{21} & z_{22}\end{bmatrix},$$

which verifies (3.40).

(2) Assume $\rho : \mathcal{M}_2 \to \mathcal{M}_{2\times 4}$, defined as $Z \mapsto ZB$. Then

$$M_\rho = B^T \otimes I_2 = \begin{bmatrix}1 & 0 & 2 & 0\\ 0 & 1 & 0 & 2\\ -1 & 0 & 1 & 0\\ 0 & -1 & 0 & 1\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\end{bmatrix}.$$

Hence

$$\begin{aligned} V_c(ZB) &= M_\rho V_c(Z)\\ &= \begin{bmatrix}1 & 0 & 2 & 0\\ 0 & 1 & 0 & 2\\ -1 & 0 & 1 & 0\\ 0 & -1 & 0 & 1\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\end{bmatrix}\begin{bmatrix}z_{11}\\ z_{21}\\ z_{12}\\ z_{22}\end{bmatrix} = \begin{bmatrix}z_{11}+2z_{12}\\ z_{21}+2z_{22}\\ -z_{11}+z_{12}\\ -z_{21}+z_{22}\\ z_{12}\\ z_{22}\\ z_{11}\\ z_{21}\end{bmatrix}. \end{aligned} \tag{3.41}$$

A direct computation shows that

$$\begin{aligned} ZB &= \begin{bmatrix}z_{11} & z_{12}\\ z_{21} & z_{22}\end{bmatrix}\begin{bmatrix}1 & -1 & 0 & 1\\ 2 & 1 & 1 & 0\end{bmatrix}\\ &= \begin{bmatrix}z_{11}+2z_{12} & -z_{11}+z_{12} & z_{12} & z_{11}\\ z_{21}+2z_{22} & -z_{21}+z_{22} & z_{22} & z_{21}\end{bmatrix}, \end{aligned}$$

which verifies (3.41).

(3) Assume $\rho : \mathcal{M}_2 \to \mathcal{M}_{3\times 2}$, defined as $Z \mapsto CZ^T$. Note that

$$W_{[2,2]} = \begin{array}{c} \begin{array}{cccc}(11) & (12) & (21) & (22)\end{array}\\ \left[\begin{array}{cccc}1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1\end{array}\right] \end{array} \begin{array}{c} \\ (11)\\ (21)\\ (12)\\ (22)\end{array}.$$

It follows that

$$M_\rho = (I_2 \otimes C)W_{[2,2]} = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 1 \end{bmatrix}.$$

Hence

$$\begin{aligned} V_c(CZ^T) &= M_\rho V_c(Z) \\ &= \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} z_{11} \\ z_{21} \\ z_{12} \\ z_{22} \end{bmatrix} = \begin{bmatrix} z_{11} + 2z_{12} \\ z_{12} \\ -z_{11} + z_{12} \\ z_{21} + 2z_{22} \\ z_{22} \\ -z_{21} + z_{22} \end{bmatrix}. \end{aligned} \tag{3.42}$$

A direct computation shows that

$$CZ^T = \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} z_{11} & z_{21} \\ z_{12} & z_{22} \end{bmatrix} = \begin{bmatrix} z_{11} + 2z_{12} & z_{21} + 2z_{22} \\ z_{12} & z_{22} \\ -z_{11} + z_{12} & -z_{21} + z_{22} \end{bmatrix},$$

which verifies (3.42).

(4) Assume $\rho : \mathcal{M}_2 \to \mathcal{M}_{2\times 4}$, defined as $Z \mapsto Z^T D$. Then

$$M_\rho = (D^T \otimes I_2)W_{[2,2]} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

Hence

$$\begin{aligned} V_c(Z^T D) &= M_\rho V_c(Z) \\ &= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} z_{11} \\ z_{21} \\ z_{12} \\ z_{22} \end{bmatrix} = \begin{bmatrix} z_{11} + z_{21} \\ z_{12} + z_{22} \\ z_{11} \\ z_{12} \\ z_{11} + 2z_{21} \\ z_{12} + 2z_{22} \\ -z_{11} + z_{21} \\ -z_{12} + z_{22} \end{bmatrix}. \end{aligned} \tag{3.43}$$

A direct computation shows that

$$\begin{aligned} Z^T D &= \begin{bmatrix} z_{11} & z_{21} \\ z_{12} & z_{22} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & -1 \\ 1 & 0 & 2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} z_{11}+z_{21} & z_{11} & z_{11}+2z_{21} & -z_{11}+z_{21} \\ z_{12}+z_{22} & z_{12} & z_{12}+2z_{22} & -z_{12}+z_{22} \end{bmatrix}, \end{aligned}$$

which verifies (3.43).

(5) Assume $\rho : \mathcal{M}_2 \to \mathcal{M}_{3\times 4}$, defined as $Z \mapsto AZB + CZ^T D$. Using (3.35), we have

$$M_\rho = (B^T \otimes A) + (D^T \otimes C)W_{[2,2]} = \begin{bmatrix} 2 & 0 & 4 & 0 \\ 2 & 1 & 5 & 3 \\ -1 & 0 & 1 & 3 \\ 0 & 1 & 3 & -1 \\ -2 & -1 & 3 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & 2 & 3 & 3 \\ 0 & 0 & 3 & 3 \\ -1 & -2 & 1 & 3 \\ 0 & 0 & -2 & 2 \\ 2 & 1 & -1 & 1 \\ 1 & 0 & -1 & 1 \end{bmatrix}.$$

Hence

$$\begin{aligned} & V_c(AZB + CZ^T D) \\ &= M_\rho V_c(Z) \\ &= \begin{bmatrix} 2 & 0 & 4 & 0 \\ 2 & 1 & 5 & 3 \\ -1 & 0 & 1 & 3 \\ 0 & 1 & 3 & -1 \\ -2 & -1 & 3 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & 2 & 3 & 3 \\ 0 & 0 & 3 & 3 \\ -1 & -2 & 1 & 3 \\ 0 & 0 & -2 & 2 \\ 2 & 1 & -1 & 1 \\ 1 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} z_{11} \\ z_{21} \\ z_{12} \\ z_{22} \end{bmatrix} = \begin{bmatrix} 2z_{11} + 4z_{12} \\ 2z_{11} + z_{21} + 5z_{12} + 3z_{22} \\ -z_{11} + z_{12} + 3z_{22} \\ z_{21} + 3z_{12} - z_{22} \\ -2z_{11} - z_{21} + 3z_{12} + z_{22} \\ -z_{11} - z_{21} + z_{12} + z_{22} \\ z_{11} + 2z_{21} + 3z_{12} + 3z_{22} \\ 3z_{12} + 3z_{22} \\ -z_{11} - 2z_{21} + z_{12} + 3z_{22} \\ -2z_{12} + 2z_{22} \\ 2z_{11} + z_{21} - z_{12} + z_{22} \\ z_{11} - z_{12} + z_{22} \end{bmatrix}. \end{aligned} \tag{3.44}$$

A direct computation shows that

$$AZB + CZ^TD = \begin{bmatrix} 2z_{11} + 4z_{12} & z_{21} + 3z_{12} - z_{22} & z_{11} + 2z_{21} + 3z_{12} + 3z_{22} & -2z_{12} + 2z_{22} \\ 2z_{11} + z_{21} + 5z_{12} + 3z_{22} & -2z_{11} - z_{21} + 3z_{12} + z_{22} & 3z_{12} + 3z_{22} & 2z_{11} + z_{21} - z_{12} + z_{22} \\ -z_{11} + z_{12} + 3z_{22} & -z_{11} - z_{21} + z_{12} + z_{22} & -z_{11} - 2z_{21} + z_{12} + 3z_{22} & z_{11} - z_{12} + z_{22} \end{bmatrix}.$$

This verifies (3.44).

Mapping ρ defined by (3.34) consists of two types of terms. In fact, it can be used for any mapping with finite such terms. A particular case is when all the matrices involved are square matrices of same dimensions. In the following we give another numerical example.

Example 3.8. Given a set of matrices as:

$$A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & -1 \\ 1 & 2 \end{bmatrix},$$

$$C = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}.$$

Consider a mapping

$$L(Z) = AZ + ZB + CZ^TD + AZB. \tag{3.45}$$

Using (3.35) and Table 3.2, the matrix expression of L can be obtained as

$$\begin{aligned} M_L^c &= I_2 \otimes A + B^T \otimes I_2 + (D^T \otimes C)W_{[2]} + B^T \otimes A \\ &= \begin{bmatrix} 1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 2 & 0 \\ 0 & -1 & 0 & 2 \end{bmatrix} \\ &\quad + \begin{bmatrix} 1 & 0 & -2 & 0 \\ 1 & -1 & -2 & 2 \\ 2 & 0 & 1 & 0 \\ 2 & -2 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & -1 & 1 \\ -1 & -1 & 2 & 2 \\ 1 & -1 & -2 & 2 \end{bmatrix} \\ &= \begin{bmatrix} 2 & -1 & 2 & 1 \\ 0 & -1 & -2 & 4 \\ 0 & 0 & 5 & 3 \\ 3 & -2 & 5 & 4 \end{bmatrix}. \end{aligned}$$

Table 3.3 Matrix expression of some mappings

Mapping	Notation	ρ	M^c_ρ
Lyapunov mapping	L_A	$Z \mapsto AZ + ZA^T$	$I \otimes A + A \otimes I$
generalized Lyapunov mapping	L_{AB}	$Z \mapsto AZ + ZB$	$I \otimes A + B^T \otimes I$
symplectic mapping	S_A	$Z \mapsto AZ + Z^T A$	$I \otimes A + (A^T \otimes I)W$
adjoint mapping	ad_A	$Z \mapsto AZ - ZA$	$I \otimes A - A^T \otimes I$
conjugate mapping	C^A_j	$Z \mapsto AZA^{-1}$	$A^{-T} \otimes A$
cogradient mapping	C^A_g	$Z \mapsto AZA^T$	$A \otimes A$

For convenience in use, we list the matrix expressions of some useful matrix mappings in Table 3.3.

Using the above matrix expressions, some useful formulas can be deduced.

Proposition 3.7. *Let $A \in \mathcal{M}_{m\times n}$, $B \in \mathcal{M}_{p\times q}$. Then*

$$(I_p \otimes A)W_{[n,p]} = W_{[m,p]}(A \otimes I_p), \tag{3.46}$$

$$W_{[m,p]}(A \otimes B)W_{[q,n]} = (B \otimes A). \tag{3.47}$$

Proof. Assume $Z \in \mathcal{M}_{p\times n}$, and consider the matrix expression of the mapping $Z \mapsto AZ^T$, which can be obtained through the following two ways:

(1) $Z \mapsto Z^T \mapsto AZ^T$: Note that here we use swap matrix $W_{[n,p]}$ first, and then use $I_p \otimes A$. It follows that the matrix expression of the mapping is $(I_p \otimes A)W_{[n,p]}$.

(2) $Z \mapsto ZA^T \mapsto (ZA^T)^T = AZ^T$: Now we first use $A \otimes I_p$, and then use $W_{[m,p]}$. Hence, the same mapping can be expressed as $W_{[m,p]}(A \otimes I_p)$.

Since the matrix expression of a linear mapping is unique, (3.46) follows.

As for (3.47), let $Z \in \mathcal{M}_{q\times n}$, and consider the matrix expression of $Z \mapsto AZ^T B^T$, which can also be obtained through two ways:

(1) $Z \mapsto ZA^T \mapsto BZA^T \mapsto (BZA^T)^T$: It is realized by $W_{[m,p]}(I_m \otimes B)(A \otimes I_q)$, which equals to $W_{[m,p]}(A \otimes B)$.

(2) $Z \mapsto Z^T \mapsto AZ^T \mapsto AZ^T B^T$: It is realized alternatively by $(B \otimes I_m)(I_q \otimes A)W_{[n,q]}$, which equals to $(B \otimes A)W_{[n,q]}$.

Hence we have

$$W_{[m,p]}(A \otimes B) = (B \otimes A)W_{[n,q]}.$$

Right multiplying both sides of the above equation by $W_{[q,n]}$ yields (3.47).

□

In fact, it is not difficult to see that (3.46) can be deduced from (3.47). Both of them will be used in the sequel.

The following proposition shows the matrix expression of a general linear matrix mapping under row stacking forms. We give them a direct proof. In fact, they can also be obtained from the expression under column stacking forms, using Proposition 3.5.

Proposition 3.8. *The matrix expression of the mapping ρ defined by (3.34) under row stacking form is*

$$M^r = (A \otimes B^T) + (C \otimes D^T)W_{[n,p]}. \tag{3.48}$$

Proof. We first also provide the matrix expressions of four fundamental linear matrix mappings, which themselves are also useful.

Table 3.4 Matrix expressions under row stacking form

ρ	M^r_ρ
$Z \mapsto AZ$	$A \otimes I_p$
$Z \mapsto ZB$	$I_n \otimes B^T$
$Z \mapsto CZ^T$	$(C \otimes I_n)W_{[n,p]}$
$Z \mapsto Z^T D$	$(I_p \otimes D^T)W_{[n,p]}$

We prove them one by one. Using Lemma 2.3, we have

$$V_r(AZ) = AV_r(Z) = (A \otimes I_p)V_r(Z),$$

which proves the first equality. As for the second one, we have

$$V_r(ZB) = V_c(B^T Z^T) = (I_n \otimes B^T)V_c(Z^T) = (I_n \otimes B^T)V_r(Z).$$

Then, for the third one, we have

$$\begin{aligned} V_r(CZ^T) &= V_c(ZC^T) = (C \otimes I_n)V_c(Z) \\ &= (C \otimes I_n)W_{[n,p]}W_{[p,n]}V_c(Z) \\ &= (C \otimes I_n)W_{[n,p]}V_r(Z). \end{aligned}$$

Finally, we have

$$\begin{aligned} V_r(Z^T D) &= V_c(D^T Z) = (I_p \otimes D^T)V_c(Z) \\ &= (I_p \otimes D^T)W_{[n,p]}W_{[p,n]}V_c(Z) \\ &= (I_p \otimes D^T)W_{[n,p]}V_r(Z). \end{aligned}$$

This is the proof for the last equality. Using Table 3.4, we have, for the compounded mapping, that

$$\begin{aligned} &V_r(AZB + CZ^T D) \\ =&[(I_m \otimes B^T)(A \otimes I_p) + (I_m \otimes D^T)(C \otimes I_n)W_{[n,p]}]V_r(Z) \\ =&[A \otimes B^T + (C \otimes D^T)W_{[n,p]}]V_r(Z). \end{aligned} \tag{3.49}$$

(3.48) follows. □

3.4 Converting Matrix Expressions

From Chapter 1 one sees that in addition to the standard form, a matrix A can also be expressed by its row stacking form $V_r(A)$ or column stacking form $V_c(A)$. Sometimes we need to convert a matrix from one form to another. These conversions are particularly important as the matrix is a matrix expression of linear mapping between vector spaces.

Let $A \in \mathcal{M}_{m\times n}$. We define two mappings

$$\pi_s^r, \pi_s^c : \mathcal{M}_{m\times n} \to \mathcal{M}_{m\times ns^2}$$

as follows:

$$\pi_s^r(A) = A[I_n \otimes V_r^T(I_s)], \tag{3.50}$$

and

$$\pi_s^c(A) = A[I_n \otimes (\delta_s^1)^T \ I_n \otimes (\delta_s^2)^T \ \cdots \ I_n \otimes (\delta_s^s)^T], \tag{3.51}$$

where $(\delta_s^i)^T = \mathrm{Row}_i(I_s)$. Then we have

Proposition 3.9.

$$\pi_s^c(A) = \pi_s^r(A)W_{[s,n]}, \quad \forall\, A \in \mathcal{M}_{m\times n}. \tag{3.52}$$

Proof. Denote by $A_i = \mathrm{Col}_i(A)$. A straightforward computation shows that

$$\begin{aligned}\pi_s^r(A) = \Big[& \underbrace{\underbrace{A_1\ 0\ \cdots\ 0}_{s}\ \underbrace{0\ A_1\ 0\ \cdots\ 0}_{s}\ \cdots\ \underbrace{0\ \cdots\ 0\ A_1}_{s}}_{s} \\ & \vdots \\ & \underbrace{\underbrace{A_n\ 0\ \cdots\ 0}_{s}\ \underbrace{0\ A_n\ 0\ \cdots\ 0}_{s}\ \cdots\ \underbrace{0\ \cdots\ 0\ A_n}_{s}}_{s}\Big],\end{aligned} \tag{3.53}$$

and

$$\begin{aligned}\pi_s^c(A) = \Big[& \underbrace{\underbrace{A_1\ 0\ \cdots\ 0}_{s}\ \underbrace{A_2\ 0\ \cdots\ 0}_{s}\ \cdots\ \underbrace{A_n\ 0\ \cdots\ 0}_{s}}_{n} \\ & \vdots \\ & \underbrace{\underbrace{0\ \cdots\ A_1}_{s}\ \underbrace{0\ \cdots\ A_2}_{s}\ \cdots\ \underbrace{0\ \cdots\ 0\ A_n}_{s}}_{n}\Big].\end{aligned} \tag{3.54}$$

Denote by

$$H_{ij} = [\underbrace{0 \cdots 0}_{j-1} A_i \underbrace{0 \cdots 0}_{s-j}], \quad i = 1, \cdots, n;\ j = 1, \cdots, s.$$

Then we have the following alternative expressions of $\pi_s^r(A)$ and $\pi_s^c(A)$ as follows:

$$\pi_s^r(A) = \Big[H_{11}\ H_{12}\ \cdots\ H_{1s}\ \cdots\ H_{n1}\ H_{n2}\ \cdots\ H_{ns}\Big];$$
$$\pi_s^c(A) = \Big[H_{11}\ H_{21}\ \cdots\ H_{n1}\ \cdots\ H_{1s}\ H_{2s}\ \cdots\ H_{ns}\Big].$$

That is, in (3.53) $\{H_{ij}\}$ are arranged in the order of $\mathrm{id}(i, j; n, s)$, while in (3.54) $\{H_{ij}\}$ are arranged in the order of $\mathrm{id}(j, i; s, n)$. Then (3.52) follows from Proposition 2.9 immediately. □

Next, we convert the product of a constant matrix with an unknown matrix as a product of a coefficient matrix with unknown vectors, which is conventional in linear algebra.

Proposition 3.10. *Assume $A \in \mathcal{M}_{m\times n}$ and $X \in \mathcal{M}_{n\times s}$, then*

$$AX = \pi_s^r(A)V_r(X), \quad or \quad AX = \pi_s^r(A)W_{[s,n]}V_c(X). \tag{3.55}$$

Alternatively,

$$AX = \pi_s^c(A)V_c(X), \quad or \quad AX = \pi_s^c(A)W_{[n,s]}V_r(X). \tag{3.56}$$

Proof. Using (3.53), a straightforward computation shows that

$$\pi_s^r(A)V_r(X) = \left[\sum_{k=1}^n A_k x_{1k}, \sum_{k=1}^n A_k x_{2k}, \cdots, \sum_{k=1}^n A_k x_{nk}\right] = AX,$$

where $A_k = \mathrm{Col}_k(A)$. This is the first equality in (3.55).

Using Proposition 3.5 to the first equality of (3.55), the second equality is obtained. The proof of (3.56) is similar. □

Combining the above formulas with some results about the linear mappings of matrices, some useful formulas can be obtained.

Proposition 3.11. *Assume $X \in \mathcal{M}_{m\times n}$ and $A \in \mathcal{M}_{n\times s}$, then*

$$XA = (I_m \otimes V_r^T(I_s))W_{[s,m]}A^TV_c(X). \tag{3.57}$$

***Proof*.** Using (3.56) and Table 3.2, we have

$$\begin{aligned} XA &= I_m(XA) = \pi_s^c(I_m)V_c(XA) \\ &= \pi_s^c(I_m)(A^T \otimes I_m)V_c(X). \end{aligned} \tag{3.58}$$

Using (3.50) and (3.52), we have

$$\pi_s^c(I_m) = \big(I_m \otimes V_r^T(I_s)\big) \otimes W_{[s,m]}.$$

Using Proposition 2.4, we have

$$(A^T \otimes I_m)V_c(X) = A^T V_c(X).$$

Plugging them into (3.58) yields (3.57). □

The following example shows how to convert a higher-order matrix product form of an unknown matrix into the standard form as coefficient matrix times power of unknowns.

Example 3.9. Let $A, B, C, Z \in \mathcal{M}_n$. Consider a mapping $p(Z) = Z \mapsto AZBZC$. We intend to express $p(Z)$ into a "quadratic form" of Z.

Using (3.35) and (3.57), we have

$$\begin{aligned} V_c(p(Z)) &= (C^T \otimes A)V_c(ZBZ) \\ &= (C^T \otimes A) \ltimes (BZ)^T \ltimes V_c(Z). \end{aligned} \tag{3.59}$$

Applying (3.47) to $(BZ)^T$ yields

$$\begin{aligned} (BZ)^T = Z^T B^T &= \big(I_n \otimes V_r^T(I_n)\big) W_{[n]} \ltimes B \ltimes V_c(Z^T) \\ &= \big(I_n \otimes V_r^T(I_n)\big) W_{[n]}(B \otimes I_n)W_{[n]}V_r(Z^T) \\ &= \big(I_n \otimes V_r^T(I_n)\big) (I_n \otimes B)V_c(Z). \end{aligned}$$

Finally, we have

$$V_c(p(Z)) = (C^T \otimes A) \big(I_n \otimes V_r^T(I_n)\big) (I_n \otimes B)V_c^2(Z). \tag{3.60}$$

The polynomial expression of a matrix product with higher order of unknown matrix, as obtained in the previous example, is very convenient in some investigations. In the following we give an example from control system.

Example 3.10. Consider a linear control system

$$\dot{x} = Ax + Bu, \quad x \in \mathbb{R}^n,\ u \in \mathbb{R}^m. \tag{3.61}$$

When a state feedback

$$u = Fx + v,$$

with F to be designed, is used for solving the decoupling problem, we have to calculate the decoupling matrix (Isidori, 1995). To this end, we have to calculate

$$(A+BF)^k. \tag{3.62}$$

$(A+BF)^k$ can be considered as a matrix product form of kth order of F. We intend to express it as a polynomial of F. First of all, using Proposition 3.10, we express $A+BF$ as

$$A+BF = A+Ef := P_0^1 + P_1^1 f,$$

where $f = V_r(F)$, and $E = \pi_m^r(B)$, which can be constructed by (3.50). Hence we have

$$\begin{aligned}
&(A+E\ltimes f)^2 = (A+E\ltimes f)(A+E\ltimes f)\\
&= A^2 + [A\ltimes E + E\ltimes (I_{mn}n\otimes A)]\ltimes f + E\ltimes (I_{mn}\otimes E)\ltimes f^2\\
&:= P_0^2 + P_1^2 f + P_2^2 f^2.
\end{aligned}$$

It is clear that we can have the following polynomial form

$$(A+BF)^k = A^k + P_1^k \ltimes f + P_2^k \ltimes f^2 + \cdots + P_k^k \ltimes f^k. \tag{3.63}$$

Hence, we have

$$\begin{aligned}
(A+BF)^{k+1} &= (A+E\ltimes f)(A^k + P_1^k\ltimes f + P_2^k\ltimes f^2 + \cdots + P_k^k \ltimes f^k)\\
&= A^{k+1} + [A\ltimes P_1^k + E\ltimes (I_{mn}\otimes A^k)]\ltimes f\\
&\quad +[A\ltimes P_2^k + E\ltimes (I_n\otimes P_1^k)]\ltimes f^2\\
&\quad +\cdots + [A\ltimes P_k^k + E\ltimes (I_n\otimes P_k^k)]\ltimes f^k\\
&\quad +E(I_{mn}\otimes P_k^k)f^{k+1}.
\end{aligned}$$

Denote $P_0^k := A^k$. Then to calculate the coefficients in (3.63) we have a recursive formula as

$$\begin{cases}
P_0^1 = A, \quad P_1^1 = E,\\
\begin{cases}
P_0^{k+1} = A^{k+1},\\
P_i^{k+1} = A\ltimes P_i^k + E\ltimes (I_{mn}\otimes P_{i-1}^k), \quad i=1,2,\cdots,k,\\
P_{k+1}^{k+1} = E(I_{mn}\otimes P_k^k), \quad k=1,2,3,\cdots.
\end{cases}
\end{cases} \tag{3.64}$$

Note that $P_i^k \in \mathcal{M}_{n\times n^{i+1}}$, which are easily computable. In fact, all the entries of the matrices in (3.62) are also functions of $\{f_{ij}\}$. The problem is, it is very difficult to calculate them. But the formulas (3.63) and (3.64) provide a convenient way to calculate the functions.

Finally, we consider how to convert a matrix into its stacking form and vice versa. The problem seems stupid because everybody knows how to do this. But what we are interested is the converting formulas, which are very useful in theoretical expressions and/or deductions.

Proposition 3.12. *Assume $A \in \mathcal{M}_{m\times n}$, then*

$$V_r(A) = A \ltimes V_r(I_n), \tag{3.65}$$

$$V_c(A) = W_{[m,n]} \ltimes A \ltimes V_c(I_n). \tag{3.66}$$

Proof. A straightforward computation yields (3.65). Applying (1.53) to (3.65) yields (3.66). □

Conversely, we can reconstruct A from its row or column stacking form.

Proposition 3.13. *Assume $A \in \mathcal{M}_{m\times n}$, then*

$$A = [I_m \otimes V_r^T(I_n)] \ltimes V_r(A) = [I_m \otimes V_r^T(I_n)] \ltimes W_{[n,m]} \ltimes V_c(A). \tag{3.67}$$

Proof. Replacing A and X in (3.55) by I_m and A respectively, then (3.50) yields the first equality of (3.67). Applying (1.53) to the first equality yields the second equality. □

We give a numerical example to depict this.

Example 3.11. Given a matrix $A \in \mathcal{M}_{3\times 2}$ as

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}.$$

Now $m = 3$, $n = 2$. Using Lemma 2.4, we have

$$V_r(I_n) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

$$A \ltimes V_r(I_n) = \begin{bmatrix} \begin{bmatrix} a_{11} \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ a_{12} \end{bmatrix} \\ \begin{bmatrix} a_{21} \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ a_{22} \end{bmatrix} \\ \begin{bmatrix} a_{31} \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ a_{32} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{12} \\ a_{21} \\ a_{22} \\ a_{31} \\ a_{31} \end{bmatrix} = V_r(A).$$

Using (3.66), we have

$$W_{[3,2]} \ltimes A = \begin{bmatrix} a_{11} & 0 & a_{12} & 0 \\ a_{21} & 0 & a_{22} & 0 \\ a_{31} & 0 & a_{32} & 0 \\ 0 & a_{11} & 0 & a_{12} \\ 0 & a_{21} & 0 & a_{22} \\ 0 & a_{31} & 0 & a_{32} \end{bmatrix}.$$

(We refer to Example 1.10 for the numerical forms of $W_{[3,2]}$ and $W_{[2,3]}$.)

Then we have

$$W_{[3,2]} \ltimes A \ltimes V_r(I_n) = \begin{bmatrix} a_{11} & 0 & a_{12} & 0 \\ a_{21} & 0 & a_{22} & 0 \\ a_{31} & 0 & a_{32} & 0 \\ 0 & a_{11} & 0 & a_{12} \\ 0 & a_{21} & 0 & a_{22} \\ 0 & a_{31} & 0 & a_{32} \end{bmatrix} \ltimes \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a_{12} \\ a_{22} \\ a_{32} \end{bmatrix} = V_c(A).$$

Next, we verify (3.67). For the first part we have

$$\begin{aligned} [I_3 \otimes V_r^T(I_2)] \ltimes V_r(A) &= [I_3 \otimes (1\ 0\ 0\ 1)] \ltimes V_r(A) \\ &= \begin{bmatrix} 1&0&0&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&0&0&1 \end{bmatrix} \ltimes \begin{bmatrix} a_{11} \\ a_{12} \\ a_{21} \\ a_{22} \\ a_{31} \\ a_{31} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} = A. \end{aligned}$$

To verify the second part of (3.67), we have

$$\begin{aligned} &[I_3 \otimes (1\ 0\ 0\ 1)] \ltimes W_{[2,3]} \\ &= \begin{bmatrix} 1&0&0&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&0&0&1 \end{bmatrix} \ltimes \begin{bmatrix} 1&0&0&0&0&0 \\ 0&0&0&1&0&0 \\ 0&1&0&0&0&0 \\ 0&0&0&0&1&0 \\ 0&0&1&0&0&0 \\ 0&0&0&0&0&1 \end{bmatrix} \\ &= \begin{bmatrix} 1&0&0&0&0&0&0&1&0&0&0&0 \\ 0&0&1&0&0&0&0&0&0&1&0&0 \\ 0&0&0&0&1&0&0&0&0&0&0&1 \end{bmatrix}. \end{aligned}$$

Hence

$$
\begin{aligned}
&[I_3 \otimes (1\ 0\ 0\ 1)] \ltimes W_{[2,3]} \ltimes V_c(A) \\
&= \begin{bmatrix} 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0 \\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0 \\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1 \end{bmatrix} \ltimes \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a_{12} \\ a_{22} \\ a_{32} \end{bmatrix} = A.
\end{aligned}
$$

3.5 Two Applications

3.5.1 *General Linear Group and Its Algebra*

We refer to Boothby (1986) for the basic concepts and properties of Lie group and Lie algebra. A reader who is not familiar with the basic concepts in differential geometry may skip this subsection.

Definition 3.4. A Lie group is an analytic manifold and a group. Moreover, the group operations (product and inverse) are analytic with respect to the manifold structure.

Consider the set

$$G := \{ A \in \mathcal{M}_n \mid \det(A) \neq 0 \}.$$

G can be naturally imbedded into $\mathbb{R}^{n^2}$. Hence as an open subset of $\mathbb{R}^{n^2}$, G is an n^2-dimensional analytic manifold. Moreover, $(G, \cdot)$ is a group, where "$\cdot$" is the conventional matrix product of matrices in $\mathcal{M}_n$. Eventually, it is easy to check that $(G, \cdot)$, with the differential structure inherited from $\mathbb{R}^{n^2}$, is a Lie group, called the general linear group and denoted by $GL(n, \mathbb{R})$.

Let M be a k-dimensional analytic manifold, $f(x)$, $x \in M$ is called a vector field if at each $x \in M$ it is a k-dimensional vector. The set of (analytic) vector fields on M is denoted by $V^\omega(M)$, which is a vector space. The Lie bracket on $V^\omega(M)$ is defined as (within each coordinate chart)

$$[f(x), g(x)] = \frac{\partial g(x)}{\partial x} f(x) - \frac{\partial f(x)}{\partial x} g(x), \quad f(x), g(x) \in C^\omega(M), \tag{3.68}$$

where $\frac{\partial g(x)}{\partial x}$ is the Jacobian matrix of $g(x)$. That is,

$$\frac{\partial g(x)}{\partial x} = J_{g(x)} := \begin{bmatrix} \frac{\partial g_1(x)}{\partial x_1} & \cdots & \frac{\partial g_1(x)}{\partial x_k} \\ \vdots & & \\ \frac{\partial g_k(x)}{\partial x_1} & \cdots & \frac{\partial g_k(x)}{\partial x_k} \end{bmatrix}.$$

Then $V^{\omega}(M)$ with this Lie bracket becomes a Lie algebra.

Let $\psi : M \to M$ be an analytic diffeomorphism (bijective mapping). Then ψ can deduce a mapping $\psi_* : V^{\omega}(M) \to V^{\omega}(M)$ as

$$\psi_*(f(X)) := J_{\psi} f(\phi^{-1}(X)), \quad f(X) \in V^{\omega}(M). \tag{3.69}$$

Now assume G is a Lie group and $X \in G$, then we define a mapping, called the left-displacement of X, defined as

$$\phi_X^L : g \to Xg, \quad g \in G. \tag{3.70}$$

Then it is easy to see that $\phi_X^L : G \to G$ is an analytic diffeomorphism.

A vector field $f(X) \in V^{\omega}(G)$ is called a left-invariant vector field, if

$$(\phi_X^L)_*(f(P)) = f(\phi_X^L(P)) = f(XP), \quad \forall\, X, P.$$

It is easy to prove that the set of left-invariant vector fields forms a Lie-subalgebra of $V^{\omega}(G)$. This Lie algebra is called the Lie algebra of the Lie group G.

In this subsection we intend to prove the following important fact:

Theorem 3.1. *The Lie algebra of the general linear group $GL(n, \mathbb{R})$ is the general linear algebra $gl(n, \mathbb{R})$.*

Proof. For notational compactness, denote by $G := GL(n, \mathbb{R})$ and $V(G) := V^{\omega}(GL(n, \mathbb{R}))$. Let the coordinates of $X \in G$ be $V_c(X)$. Then a vector field $F(X) \in V(G)$ is also considered as an $n \times n$ matrix. In conventional way, the vector fields should be expressed as $V_c(F(X))$.

Now assume $F(X) \in V(G)$ is a left invariant vector field, and $F(I_n) = A$. Fix an $X \in G$, we define the left displacement as

$$\phi_X^L : P \mapsto XP, \quad \forall\, P \in GL(n, \mathbb{R}).$$

Since $X \in \mathbb{R}^{n \times n}$, according to Table 3.2 we know that the matrix expression of ϕ_X^L under column stacking form is $I_n \otimes X$. Hence

$$V_c\left(\phi_X^L(P)\right) = (I_n \otimes X)p,$$

where $p = V_c(P)$. Denote by $a = V_c(A)$. Now consider the left displacement as a diffeomorphism on $GL(n, \mathbb{R})$, it can drive the tangent vector A at $I \in G$ to the point $X \in G$, precisely, it is

$$\frac{\partial \phi_X^L(P)}{\partial p} a = (I_n \otimes X)a.$$

Now since $F(X)$ is left invariant, when A is driven to X the vector is exactly $F(X)$. Denoting $x = V_c(X)$, we, therefore, have

$$V_c(F(X)) = J_{\phi_X^L} a = \frac{\partial (I_n \otimes X)x}{\partial x} a = (I_n \otimes X)a = V_c(XA).$$

Again, the last equality comes from Table 3.2. We conclude that the left invariant vector field $F(X)$, when expressed as a matrix, is

$$F(X) = XA. \tag{3.71}$$

Now let $F(X)$ and $W(X)$ be two left invariant vector fields with $F(I_n) = A$ and $W(I_n) = B$. Using (3.71), we have that the matrix expressions of $F(X)$ and $G(X)$ are $F(X) = XA$ and $W(X) = XB$ respectively. Back to conventional vector form, we have

$$F(X) = (A^T \otimes I_n)x, \quad W(x) = (B^T \otimes I_n)x.$$

Using formula (3.68), we have

$$\begin{aligned}
[F(X), W(X)] &= (B^T \otimes I_n)(A^T \otimes I_n)x - (A^T \otimes I_n)(B^T \otimes I_n)x \\
&= ((B^T \otimes I_n)(A^T \otimes I_n) - (A^T \otimes I_n)(B^T \otimes I_n))x \\
&= (B^T A^T \otimes I_n - A^T B^T \otimes I_n)x = ((AB - BA) \otimes I_n)x.
\end{aligned}$$

Checking Table 3.2 again, we know that the matrix expression of $U(X) := [F(X), W(X)]$ is $(AB - BA)X$, which is a left invariant vector field with $U(I_n) = AB - BA$.

We conclude that the generalized linear algebra $gl(n, \mathbb{R})$, as a Lie algebra, is the Lie algebra of the generalized linear group $GL(n, \mathbb{R})$, as a Lie group. □

Definition 3.5. Let G be a Lie group, and $H < G$ a subgroup. H is called a Lie subgroup of G, if H is a regular submanifold of G.

Checking a regular submanifold is in general a difficult job. The following result is very convenient in use.

Theorem 3.2 (Boothby, 1986). *Let G be a Lie group. $H < G$ is a subgroup. If H is closed under the topology of G, H is a Lie subgroup.*

Proposition 3.14.

(1) Define

$$O(n, \mathbb{R}) = \left\{ A \in GL(n, \mathbb{R}) \mid A^T = A^{-1} \right\}.$$

Then $O(n, \mathbb{R})$ is a Lie subgroup of $GL(n, \mathbb{R})$, called the orthogonal group.

(2) Define

$$SO(n, \mathbb{R}) = \{ A \in O(n, \mathbb{R}) \mid \det(A) = 1 \}.$$

Then $SO(n, \mathbb{R})$ is a Lie subgroup of $O(n, \mathbb{R})$, called the special orthogonal group.

(3) Define

$$SP(2n,\mathbb{R}) = \left\{ A \;\middle|\; \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix} A + A^T \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix} = 0 \right\}.$$

Then $SP(2n,\mathbb{R})$ is a Lie subgroup of $GL(2n,\mathbb{R})$, called the symplectic group.

(4) $o(n,\mathbb{R})$ is the Lie algebra of $O(n,\mathbb{R})$.

(5) $o(n,\mathbb{R})$ is also the Lie algebra of $SO(n,\mathbb{R})$.

(6) $sp(2n,\mathbb{R})$ is the Lie algebra of $SP(2n,\mathbb{R})$.

We leave the proof of Proposition 3.14 to the reader.

3.5.2 *Hautus and Sylvester Equations*

As another application, we consider Hautus equation. Let $A_i \in \mathcal{M}_{n\times m}$, $i = 1,\cdots,k$; $q_i(t)$, $i = 1,\cdots,k$ be some polynomials, $S \in \mathcal{M}_{p\times p}$, $R \in \mathcal{M}_{n\times p}$, $X \in \mathcal{M}_{m\times p}$. The following equation about unknown matrix X is called the Hautus equation:

$$A_1Xq_1(S) + \cdots + A_kXq_k(S) = R. \tag{3.72}$$

To understand the importance of Hautus equation, we consider some of its special cases.

Let $A \in \mathcal{M}_{n\times n}$, $S \in \mathcal{M}_{p\times p}$, and $R \in \mathcal{M}_{n\times p}$. The following equation is called the Sylvester equation, which is very useful in control theory.

$$AX - XS = R. \tag{3.73}$$

It is easy to see that (3.73) is obtained from (3.72) by setting $A_1 = A$, $A_2 = I$, $q_1(t) = 1$, $q_2(t) = -t$.

Assume A and S are square matrices, and B, P, C, Q are matrices with proper dimensions. The following equation is called the regulation equation, which is proposed for investigating regulation problem of control systems.

$$\begin{cases} \Pi S = A\Pi + B\Gamma + P, \\ 0 = C\Pi + Q. \end{cases} \tag{3.74}$$

(3.74) can be converted into a Hautus equation as

$$A_1X - A_2XS = R,$$

where

$$A_1 = \begin{bmatrix} A & B \\ C & 0 \end{bmatrix}; A_2 = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}; R = \begin{bmatrix} -P \\ -Q \end{bmatrix}; X = \begin{bmatrix} \Pi \\ \Gamma \end{bmatrix}.$$

Theorem 3.3. *The Hautus equation has solution for each R, if and only if, the n rows of matrix*

$$A(\lambda) = A_1 q_1(\lambda) + \cdots + A_k q_k(\lambda) \tag{3.75}$$

are linearly independent for each eigenvalue of S. Moreover, if $n = m$, the solution is unique.

Proof. Assume $T \in O(p, \mathbb{R})$, and let

$$\tilde{X} = XT, \quad \tilde{S} = T^{-1}ST, \quad \tilde{R} = R.$$

Then equation (3.72) can be converted to

$$A_1 \tilde{X} q_1(\tilde{S}) + \cdots + A_k \tilde{X} q_k(\tilde{S}) = \tilde{R}. \tag{3.76}$$

It is clear that the existence of solution X with respect to each R is equivalent to the existence of $\tilde{X}$ with respect to $\tilde{R}$. Moreover, $X = \tilde{X}T^{-1}$. Hence, without loss of generality, we can assume S is in its Jordan canonical form.

Using Proposition 3.6, (3.72) can be converted to

$$\left[q_1(S^T) \otimes A_1 + \cdots + q_k(S^T) \otimes A_k\right] x = r, \tag{3.77}$$

where $x = V_c(X)$, $r = V_c(R)$. Since S has Jordan canonical form

$$S = \begin{bmatrix} \lambda_1 & * & \cdots & * \\ 0 & \lambda_2 & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & \lambda_p \end{bmatrix},$$

then (3.77) can be expressed as $Ex = r$, where

$$E = \begin{bmatrix} Q(\lambda_1) & 0 & \cdots & 0 \\ * & Q(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ * & \cdots & * & Q(\lambda_p) \end{bmatrix},$$

where

$$Q(t) = q_1(t)A_1 + \cdots + q_k(t)A_k.$$

The conclusion follows. □

Corollary 3.1. *Sylvester equation (3.73) has solution with respect to each R, if and only if, A and S have no common eigenvalue. Moreover, in this case the solution is unique.*

Proof. For Sylvester equation, we have

$$A(\lambda) = A - I\lambda.$$

Then it is clear that for each $\lambda \in \sigma(S)$, $A(\lambda)$ is nonsingular, if and only if, A and S have no common eigenvalue. □

Remark 3.1.

(1) Lyapunov mapping is closely related to the stability of linear systems. There are many papers investigating it. For instance, Its many properties have been discussed in Cheng (2001). Its applications to switched linear systems have been discussed in Cheng *et al.* (2003). The following is an interesting problem: Denote by S the set of $n \times n$ symmetric matrices and K the set of $n \times n$ skew-symmetric matrices. Both of them are subspaces of $\mathcal{M}_{n\times n}$. In addition, it is easy to prove that they are invariant subspaces with respect to the Lyapunov mapping L_A. Hence, we can restrict L_A on these two invariant subspaces respectively, and denote them as L_A^S, L_A^K. It was proved in Cheng (2001) that

$$\|L_A\| = \max\{\|L_A^S\|, \|L_A^K\|\}.$$

Based on this we propose a conjecture (Cheng *et al.*, 2009)

$$\|L_A\| = \|L_A^S\|. \tag{3.78}$$

It is still an open conjecture.

(2) Hautus equation and Sylvester equation have many applications in control theory, particularly, in the investigation of output regulation problem. More discussion about Hautus equation and Sylvester equation can be found in Knobloch *et al.* (1993). We refer the reader, who is interested in output regulation, to Huang (2004).

Exercises

3.1 Prove $(\mathbb{R}^3, \bowtie)$ is a Lie algebra.

3.2 Prove $(\mathcal{M}_n, [\cdot,\cdot])$ is a Lie algebra, where the bracket is defined in (3.16).

3.3 Prove Proposition 3.2.

3.4 Let V be a finite-dimensional vector space over $\mathbb{R}$. Denote by $End(V)$ the set of linear mappings $\phi : V \to V$. (In general this set is called the endomorphism of V.) Let $\mathcal{G}$ be a Lie algebra. A mapping $\pi : \mathcal{G} \to End(V)$ is called a representation of $\mathcal{G}$ in V, if π satisfies (Varadarajan, 1984) (i) π is linear, (ii) $\pi([X,Y]) = \pi(X)\pi(Y) - \pi(Y)\pi(X)$, $X, Y \in \mathcal{G}$. Prove that the adjoint representation ad_X is a representation.

3.5 Consider $o(3,\mathbb{R})$.

(i) Find a basis of $o(3,\mathbb{R})$.

(ii) Find the structure matrix of $o(3,\mathbb{R})$ with respect to the basis you obtained.

(iii) Let

$$X = \begin{bmatrix} 0 & 2 & 1 \\ -2 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix} \in o(3, \mathbb{R}).$$

Find the matrix expression of ad_X with respect to your basis.

3.6 Consider $sp(4, \mathbb{R})$.

(i) Find a basis of $sp(4, \mathbb{R})$.

(ii) Find the structure matrix of $sp(4, \mathbb{R})$ with respect to the basis you obtained.

(iii) Let

$$X = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 1 & 1 & 1 & 0 \\ 0 & -1 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

Check whether $X \in sp(4, \mathbb{R})$? If "yes", find the matrix expression of ad_X with respect to the basis you proposed.

3.7 Assume $(V, [\cdot, \cdot])$ is a Lie algebra and $e = \{e_1, \cdots, e_n\}$ is a basis of V. Moreover, $X \in V$ and the adjoint representation ad_X has matrix expression as $M_e \in \mathcal{M}$ with respect to a basis e. Let $\alpha = \{\alpha_1, \cdots, \alpha_n\}$ be another basis of V and

$$\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = T \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}.$$

Find the matrix expression of ad_X with respect to the basis α.

3.8 Prove that

$$\mathcal{G}_N := \{ X \in gl(n, \mathbb{R}) \mid L_N(X) = 0 \},$$

where L_N is defined in (3.30), is a Lie subalgebra of $gl(n, \mathbb{R})$.

3.9 Prove Proposition 3.4.

3.10 Let $A \in \mathcal{M}_{m\times n}$, $X \in \mathcal{M}_{n\times q}$, $Y \in \mathcal{M}_{p\times m}$. Prove the following two formulas:

$$V_c(AX) = W_{[m,q]} A W_{[q,n]} V_c(X). \tag{3.79}$$

$$V_r(YA) = W_{[n,p]} A^T W_{[p,m]} V_r(Y). \tag{3.80}$$

3.11 (i) Consider $\mathcal{M}_n$ and define a product as

$$\langle A, B \rangle = BA - AB.$$

Prove that $(\mathcal{M}_n, \langle\cdot,\cdot\rangle)$ is a Lie algebra.

(ii) Assume the Lie algebra of a Lie group is generalized by right invariant vector fields. Then prove that the Lie algebra of $GL(n,\mathbb{R})$ is $(\mathcal{M}_n, \langle\cdot,\cdot\rangle)$.

(iii) Prove Proposition 3.14. (Ignore this if you are not familiar with differential geometry.)

3.12 Given

$$A = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 2 \\ 2 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 2 & -1 \\ 1 & 0 & 3 \\ 2 & -1 & 2 \end{bmatrix},$$

and let $Y = AXB + CX^T$, where $X \in \mathcal{M}_{3\times 3}$.

(i) Find M^r such that $V_r(Y) = M^r V_r(X)$.

(ii) Find M^c such that $V_c(Y) = M^c V_c(X)$.

3.13 Assume $M, N \in \mathcal{M}_{m\times n}$, $Z \in \mathbb{R}^n$. Prove that

$$MZNZ = M(I_n \otimes N)Z^2. \tag{3.81}$$

3.14 Assume $A \in \mathcal{M}_{m\times n}$, $B \in \mathcal{M}_{p\times q}$, $C \in \mathcal{M}_{q\times p}$, $D \in \mathcal{M}_{n\times s}$, and $Z \in \mathcal{M}_{n\times p}$.

$$Y = AZBCZ^T D.$$

(i) Find $M^r \in \mathcal{M}_{m\times sn^2p^2}$, such that

$$V_r(Y) = M^r V_r^2(X).$$

(ii) Find $M^c \in \mathcal{M}_{m\times sn^2p^2}$, such that

$$V_c(Y) = M^c V_c^2(X).$$

(Hint: Use (3.81).)

3.15 (i) Solve X, where

$$X^2 = (1\ -2\ 1\ -2\ 4\ 2\ 1\ -2\ 1)^T.$$

(ii) Solve X, where

$$X^2 + MXNX = (2.5\ -0.5\ -3.5\ -0.25\ 0.75\ 0.75\ -4\ 1.5\ 6)^T,$$

and

$$M = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 1 \\ -1 & 0 & 1 \end{bmatrix}, \quad N = \begin{bmatrix} 2 & -1 & 0 \\ 1 & 0 & 2 \\ -1 & 1 & 2 \end{bmatrix}.$$

(iii) Solve $X \in \mathbb{R}^3$, $Y \in \mathbb{R}^2$, and $X \in \mathbb{R}^3$, where

$$XYZ = (-1\ 1\ 0\ -2\ 2\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ -1\ 0\ 2\ -2\ 0)^T.$$

Is the solution unique? If not, what is the general form?

3.16 Consider the Riccati equation

$$A^T X + XA + Q + XRX = 0,$$

where $A, Q, R \in \mathcal{M}_n$. Express it into a polynomial form as

$$C_0 x^2 + C_1 x + C_2 = 0,$$

where $x = V_c(X)$.

3.17 Given a matrix equation

$$AXBX^T CX = XDX^T,$$

where $A, B, C, D, X \in \mathcal{M}_n$. Express it into a polynomial form as

$$Fx^3 - Gx^2 = (Fx - G)x^2 = 0,$$

where (i) $x = V_r(X)$, (ii) $x = V_c(X)$.

3.18 Given

$$X = \begin{bmatrix} 1 & 0 \\ -3 & 4 \\ 2 & -1 \end{bmatrix}.$$

(i) Find $V_r(X)$ and $V_c(X)$.

(ii) Find $\Pi_r \in \mathcal{M}_{3\times 12}$ such that

$$X = \Pi_r V_r(X).$$

Check your result.

(ii) Find $\Pi_c \in \mathcal{M}_{3\times 12}$ such that

$$X = \Pi_c V_c(X).$$

Check your result.

3.19 Recall Example 3.10. Let $f := V_c(F)$. Find the recursive formula, corresponding to (3.64).

3.20 Given

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \\ -1 & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -1 & 1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 1 & 1 \\ 0 & -1 \\ 2 & 3 \end{bmatrix};$$

$$S = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & -1 \\ 0 & 2 \\ -1 & 3 \end{bmatrix};$$

$$q_1(t) = t, \quad q_2(t) = 1 - t, \quad q_3(t) = t^2.$$

Check if the corresponding Hautus equation has solution? Find a solution if it exists.

3.21 Consider the Lyapunov mapping $L_A : \mathcal{M}_n \to \mathcal{M}_n$ defined as $X \mapsto AX + XA^T$. Denote by S and K the sets of symmetric matrices and skew-symmetric matrices respectively.

(i) Show that both S and K are linear subspace of $\mathcal{M}_n$, $S \perp K$, and

$$\mathcal{M}_n = S + K := \{s + k \mid s \in S \text{ and } k \in K\}.$$

(ii) Both S and K are invariant subspace of L_A, which are denoted by L_A^S and L_A^K respectively.

(iii) Prove

$$\|L_A\| = \max\{\|L_A^S\|, \|L_A^K\|\}.$$

3.22 Show that

$$G = \{A \in \mathcal{M}_n \mid \det(A) \neq 0\}$$

is an open subset of $\mathbb{R}^{n^2}$ under conventional imbedding.

Chapter 4

Right and General Semi-Tensor Products

In the previous two chapters the left STP has been discussed. A natural question is: Can we define the right STP? If yes, what is its relationship with left STP? Secondly, so far the left STP has only been defined over two factor matrices satisfying the multiplier dimension case. Another natural question is: Can we define the STP for two arbitrary matrices? This chapter is devoted to the right STP and both left and right STPs over arbitrary matrices. The right STP and the generalized STP were firstly proposed by Cheng and Qi (2007) or Cheng (2007).

4.1 Right STP

First, we recall the Kronecker product of matrices. Observe Proposition 1.4, which says that assume $A \in \mathcal{M}_{m\times n}$ and $B \in \mathcal{M}_{p\times q}$, then

$$A \otimes B = (A \otimes I_p)(I_n \otimes B). \tag{4.1}$$

According to Proposition 2.4, the left semi-tensor product has an alternative definition as

$$A \ltimes B = \begin{cases} (A \otimes I_t)B, & A \prec_t B, \\ A(B \otimes I_t), & A \succ_t B. \end{cases} \tag{4.2}$$

Compared to (4.1), (4.2) seems to be obtained by "only making a left identity matrix matching". This is also a reason for calling this new matrix product the semi-tensor product.

Now assume two matrices satisfy the requirement of multiplier dimension, then it is obvious that we can also make a right identity matrix matching. Hence the following definition is very natural.

Definition 4.1. Given two matrices A and B. Assume either $A \prec_t B$ or

$A \succ_t B$. Then the right STP of A and B, denoted by $A \rtimes B$, is defined as

$$A \rtimes B = \begin{cases} (I_t \otimes A)B, & A \prec_t B, \\ A(I_t \otimes B), & A \succ_t B. \end{cases} \tag{4.3}$$

Many properties of right semi-tensor product are paralleled to its left counterpart. We state the following properties, and leave the proves to the reader.

Proposition 4.1. *The right STP has the following properties.*

(1) (Associative Law)

$$(A \rtimes B) \rtimes C = A \rtimes (B \rtimes C). \tag{4.4}$$

(2) (Distributive Law)

$$(A + B) \rtimes C = A \rtimes C + B \rtimes C, \quad C \rtimes (A + B) = C \rtimes A + C \rtimes B. \tag{4.5}$$

(3) Let X and Y be two column vectors. Then

$$X \rtimes Y = Y \otimes X. \tag{4.6}$$

Let X and Y be two row vectors. Then

$$X \rtimes Y = X \otimes Y. \tag{4.7}$$

(4) Assume $A \rtimes B$ is well defined. Then

$$(A \rtimes B)^T = B^T \rtimes A^T. \tag{4.8}$$

(5) If $M \in \mathcal{M}_{m \times pn}$, then

$$M \rtimes I_n = M; \tag{4.9}$$

If $M \in \mathcal{M}_{m \times n}$, then

$$M \rtimes I_{pn} = I_p \otimes M; \tag{4.10}$$

If $M \in \mathcal{M}_{pm \times n}$, then

$$I_p \rtimes M = M; \tag{4.11}$$

If $M \in \mathcal{M}_{m \times n}$, then

$$I_{pm} \rtimes M = I_p \otimes M. \tag{4.12}$$

Proposition 4.2. *Assume A and B are two square matrices with proper dimensions such that $A \rtimes B$ is well defined. Then*

(1) $A \rtimes B$ and $B \rtimes A$ have the same characteristic functions.

(2)

$$\mathrm{tr}(A \rtimes B) = \mathrm{tr}(B \rtimes A). \tag{4.13}$$

(3) If both A and B are orthogonal matrices (upper triangular matrices, lower triangular matrices, or diagonal matrices), then so is $A \rtimes B$.

(4) If either A or B is (or both are) invertible , then $A \rtimes B \sim B \rtimes A$ (here "$\sim$" means "be similar to").

(5) If both A and B are invertible, then

$$(A \rtimes B)^{-1} = B^{-1} \rtimes A^{-1}. \tag{4.14}$$

(6) If $A \prec_t B$, then

$$\det(A \rtimes B) = [\det(A)]^t \det(B); \tag{4.15}$$

If $A \succ_t B$, then

$$\det(A \rtimes B) = \det(A)[\det(B)]^t. \tag{4.16}$$

Recall that in Chapter 2 we define the left semi-tensor product in two steps: First, we define the left STP of two vectors in Definition 2.2, and then use it to define the STP for two matrices as in Definition 2.3. Can the right STP also been defined in this way? We first consider the vector case.

Definition 4.2. Let $X = [x_1, \cdots, x_s]$ be a row vector and $Y = [y_1, \cdots, y_t]^T$ be a column vector.

Case 1: If $s = t \times n$, where $n \in \mathbb{N}$, then we define

$$X \rtimes Y := \left[X^1 Y, X^2 Y, \cdots, X^t Y\right] \in \mathbb{R}^n, \tag{4.17}$$

where $X = [X^1, X^2, \cdots, X^t]$, $X^i \in \mathbb{R}^n,\ i = 1, \cdots, t$.

Case 2: If $t = s \times n$, where $n \in \mathbb{N}$, then we define

$$X \rtimes Y := \begin{bmatrix} XY^1 \\ XY^2 \\ \vdots \\ XY^n \end{bmatrix} \in \mathbb{R}^n. \tag{4.18}$$

It is easy to verify that when both X and Y are vectors, Definition 4.2 coincides with Definition 4.1. Unfortunately, Definition 4.2 cannot be extended to the right STP of matrices.

In fact, the row-column multiplication rule is a particular case of the block multiplication rule, as demonstrated in Proposition 2.2. The following example shows that the right STP does not satisfy the block multiplication rule.

Example 4.1. Let $A = [a_1, a_2, a_3, a_4]$, $B = [b_1, b_2]^T$. Then by definition

$$A \rtimes B = [a_1 b_1 + a_2 b_2 \ , \ a_3 b_1 + a_4 b_2]. \tag{4.19}$$

If we split A and B into blocks as $A = [A_1, A_2]$ and $B = [B_1, B_2]^T$, then according to the block multiplication rule, we have

$$\begin{aligned} A_1 \rtimes B_1 + A_2 \rtimes B_2 &= [a_1, a_2] \rtimes b_1 + [a_3, a_4] \rtimes b_2 \\ &= [a_1 b_1 + a_3 b_2, a_2 b_1 + a_4 b_2]. \end{aligned} \tag{4.20}$$

(4.20) differs from (4.19), which means (4.20) is incorrect.

Though both left STP and right STP are generalizations of conventional matrix product, but in most cases the left STP is more useful than the right STP. One of the reasons is that the right STP does not satisfy block multiplication rule. Another basic reason is that the left STP has very clear physical meaning in representing multi-dimensional data. We, therefore, consider the left STP as the default STP in the rest of this book.

The following proposition shows how to convert one STP to the other.

Proposition 4.3. *Given $A \in \mathcal{M}_{m\times n}$ and $B \in \mathcal{M}_{p\times q}$. If $A \succ_t B$, then*

$$A \rtimes B = A \ltimes W_{[p,t]} \ltimes B \ltimes W_{[t,q]}. \tag{4.21}$$

Conversely,

$$A \ltimes B = A \rtimes W_{[t,p]} \rtimes B \rtimes W_{[q,t]}. \tag{4.22}$$

If $A \prec_t B$, then

$$A \rtimes B = W_{[m,t]} \ltimes A \ltimes W_{[t,n]} \ltimes B. \tag{4.23}$$

Conversely,

$$A \ltimes B = W_{[t,m]} \rtimes A \rtimes W_{[n,t]} \rtimes B. \tag{4.24}$$

Proof. We prove (4.21) only. The proves of (4.22)−(4.24) are similar. Using (2.60), we have

$$\begin{aligned} A \rtimes B = A(I_t \otimes B) &= AW_{[p,t]}(B \otimes I_t)W_{[t,q]} \\ &= A \ltimes W_{[p,t]} \ltimes B \ltimes W_{[t,q]}. \end{aligned}$$

□

Particularly, when the STP of vectors are considered we have the following corollary.

Corollary 4.1.

(1) Let X be a row vector of dimension np and Y a column vector of dimension p. Then

$$X \rtimes Y = \left(XW_{[p,n]}\right) \ltimes Y. \tag{4.25}$$

Conversely,

$$X \ltimes Y = \left(XW_{[n,p]}\right) \rtimes Y. \tag{4.26}$$

(2) Let X be a row with $\dim(X) = p$ and Y a column with $\dim(Y) = pn$. Then

$$X \rtimes Y = X \ltimes \left(W_{[n,p]}Y\right). \tag{4.27}$$

Conversely,

$$X \ltimes Y = X \ltimes \left(W_{[p,n]}Y\right). \tag{4.28}$$

Proof. (4.25)–(4.28) are particular cases of (4.21)–(4.24) respectively. It is easy to see that the conclusions follow from Proposition 4.3. □

We can also reveal the physical meaning of right STP through the multi-dimensional mappings. The right STP can also search the hierarchical structure of data, or find the "pointer", "pointer to pointer" etc. Let $\sigma : \underbrace{\mathbb{R}^n \times \cdots \times \mathbb{R}^n}_{s} \to \mathbb{R}^s$ be a multilinear mapping. For each basis of $\mathbb{R}^n$ denoted by $\{e_1, \cdots, e_n\}$ and a basis of $\mathbb{R}^s$ denoted by $\{d_1, \cdots, d_s\}$, assume

$$\sigma(e_{i_1}, e_{i_2}, \cdots, e_{i_s}) = \sum_{k=1}^{s} \alpha^k_{i_1,\cdots,i_s} d_k.$$

Then the structure matrix of M_σ is defined as

$$M_\sigma = \begin{bmatrix} \alpha^1_{1\cdots11} & \alpha^1_{1\cdots12} & \cdots & \alpha^1_{n\cdots nn} \\ \alpha^2_{1\cdots11} & \alpha^2_{1\cdots12} & \cdots & \alpha^2_{n\cdots nn} \\ \vdots & \vdots & & \vdots \\ \alpha^n_{1\cdots11} & \alpha^n_{1\cdots12} & \cdots & \alpha^n_{n\cdots nn} \end{bmatrix}. \tag{4.29}$$

For $X_1, \cdots, X_s \in \mathbb{R}^n$, express each X_i by its coefficient vector, that is, if $X_i = \sum_{j=1}^n x^i_j e_j$, then its vector form is $X_i = [x^i_1, \cdots, x^i_n]^T$. Then we have

$$\begin{aligned} \sigma(X_1, \cdots, X_s) &= M_\sigma \ltimes X_1 \ltimes X_2 \ltimes \cdots \ltimes X_s \\ &= M_\sigma \rtimes X_k \rtimes X_{k-1} \rtimes \cdots \rtimes X_1. \end{aligned} \tag{4.30}$$

In applications, the right STP is sometimes convenient. Corresponding to (2.49) and (2.50), we have the following conclusions.

Proposition 4.4. *Assume $A \in \mathcal{M}_{m\times n}$, $X \in \mathcal{M}_{n\times q}$, $Y \in \mathcal{M}_{p\times m}$, then the stacking forms of the products are*

$$V_r(YA) = (I_p \otimes A^T)V_r(Y) = A^T \rtimes V_r(Y); \tag{4.31}$$

$$V_c(AX) = (I_q \otimes A)V_c(X) = A \rtimes V_c(X). \tag{4.32}$$

Proof. To prove (4.31), we use Table 3.2 to get that

$$\begin{aligned} V_r(YA) &= V_c(A^TY^T) \\ &= \left(I_p \otimes A^T\right) V_c(Y^T) \left(I_p \otimes A^T\right) V_r(Y) \\ &= A^T \rtimes V_r(Y). \end{aligned}$$

As for (4.32), using Table 3.4 and equation (2.51), we have

$$\begin{aligned} V_c(AX) &= W_{[m,q]}V_r(AX) = W_{[m,q]} \ltimes A \ltimes V_r(X) \\ &= W_{[m,q]} \ltimes A \ltimes W_{[q,p]} \ltimes V_c(X) \\ &= (I_q \otimes A) \ltimes V_c(B) = A \rtimes V_c(B). \end{aligned}$$

□

Using Proposition 4.4, we can deduce some formulas for the product of three matrices.

Proposition 4.5.

$$V_r(ABC) = (A \otimes C^T)V_r(B); \tag{4.33}$$

$$V_c(ABC) = (C^T \otimes A)V_c(B). \tag{4.34}$$

Proof. Using (2.49) and (4.30), we have

$$\begin{aligned} V_r(ABC) &= (A \otimes I_q)V_r(BC) = (A \otimes I_q)(I_n \otimes C^T)V_r(B) \\ &= (A \otimes C^T)V_r(B). \end{aligned}$$

This proves (4.33).

Using (2.50) and (4.30), we have

$$\begin{aligned} V_c(ABC) &= (I_q \otimes A)V_c(BC) = (I_q \otimes A)(C^T \otimes I_n)V_c(B) \\ &= (C^T \otimes A)V_c(B). \end{aligned}$$

This proves (4.34). □

The following proposition is an immediate consequence of the definition.

Proposition 4.6. *Assume $A \in \mathcal{M}_{m\times n}$ and $B \in \mathcal{M}_{n\times m}$. Then*

$$\operatorname{tr}(AB) = V_c^T(A)V_r(B) = V_r^T(B)V_c(A) = V_c^T(B)V_r(A) = V_r^T(A)V_c(B). \tag{4.35}$$

Combining Proposition 4.6 with (4.33) and (4.34) yields

Proposition 4.7. *Let A, B, C, D be of proper dimensions such that the conventional matrix product of $ABCD$ is well defined and is a square matrix. Then*

$$\operatorname{tr}(ABCD) = V_c^T(A)(B \otimes D^T)V_r(C) = V_r^T(C)(B^T \otimes D)V_c(A). \tag{4.36}$$

Finally, we give two propositions, which are convenient when we convert a matrix polynomial into a polynomial of its entries.

First we define the power of right semi-tensor product: Let $A \in \mathcal{M}_{m\times n}$, and $n|m$ or $m|n$. Then $A \rtimes A$ is well defined, and hence the power of right-tensor product can also be defined as

$$A^{\rtimes k} = \underbrace{A \rtimes A \rtimes \cdots \rtimes A}_{k}.$$

Note that as a convention, we define

$$A^k = \underbrace{A \ltimes A \ltimes \cdots \ltimes A}_{k}.$$

Hence for the power of right-tensor product the symbol $\rtimes$ on the power should not be omitted. It is obvious that when A is a square matrix the powers of both the left and the right semi-tensor products are the same.

Now we are ready to state the propositions.

Proposition 4.8. *Let $Z \in \mathcal{M}_n$. Then*

$$Z^k = (F^{\rtimes k}) \ltimes V_c^k(Z), \quad k \geq 1, \tag{4.37}$$

where

$$F = \big(I_n \otimes V_c^T(I_n)\big) \ltimes W_{[n]}. \tag{4.38}$$

Proof. We prove (4.37) by mathematical induction. Using (3.56), we have

$$Z = \pi_n^c(I_n) \ltimes V_c(Z).$$

Let $F = \pi_n^c(I_n)$. Using (7.43) and (7.45), it is easy to see that F has the form as in (4.38). Note that $F \in \mathcal{M}_{n\times n^3}$, it follows that $F^{\rtimes k} \in \mathcal{M}_{n\times n^{2k+1}}$. Next, we assume (4.37) holds for k. Then using (2.55), we have

$$\begin{aligned} Z^{k+1} &= Z \ltimes Z^k = F \ltimes V_c(Z) \ltimes F^{\rtimes k} \ltimes V_c^k(Z) \\ &= F \ltimes (I_{n^2} \otimes F^{\rtimes k}) \ltimes V_c(Z) \ltimes V_c^k(Z) \\ &= \big(F(I_{n^2} \otimes F^{\rtimes k})\big) \ltimes V_c^{k+1}(Z) \\ &= F^{\rtimes (k+1)} \ltimes V_c^{k+1}(Z). \end{aligned}$$

□

Proposition 4.9. *Let $Z \in \mathcal{M}_n$. Then*

$$V_r(Z^k) = \left(E^{\rtimes(k-1)}\right) V_r^k(Z), \quad k \geq 2, \tag{4.39}$$

$$V_c(Z^k) = W_{[n]} \left(E^{\rtimes(k-1)}\right) (W_{[n]})^{\otimes k} V_c^k(Z), \quad k \geq 2, \tag{4.40}$$

where

$$E = I_n \otimes V_c^T(I_n).$$

Proof. First, we prove (4.39). For $k = 1$, we have

$$V_r(Z^2) = Z \ltimes V_r(Z) = F V_c(Z) V_r(Z) = F W_{[n]} V_r^2(Z) = E V_r^2(Z).$$

Now we assume (4.39) is true for k, then from (2.38) we have

$$\begin{aligned} V_r(Z^{k+1}) &= Z \ltimes V_r(Z^k) = E \ltimes V_r(Z) \ltimes E^{\rtimes(k-1)} \ltimes V_r^k(Z) \\ &= \big(E \ltimes (I_{n^2} \otimes E^{\rtimes(k-1)})\big) \ltimes V_r(Z) \ltimes V_r^k(Z) \\ &= \big(E^{\rtimes k}\big) \ltimes V_r^{k+1}(Z). \end{aligned}$$

This proves (4.39).

Note that

$$V_r^k(Z) = (\underbrace{W_{[n]} \otimes \cdots \otimes W_{[n]}}_{k}) V_c^k(Z) = (W_{[n]})^{\otimes k} V_c^k(Z).$$

Left multiplying $W_{[n]}$ to both sides of (4.39) yields (4.40). □

4.2 Semi-Tensor Product of Arbitrary Matrices

This section considers the left and right semi-tensor products for two arbitrary matrices. For statement ease, we call them the general semi-tensor product of matrices.

Let $a, b \in \mathbb{Z}^+$. Denote the least common multiple of a and b by $\text{lcm}\{a, b\}$.

Definition 4.3. Let $A \in \mathcal{M}_{m\times n}$, $B \in \mathcal{M}_{p\times q}$ and $\alpha = \text{lcm}\{n, p\}$.

(1) The general left STP of A and B is defined as as

$$A \ltimes B = \big(A \otimes I_{\alpha/n}\big)\big(B \otimes I_{\alpha/p}\big). \tag{4.41}$$

(2) The general right STP of A and B is defined as

$$A \rtimes B = \big(I_{\alpha/n} \otimes A\big)\big(I_{\alpha/p} \otimes B\big). \tag{4.42}$$

Note that when $n = p$ the general left (right) STP of matrices becomes the conventional matrix product.

If $\mathrm{lcm}\{n,p\} = n$ or $\mathrm{lcm}\{n,p\} = p$, then the general left (right) STP of matrices becomes the previously defined left (right) STP. Unless otherwise stated, throughout this book the STP is defined for multiplier dimensional case except this section.

The reason that we do not pay much attention to the general STP is that unlike the multiplier dimensional case, we could not find a reasonable physical explanation for the general STP and did not find meaningful applications so far.

In the following we consider some basic properties of general left (right) STP.

Proposition 4.10. *The general left (right) STP satisfies*

(1) (Distributive Law)

$$(A + B) \ltimes C = (A \ltimes C) + (B \ltimes C), \tag{4.43}$$

$$(A + B) \rtimes C = (A \rtimes C) + (B \rtimes C), \tag{4.44}$$

$$C \ltimes (A + B) = (C \ltimes A) + (C \ltimes B), \tag{4.45}$$

$$C \rtimes (A + B) = (C \rtimes A) + (C \rtimes B). \tag{4.46}$$

(2) (Associative Law)

$$(A \ltimes B) \ltimes C = A \ltimes (B \ltimes C), \tag{4.47}$$

$$(A \rtimes B) \rtimes C = A \rtimes (B \rtimes C). \tag{4.48}$$

Proof. Distributive law is easily verifiable. We prove the associative law.

Let $A \in \mathcal{M}_{m\times n}$, $B \in \mathcal{M}_{p\times q}$, $C \in \mathcal{M}_{s\times t}$. We first show that both sides of (4.47) (similar for (4.48)) can be expressed as

$$(A \otimes I_\alpha)(B \otimes I_\beta)(C \otimes I_\gamma). \tag{4.49}$$

This equality is obtained from the equality that

$$[(A \otimes I_m)(B \otimes I_n)] \otimes I_s = (A \otimes I_{ms})(B \otimes I_{ns}).$$

From the definition of the general semi-tensor product one sees that no matter what is the order for the product to be executed, we finally need to find the smallest natural numbers α, β, γ, such that

$$\begin{cases} \alpha n = \beta p \\ \beta q = \gamma s. \end{cases} \tag{4.50}$$

If we can show that the smallest solution of (4.50) is unique, the conclusion follows. Since $\alpha nq = \gamma sp$, we can assume the greatest common divisor of nq and sp is h, i.e., $h = \gcd\{nq, sp\}$. Then

$$\alpha = \mu \frac{sp}{h}, \quad \gamma = \mu \frac{nq}{h}.$$

It follows that

$$\beta = \frac{\mu sn}{h} = \frac{sn}{h/\mu}.$$

Now we have to find the smallest μ, which makes β be an integer. It is clear that the μ should be

$$\mu = \frac{h}{\gcd\{h, sm\}} = \frac{h}{\gcd\{nq, sp, sm\}}.$$

Define

$$c = \frac{h}{\mu} = \gcd\{nq, sp, sn\}.$$

Then it is verified that both sides of (4.47) equal to (4.49) with

$$\alpha = \frac{sp}{c}, \quad \beta = \frac{sn}{c}, \quad \gamma = \frac{nq}{c}.$$

□

Almost all the major properties of the conventional matrix product remain true for generalized left (right) semi-tensor product of matrices. We list the major properties as follows:

Proposition 4.11.

(1)

$$\begin{cases} (A \ltimes B)^T = B^T \ltimes A^T, \\ (A \rtimes B)^T = B^T \rtimes A^T. \end{cases} \tag{4.51}$$

(2) If $M \in \mathcal{M}_{m\times pn}$, then

$$\begin{cases} M \ltimes I_n = M, \\ M \rtimes I_n = M; \end{cases} \tag{4.52}$$

If $M \in \mathcal{M}_{pm\times n}$, then

$$\begin{cases} I_m \ltimes M = M, \\ I_m \rtimes M = M. \end{cases} \tag{4.53}$$

In the following items, A and B are two square matrices.

(3) $A \ltimes B$ and $B \ltimes A$ ($A \rtimes B$ and $B \rtimes A$) have the same characteristic function.

(4)

$$\begin{cases} \mathrm{tr}(A \ltimes B) = \mathrm{tr}(B \ltimes A), \\ \mathrm{tr}(A \rtimes B) = \mathrm{tr}(B \rtimes A). \end{cases} \tag{4.54}$$

(5) If both A and B are orthogonal matrices (upper triangular matrices, down triangular matrices, or diagonal matrices), then so is $A \ltimes B$ ($A \rtimes B$).

(6) If at least one of A and B is invertible, then $A \ltimes B \sim B \ltimes A$ ($A \rtimes B \sim B \rtimes A$).

(7) If both A and B are invertible, then

$$\begin{cases} (A \ltimes B)^{-1} = B^{-1} \ltimes A^{-1}, \\ (A \rtimes B)^{-1} = B^{-1} \rtimes A^{-1}. \end{cases} \tag{4.55}$$

(8) The determinant of the product satisfies

$$\begin{cases} \det(A \ltimes B) = [\det(A)]^{\alpha/n}[\det(B)]^{\alpha/p}, \\ \det(A \rtimes B) = [\det(A)]^{\alpha/n}[\det(B)]^{\alpha/p}, \end{cases} \tag{4.56}$$

where $\alpha = \mathrm{lcm}\{n, p\}$.

The following property is for the general left STP only.

Proposition 4.12. *Let $A \in \mathcal{M}_{m\times n}$ and $B \in \mathcal{M}_{p\times q}$. Then*

$$C = A \ltimes B = (C^{ij}), \quad i = 1, \cdots, m,\ j = 1, \cdots, q, \tag{4.57}$$

where

$$C^{ij} = A^i \ltimes B_j.$$

Here $A^i = \mathrm{Row}_i(A)$ and $B_j = \mathrm{Col}_j(B)$.

Note that (4.57) can also be considered as the definition of general left STP. In equal dimensional case, C^{ij} is a number, in multiplier case it is a vector, and in general case it is a block.

Remark 4.1. As a convention, for any two matrices $A \in \mathcal{M}_{m\times n}$ and $B \in \mathcal{M}_{p\times q}$, the default matrix product is assumed to be the left STP. That is,

$$AB = A \ltimes B. \tag{4.58}$$

When $n = p$, it is the conventional matrix product; when $\mathrm{lcm}\{n, p\} = \max\{n, p\}$, it is the left STP defined in Chapter 2; and otherwise, it is the general left STP. Under this convention, the symbol $\ltimes$ is usually omitted, unless we would like to emphasize it is STP.

The reader may be convinced by the discussion so far that the concept of conventional matrix product can be replaced by the left STP.

Let $A \in \mathcal{M}_{m\times n}$. We can define the power of general left STP of A as

$$\begin{cases} A^1 = A, \\ A^{k+1} = A \ltimes A^k, \quad k \geq 1. \end{cases}$$

Similarly, we can also define the power of general left STP of A as

$$\begin{cases} A^{\rtimes 1} = A, \\ A^{\rtimes (k+1)} = A \rtimes A^{\rtimes k}, \quad k \geq 1. \end{cases}$$

To consider the dimension of A^k (or $A^{\rtimes k}$), let $\mathrm{lcm}\{m, n\} = t$. Set $m = m_0 t$ and $n = n_0 t$, then m_0 and n_0 are coprime. It is easy to prove by mathematical induction that $A^k \in \mathcal{M}_{m_0^k t\times n_0^k t}$ ($A^{\rtimes k} \in \mathcal{M}_{m_0^k t\times n_0^k t}$).

Exercises

4.1

$$A = \begin{bmatrix} 2 & 1 \\ 0 & -1 \\ 3 & -3 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & 1 \\ 0 & 2 \\ 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 2 \\ 1 & -3 \\ 2 & -1 \end{bmatrix}.$$

(i) Calculate $(A \ltimes B) \rtimes C$.

(ii) Calculate $A \ltimes (B \rtimes C)$.

(iii) Comparing the results in (i) and (ii) to see whether there is an associativity in mixed left and right STPs.

4.2 Given $A = [A_1\ A_2]$, and $B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$, where

$$A_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ -1 & 2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & -3 \end{bmatrix},$$

$$B_1 = \begin{bmatrix} 1 & 0 & 2 \\ 0 & -1 & 0 \\ -1 & 1 & 3 \\ 0 & -2 & 2 \end{bmatrix}, \quad B_2 = \begin{bmatrix} -1 & 1 & 0 \\ 3 & -1 & 2 \\ 0 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix}.$$

(i) Calculate $A_1 \rtimes B_1 + A_2 \rtimes B_2$.

(ii) Calculate $A \rtimes B$.

(iii) Comparing the results to see whether the right STP satisfies block multiplication rule.

4.3 Let $A \in \mathcal{M}_{m,n}$, $B \in \mathcal{M}_{p,q}$, and $C \in \mathcal{M}_{r,s}$. Find the dimension of $A \ltimes B \ltimes C$.

4.4 Prove (4.6) and (4.7).

4.5 Prove (4.8).

4.6 Prove (4.10)–(4.12).

4.7 In cubic matrix theory, let

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix},$$

where X_i, $i = 1, \cdots, n$ are $n \times n$ matrices. Assume $b \in \mathbb{R}^n$, the quadratic form of X is defined as (Wang, 2002)

$$b^T X b := \begin{bmatrix} b^T X_1 b \\ \vdots \\ b^T X_m b \end{bmatrix}.$$

Prove that it can be expressed as

$$b^T X b = b^T \rtimes (Xb) = (b^T \rtimes X)b.$$

4.8 Give the matrix-vector expression of $V_r(AX)$ using the right semi-tensor product.

4.9 Given $A \in \mathcal{M}_{m \times n}$ and $B \in \mathcal{M}_{p \times q}$.

(i) When $A \ltimes B = A \rtimes B$?

(ii) When $A \ltimes B = A \otimes B$?

(iii) When $A \rtimes B = A \otimes B$?

4.10 (i) Let $A \in \mathcal{M}_{m\times n}$, $X \in \mathcal{M}_{p\times q}$, and $\operatorname{lcm}\{n,p\} = t$. Discuss the solution of

$$A \ltimes X = B, \quad \text{where } B \in \mathcal{M}_{mt/n\times qt/p}.$$

(ii)

$$A = \begin{bmatrix} 2 & 1 \\ 0 & -1 \\ 3 & -3 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 1 & 0 & 0 & -1 & 0 & 3 & -3 & 0 \\ 0 & 2 & 1 & 0 & 0 & -1 & 0 & 3 & -3 \\ 0 & 1 & -2 & 0 & -1 & 0 & 0 & -3 & -3 \\ -1 & 0 & 1 & 1 & 0 & -1 & 3 & 0 & -3 \end{bmatrix}^T.$$

Solve the equation

$$A \ltimes X = B, \quad \text{where } X \in \mathcal{M}_{3\times 2}.$$

4.11 Assume $A \prec_t X$ (or $A \succ_t X$).

(i) Find M^r such that

$$V_r(A \rtimes X) = M^r V_r(X).$$

(ii) Find M^c such that

$$V_c(A \rtimes X) = M^c V_c(X).$$

4.12 Assume $A \in \mathcal{M}_{m\times n}$, $X \in \mathcal{M}_{p\times q}$, and $B \in \mathcal{M}_{s\times t}$. $\pi : X \mapsto A \ltimes X \ltimes B$.

(i) Find M^r_π such that

$$V_r(A \ltimes X \ltimes B) = M^r_\pi V_r(X).$$

(ii) Find M^c_π such that

$$V_c(A \ltimes X \ltimes B) = M^c_\pi V_c(X).$$

(iii) When $\pi : X \mapsto A \ltimes (X \rtimes B)$. Find the corresponding M^r_π and M^c_π.

4.13 Given two matrices $A \in \mathcal{M}_{m\times n}$ and $B \in \mathcal{M}_{p\times q}$. What can you say about (i) $\operatorname{rank}(A \ltimes B)$, or (ii) $\operatorname{rank}(A \rtimes B)$?

4.14 Consider

$$GL(\mathbb{R}) := \cup_{n=1}^{\infty} GL(n, \mathbb{R}).$$

Define an equivalence on $GL(\mathbb{R})$ as

$$A \sim A \otimes I_k, \quad k \in \mathbb{N}.$$

Denote the equivalent class of A by $[A]$. Consider the quotient set

$$G := GL(\mathbb{R})/\sim = \left\{ [A] \,|\, A \in GL(\mathbb{R}) \right\}.$$

Define the product over G as

$$[A] \ltimes [B] := [A \ltimes B].$$

(i) Prove that $(G, \ltimes)$ is a group.

(ii) Let

$$O(\mathbb{R}) := \cup_{n=1}^{\infty} O(n, \mathbb{R}).$$

Define $G_O := O(\mathbb{R})/\sim$. Prove that G_O is a subgroup of G.

This page intentionally left blank

Chapter 5

Rank, Pseudo-Inverse, and Positivity of STP

This chapter considers some further properties of left semi-tensor product. The left STP might be either general or multi-dimensional one. The rank, the pseudo-inverse, and the positivity of STP are considered. This chapter is mainly based on the works of a research group of Liaocheng University. We refer to Li *et al.* (2008); Song *et al.* (2008); Wang *et al.* (2009a,b); Liu *et al.* (2008) for details. We also refer to two theses of Liaocheng University, Song (2009); Wang (2010), and the references therein for systematic introduction to their works.

5.1 Rank of Products

Proposition 5.1. *Assume $A \in \mathcal{M}_{m\times n}$, $B \in \mathcal{M}_{p\times q}$, $\operatorname{rank}(A) = r_A$, and B is of full row rank. Then*

$$\begin{aligned} \operatorname{rank}(A \ltimes B) &= r_A, \quad \textit{if} \;\; A \succ_t B, \\ \operatorname{rank}(A \ltimes B) &= r_A t, \quad \textit{if} \;\; A \prec_t B. \end{aligned} \tag{5.1}$$

Proof. When $A \succ_t B$, we have

$$A \ltimes B = A(B \otimes I_t).$$

Since $(B\otimes I_t)$ has full row rank, which is pt, it follows that $\operatorname{rank}(A\ltimes B) = r_A$. Similarly, when $A \prec_t B$, we have

$$A \ltimes B = (A \otimes I_t)B.$$

Since B is of full row rank, we have

$$\operatorname{rank}(A \ltimes B) = \operatorname{rank}(A \otimes I_t) = r_A t.$$

□

Similarly, we have the following propositions. We leave their proves to the reader.

Proposition 5.2. *Assume* $A \in \mathcal{M}_{m\times n}$*,* $B \in \mathcal{M}_{p\times q}$*,* $\text{rank}(B) = r_B$*, and* A *is of full column rank. Then*

$$\begin{aligned} \text{rank}(A \ltimes B) &= r_B t, \quad \textit{if} \ \ A \succ_t B, \\ \text{rank}(A \ltimes B) &= r_B, \quad \textit{if} \ \ A \prec_t B. \end{aligned} \tag{5.2}$$

Proposition 5.3. *Assume* $A \in \mathcal{M}_{m\times n}$*,* $B \in \mathcal{M}_{p\times q}$*,* A *is of full column rank and* B *is of full row rank. Then*

$$\begin{aligned} \text{rank}(A \ltimes B) &= n = pt, \quad \textit{if} \ \ A \succ_t B, \\ \text{rank}(A \ltimes B) &= nt = p, \quad \textit{if} \ \ A \prec_t B. \end{aligned} \tag{5.3}$$

Proposition 5.4. *If* A *and* B *are of multi-dimensional case. Then*

$$\begin{aligned} \text{rank}(A \ltimes B) &\le \min\{r_A, t r_B\}, \quad \textit{if} \ \ A \succ_t B, \\ \text{rank}(A \ltimes B) &\le \min\{t r_A, r_B\}, \quad \textit{if} \ \ A \prec_t B. \end{aligned} \tag{5.4}$$

We give an example to depict this.

Example 5.1. Let

$$A = \begin{bmatrix} 1 & 2 & 2 & 2 \\ 1 & 2 & 0 & 2 \\ 1 & 1 & 1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 3 & 0 \\ 4 & 2 & 2 \end{bmatrix}.$$

We know that $\text{rank}(A) = 2$, B is of full row rank. By straightforward computation, we have

$$A \ltimes B = \begin{bmatrix} 9 & 10 & 7 & 10 & 4 & 4 \\ 1 & 10 & 3 & 10 & 0 & 4 \\ 5 & 10 & 5 & 10 & 2 & 4 \end{bmatrix},$$

and $\text{rank}(A \ltimes B) = 2$ which satisfies (5.1).

5.2 Pseudo-Inverse of STP

In this section two kinds of pseudo-inverse of the STP of two matrices are considered. Pseudo-inverse, as a generalized inverse, is an important topic in both theoretical and application aspects. We refer to Wang *et al.* (2004) as a general reference. In this section the matrices are assumed to be complex. We first give some notations:

- $\mathbb{C}_{m\times n}$: the set of $m \times n$ complex matrices.
- A^*: the conjugate and transpose of A, i.e., $A^* = \bar{A}^T$.
- $M \in \mathbb{C}_{n\times n}$ is called a Hermitian matrix, if $M^* = M$.

5.2.1 *Moore-Penrose Inverse*

Definition 5.1. Let $A \in \mathbb{C}_{m\times n}$, and $M \in \mathbb{C}_{m\times m}$ and $N \in \mathbb{C}_{n\times n}$ be two positive definite Hermitian matrices. Assume there exists a matrix X, satisfying

(i)

$$AXA = A; \tag{5.5}$$

(ii)

$$XAX = X; \tag{5.6}$$

(iii)

$$(MAX)^* = MAX; \tag{5.7}$$

(iv)

$$(NXA)^* = NXA, \tag{5.8}$$

then X is called the weighted Moore-Penrose inverse of A, denoted by $X = A^+_{MN}$.

Particularly, if $M = I_m$ and $N = I_n$, the matrix $X = A^+_{MN}$ becomes the Moore-Penrose inverse of A, denoted by A^+.

Moore-Penrose inverse of A is also called the pseudo-inverse in literature. It is very useful. We give its basic properties as follows:

Theorem 5.1 (Rohde, 1966). *For any $A \in \mathbb{C}_{m\times n}$, there exists unique Moore-Penrose inverse A^+, satisfying*

(i)

$$(A^*)^+ = (A^+)^*; \tag{5.9}$$

(ii)

$$A^+AA^* = A^*AA^+ = A^*; \tag{5.10}$$

(iii)

$$AA^*(A^*)^+ = (A^*)^+A^*A = A; \tag{5.11}$$

(iv) AA^+, A^+A, $(I - AA^+)$, $(I - A^+A)$ are all idempotent.

Note that a matrix A is idempotent means $A^2 = A$.

The following result shows that the weighted Moore-Penrose inverse A^+_{MN} can be calculated via non-weighted inverse (Wang *et al.*, 2004)

$$A^+_{MN} = N^{-\frac{1}{2}} \left(M^{\frac{1}{2}} A N^{-\frac{1}{2}} \right)^+ M^{\frac{1}{2}}. \tag{5.12}$$

Note that since

$$(A \ltimes B)^{-1} = B^{-1} \ltimes A^{-1}, \tag{5.13}$$

it is natural to ask if

$$(A \ltimes B)^{+} = B^{+} \ltimes A^{+}, \tag{5.14}$$

or, if it is true for weighted Moore-Penrose inverse. In general, it is incorrect. Hence, the question becomes: when are they correct? For conventional matrix product the answers are known. To present it, we need a new concept: the weighted conjugate transpose matrix of $A \in \mathbb{C}_{m\times n}$ is defined as

$$A^{\#}_{MN} = N^{-1} A^* M, \tag{5.15}$$

where M and N are m, n dimensional positive definite matrices.

Theorem 5.2. *Assume $A \in \mathbb{C}_{m\times n}$, $B \in \mathbb{C}_{n\times p}$, and M, N, P are m-, n-, p-dimensional positive definite matrices. Then the followings are equivalent:*

(1)

$$(AB)^+_{MP} = B^+_{NP} A^+_{MN}. \tag{5.16}$$

(2)

$$\begin{cases} BB^+_{NP} A^{\#}_{MN} AB = A^{\#}_{MN} AB, \\ A^+_{MN} ABB^{\#}_{NP} A^{\#}_{MN} = BB^{\#}_{NP} A^{\#}_{MN}. \end{cases} \tag{5.17}$$

(3)

$$\mathcal{R}\left(A^{\#}_{MN} ABB^{\#}_{NP} \right) = \mathcal{R}\left(BB^{\#}_{NP} A^{\#}_{MN} A \right). \tag{5.18}$$

(4)

$$\begin{cases} \mathcal{R}\left(A^{\#}_{MN} AB \right) \subset \mathcal{R}(B), \\ \mathcal{R}\left(BB^{\#}_{NP} A \right) \subset \mathcal{R}(A^{\#}_{MN}). \end{cases} \tag{5.19}$$

The notation $\mathcal{R}(\cdot)$ in the above means $\mathcal{R}(S) = \mathrm{Span}\,\mathrm{Col}(S)$.

Then we have the following result, where we only consider the case that $A \succ_t B$. We leave to the reader to find similar results for the case of $A \prec_t B$.

Theorem 5.3. *Assume $A \in \mathbb{C}_{m\times nt}$, $B \in \mathbb{C}_{n\times p}$, and M, N, P are m-, n-, p-dimensional positive definite matrices. Then the followings are equivalent:*

(1)

$$(A \ltimes B)^{+}_{M(P\otimes I_t)} = B^{+}_{NP} \ltimes A^{+}_{M(N\otimes I_t)}. \tag{5.20}$$

(2)

$$\begin{cases} \left(BB^{+}_{NP} \ltimes A^{\#}_{M(N\otimes I_t)}\right)(A \ltimes B) = A^{\#}_{M(N\otimes I_t)}(A \ltimes B), \\ A^{+}_{M(N\otimes I_t)}(A \ltimes B)\left(B^{\#}_{NP} \ltimes A^{\#}_{M(N\otimes I_t)}\right) = BB^{\#}_{NP} \ltimes A^{\#}_{M(N\otimes I_t)}. \end{cases} \tag{5.21}$$

(3)

$$\mathcal{R}\left[A^{\#}_{M(N\otimes I_t)}(A \ltimes B \ltimes B^{\#}_{NP})\right] = \mathcal{R}\left[\left(BB^{\#}_{NP} \ltimes A^{\#}_{M(N\otimes I_t)}\right)A\right]. \tag{5.22}$$

(4)

$$\begin{cases} \mathcal{R}\left[A^{\#}_{M(N\otimes I_t)}(A \ltimes B)\right] \subset \mathcal{R}(B), \\ \mathcal{R}\left(BB^{\#}_{NP} \ltimes A\right) \subset \mathcal{R}(A^{\#}_{M(N\otimes I_t)}). \end{cases} \tag{5.23}$$

Proof. Note that for any matrices $C \in \mathbb{C}_{m\times n}$ and $D \in \mathbb{C}_{n\times p}$,

$$(C \otimes I_t)(D \otimes I_t) = (CD \otimes I_t). \tag{5.24}$$

Using (5.24), we can verify that

$$(B \otimes I_t)^{+}_{(N\otimes I_t)(P\otimes I_t)} = (B^{+}_{NP} \otimes I_t), \tag{5.25}$$

$$(B \otimes I_t)^{\#}_{(N\otimes I_t)(P\otimes I_t)} = (B^{\#}_{NP} \otimes I_t). \tag{5.26}$$

Since

$$\begin{aligned} (A \ltimes B)^{+}_{M(P\otimes I_t)} &= (A(B \otimes I_t))^{+}_{M(P\otimes I_t)}; \\ B^{+}_{NP} \ltimes A^{+}_{M(N\otimes I_t)} &= (B^{+}_{NP} \otimes I_t)A^{+}_{M(N\otimes I_t)} \\ &= (B \otimes I_t)^{+}_{(N\otimes I_t)(P\otimes I_t)}A^{+}_{M(N\otimes I_t)}, \end{aligned}$$

we only need to find the equivalent conditions for

$$(A(B \otimes I_t))^{+}_{M(P\otimes I_t)} = (B \otimes I_t)^{+}_{(N\otimes I_t)(P\otimes I_t)}A^{+}_{M(N\otimes I_t)}.$$

Hence, replacing B by $B \otimes I_t$ in Theorem 5.2, a straightforward verification leads to the conclusion. □

Remark 5.1.

(1) In fact, the symbol "$\ltimes$" in Theorem 5.3 can be omitted, then we can see that it is of almost the same form as Theorem 5.2.

(2) For non-weighted Moore-Penrose inverse, we can just omit the M, N-related subscripts and replace the superscript $\#$ by $*$.

We give an example to depict this.

Example 5.2. Let

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix},$$

and we only consider Moore-Penrose inverse without the weights M, N, P. We can get the Moore-Penrose inverses of A and B respectively as

$$A^+ = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ -2 & 1 \\ 1 & 0 \end{bmatrix}, \quad B^+ = B^{-1} = \begin{bmatrix} 0 & 1 \\ 0.5 & 0 \end{bmatrix}.$$

Since

$$A \ltimes B = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \end{bmatrix},$$

its Moore-Penrose inverse can be calculated as

$$(A \ltimes B)^+ = \begin{bmatrix} -2 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Then it is ready to verify that

$$(A \ltimes B)^+ = B^+ \ltimes A^+.$$

We can check the other equivalent conditions of Theorem 5.3. For instance, consider the second equation (5.21). By direct computation, we know that

$$(B \ltimes B^+ \ltimes A^T)(A \ltimes B) = A^T(A \ltimes B) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 2 & 5 & 0 & 0 \end{bmatrix},$$

$$A^+(A \ltimes B)(B^T \ltimes A^T) = (B \ltimes B^T \ltimes A^T) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 1 & 2 \end{bmatrix}.$$

We leave to the reader to check the third and fourth equations (5.22), and (5.23).

5.2.2 *Drazin Inverse*

Definition 5.2.

(1) Let $A \in \mathbb{C}_{n\times n}$. The Index of A, denoted by $\mathrm{Ind}(A)$, is defined as the smallest nonnegative k such that

$$\mathrm{rank}(A^k) = \mathrm{rank}(A^{k+1}).$$

(2) Let $A \in \mathbb{C}_{m\times n}$ and $W \in \mathbb{C}_{n\times m}$ with $\mathrm{Ind}(AW) = k$. Assume there exists an $X \in \mathbb{C}_{m\times n}$, satisfying

(i)

$$(AW)^{k+1}XW = (AW)^k; \tag{5.27}$$

(ii)

$$XWAWX = X; \tag{5.28}$$

(iii)

$$AWX = XWA, \tag{5.29}$$

then X is called the weighted Drazin inverse of A with respect to W, denoted by $X = A_{d,W}$.

If $A \in \mathbb{C}_{n\times n}$ and $W = I_n$, then X is called the Drazin inverse of A, denoted by $X = A_d$.

It is well known that for a given W the weighted drazin inverse uniquely exists.

We give an example to depict it.

Example 5.3. Consider a matrix

$$A = \begin{bmatrix} \frac{30}{7} & \frac{16}{7} & \frac{1}{7} & \frac{1}{7} \\ \frac{16}{7} & \frac{30}{7} & \frac{1}{7} & \frac{1}{7} \\ -2 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \end{bmatrix}.$$

Its Drazin inverse is

$$A_d = \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{4} & \frac{1}{4} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{13}{8} & \frac{15}{8} & \frac{23}{2} & \frac{23}{2} \\ \frac{15}{8} & \frac{13}{8} & \frac{23}{2} & \frac{23}{2} \end{bmatrix}.$$

We leave to the reader to verify that A and $X = A_d$ satisfy (5.27)-(5.29) with $m = n = 4$ and $W = I_4$.

Similar to Theorem 5.3, we can obtain the following result by replacing A_2 by $A_2 \otimes I_t$ of the corresponding result for conventional matrices product in Wang and Xu (2006). We only state the following result for STP.

Theorem 5.4. *Assume* $A_1, W_1^T \in \mathbb{C}_{m\times nt}$, $A_2, W_2^T \in \mathbb{C}_{n\times p}$, $A = A_1 \ltimes A_2$, $W = W_2 \ltimes W_1$, $\mathrm{Ind}(A_\ell W_\ell) = i_\ell$, $\mathrm{Ind}(W_\ell A_\ell) = j_\ell$, $h_\ell = \max\{i_\ell, j_\ell\}$ *for* $\ell = 1, 2$, $\mathrm{Ind}(AW) = k_1$, $\mathrm{Ind}(WA) = k_2$ *and* $k = \max\{k_1, k_2\}$. *Then*

$$W(A_1 \ltimes A_2)_{d,W} W = (W_2 (A_2)_{d,W_2} W_2) \ltimes (W_1 (A_1)_{d,W_1} W_1), \tag{5.30}$$

if and only if

$$\mathrm{rank}\begin{bmatrix} (AW)^{2k+1} & 0 & 0 & (AW)^k \\ 0 & 0 & (A_1W_1)^{2h_1+1} & (A_1W_1)^{h_1} \\ 0 & (A_2W_2)^{2h_2+1} \otimes I_t & (A_2W_2)^{h_2} \ltimes (W_1(A_1W_1)^{h_1}) & 0 \\ W(AW)^k & W_2(A_2W_2)^{h_2} \otimes I_t & 0 & 0 \end{bmatrix}$$
$$= \mathrm{rank}\left[(A_1W_1)^{i_1}\right] + \mathrm{rank}\left[(A_2W_2)^{i_2}\right] + \mathrm{rank}\left[(AW)^{k_1}\right]. \tag{5.31}$$

5.3 Positivity of Products

In this section, the matrices considered are assumed to be real.

First, we briefly review normal matrix. $A \in \mathcal{M}_{n\times n}$ is called a normal matrix if

$$A^T A = AA^T.$$

Lemma 5.1. *Assume* $A \in \mathcal{M}_{n\times n}$ *is a normal matrix. Then there exists a unitary matrix* $U \in \mathbb{C}_{n\times n}$, *and* $\lambda_i \in \mathbb{C}, 1 \le i \le n$, *such that*

$$U^* A U = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n). \tag{5.32}$$

Throughout this section we do not require a positive definite matrix to be symmetric, that is

A1. $A \in \mathcal{M}_{n\times n}$ is positive definite, if

$$x^T A x > 0, \quad 0 \neq x \in \mathbb{R}^n.$$

Hence A is positive definite, if and only if $A + A^T$ is symmetric and positive definite.

Remark 5.2.

(1) Since U is an unitrary matrix, i.e., $U^{-1} = U^*$, we know that in (5.32) $\lambda_i \in \sigma(A)$.

(2) Under A1, we can see that a normal matrix A is positive definite, if and only if all the eigenvalues of A are positive real numbers, denoted by $\sigma(A) > 0$.

Theorem 5.5. *Assume A is a symmetric and positive definite matrix and*

$$(A \ltimes B)(B \ltimes A^{-1})^T = (A \ltimes B)^T (B \ltimes A^{-1}). \tag{5.33}$$

Then $A \ltimes B$ is positive definite, if and only if $\sigma(B) > 0$.

Proof. We prove it for the case that $A \succ_t B$. The prove for $A \prec_t B$ is similar. In fact, this is also true for general dimensional case.

(Necessity:) Since $A \ltimes B$ is positive definite, then

$$A^{-1/2} A (B \otimes I_t) A^{-1/2} = A^{1/2} (B \otimes I_t) A^{-1/2}$$

is also a positive definite matrix. Hence, $\sigma(B \otimes I_t) > 0$, and it follows that $\sigma(B) > 0$.

(Sufficiency:) Assume $\sigma(B) > 0$, then $\sigma\left[A^{1/2}(B \otimes I_t)A^{-1/2}\right] > 0$. According to Remark 5.2, it suffices to prove that $A^{1/2}(B \otimes I_t)A^{-1/2}$ is a normal matrix. Using (5.32), the normality can be proved by a straightforward computation. □

We have several corollaries, which are convenient in use.

Corollary 5.1. *Assume A is symmetric and positive definite, B is symmetric, and $A \ltimes B = B \ltimes A$. Then $A \ltimes B$ is positive definite, if and only if $\sigma(B) > 0$.*

Corollary 5.2. *Assume both A and B are symmetric and positive definite. Then $A \ltimes B$ is positive definite, if and only if $A \ltimes B = B \ltimes A$.*

Corollary 5.3. *Assume A is symmetric and positive definite, B is positive definite, and $A \ltimes B = B \ltimes A$. Then $A \ltimes B$ is a positive definite matrix.*

Example 5.4. Let

$$A = \begin{bmatrix} 4 & 1 & 2 & 1 \\ 1 & 4 & 1 & 2 \\ 2 & 1 & 4 & 1 \\ 1 & 2 & 1 & 4 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

We know that A and B are all symmetric and positive definite. Meanwhile,

$$A \ltimes B = \begin{bmatrix} 10 & 3 & 8 & 3 \\ 3 & 10 & 3 & 8 \\ 8 & 3 & 10 & 3 \\ 3 & 8 & 3 & 10 \end{bmatrix} = B \ltimes A.$$

By Corollary 5.2, it is positive definite. And we can check that its eigenvalues are 2, 12, and 24.

If we change B to

$$B = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix},$$

B is not positive definite. However, $A \ltimes B = B \ltimes A$ also holds, since

$$A \ltimes B = B \ltimes A = \begin{bmatrix} 8 & 3 & 10 & 3 \\ 3 & 8 & 3 & 10 \\ 10 & 3 & 8 & 3 \\ 3 & 10 & 3 & 8 \end{bmatrix}.$$

Thus, from Corollary 5.1, we know that it is not positive definite. It is easy to check that its eigenvalues as -2, 12, and 24.

We give some other results concerning the positivity of STP of matrices.

Theorem 5.6. *Assume A is a symmetric and positive definite matrix, and $A \ltimes B$ is a symmetric matrix. Then $A \ltimes B$ is positive definite, if and only if $\sigma(B) > 0$.*

Proof. We prove it for $A \succ_t B$.

(Necessity) Since A is symmetric and positive definite and $A \ltimes B$ is symmetric, there exists a nonsingular matrix P such that

$$P^T A P = I_m, \quad P^T(A \ltimes B)P = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_m). \tag{5.34}$$

Since $A \ltimes B$ is positive definite, $\sigma\left(A \ltimes B\right) > 0$. Note that

$$\begin{aligned} P^T(A \ltimes B)P = P^T(A(B \otimes I_t))P &= P^T A P P^{-1}(B \otimes I_t)P \\ = I_m P^{-1}(B \otimes I_t)P &= P^{-1}(B \otimes I_t)P, \end{aligned} \tag{5.35}$$

which is a positive definite matrix. It follows that $\sigma(B) > 0$.

(Sufficiency) From (5.34) and (5.35) one sees that the $\{\lambda_i | i = 1, \cdots, m\}$ in (5.34) are eigenvalues of $B \otimes I_t$, which are all positive. Hence, $P^T(A \ltimes B)P$ is a symmetric and positive definite matrix. It follows that $A \ltimes B$ is positive definite. □

Theorem 5.7. *Assume A is a normal matrix, B is a symmetric and positive definite matrix, and $A \ltimes B = B \ltimes A$. Then $A \ltimes B$ is positive definite, if and only if A is positive definite.*

Proof. We prove it for the case of $A \succ_t B$. The other case and the generalization can also be proved similarly.

Before proving it, we need some preparations. First, we claim that if A is a normal matrix, B is a symmetric and positive definite matrix, and $A \ltimes B = B \ltimes A$, then $A \ltimes B$ is normal. In fact, we have

$$\begin{aligned}(A \ltimes B)(A \ltimes B)^T &= (B \ltimes A)(B \ltimes A)^T \ = B \ltimes A \ltimes A^T \ltimes B^T \\ &= B \ltimes A^T \ltimes A \ltimes B^T = (B \ltimes A^T)(A \ltimes B) \\ &= (A \ltimes B)^T(A \ltimes B).\end{aligned}$$

The claim is proved.

Next, we set $C = (B \otimes I_t)^{1/2}$, which is symmetric and positive definite. Then

$$C(A \ltimes B)C^{-1} = CA(B \otimes I_t)C^{-1} = CAC = C^TAC. \tag{5.36}$$

Now we are ready to prove the theorem.

(Necessity) Since $A \ltimes B$ is positive definite, $\sigma(A \ltimes B) > 0$. It follows from (5.36) that $\sigma(C^TAC) > 0$. It is also easy to verify that CAC^T is a normal matrix, it follows that C^TAC is a positive definite matrix, and so is A.

(Sufficiency) Since A is a positive definite matrix, if follows from (5.36) again that $\sigma(A \ltimes B) > 0$. According to the claim, $A \ltimes B$ is a normal matrix, and hence $A \ltimes B$ is positive definite. □

Exercises

5.1 Assume $A \in \mathcal{M}_{m\times n}$, $B \in \mathcal{M}_{p\times q}$, $\text{rank}(A) = r_A$, $\text{rank}(B) = r_B$, $\text{rank}(C) = r_c$. Prove that

(i) if $A \succ_t B$, then

$$\text{rank}\,[C + (A \ltimes B)] \le r_A + r_C.$$

(ii) if $A \prec_t B$, then

$$\text{rank}\,[C + (A \ltimes B)] \le r_At + r_C.$$

(iii) if $A \succ_t B$, and $C \in \mathcal{M}_{s\times m}$, then

$$\text{rank}\,[C(A \ltimes B)] \ge r_A + r_C - m.$$

(iv) if $A \prec_t B$, and $C \in \mathcal{M}_{s\times mt}$, then

$$\operatorname{rank}\left[C(A \ltimes B)\right] \geq r_A t + r_C - tm.$$

5.2 Prove Proposition 5.2.

5.3 Prove Proposition 5.3.

5.4 Prove Proposition 5.4.

5.5 Prove that $x = A^{+}b$ is a least square solution to $Ax = b$.

5.6 Use Theorem 5.1 to prove the weighted Moore-Penrose inverse uniquely exists, and satisfies (5.12). (Hint: set $\tilde{A} = M^{\frac{1}{2}}AN^{-\frac{1}{2}}$.)

5.7 Prove equations (5.24) and (5.25).

5.8 Assume $A \prec_t B$. Then state and prove the corresponding results parallel to those in Theorem 5.3.

5.9 Consider Example 5.2. Check the third and fourth, i.e., (5.22) and (5.23), of Theorem 5.3.

5.10 Consider Example 5.3. Check that the obtained $X = A_d$ satisfies (5.27)–(5.29) with $m = n = 4$ and $W = I_4$.

5.11 Assume $A \in \mathbb{C}_{n\times n}$ is nonsingular. Then show that both (weighted) Moore-Penrose inverse and (weighted) Drazin inverse coincide the convention inverse A^{-1}.

5.12 Let

$$A = \begin{bmatrix} 1 & 0 & 0 & 3 \\ 2 & 1 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & -1 & 1 \\ 1 & 2 & -1 \end{bmatrix}.$$

Check whether $(A \ltimes B)_d = B_d \ltimes A_d$?

5.13 Prove that a normal matrix A is positive definite, if and only if $\sigma(A) > 0$.

5.14 Use equation (5.33) to prove $A^{\frac{1}{2}}(B \otimes I_t)A^{-\frac{1}{2}}$ is normal.

5.15 Prove Corollary 5.1.

5.16 Prove Corollary 5.2.

5.17 Prove Corollary 5.3.

5.18 Given two matrices $A \in \mathcal{M}_{m\times n}$ and $B \in \mathcal{M}_{p\times q}$. Is the following equation (5.37) correct?

$$\operatorname{rank}(A \ltimes B) = \operatorname{rank}(A \rtimes B). \tag{5.37}$$

If "yes", prove it; if "no" give a counterexample.

5.19 Assume A and B are of arbitrary dimensions. Prove that Theorem 5.5 remains true.

Chapter 6

Matrix Expression of Logic

Mathematical logic is a discipline of both philosophy and natural science. From natural science point of view, it is developed by the efforts of mathematicians to reveal the essence of mathematical thinking and mathematical deduction. The basic concepts and results of mathematical logic can be found in any standard textbooks, e.g., Hamilton (1988).

The purpose of this chapter is to express logical variables, logical operators, and logical equations into matrix forms by using STP. We first introduce the matrix expression of logic. Using this form, many fundamental properties of logic can be discovered. These results are then extended to the multi-valued logic and its operations. Finally, we consider some of its applications, including logical inference.

6.1 Logic and Its Expression

A logical variable means a proposition. Usually, a proposition can either be "true" or "false". When the proposition is true, we say that the logical variable takes value "T" or "1", and when it is false, the logical variable takes value "F" or "0". We consider some simple examples.

Example 6.1. Consider the following propositions.

A: A dog has 4 legs;
B: The snow is black;
C: There is another human in the universe.

It is obvious that $A = 1$, $B = 0$. As for C, it could be either 1 or 0. But C should be one of them, though nowadays we still do not know the answer.

In classical logic a logical variable can only take value from $\{0, 1\}$. But in real world a proposition may not be described precisely by only "true" or "false". For instance, "Mr. Smith is an old man". If Mr. Smith is 20 years old, the statement is obviously "False". If Mr. Smith is 80 years old, the statement is surely "True". But if this person is 40 or 50 years old, then what can we say? It seems that we need some values between 0 ("False") and 1 ("True") to describe this statement, and hence the classical logic is not enough for analyzing such problems. Fuzzy logic allows a logical variable to take any value from internal $[0, 1]$.

Usually, we use a membership function to describe the value of a fuzzy logical variable. For instance, we may use the following membership functions to describe the statement x : "Somebody is old".

$$f(x) = \begin{cases} 0, & x \leq 20 \\ 0.01 \times (x - 20), & 20 < x \leq 40 \\ 0.2 + 0.04 \times (x - 40), & 40 < x \leq 60 \\ 0.8 + 0.01 \times (x - 60), & 60 < x \leq 80 \\ 1, & x > 80. \end{cases} \tag{6.1}$$

This function is depicted in Fig. 6.1. We refer to Liu and Liu (1996) for more details about fuzzy logic. Its applications in fuzzy control systems can be found in Zadeh and Kacprzyk (1992); Verbruggen and Babuška (1999); Wang (1996). Fuzzy logic and fuzzy control will be re-considered from the STP point of view in Chapters 8–10.

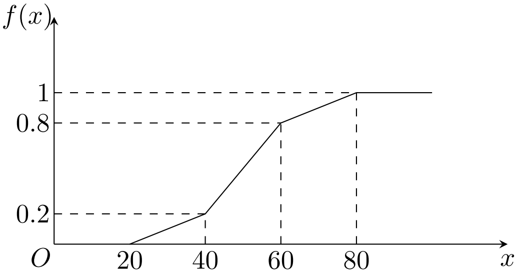


Fig. 6.1 Membership function of x

In classical logic a logical variable can take only one from two possible values 0 or 1, while in fuzzy logic a logical variables can take continuous values between 0 and 1. It is obvious that under certain circumstances using continuous values can describe a logical statement more precise than the classical two value case. But in many cases it may be too diverse and complicated to consider the continuous logic values. For instance, consider the statement x : "Mr. Smith is an old man". It is hard to tell what is the difference between $x = 0.41$ and $x = 0.42$. Hence, a precise value may not have much sense when it is used to describe a proposition. Then we may consider to quantize the continuous membership function. For instance, in the age problem, we may classify different ages into three categories: "young", "middle aged", and "old" and use "0", "0.5", and "1" for them respectively. A quantized membership function of $f(x)$ in (6.1) becomes

$$q(x) = \begin{cases} 0, & x \leq 40 \\ 0.5, & 40 < x \leq 60 \\ 1, & x > 60, \end{cases} \tag{6.2}$$

which is depicted in Fig. 6.2.

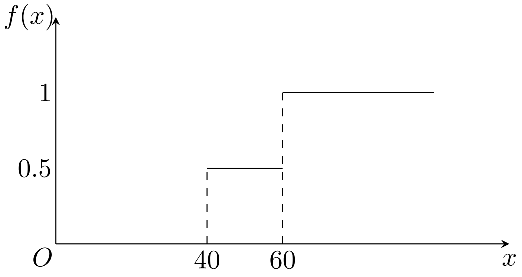


Fig. 6.2 Quantized membership function of x

In general, a logic, where a logical variable can take k different values between 0 and 1, is called the k-valued logic. When $k = 2$ it is classical logic, and when $k > 2$ it is called a multi-valued logic. Readers who are interested in multi-valued logic may refer to Gerla (2001) for more details.

Definition 6.1.

(1) The domain of (classical) logic is denoted by

$$\mathcal{D} := \{T = 1, F = 0\}. \tag{6.3}$$

A logical variable x takes value from $\mathcal{D}$, that is, $x \in \mathcal{D}$.

(2) The domain of k-valued logic is denoted by

$$\mathcal{D}_k := \left\{T = 1, \frac{k-2}{k-1}, \frac{k-3}{k-1}, \cdots, \frac{1}{k-1}, F = 0\right\}. \tag{6.4}$$

A k-valued logical variable x takes value from $\mathcal{D}_k$, that is, $x \in \mathcal{D}_k$.

(3) The domain of fuzzy logic is denoted by

$$\mathcal{D}_\infty := [0, 1]. \tag{6.5}$$

A fuzzy logical variable x takes value from $\mathcal{D}_\infty$, that is, $x \in [0, 1]$.

Next, we define the logical operators.

Definition 6.2 (Barnes and Mack, 1975). An r-ary (multi-valued, fuzzy) logical operator is a mapping $\sigma : \underbrace{D \times D \times \cdots \times D}_{r} \to D$ (correspondingly, $\underbrace{D_k \times D_k \times \cdots \times D_k}_{r} \to D_k$, $\underbrace{D_\infty \times D_\infty \times \cdots \times D_\infty}_{r} \to D_\infty$).

An r-ary logical operator can also be called a logical function with r arguments. A classical logical function is also called a Boolean function. Mathematically, they are the same. But in applications they have a mild difference. Conventionally, "operator" is mostly used for $r = 1, 2$, and an operator has its obvious logical meaning, such as "conjunction", "disjunction" etc. Meanwhile, "function" is used for more general case. Most likely, a logical function is composed of its arguments connected by operators. You may consider that a logical function is a "compounded operator".

In the rest of this section we consider the classical logic only. We first introduce some fundamental operators.

(i) Negation: A unary logical operator, denoted by $\neg$. Negation is defined as

$$\neg x = \begin{cases} 0, & x = 1 \\ 1, & x = 0. \end{cases} \tag{6.6}$$

(ii) Conjunction: A binary logical operator, denoted by $\wedge$. Conjunction is defined as

$$x \wedge y = \begin{cases} 1, & x = 1 \text{ and } y = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{6.7}$$

(iii) Disjunction: A binary logical operator, denoted by $\vee$. Disjunction is defined as

$$x \vee y = \begin{cases} 0, & x = 0 \text{ and } y = 0 \\ 1, & \text{otherwise.} \end{cases} \tag{6.8}$$

(iv) Conditional: A binary logical operator, denoted by →. Conditional is defined as

$$x \to y = \begin{cases} 0, & x = 1 \text{ and } y = 0 \\ 1, & \text{otherwise.} \end{cases} \tag{6.9}$$

(v) Biconditional: A binary logical operator, denoted by ↔. Biconditional is defined as

$$x \leftrightarrow y = \begin{cases} 1, & x = y \\ 0, & \text{otherwise.} \end{cases} \tag{6.10}$$

A conventional way to depict the values of an operator is using a table, called the truth table. For instance, for "negation", we have Table 6.1.

Table 6.1 Truth table for "negation"

x	$\neg x$
1	0
0	1

Similarly, we can have the truth table for "conjunction", "disjunction", "conditional", and "biconditional" etc. respectively as in Table 6.2. In Table 6.2 "$\bar{\vee}$" is "exclusive or", "↑" is "not and","↓" is "not or" (Rade and Westergren, 1998).

Table 6.2 Truth table for $\wedge$, $\vee$, $\to$, $\leftrightarrow$, $\bar{\vee}$, $\uparrow$, $\downarrow$

x	y	$x \wedge y$	$x \vee y$	$x \to y$	$x \leftrightarrow y$	$x \bar{\vee} y$	$x \uparrow y$	$x \downarrow y$
1	1	1	1	1	1	0	0	0
1	0	0	1	0	0	1	1	0
0	1	0	1	1	0	1	1	0
0	0	0	0	1	1	0	1	1

The truth values of a logical function can easily be obtained from the truth tables of basic operators. We use an example to depict this.

Example 6.2.

(1) Let $p = x \vee (\neg y)$. Then the truth table of p is shown in Table 6.3.
(2) Let $q = (x \wedge y) \leftrightarrow (\neg z)$. Then the truth table of q is shown in Table 6.4.

For statement ease, we define the truth vector of a Boolean function.

Definition 6.3. Let $f(x_1, \cdots, x_k)$ be a Boolean function. Denote the column of f in its truth table by T_f, and call it the truth vector of f.

Table 6.3 Truth table for p

x	y	$\neg y$	$p = x \vee (\neg y)$
1	1	0	1
1	0	1	1
0	1	0	0
0	0	1	1

Table 6.4 Truth table for q

x	y	z	$x \wedge y$	$\neg z$	$q = x \wedge y \leftrightarrow (\neg z)$
1	1	1	1	0	0
1	1	0	1	1	1
1	0	1	0	0	1
1	0	0	0	1	0
0	1	1	0	0	1
0	1	0	0	1	0
0	0	1	0	0	1
0	0	0	0	1	0

Example 6.3.

(1) Consider $f(x,y) = x \vee (\neg y)$. According to Table 6.3,

$$T_f = [1\ 1\ 0\ 1]^T.$$

(2) Consider $g(x,y,z) = x \wedge y \leftrightarrow (\neg z)$. According to Table 6.4,

$$T_g = [0\ 1\ 1\ 0\ 1\ 0\ 1\ 0]^T.$$

We introduce some new notations:

(i)

$$\delta_n^i := \mathrm{Col}_i(I_n), \quad i = 1, \cdots, n.$$

(ii)

$$\Delta_n := \mathrm{Col}(I_n), \text{ when } n = 2, \ \Delta := \Delta_2.$$

(iii) $L \in \mathcal{M}_{m\times n}$ is called a logical matrix if $\mathrm{Col}(L) \subset \Delta_m$. The set of $m \times n$ logical matrices is denoted by $\mathcal{L}_{m\times n}$.

(vi) Let $L \in \mathcal{L}_{m\times n}$. Then L can be expressed as

$$L = \left[\delta_m^{i_1}\ \delta_m^{i_2}\ \cdots\ \delta_m^{i_n}\ \right].$$

For notational compactness, we denote L briefly by

$$L = \delta_m[i_1\ i_2\ \cdots\ i_n].$$

To use matrix expression of logic, we identify

$$T = 1 \sim \delta_2^1, \quad F = 0 \sim \delta_2^2,$$

and call it the vector form of logic. Then in vector form an r-ary Boolean function f becomes a mapping $f : \Delta^r \to \Delta$.

Definition 6.4. Let $f(x_1, \cdots, x_r)$ be an r-ary Boolean function. $L_f \in \mathcal{L}_{2\times 2^r}$ is called the structure matrix of f, if in vector form we have

$$f(x_1, \cdots, x_r) = L_f \ltimes_{i=1}^r x_i. \tag{6.11}$$

Proposition 6.1. *Let $f(x_1, \cdots, x_r)$ be an r-ary Boolean function. Then there exists a unique structure matrix $L_f \in \mathcal{L}_{2\times 2^r}$ such that (6.11) holds.*

Proof. Assume the truth vector of f is T_f. Construct L_f as follows:

$$\mathrm{Row}_1(L_f) := T_f^T; \quad \mathrm{Row}_2 = \neg(\mathrm{Row}_1).$$

Here we use $\neg\,\mathrm{Row}_1$ for taking negation on every elements of Row_1. It follows from the construction of truth table and the definition of semi-tensor product that the constructed L_f satisfies (6.11) and such structure matrix must be unique. □

Using Proposition 6.1, the structure matrices of some fundamental operators are obtained as

$$\begin{aligned} M_\neg &:= M_n = \delta_2[2\ 1];\\ M_\vee &:= M_d = \delta_2[1\ 1\ 1\ 2];\\ M_\wedge &:= M_c = \delta_2[1\ 2\ 2\ 2];\\ M_\rightarrow &:= M_i = \delta_2[1\ 2\ 1\ 1];\\ M_\leftrightarrow &:= M_e = \delta_2[1\ 2\ 2\ 1]. \end{aligned} \tag{6.12}$$

Remark 6.1.

(1) Note that the structure matrix of a logical function is unique. In fact, it is clear from the construction that there is a one-to-one correspondence between logical functions and their structure matrices. Hence determining a logical function is equivalent to determining its structure matrix.

(2) From Proposition 6.1 one also sees that there is a one-to-one correspondence between truth vectors and structure matrices.

In most applications it is not convenient to construct the structure matrix of a logical function by using its truth table. We give a method to construct it. To begin with, we need a tool, called the power-reducing matrix, which is defined as

$$M_r := \delta_4[1\ 4]. \tag{6.13}$$

The following lemma shows that the power-reducing matrix can reduce the power of a logical variable. It can be proved by a straightforward computation.

Lemma 6.1. *Given a logical variable $x \in \Delta$. Then*

$$x^2 = M_r x. \tag{6.14}$$

Lemma 6.2. *A logical function $f(x_1, \cdots, x_k)$ can be expressed in vector form as*

$$f(x_1, \cdots, x_k) = L x_1^{n_1} \cdots x_k^{n_k}. \tag{6.15}$$

Proof. First, using the structure matrices of $\neg$, $\wedge$, $\vee$, etc., we can express the function into a product as

$$f(x_i, \cdots, x_k) = \ltimes_{j=1}^{r} \xi_j, \tag{6.16}$$

where ξ_j is either a structure matrix of a unary or binary logical operator, or an argument x_i. Assume there are two adjacent factors as $x_i M_\sigma$, where x_i is an argument and M_σ is the structure matrix of operator σ. Using Theorem 2.10, we can swap two factors as

$$x_i M_\sigma = [I_2 \otimes M_\sigma]\, x_i.$$

Using this technique, we can move all the arguments to the rear of the product. Then use the swap matrix

$$x_i^p x_j^q = W_{[2^q, 2^p]} x_j^q x_i^p$$

we can re-arrange the order of arguments into the required order. □

We give a simple example to depict this:

Example 6.4. let $f(x, y) = (x \vee y) \to (x \wedge y)$. Then in vector form we have

$$\begin{aligned} f(x,y) &= M_i (x \vee y)(x \wedge y) \\ &= M_i M_d x y M_c x y \\ &= M_i M_d (I_4 \otimes M_c) x y x y \\ &= M_i M_d (I_4 \otimes M_c) x W_{[2]} x y^2 \\ &= M_i M_d (I_4 \otimes M_c) \left(I_2 \otimes W_{[2]}\right) x^2 y^2. \end{aligned}$$

Using Lemmas 6.1 and 6.2, we can give an alternative proof for Proposition 6.1. In fact, starting from (6.15), we can use (6.14) to reduce the powers of each x_i to 1. After some additional swaps, (6.11) can be obtained.

Example 6.5. (Continuing Example 6.4)

$$\begin{aligned} f(x,y) &= M_i M_d (I_4 \otimes M_c)\left(I_2 \otimes W_{[2]}\right) x^2 y^2 \\ &= M_i M_d (I_4 \otimes M_c)\left(I_2 \otimes W_{[2]}\right) M_r x M_r y \\ &= M_i M_d (I_4 \otimes M_c)\left(I_2 \otimes W_{[2]}\right) M_r \left(I_2 \otimes M_r\right) xy. \end{aligned}$$

Finally, we conclude that

$$f(x,y) = (x \vee y) \rightarrow (x \wedge y) = Lxy,$$

where

$$L = M_i M_d (I_4 \otimes M_c)\left(I_2 \otimes W_{[2]}\right) M_r \left(I_2 \otimes M_r\right) = \delta_2[1\ 2\ 2\ 1].$$

6.2 General Structure of Logical Operators

According to Proposition 6.1 and Remark 6.1, the number of r-ary logical functions is the same as the number of truth vectors. It is obvious that if there are r variables and each variable can only take two possible values, then there are 2^r different value combinations of variables. Moreover, each variable value combination may correspond to two function values. Hence there are 2^{2^r} different truth vectors. That is, there are 2^{2^r} different r-ary logical functions.

Remark 6.2.

(1) Let $s < r$. Then an s-ary logical function can be considered as a special r-ary logical function, which is independent of $r - s$ logical variables. In constructing logical functions this observation should be taken into consideration.

(2) Consider k-valued logic. The number of r-ary functions is k^{k^r}.

The following theorem is very useful in recovering the logical form of a logical function from its structure matrix.

Theorem 6.1. *Assume $f(x_1, \cdots, x_r)$ is an r-ary operator with its structure matrix $M_f \in \mathcal{L}_{2\times 2^r}$, $r \geq 2$. Split M_f into two equal-size blocks as*

$$M_f = [B_1, B_2].$$

Then

$$f(x_1,\cdots,x_r)=(x_1\wedge f_1(x_2,\cdots,x_r))\vee(\neg x_1\wedge f_2(x_2,\cdots,x_r)), \tag{6.17}$$

where $f_i(x_2,\cdots,x_r)$ *has* B_i *as its structure matrix,* $i=1,2$.

Proof. First, we prove

$$f(x_1,\cdots,x_r)=(x_1\wedge f(1,x_2,\cdots,x_r))\vee(\neg x_1\wedge f(0,x_2,\cdots,x_r)). \tag{6.18}$$

If $x_1=1$,

$$\begin{aligned} RHS &= (1\wedge f(1,x_2,\cdots,x_r))\vee(0\wedge f(0,x_2,\cdots,x_r))\\ &= f(1,x_2,\cdots,x_r)\vee 0\\ &= f(1,x_2,\cdots,x_r)=LHS, \end{aligned}$$

if $x_1=0$,

$$\begin{aligned} RHS &= (0\wedge f(1,x_2,\cdots,x_r))\vee(1\wedge f(0,x_2,\cdots,x_r))\\ &= 0\vee f(0,x_2,\cdots,x_r)\\ &= f(0,x_2,\cdots,x_r)=LHS, \end{aligned}$$

thus (6.18) follows.

Denote by

$$\begin{aligned} f_1(x_2,\cdots,x_r) &= f(1,x_2,\cdots,x_r),\\ f_2(x_2,\cdots,x_r) &= f(0,x_2,\cdots,x_r), \end{aligned}$$

it is easy to verify that $f_i(x_2,\cdots,x_r)$ has B_i as its structure matrix. □

Using (6.17) repetitively, we finally can express any logical function as an expression compounded by ¬, ∧, and ∨ with logical variables. To represent this form in a condensed form we introduce a new notation, which comes from cryptograph (refer to Chapter 11):

$$x^1:=x,\quad x^0:=\neg x,\quad x\in\mathcal{D}.$$

Then we state this expression as a corollary.

Corollary 6.1. *Assume* $f(x_1,\cdots,x_r)$ *is an* r*-ary Boolean function. Then* f *can be expressed as*

$$\begin{aligned} f(x_1,\cdots,x_r) = &\vee_{i_1=0}^{1}\vee_{i_2=0}^{1}\cdots\vee_{i_{r-1}=0}^{1}\\ &\left[x_1^{i_1}\wedge x_2^{i_2}\wedge\cdots\wedge x_{r-1}^{i_{r-1}}\wedge\sigma_{i_1,i_2,\cdots,i_{r-1}}(x_r)\right], \end{aligned} \tag{6.19}$$

where $\sigma_{i_1,i_2,\cdots,i_{r-1}}$ *are certain one-ary logical operators.*

Table 6.5 Unary operators

P	σ_0^1	σ_1^1	σ_2^1	σ_3^1
	F	$\neg$	$\equiv$	T
1	0	0	1	1
0	0	1	0	1

Table 6.6 Binary operators

x	y	σ_0^2	σ_1^2	σ_2^2	σ_3^2	σ_4^2	σ_5^2	σ_6^2	σ_7^2
		F	$\downarrow$	$-^*$	$\neg_1$	$-$	$\neg_2$	$\bar{\vee}$	$\uparrow$
1	1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	1	1	1	1
0	1	0	0	1	1	0	0	1	1
0	0	0	1	0	1	0	1	0	1
x	y	σ_8^2	σ_9^2	σ_{10}^2	σ_{11}^2	σ_{12}^2	σ_{13}^2	σ_{14}^2	σ_{15}^2
		$\wedge$	$\leftrightarrow$	y	$\rightarrow$	x	$\rightarrow^*$	$\vee$	T
1	1	1	1	1	1	1	1	1	1
1	0	0	0	0	0	1	1	1	1
0	1	0	0	1	1	0	0	1	1
0	0	0	1	0	1	0	1	0	1

(6.19) is called the disjunctive normal form of Boolean function f.

Next, we consider all the logical operators with ary $r = 1$ or $r = 2$.

Assume $r = 1$. In general, we have 4 logical operators, which are listed in Table 6.5.

Here "F" is the constant "False" operator, and "T" is the constant "True" operator, "$\neg$" is the negation, and "$\equiv$" is the identity operator.

The structure matrices of these four unary operators are as follows.

$$M_F = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}; \quad M_n = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}; \quad M_{id} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad M_T = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}. \tag{6.20}$$

Next, we assume $r = 2$. In addition to the four operators ($\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$) the following 3 are also commonly used (Rade and Westergren, 1998).

(i) EOR (exclusive or), $x\bar{\vee}y$, it is true whenever either x or y, but not both are true;

(ii) NAND (not and), $x \uparrow y$, defined by $x \uparrow y = \neg(x \wedge y)$.

(iii) NOR (not or), $x \downarrow y$, defined by $x \downarrow y = \neg(x \vee y)$.

In addition to these 7 commonly used operators, we still have $2^{2^2} - 7 = 9$ other 2-ary (binary) operators. We listed these 16 operators in Table 6.6.

Here $\sigma_0^2 = F$, $\sigma_{15}^2 = T$ are two constant operators; $\sigma_3^2 = \neg x$, $\sigma_5^2 = \neg y$, $\sigma_{10}^2 = x$ and $\sigma_{12}^2 = y$ are essentially unary operators. The following is some new binary operators.

- EOR (σ_6^2, exclusive or) ($x\bar{\vee}y$):

$$x\bar{\vee}y = \neg(x \leftrightarrow y); \tag{6.21}$$

- NAND (σ_7^2, not and) ($x \uparrow y$):

$$x \uparrow y = \neg(x \wedge y); \tag{6.22}$$

- NOR (σ_1^2, not or) ($x \downarrow y$):

$$x \downarrow y = \neg(x \vee y); \tag{6.23}$$

- NC (σ_4^2, not conditional) ($x - y$):

$$x - y = \neg(x \rightarrow y); \tag{6.24}$$

- NINVC (σ_2^2, not inverse conditional) ($x -^* y$):

$$x -^* y = \neg(y \rightarrow x); \tag{6.25}$$

- INVC (σ_{13}^2, inverse conditional) $x \rightarrow^* y$:

$$x \rightarrow^* y = y \rightarrow x. \tag{6.26}$$

The equivalence of (6.26) comes from its definition. The equivalences (6.21)–(6.25) will be proved later.

For convenience, we use σ_j^i to represent all the operators, where the superscript i is the ary of the operator, and the subscript j is the order of the operator. It is interesting that when the order, counting from zero, is converted to binary number, it is exactly the truth vector of the operator. Hence the structure matrix of s_j^i can easily be constructed.

Example 6.6. Consider $\neg_2$, which is the negation of the second variable. Its alternative notation is σ_5^2. Since $5 = 101 = 0101$, we have

$$V_{\neg_2} = [0\ 1\ 0\ 1], \quad \text{or} \ \ M_{\neg_2} = \delta_2[2\ 1\ 2\ 1].$$

6.3 Fundamental Properties of Logical Operators

According to the structure of truth tables, we have the following conjugate properties.

Proposition 6.2. *Given an r-ary operator σ_a^r, its negation operator is $\sigma_{2^{2^r}-a-1}^r$. That is,*

$$\neg\sigma_a^r(x, y) = \sigma_{2^{2^r}-a-1}^r(x, y). \tag{6.27}$$

***Proof*.** Since $2^{2^r} - 1 = \underbrace{1\ 1\ \cdots\ 1}_{2^r}$. Expressing a positive integer a into 2^r digital binary number (adding some zeros in the front if necessary), say, it is

$$a = a_1 a_2 \cdots a_{2^r}.$$

Then $[a_1, a_2, \cdots, a_{2^r}]^T$ is the truth vector of σ_a^r. In binary form we have $2^{2^r} - 1 - a = [(1-a_1), (1-a_2), \cdots, (1-a_r)]$. Hence we know that the truth vector of $\sigma_{2^{2^r}-a-1}^r$ is $(1-a_1, 1-a_2, \cdots, 1-a_r)^T$. That implies (6.27). □

Remark 6.3. Using Proposition 6.2, we can prove (6.21)–(6.25) immediately.

Definition 6.5.

(1) Two logical expressions are said to be logically equivalent, if for any particulary chosen values of logical variables from $\mathcal{D}$ the two expressions have the same value. If $f(x_1, \cdots, x_k)$ and $g(x_1, \cdots, x_k)$ are logically equivalent, it is denoted as

$$f(x_1, \cdots, x_k) \Leftrightarrow g(x_1, \cdots, x_k).$$

(2) Two logical expressions are said to be absolutely logically equivalent, if for any particulary chosen values of logical variables from $\mathcal{D}_\infty$ the two expressions have the same value. If $f(x_1, \cdots, x_k)$ and $g(x_1, \cdots, x_k)$ are absolutely logically equivalent, it is denoted as

$$f(x_1, \cdots, x_k) \tilde{\Leftrightarrow} g(x_1, \cdots, x_k).$$

Proposition 6.3. *Assume two logical expressions f and g have same arguments, moreover, every argument appears to f (or g) precisely once. Then the logical equivalence of f and g is equivalent to absolutely logically equivalence.*

***Proof*.** Assume f and g are logically equivalent. Recall the proof of Lemma 6.2, one sees easily that for $x_i \in \mathcal{D}_\infty$ (6.15) remains true. Denote the L in (6.15) for f and g by L_f and L_g respectively. Now by the assumption, $n_i = 1$, $i = 1, \cdots, k$. It follows that L_f and L_g are the structure matrices of f and g as $x_i \in \mathcal{D}$. Since f and g are logically equivalent, we have $L_f = L_g$. The conclusion follows. □

Using Proposition 6.3, we can have the following absolutely logically equivalent expressions.

Proposition 6.4. *The followings are absolutely logically equivalent.*

$$\neg\neg x \tilde{\Leftrightarrow} x; \tag{6.28}$$

$$(x \wedge y) \wedge z \tilde{\Leftrightarrow} x \wedge (y \wedge z); \tag{6.29}$$

$$(x \vee y) \vee z \tilde{\Leftrightarrow} z \vee (y \vee z); \tag{6.30}$$

$$\neg(x \wedge y) \tilde{\Leftrightarrow} \neg x \vee \neg y; \tag{6.31}$$

$$\neg(x \vee y) \tilde{\Leftrightarrow} \neg x \wedge \neg y; \tag{6.32}$$

$$x \rightarrow y \tilde{\Leftrightarrow} \neg x \vee y; \tag{6.33}$$

$$\neg(x \rightarrow y) \tilde{\Leftrightarrow} x \wedge \neg y; \tag{6.34}$$

$$x \rightarrow y \tilde{\Leftrightarrow} \neg y \rightarrow \neg x; \tag{6.35}$$

$$x \rightarrow (y \rightarrow z) \tilde{\Leftrightarrow} (x \wedge y) \rightarrow z; \tag{6.36}$$

$$\neg(x \leftrightarrow y) \tilde{\Leftrightarrow} x \leftrightarrow \neg y. \tag{6.37}$$

Proof. We prove (6.35) only. The proves of others are similar and we leave them to the reader. According to Proposition 6.3, we have only to prove they are logically equivalent.

$$\begin{aligned} RHS &= M_i M_n y M_n x = M_i M_n (I_2 \otimes M_n) yx \\ &= M_i M_n (I_2 \otimes M_n) W_{[2]} xy. \end{aligned}$$

Since

$$\begin{aligned} & M_i M_n (I_2 \otimes M_n) W_{[2]} \\ &= \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} = M_i, \end{aligned}$$

(6.35) follows. □

Definition 6.6. An r-ary operator is said to be symmetric if

$$M_\sigma x_1 x_2 \cdots x_k = M_\sigma x_{\lambda(1)} x_{\lambda(2)} \cdots x_{\lambda(k)}, \quad \forall\, \lambda \in \mathbf{S}_k. \tag{6.38}$$

Recall that $\mathbf{S}_k$ is the kth order symmetric group.

Proposition 6.5. *A binary operator, σ is symmetric, if and only if in its truth vector $T_\sigma = [s_1\ s_2\ s_3\ s_4]^T$ satisfies*

$$s_2 = s_3.$$

Proof. Note that the structure matrix of σ is

$$M_\sigma = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 \\ 1-s_1 & 1-s_2 & 1-s_3 & 1-s_4 \end{bmatrix}.$$

Then

$$\begin{aligned} \sigma(x,y) =& M_\sigma xy = M_\sigma W_{[2]} yx \\ =& \begin{bmatrix} s_1 & s_3 & s_2 & s_4 \\ 1-s_1 & 1-s_3 & 1-s_2 & 1-s_4 \end{bmatrix} yx \\ =& M_\sigma yx = \sigma(y,x). \end{aligned}$$

□

Example 6.7. Consider the binary operators in Table 6.6. According to Proposition 6.5, we have that F, $\downarrow$, $\bar{\vee}$, $\uparrow$, $\wedge$, $\leftrightarrow$, $\vee$, and T are symmetric, and the others are not.

We give some useful logical equivalences as follows.

Proposition 6.6. *The followings are logically equivalent.*

$$x \vee x \Leftrightarrow x; \tag{6.39}$$

$$x \wedge x \Leftrightarrow x; \tag{6.40}$$

$$y \vee (x \wedge \neg x) \Leftrightarrow y; \tag{6.41}$$

$$y \wedge (x \vee \neg x) \Leftrightarrow y; \tag{6.42}$$

$$x \wedge (y \vee z) \Leftrightarrow (x \wedge y) \vee (x \wedge z); \tag{6.43}$$

$$x \vee (y \wedge z) \Leftrightarrow (x \vee y) \wedge (x \vee z); \tag{6.44}$$

$$x \leftrightarrow y \Leftrightarrow (x \rightarrow y) \wedge (y \rightarrow x); \tag{6.45}$$

$$x \leftrightarrow y \Leftrightarrow (x \wedge y) \vee (\neg x \wedge \neg y). \tag{6.46}$$

Proof. We prove (6.41) only. Assume $y = [\delta, 1-\delta]^T$, $x = [\mu, 1-\mu]^T$. Then

$$\begin{aligned} LHS =& M_\vee y M_\wedge x M_\neg x \\ =& \begin{bmatrix} 1\,1\,1\,0 \\ 0\,0\,0\,1 \end{bmatrix} \begin{bmatrix} \delta \\ 1-\delta \end{bmatrix} \begin{bmatrix} 1\,0\,0\,0 \\ 0\,1\,1\,1 \end{bmatrix} \begin{bmatrix} \mu \\ 1-\mu \end{bmatrix} \begin{bmatrix} 0\,1 \\ 1\,0 \end{bmatrix} \begin{bmatrix} \mu \\ 1-\mu \end{bmatrix} \\ =& \begin{bmatrix} 1\,1\,1\,0 \\ 0\,0\,0\,1 \end{bmatrix} \begin{bmatrix} \delta \\ 1-\delta \end{bmatrix} \begin{bmatrix} \mu(1-\mu) \\ \mu^2 + \mu(1-\mu) + (1-\mu)^2 \end{bmatrix} \\ =& \begin{bmatrix} 1\,1\,1\,0 \\ 0\,0\,0\,1 \end{bmatrix} \begin{bmatrix} \delta[\mu(1-\mu)] \\ \delta[\mu^2 + \mu(1-\mu) + (1-\mu)^2] \\ (1-\delta)[\mu(1-\mu)] \\ (1-\delta)[\mu^2 + \mu(1-\mu) + (1-\mu)^2] \end{bmatrix} \\ =& \begin{bmatrix} \delta + (1-\delta)\mu(1-\mu) \\ (1-\delta)[\mu^2 + \mu(1-\mu) + (1-\mu)^2] \end{bmatrix}. \end{aligned}$$

Since $\mu \in \mathcal{D}$, we have $LHS = [\delta, 1-\delta]^T = y$. □

It is easy to verify that (6.39)–(6.46) are not absolutely logically equivalent. For instance, consider (6.41), in the above proof taking $\mu = 0.5$, then $LHS = \begin{bmatrix} 0.25 + 0.75\delta \\ 0.75 - 0.75\delta \end{bmatrix} \neq y$.

The following proposition is very useful. We leave the proof to the reader.

Proposition 6.7 (De Morgan's Law).

1.
$$\neg(x \wedge y) = (\neg x) \vee (\neg y). \tag{6.47}$$

2.
$$\neg(x \vee y) = (\neg x) \wedge (\neg y). \tag{6.48}$$

Definition 6.7.

(1) A logical expression is called a tautology, if it is always "true" no matter what values the arguments take.

(2) A logical expression is called a contradiction, if it is always "false" no matter what values the arguments take.

(3) Let L_1 and L_2 be two logical expressions. If $L_1 \to L_2$ is a tautology, then we say that L_1 tautologically implicates L_2, denoted by $L_1 \Rightarrow L_2$.

Proposition 6.8. *$L_1 \Rightarrow L_2$, if and only if, when $L_2 = \delta_2^2$, $L_1 = \delta_2^2$.*

Proof. (Sufficiency) Assume $L_2 = \delta_2^1$, and $L_1 = \begin{bmatrix} \alpha \\ 1-\alpha \end{bmatrix}$. Then

$$L_1 \to L_2 = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ 1-\alpha \end{bmatrix} \delta_2^1 = \delta_2^1.$$

Assume $L_2 = \delta_2^2$. Then according to the assumption, $L_1 = \delta_2^2$. Hence

$$L_1 \to L_2 = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \delta_2^2 \delta_2^2 = \delta_2^1.$$

(Necessary) Assume $L_2 = \delta_2^2$ but $L_1 = \delta_2^1$. Then

$$L_1 \to L_2 = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \delta_2^1 \delta_2^2 = \delta_2^2,$$

which means $L_1 \to L_2$ is not a tautology. □

Note that the physical meaning of Proposition 6.8 is clear: As long as L_1 is "true", L_2 must be "true".

The following tautological implications can be proved by using Proposition 6.8.

Proposition 6.9. *The followings are tautological implications.*

$$x \wedge y \Rightarrow x; \tag{6.49}$$

$$x \wedge y \Rightarrow y; \tag{6.50}$$

$$x \Rightarrow x \vee y; \tag{6.51}$$

$$y \Rightarrow x \vee y; \tag{6.52}$$

$$\neg x \Rightarrow x \rightarrow y; \tag{6.53}$$

$$y \Rightarrow x \rightarrow y; \tag{6.54}$$

$$\neg(x \rightarrow y) \Rightarrow x; \tag{6.55}$$

$$\neg(x \rightarrow y) \Rightarrow \neg y; \tag{6.56}$$

$$\neg x \wedge (x \vee y) \Rightarrow y; \tag{6.57}$$

$$x \wedge (x \rightarrow y) \Rightarrow y; \tag{6.58}$$

$$\neg y \wedge (x \rightarrow y) \Rightarrow \neg x; \tag{6.59}$$

$$(x \rightarrow y) \wedge (y \rightarrow z) \Rightarrow x \rightarrow z; \tag{6.60}$$

$$(x \vee y) \wedge (x \rightarrow z) \wedge (y \rightarrow z) \Rightarrow z. \tag{6.61}$$

***Proof*.** We prove (6.61) only. Using Proposition 6.8 and assuming the right hand side is "false", i.e., $z = \delta_2^2$, we check the left hand side.

$$\begin{aligned}
&(x \vee y) \wedge (x \rightarrow z) \wedge (y \rightarrow z)\\
&= M_\wedge M_\vee xy M_\wedge M_\rightarrow xz M_\rightarrow yz\\
&= \begin{bmatrix} 1\ 0\ 0\ 0 \\ 0\ 1\ 1\ 1 \end{bmatrix} \begin{bmatrix} 1\ 1\ 1\ 0 \\ 0\ 0\ 0\ 1 \end{bmatrix} \begin{bmatrix} p \\ 1-p \end{bmatrix} \begin{bmatrix} q \\ 1-q \end{bmatrix} \begin{bmatrix} 1\ 0\ 0\ 0 \\ 0\ 1\ 1\ 1 \end{bmatrix}\\
&\quad \begin{bmatrix} 1\ 0\ 1\ 1 \\ 0\ 1\ 0\ 0 \end{bmatrix} \begin{bmatrix} p \\ 1-p \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1\ 0\ 1\ 1 \\ 0\ 1\ 0\ 0 \end{bmatrix} \begin{bmatrix} q \\ 1-q \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}\\
&= \begin{bmatrix} (p+q-pq)(1-p)(1-q) \\ (p+q-pq)^2 + (1-p)^2(1-q)^2 + (p+q-pq)(1-p)(1-q) \end{bmatrix}.
\end{aligned}$$

We need to check four cases: (i) $p = 0, q = 0$, (ii) $p = 0, q = 1$, (iii) $p = 1, q = 0$, (iv) $p = 1, q = 1$. No matter which case was chosen. The result is the same as δ_2^2. □

Finally, we can use their structure matrices to prove some logical equivalences on "EOR", "NAND", and "NOR". The proves are similar, and we leave them to the reader.

Proposition 6.10. *The followings are logical equivalences.*

$$x\bar{\vee}y \Leftrightarrow y\bar{\vee}x; \tag{6.62}$$

$$(x\bar{\vee}y)\bar{\vee}z \Leftrightarrow x\bar{\vee}(y\bar{\vee}z); \tag{6.63}$$

$$x \wedge (y\bar{\vee}z) \Leftrightarrow (x \wedge y)\bar{\vee}(x \wedge z); \tag{6.64}$$

$$x\bar{\vee}y \Leftrightarrow (x \wedge \neg y) \vee (\neg x \wedge y); \tag{6.65}$$

$$x\bar{\vee}y \Leftrightarrow \neg(x \leftrightarrow y); \tag{6.66}$$

$$x \uparrow y \Leftrightarrow y \uparrow x; \tag{6.67}$$

$$x \downarrow y \Leftrightarrow y \downarrow x; \tag{6.68}$$

$$x \uparrow (y \uparrow z) \Leftrightarrow \neg x \vee (y \wedge z); \tag{6.69}$$

$$(x \uparrow y) \uparrow z \Leftrightarrow (x \wedge y) \vee \neg z; \tag{6.70}$$

$$x \downarrow (y \downarrow z) \Leftrightarrow \neg x \wedge (y \vee z); \tag{6.71}$$

$$(x \downarrow y) \downarrow z \Leftrightarrow (x \vee y) \wedge \neg z. \tag{6.72}$$

6.4 Logical System and Logical Inference

Assume two logical expressions $f(x_1, \cdots, x_k)$ and $g(x_1, \cdots, x_k)$ are logically equivalent. That is,

$$f(x_1, \cdots, x_k) \Leftrightarrow g(x_1, \cdots, x_k).$$

We can simply use the conventional expression as

$$f(x_1, \cdots, x_k) = g(x_1, \cdots, x_k).$$

So for logical expressions "$\Leftrightarrow$" is the same as "$=$".

Definition 6.8. A static logical system (briefly, logical system) is expressed as

$$\begin{cases} f_1(x_1, x_2, \cdots, x_k) = c_1, \\ f_2(x_1, x_2, \cdots, x_k) = c_2, \\ \vdots \\ f_m(x_1, x_2, \cdots, x_k) = c_m, \end{cases} \tag{6.73}$$

where f_i, $i = 1, \cdots, m$ are logical functions, x_i, $i = 1, \cdots, n$ are logical arguments (unknowns), and c_i, $i = 1, \cdots, m$ are logical constants. A set of logical constants d_i, $i = 1, \cdots, n$, which make

$$x_i = d_i, \quad i = 1, \cdots, n, \tag{6.74}$$

satisfy (6.73), is said to be a solution of the logical system (6.73).

In this section we first consider how to solve the logical system (6.73). Assume the structure matrix of f_i is M_i, $i = 1, \cdots, m$. Then in vector form (6.73) can be expressed as

$$\begin{cases} M_1 \ltimes_{i=1}^n x_i = \delta_2^{r_1}, \\ M_2 \ltimes_{i=1}^n x_i = \delta_2^{r_2}, \\ \vdots \\ M_m \ltimes_{i=1}^n x_i = \delta_2^{r_m}, \end{cases} \tag{6.75}$$

where $r_i \in \{1, 2\}$, $i = 1, \cdots, m$.

To further simplify (6.75), we need some preparations.

Proposition 6.11. *Assume*

$$\begin{cases} Y = M_y \ltimes_{i=1}^n x_i \\ Z = M_z \ltimes_{i=1}^n x_i, \end{cases} \tag{6.76}$$

where $M_y \in \mathcal{L}_{p\times 2^n}$ and $M_z \in \mathcal{L}_{q\times 2^n}$. Then

$$YZ = (M_y * M_z) \ltimes_{i=1}^n x_i. \tag{6.77}$$

(Here $*$ is the Khatri-Rao product. We refer to Chapter 1 for the definition.)

Proof. First of all, using the properties of semi-tensor product and the power-reducing matrix, we can prove that

$$YZ = M_y \ltimes_{i=1}^n x_i M_z \ltimes_{i=1}^n x_i = M_{yz} \ltimes_{i=1}^n x_i, \tag{6.78}$$

where $M_{yz} \in \mathcal{L}_{pq\times 2^n}$. In fact, using Lemmas 6.1 and 6.2, the same computation process as for structure matrix of a logical function yields M_{yz}.

Assume $\ltimes_{i=1}^n x_i = \delta_{2^n}^r$. Then $Y = \mathrm{Col}_r(M_y)$ and $Z = \mathrm{Col}_r(M_z)$. That is,

$$\mathrm{Col}_r(M_{yz}) = \mathrm{Col}_r(M_y) \ltimes \mathrm{Col}_r(M_z) = \mathrm{Col}_r(M_y) \otimes \mathrm{Col}_r(M_z).$$

Now $1 \le r \le 2^n$ is arbitrary, the conclusion follows. □

Using Proposition 6.11, system (6.75) can be converted into the following form:

$$M_1 * M_2 * \cdots * M_m \ltimes_{i=1}^{n} x_i = \delta_{2^n}^{r}, \tag{6.79}$$

where

$$r = \sum_{i=1}^{n-1} (r_i - 1) 2^{n-i} + r_n.$$

In other words,

$$[r_1 - 1,\ r_2 - 1,\ \cdots,\ r_n - 1]$$

is the binary form of $r - 1$. From this observation one sees easily that system (6.75) is equivalent to

$$Mx = \delta_{2^n}^{r}, \tag{6.80}$$

where

$$M = M_1 * M_2 * \cdots * M_m, \quad x = \ltimes_{i=1}^{n} x_i.$$

Note that $x = \ltimes_{i=1}^{n} x_i \in \Delta_{2^n}$. System (6.80) can be solved immediately. We summarize the above arguments as the following proposition.

Proposition 6.12. *$x = \delta_{2^n}^{k}$ is the solution of (6.80), if and only if,*

$$\mathrm{Col}_k(M) = \delta_{2^n}^{r}. \tag{6.81}$$

Finally, we need one more tool in solving a logical system. In general, an equation of f_i in system (6.73) may not involve some arguments. Then how to get (6.75)? For instance, we consider the following system

$$\begin{cases} x_1 \wedge x_2 = 0 \\ x_2 \vee x_3 = 1 \\ x_3 \leftrightarrow x_1 = 1. \end{cases} \tag{6.82}$$

To get the component-wise algebraic form (6.75), we have to add some fabricated arguments to each equation. We introduce the following dummy matrix.

$$M_u = \delta_2[1\ 1\ 2\ 2]. \tag{6.83}$$

A straightforward computation shows the following proposition.

Proposition 6.13. *In vector form we have*

$$M_u x y = x. \tag{6.84}$$

It is obvious that the dummy matrix can be used to add fabricated arguments to each equation if necessary. We give an example to depict this.

Example 6.8. Consider system (6.82). To convert it into the form of (6.75), we have the algebraic form of the first equation as

$$M_c x_1 x_2 = \delta_2^2. \tag{6.85}$$

The missing x_3 can be plugged in as

$$M_c x_1 M_u x_2 x_3 = \delta_2^2.$$

Equivalently, we have

$$M_c [I_2 \otimes M_u] x_1 x_2 x_3 = \delta_2^2.$$

Setting $x = x_1 x_2 x_3$, (6.85) becomes

$$\delta_2[1\ 2\ 1\ 2\ 2\ 2\ 2\ 2]x = \delta_2^2. \tag{6.86}$$

Similarly, for the second and third equations we have

$$\begin{aligned} \delta_2[1\ 1\ 1\ 1\ 1\ 2\ 1\ 2]x &= \delta_2^1; \\ \delta_2[1\ 2\ 2\ 1\ 1\ 2\ 2\ 1]x &= \delta_2^1. \end{aligned} \tag{6.87}$$

Using (6.86), (6.87) and Proposition 6.11, the system can further be converted into the form (6.80) as

$$\delta_8[1\ 2\ 5\ 8\ 6\ 5\ 6\ 7]x = \delta_8^5. \tag{6.88}$$

Finally, according to Proposition 6.12, the solutions of system (6.82) are $\delta_8^3 \sim (1, 0, 1)$ and $\delta_8^6 \sim (0, 1, 0)$.

Next, we consider the problem of logical inference by solving logical equations. We refer to Truemper (2004) for logical inference through intelligent systems.

Our basic technique is to convert the problem into a logical system. Solving the system provides the solution of the logical problem. We use some simple examples to demonstrate this.

Example 6.9. A says: "B is a liar", B says: "C is a liar", C says: "both A and B are liars". Who is a liar?

To solve this problem we define three logical variables as

- x: A is honest;
- y: B is honest;
- z: C is honest.

Then the three statements can be expressed in logical version as

$$\begin{cases} x \Leftrightarrow \neg y \\ y \Leftrightarrow \neg z \\ z \Leftrightarrow \neg x \wedge \neg y. \end{cases} \tag{6.89}$$

Let $c = \delta_2^1$. Then equation (6.89) can be converted into an algebraic form as

$$\begin{cases} M_e x M_n y = c \\ M_e y M_n z = c \\ M_e z M_c M_n x M_n y = c. \end{cases} \tag{6.90}$$

It is easy to convert (6.90) into an algebraic equation as

$$L\xi = b, \tag{6.91}$$

where $\xi = xyz$, $b = c^3 = \delta_8^1$, and

$$L = \delta_8[8, 5, 2, 3, 4, 1, 5, 8].$$

Since only $\mathrm{Col}_6(L) = b$, we have unique solution of (6.91) as

$$\xi = \delta_8^6,$$

which implies that

$$x = 0, \quad y = 1, \quad z = 0.$$

We conclude that only B is honest.

Next, we consider the logical minimization. Consider system (6.73) again. Assume the set of its solutions is Z and we have a performance criteria as

$$J = J(x_1, \cdots, x_n).$$

The purpose is to find a best feasible solution $x^* = (x_1^*, \cdots, x_n^*)$, such that

$$J(x^*) = \min_{x \in Z} J(x).$$

We give an example to depict this.

Example 6.10 (Truemper, 2004). Consider the following problem: The weather is either sunshine or rain. We take either a bus or a taxi to go to work. Suppose it is raining. If we take a bus, we must walk to the bus stop and hence use an umbrella. If we take a taxi, we do not need an umbrella.

We define some variables as: S (sunshine), R (rain), B (bus), T (taxi), and U (umbrella). Then we have the following equations:

$$\begin{cases} S \leftrightarrow \neg R = 1 \\ B \leftrightarrow \neg T = 1 \\ (R \wedge B) \leftrightarrow U = 1. \end{cases}$$

We may consider two cases: Case 1: $R = 1$. That is, it is rain. Case 2: $R = 0$.

Suppose that the fare for bus is \$3 and for taxi is \$4, and that we view the inconvenience of handling an umbrella to be equivalent to a cost of \$2. Using vector form, the cost function becomes

$$J = 3 \times (\delta_2^1)^T B + 4 \times (\delta_2^1)^T T + 2 \times U.$$

Case 1: The component-wise algebraic form of the system becomes

$$\begin{cases} M_e S M_n R = \delta_2^1 \\ M_e B M_n T = \delta_2^1 \\ M_e M_c R B U = \delta_2^1 \\ R = \delta_2^1. \end{cases}$$

The algebraic form is

$$L\xi = b_1,$$

where $\xi = SRBTU$, $b_1 = \delta_{16}^1$, and

$$\begin{array}{rl} L = [13\ 15\ \ 9\ 11\ 11\ \ 9\ 15\ 13\ \ 8\ \ 6\ \ 4\ \ 2\ \ 4\ \ 2\ \ 8\ \ 6 \\ 5\ \ 7\ \ 1\ \ 3\ \ 3\ \ 1\ \ 7\ \ 5\ 16\ 14\ 12\ 10\ 12\ 10\ 16\ 14]. \end{array} \tag{6.92}$$

The set of solutions is

$$Z = \left\{ \delta_{32}^{19} \sim (0,1,1,0,1), \quad \delta_{32}^{22} \sim (0,1,0,1,0) \right\}.$$

Then it is easy to figure out that he optimal solution is

$$\delta_{32}^{22} \sim (0,1,0,1,0).$$

Case 2: The component-wise algebraic form of the system becomes

$$\begin{cases} M_e S M_n R = \delta_2^1 \\ M_e B M_n T = \delta_2^1 \\ M_e M_c R B U = \delta_2^1 \\ R = \delta_2^2. \end{cases}$$

The algebraic form is

$$L\xi = b_2,$$

where $\xi = SRBTU$, $b_2 = \delta_{16}^2$, and L is the same as (6.92).

The set of solutions is

$$Z = \left\{\delta_{32}^{12} \sim (1,0,1,0,0), \quad \delta_{32}^{14} \sim (1,0,0,1,0)\right\}.$$

Then the optimal solution can be found as

$$\delta_{32}^{12} \sim (1,0,1,0,0).$$

6.5 Multi-Valued Logic

One of the advantages of the matrix expression of logic is that it can easily be extended to multi-valued logic. Let x be a k-valued logical variable. That is, $x \in \mathcal{D}_k$. To use the matrix approach, we identify

$$\frac{i}{k-1} \sim \delta_k^{k-i}, \quad i = 1, 2, \cdots, k-1.$$

That is,

$$1 \sim \delta_k^1, \quad \frac{k-2}{k-1} \sim \delta_k^2, \quad \cdots, \quad \frac{1}{k-1} \sim \delta_k^{k-1}, 0 \sim \delta_k^k.$$

Then in vector form we have $x \in \Delta_k$.

Definition 6.9. Let x and y be two k-valued logical variables. Define

(i) (Negation)

$$\neg x = 1 - x. \tag{6.93}$$

(ii) (Disjunction)

$$x \vee y = \max\{x, y\}. \tag{6.94}$$

(iii) (Conjunction)

$$x \wedge y = \min\{x, y\}. \tag{6.95}$$

Using vector form, it is easy to calculate the structure matrices of the k-valued logical operators in Definition 6.9.

For notational ease, we introduce a set of k-dimensional vectors as:

$$U_s = (1\ 2\ \cdots\ s-1\ \underbrace{s\ \cdots\ s}_{k-s+1})$$
$$V_s = (\underbrace{s\ \cdots\ s}_{s}\ s+1\ s+2\ \cdots\ k), \quad s = 1, 2, \cdots, k.$$

Proposition 6.14.

(1) For k-valued negation, its structure matrix is

$$M_{n,k} = \delta_k[k\ k-1\ \cdots\ 1]. \tag{6.96}$$

When $k = 3$ we have

$$M_{n,3} = \delta_3[3\ 2\ 1]. \tag{6.97}$$

(2) For k-valued disjunction, its structure matrix is

$$M_{d,k} = \delta_k[U_1\ U_2\ \cdots\ U_k]. \tag{6.98}$$

When $k = 3$ we have

$$M_{d,3} = \delta_3[1\ 1\ 1\ 1\ 2\ 2\ 1\ 2\ 3]. \tag{6.99}$$

(3) For k-valued conjunction, its structure matrix is

$$M_{c,k} = \delta_k[V_1\ V_2\ \cdots\ V_k]. \tag{6.100}$$

When $k = 3$ we have

$$M_{c,3} = \delta_3[1\ 2\ 3\ 2\ 2\ 3\ 3\ 3\ 3]. \tag{6.101}$$

Definition 6.9 is a natural extension of the corresponding objects in classical logic. When $k = 2$ they coincide with the objects in classical logic. Next, we consider "conditional" and "biconditional". They not so natural. There are many different definitions. In the following we consider the case of $k = 3$, and give 3 different definitions: (i) the type of Kleene-Dienes (KD), (ii) the type of Luekasiewic (L), (iii) the type of Bochvar (B). They are listed in Table 6.7 (where $T = 1$, $U = 0.5$, $F = 0$) (Liu and Liu, 1996).

Table 6.7 3-valued logics

		KD		L		B	
P	Q	$\to$	$\leftrightarrow$	$\to$	$\leftrightarrow$	$\to$	$\leftrightarrow$
T	T	T	T	T	T	T	T
T	U	U	U	U	U	U	U
T	F	F	F	F	F	F	F
U	T	T	U	T	U	U	U
U	U	U	U	T	T	U	U
U	F	U	U	U	U	U	U
F	T	T	F	T	F	T	F
F	U	T	U	T	U	U	U
F	F	T	T	T	T	T	T

Next, we consider a natural way to define them. Since we have already defined the "negation", "disjunction", and "conjunction" for k-valued logic. We may use them to define the "conditional" and "biconditional" as follows.

Definition 6.10. For k-valued logic, we define

(vi) (Conditional)

$$x \to y \Leftrightarrow \neg x \vee y. \tag{6.102}$$

(v) (Biconditional)

$$x \leftrightarrow y \Leftrightarrow (x \to y) \wedge (y \to x). \tag{6.103}$$

Note that (6.102) is from property (6.33) of classical logic, and (6.103) is from (6.45). So Definition 6.10 is called a natural extension of the classical logic.

Using (6.102), we have

$$M_{i,k}xy = M_{d,k}M_{n,k}xy.$$

Hence the structure matrix of → can be calculated as

$$M_{i,k} = M_{d,k}M_{n,k}. \tag{6.104}$$

It is easy to calculate that when $k = 3$ we have

$$M_{i,3} = \delta_3[1\ 2\ 3\ 1\ 2\ 2\ 1\ 1\ 1]. \tag{6.105}$$

To calculate the structure matrix of biconditional, we need the k-valued power-reducing matrix. Define the k-valued power-reducing matrix as

$$M_{r,k} = \begin{bmatrix} \delta_k^1 & 0 & \cdots & 0 \\ 0 & \delta_k^2 & \cdots & 0 \\ & & & \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \delta_k^k \end{bmatrix}. \tag{6.106}$$

Then it is easy to prove the following result.

Proposition 6.15. *Let $x \in \Delta_k$. Then we have*

$$x^2 = M_{r,k}x. \tag{6.107}$$

Now we are ready to calculate the structure matrix of the natural biconditional. Using (6.103), we have

$$\begin{aligned} M_{e,k}xy &= M_{c,k}M_{i,k}xyM_{i,k}yx \\ &= M_{c,k}M_{d,k}M_{n,k}xyM_{d,k}M_{n,k}yx \\ &= M_{c,k}M_{d,k}M_{n,k}[I_{k^2} \otimes M_{d,k}M_{n,k}]xy^2x \\ &= M_{c,k}M_{d,k}M_{n,k}[I_{k^2} \otimes M_{d,k}M_{n,k}]xW_{[k,k^2]}xy^2 \\ &= M_{c,k}M_{d,k}M_{n,k}[I_{k^2} \otimes M_{d,k}M_{n,k}][I_k \otimes W_{[k,k^2]}]x^2y^2 \\ &= M_{c,k}M_{d,k}M_{n,k}[I_{k^2} \otimes M_{d,k}M_{n,k}][I_k \otimes W_{[k,k^2]}]M_{r,k}xM_{r,k}y \\ &= M_{c,k}M_{d,k}M_{n,k}[I_{k^2} \otimes M_{d,k}M_{n,k}][I_k \otimes W_{[k,k^2]}]M_{r,k}[I_k \otimes M_{r,k}]xy. \end{aligned}$$

Hence we have

$$M_{e,k} = M_{c,k}M_{d,k}M_{n,k}[I_{k^2} \otimes M_{d,k}M_{n,k}][I_k \otimes W_{[k,k^2]}]M_{r,k}[I_k \otimes M_{r,k}]. \tag{6.108}$$

Using this formula, we can calculate that when $k = 3$ we have

$$M_{e,3} = \delta_3[1\ 2\ 3\ 2\ 2\ 2\ 3\ 2\ 1]. \tag{6.109}$$

It is easy to check that when $k = 3$ Definition 6.10 coincides with the type of Kleene-Dienes logic.

Throughout this book we assume the default conditional and biconditional operators of k-valued logic are defined by Definitions 6.9 and 6.10, unless elsewhere stated.

Next, we consider the k-valued logical system. Recall system (6.73), assume the unknowns $x_1, \cdots, x_n \in \mathcal{D}_k$ and the constants $c_1, \cdots, c_m \in \mathcal{D}_k$. Then (6.73) becomes a k-valued logical system. A step-by-step verification shows that the deductions there for standard logical systems remain true for k-valued case, and we can obtain the algebraic form of a k-valued logical system as follows, which is similar to (6.80):

$$Mx = \delta_{k^m}^r, \tag{6.110}$$

where $M \in \mathcal{L}_{k^m \times k^n}$. Similar to Proposition 6.12, we have

Proposition 6.16. *$x = \delta_{k^n}^p$ is a solution of (6.110), if and only if,*

$$\mathrm{Col}_p(M) = \delta_{k^m}^r. \tag{6.111}$$

In the following we give an example to show how to use multi-valued logic to carry out a logical inference.

Example 6.11. A detective is investigating a murder case. He has the following clues:

- 80% for sure that either A or B is the murderer;
- if A is the murderer, it is very likely that the murder happened after midnight;
- if B's confession is true, the light at midnight was on;
- if B's confession is false, it is very likely that the murder happened before midnight;
- there is an evidence that the light in the room of the murder was off at midnight.

What conclusion he can get?

First, we have to figure out the levels of logical values. Say, “very likely” is more possible that “80%”, then we may quantize the logical values into six levels as: “T”, “very likely”, “80%”, “1-80%”, “very unlikely”, and “F”. Hence we may consider the problem as a problem of 6-valued logical inference.

Define the logical variables (unknowns) as

- A: A is the murderer;
- B: B is the murderer;
- M: the murder happened before the midnight;
- S: B's confession is true;
- L: the light in the room was on at midnight.

Then we can convert the statements into logical equations as

$$\begin{cases} A \vee B = \delta_6^3 \\ A \to \neg M = \delta_6^2 \\ S \to L = \delta_6^1 \\ \neg S \to M = \delta_6^2 \\ \neg L = \delta_6^1. \end{cases} \tag{6.112}$$

We may use general method, described in Proposition 6.16, to solve that 6-valued system. But since this system has certain special simple form, we may use “substitution” to solve it. Here the so called “substitution” is exactly the same as the one in solving linear system in high school algebra: We use some equations to solve some unknowns first, and then substitute the solved unknowns into the other equations to solve the other unknowns.

First, since $\neg L = \delta_6^1$, we have

$$L = \delta_6^6.$$

Next, the equation $S \to L = \delta_6^1$ provides the following matrix equation:

$$M_{i,6}SL = M_{i,6}W_{[6]}LS := \Psi_1 S = \delta_6^1.$$

It is easy to calculate that

$$\Psi_1 = M_{i,6}W_{[6]}L = \delta_6[6\ 5\ 4\ 3\ 2\ 1].$$

It follows that $S = \delta_6^6$. Similarly, from $\neg S \to M = \delta_6^2$ we have

$$M_{i,6}M_{n,6}SM = \delta_6^2.$$

M can be solved as

$$M = \delta_6^2.$$

Next, consider $A \to \neg M = M_{i,6}AM_{n,6}M = \delta_6^2$. Using the properties of semi-tensor product, we have

$$M_{i,6}AM_{n,6}M = M_{i,6}(I_6 \otimes M_{n,6})AM = M_{i,6}(I_6 \otimes M_{n,6})W_{[6]}MA := \psi_2 A.$$

Since

$$\psi_2 = M_{i,6}(I_6 \otimes M_{n,6})W_{[6]}M = \delta_6[5\ 5\ 4\ 3\ 2\ 1],$$

we have that

$$A = \delta_6^5.$$

Finally, from $A \vee B = M_d AB = \delta_6^3$ we have

$$B = \delta_6^3.$$

We conclude that, A is "very unlikely" the murderer, and 80% that B is the murderer.

Exercises

6.1 Consider membership function (6.1). Now assume we want to convert it into four-valued logic in the following ways: Let $a_1 = 1$, $a_2 = \frac{2}{3}$, $a_3 = \frac{1}{3}$, and $a_4 = 0$ be four possible logical values. Set $f(x) = a_i$, if and only if

$$|f(x) - a_i| = \min_{1 \le j \le 4} |f(x) - a_j|.$$

Draw the quantized membership function.

6.2 Let x, y, z be logical variables. Find the truth vector of the following logical functions with respect to x, y, z:

(i) $f(x, y, z) = (z \wedge x) \vee y$.

(ii) $f(x, y, z) = (x \vee y) \leftrightarrow (x \vee z)$.

(iii) $f(x, y, z) = (x \wedge y) \vee (y \wedge z) \vee (z \wedge x)$.

6.3 Prove Lemma 6.1. Show that when $x \in \mathcal{D}_\infty$ formula (6.14) is incorrect.

6.4 Using Theorem 6.1, we can express a logical expression into a pure disjunctive form as

$$f = f_1 \vee f_2 \vee \cdots \vee f_s, \tag{6.113}$$

where each f_i has pure conjunctive form as

$$f_i = a_1^i \wedge a_2^i \wedge \cdots \wedge a_{t_i}^i, \quad i = 1, \cdots, s,$$

with $a_j^i = x_k$ or $a_j^i = \neg x_k$. (6.113) is called the disjunctive normal form of f. Find the disjunctive normal form for the following logical functions:

(i) $f(x, y) = (y \wedge x) \vee y$.

(ii) $f(x, y, z) = (x \vee y) \leftrightarrow (x \vee z)$.

6.5 We can also express a logical expression into a pure conjunctive form as

$$f = f_1 \wedge f_2 \wedge \cdots \wedge f_s, \tag{6.114}$$

where each f_i has pure disjunctive form as

$$f_i = a_1^i \vee a_2^i \vee \cdots \vee a_{t_i}^i, \quad i = 1, \cdots, s,$$

with $a_j^i = x_k$ or $a_j^i = \neg x_k$. (6.114) is called the conjunctive normal form of f. Find the conjunctive normal form for the following logical functions:

(i) $f(x, y) = (x \to y) \wedge x$.

(ii) $f(x, y, z) = (z \bar{\vee} y) \leftrightarrow (y \wedge z)$.

(Hint: Express $\neg f$ into disjunctive normal form and then use De Morgan's Law.)

6.6 Find two equivalent expressions which are not absolutely equivalent.

6.7 Prove the other equivalences (except (6.35)) in Proposition 6.4.

6.8 Prove the other equivalences (except (6.41)) in Proposition 6.6.

6.9 Prove the tautological implications (6.49)–(6.60) in Proposition 6.9.

6.10 Prove the logical equivalences (6.62)–(6.72) in Proposition 6.10.

6.11 (i) Prove De Morgan's Law (Proposition 6.7). (ii) Prove the following general De Morgan's Law:

$$\neg(x_1 \wedge \cdots \wedge x_n) = (\neg x_1) \vee \cdots \vee (\neg x_n). \tag{6.115}$$

$$\neg(x_1 \vee \cdots \vee x_n) = (\neg x_1) \wedge \cdots \wedge (\neg x_n). \tag{6.116}$$

6.12 Show that $L_1 \Leftrightarrow L_2$, if and only if $L_1 \leftrightarrow L_2$ is a tautology, if and only if $L_1 \bar{\vee} L_2$ is a contradiction, if and only if $L_1 \Rightarrow L_2$ and $L_2 \Rightarrow L_1$.

6.13 What n can make the following expression a tautology?

$$\underbrace{(\cdots((A \to A) \to A) \to \cdots \to A)}_{n}.$$

6.14 Give an equivalent condition for a 3-ary logical operator being symmetric.

6.15 Define a binary logical operator "|" whose structure matrix is

$$M_{|} = \delta_2[2\ 1\ 1\ 1].$$

Show that

(i) "|" can be expressed by "¬" and "∧".

(ii) "$\neg$", "$\wedge$" and "$\vee$" can be expressed by "$|$". (In fact, this implies that all other logical operators can be expressed by "$|$" referring to the concept of "adequate set" in next chapter.)

6.16 Solve the following logical equations.

(i)

$$\begin{cases} x_1 \wedge x_2 \wedge x_3 = 0 \\ (x_1 \to x_2) \vee x_3 = 1 \\ x_1 \leftrightarrow x_3 = 1. \end{cases}$$

(ii)

$$\begin{cases} x_1 \wedge x_2 = x_2 \to x_3 \\ x_1 = \neg(x_2 \vee x_3) \\ x_2 \vee x_3 = 1. \end{cases}$$

6.17 Given a logical equation as

$$\begin{cases} f_1(x_1, \cdots, x_n) = g_1(x_1, \cdots, x_n) \\ f_2(x_1, \cdots, x_n) = g_2(x_1, \cdots, x_n) \\ \vdots \\ f_m(x_1, \cdots, x_n) = g_m(x_1, \cdots, x_n). \end{cases} \tag{6.117}$$

Give a general procedure to solve this equation.

6.18 Consider the following logical system

$$\begin{cases} x_1 = x_2 \sigma x_3 \\ x_2 = x_3 \sigma x_4 \\ \vdots \\ x_{n-1} = x_n \sigma x_1 \\ x_n = x_1 \sigma x_2. \end{cases} \tag{6.118}$$

(i) Assume $\sigma = \wedge$. Solve system (6.118).

(ii) Assume $\sigma = \vee$. Solve system (6.118).

6.19 Is De Morgan's Law true for k-valued logic?

6.20 Are the equivalences in Proposition 6.6 true for 3-valued logic?

6.21 A says: "we are all honest", B says: "we are all liars", C says: "only one of us is honest", D says: "only two of us are liars". Construct the logical system of equations for this, and solve it.

This page intentionally left blank

Chapter 7

Mix-Valued Logic

In this chapter we first introduce the normal form of logic and k-valued logic. Using the normal form a generalized new logic, called the mix-valued logic, is proposed. Then the properties of general logical mappings are explored. As practical background, the fuzzy control and the game of finite strategies are formulated by mix-valued logic.

7.1 Normal Form of Logical Operators

Consider the set of r-ary logical operators (i.e., logical functions). Since there are r arguments and each argument can take 2 different values, so an r-ary logical operator is a mapping from a set of cardinality 2^r to a set of cardinality 2. Hence, there are 2^{2^r} different operators.

When $r = 0$, there are two nullary (0-ary) operators, which are $f \equiv 0$ and $f \equiv 1$. Where $r = 1$ there are 4 unary (1-ary) operators, which are $f(x) = x$, $f(x) = \neg x$, and the two nullary operators as its particular cases. When $r = 2$, there are 16 binary (2-ary) operators, which are listed in Table 6.6, including 4 unary operators as its particular cases. When $r = 3$ we have $2^{2^3} = 256$ ternary (3-ary) operators, and so on.

If we need all of these different forms to express all possible logical functions, it will be a terrible mess. Hence a natural question is: Is it possible to find a finite set of operators, which can be used to describe all the operators?

Definition 7.1. A set of logical operators is said to be an adequate set, if any operator can be expressed as a combination of them.

Proposition 7.1 (Hamilton, 1988). *The pairs* $\{\neg, \wedge\}$, $\{\neg, \vee\}$ *are ade-*

quate sets.

Using Theorem 6.1 repeatedly, we have shown that $\{\neg, \wedge, \vee\}$ is an adequate set. From Proposition 7.1 this conclusion is obvious and we know that there is a redundant operator. But in application $\{\neg, \wedge, \vee\}$ is a commonly used adequate set, because it is very convenient in use. In fact, since $\{\neg, \wedge, \vee\}$ is an adequate set, then using De Morgan's law, Proposition 7.1 can be proved easily.

Consider the k-valued logic. Similar to $k = 2$ case, it is easy to see that there are k^{k^r} r-ary logical operators. Particularly, there are k trivial nullary operators. We consider unary operators. There are k^k unary operators. We name some of them as follows: (Cheng *et al.*, 2011b) (The following operators are defined by their scalar values.)

(i) Negation

$$\neg x := 1 - x. \tag{7.1}$$

Its structure matrix is

$$M_{n,k} = \delta_k[k \;\; k-1 \;\; \cdots \;\; 1]. \tag{7.2}$$

(ii) Rotator $\oslash_k$ is defined as

$$\oslash_k(x) := \begin{cases} x - \frac{1}{k-1}, & x \neq 0, \\ 1, & x = 0. \end{cases} \tag{7.3}$$

Its structure matrix, $M_{o,k}$, is

$$M_{o,k} = \delta_k \left[2 \;\; 3 \;\; \cdots \;\; k \;\; 1\right]. \tag{7.4}$$

For instance, we have

$$M_{o,3} = \delta_3[2\ 3\ 1], \quad M_{o,4} = \delta_4[2\ 3\ 4\ 1]. \tag{7.5}$$

(iii) i-confirmor, $\triangledown_{i,k}$, $i = 1, \cdots, k$, are defined as

$$\nabla_{i,k}(x) = \begin{cases} 1, & x = \frac{k-i}{k-1}, \text{ (equivalently } x = \delta_k^i) \\ 0, & \text{otherwise.} \end{cases} \tag{7.6}$$

Its structure matrix (using same notation)

$$\nabla_{i,k} = \delta_k[\underbrace{k \cdots k}_{i-1}\ 1\ \underbrace{k \cdots k}_{k-i}], \quad i = 1, 2, \cdots, k. \tag{7.7}$$

For instance, we have

$$\triangledown_{2,3} = \delta_3[3\ 1\ 3], \quad \triangledown_{2,4} = \delta_4[4\ 1\ 4\ 4], \quad \triangledown_{3,4} = \delta_4[4\ 4\ 1\ 4]. \tag{7.8}$$

(iv) Conjunction

$$x \wedge y := \min\{x, y\}. \tag{7.9}$$

Its structure matrix is (to save space let $n = 3$)

$$M_{c,3} = \delta_3[1\ 2\ 3\ 2\ 2\ 3\ 3\ 3\ 3]. \tag{7.10}$$

(v) Disjunction

$$x \vee y := \max\{x, y\}. \tag{7.11}$$

Its structure matrix is ($n = 3$)

$$M_{d,3} = \delta_3[1\ 1\ 1\ 1\ 2\ 2\ 1\ 2\ 3]. \tag{7.12}$$

In fact, we did not name all and the above set of operators are also not enough to construct all the unary operators. For statement ease, we use a general notation for all unary operators.

Definition 7.2. Let $i_1, \cdots, i_k \in \{1, 2, \cdots, k\}$. The operator $\oslash_{i_1,i_2,\cdots,i_k}$ is a unary operator, defined by (in vector form)

$$\oslash_{i_1,i_2,\cdots,i_k}(\delta_k^j) = \delta_k^{i_j}, \quad j = 1, \cdots, k. \tag{7.13}$$

If $i_s = s$, which means when $x = \delta_k^s$, then $\oslash_{i_1,i_2,\cdots,i_k}(\delta_k^s) = \delta_k^s$. That is, x is invariant with respect to this operator. In this case, we replace i_s by $*$, which makes the operators more clear. For instance, $\oslash_{*,3,2} = \oslash_{1,3,2}$, $\oslash_{*,*,5,*,*} = \oslash_{1,2,5,4,5}$.

We give some examples to illustrate these.

Example 7.1.

(1)

$$\oslash_{2,3,3}(x) = \oslash_{23*}(x) = \begin{cases} \delta_3^2, & x = \delta_3^1 \\ \delta_3^3, & x = \delta_3^2 \\ \delta_3^3, & x = \delta_3^3. \end{cases}$$

(2)

$$\oslash_{2,*,1,*}(x) = \begin{cases} \delta_4^2, & x = \delta_4^1 \\ \delta_4^2, & x = \delta_4^2 \\ \delta_4^1, & x = \delta_4^3 \\ \delta_4^4, & x = \delta_4^4. \end{cases}$$

Example 7.2. Using the general expression, we have

$$\begin{aligned} \neg_k &= \oslash_{k,k-1,\cdots,1}; \\ \oslash_k &= \oslash_{2,3,\cdots,k,1}; \\ \nabla_{i,k} &= \oslash_{\underbrace{k,\cdots,k}_{i-1},1,\underbrace{k,\cdots,k}_{k-i}}. \end{aligned}$$

Next, we consider whether there is an expression of k-valued logical function similar to Theorem 6.1. In fact, we have the following.

Theorem 7.1. *Assume $f(x_1,\cdots,x_r)$ is an r-ary k-valued operator with its structure matrix $M_f \in \mathcal{L}_{k\times k^r}$, $r \geq 2$. Split M_f into k equal-size parts as*

$$M_f = [B_1, B_2, \cdots, B_k].$$

Then

$$\begin{aligned} f(x_1,\cdots,x_r) = {} & (\nabla_{1,k}(x_1)\wedge f_1(x_2,\cdots,x_r)) \vee (\nabla_{2,k}(x_1)\wedge f_2(x_2,\cdots,x_r)\cdots \\ & \nabla_{k,k}(x_1)\wedge f_k(x_2,\cdots,x_r)), \end{aligned} \tag{7.14}$$

where $f_i(x_2,\cdots,x_r)$ has B_i as its structure matrix, $i = 1,2,\cdots,k$.

Proof. Similar to the proof of Theorem 6.1, we only need to prove

$$f(x_1,\cdots,x_r) = \vee_{i=1}^{k}\left(\nabla_{i,k}(x_1)\wedge f(\frac{k-i}{k-1},x_2,\cdots,x_r)\right). \tag{7.15}$$

If $x_1 = \frac{k-j}{k-1}$,

$$\begin{aligned} RHS = {} & \vee_{i=1}^{j-1}\left(0\wedge f(\frac{k-j}{k-1},x_2,\cdots,x_r)\right)\vee\left(1\wedge f(\frac{k-j}{k-1},x_2,\cdots,x_r)\right)\vee \\ & \vee_{i=j+1}^{k}\left(0\wedge f(\frac{k-j}{k-1},x_2,\cdots,x_r)\right) \\ = {} & f(\frac{k-j}{k-1},x_2,\cdots,x_r) = LHS. \end{aligned}$$

Since j can run from 1 to k, which covers every case, (7.15) holds, and then Theorem 7.1 follows. □

Denote by $\mathcal{U}_k$ the set of unary k-valued operators. That is,

$$\mathcal{U}_k = \{\oslash_{i_1,\cdots,i_k} | 1 \leq i_j \leq k,\ j = 1,\cdots,k\}.$$

Then it is easy to see the following corollary.

Corollary 7.1. *For k-valued logic, the set*

$$\mathcal{U}_k \cup \{\vee\} \cup \{\wedge\}$$

is an adequate set.

We use the following example to depict this.

Example 7.3. Assume $k = 3$ and

$$f(x,y) = \delta_3[3\ 2\ 1\ 2\ 2\ 3\ 3\ 1\ 2]xy.$$

Find the logical expression of $f(x,y)$.

Using Theorem 7.1, it is easy to decompose f as

$$f(x,y) = (\nabla_{1,3}(x) \wedge \sigma_1(y)) \vee (\nabla_{2,3}(x) \wedge \sigma_2(y)) \vee (\nabla_{3,3}(x) \wedge \sigma_3(y))\,.$$

Moreover, the structure matrices of σ_i are

$$M_{\sigma_1} = [3\ 2\ 1], \quad M_{\sigma_2} = [2\ 2\ 3], \quad M_{\sigma_3} = [3\ 1\ 2].$$

Hence, we have

$$\begin{aligned}
&\sigma_1(y) = \oslash_{3,2,1}(y) = \neg y;\\
&\sigma_2(y) = \oslash_{2,2,3}(y) = \oslash_{2,*,*}(y);\\
&\sigma_3(y) = \oslash_{3,1,2}(y) = \oslash^2(y)(\text{or } \oslash^{-1}(y)).
\end{aligned}$$

Similar to classical logic, using Theorem 7.1 repeatedly, we can have a normal form. We state it as a corollary.

Corollary 7.2. *Assume $y = f(x_1, \cdots, x_r)$ is an r-ary k-valued logical form. It can be expressed into the following normal form, which is called the k-valued disjunctive normal form*

$$y = \vee_{j=1}^{s} \left(\nabla_{i_1^j,k}(x_1) \wedge \nabla_{i_2^j,k}(x_2) \wedge \cdots \wedge \nabla_{i_{r-1}^j,k}(x_{r-1}) \wedge \sigma_j(x_r)\right), \quad (7.16)$$

where σ_j, $j = 1, \cdots, s$ are some unary operators.

7.2 Mix-Valued Logic

Assume we have a set of logical variables $x_0, x_1, \cdots, x_n$, where

$$x_i \in \mathcal{D}_{k_i}, \quad k_i \geq 2, \quad i = 0, 1, \cdots, n. \quad (7.17)$$

The mix-valued logic considers the operators over logical variables which belong to different logical regions. We first give a rigorous definition.

Definition 7.3. Let x_i, $i = 0, 1, \cdots, n$ be as in (7.17). An n-ary mix-valued logical operator (function) f is a mapping $f : \mathcal{D}_{k_1} \times \cdots \times \mathcal{D}_{k_n} \to \mathcal{D}_{k_0}$.

Fixing $\mathcal{D}_{k_s}$ for an s satisfying $1 \le s \le n$, we consider unary mix-valued logical operators, which are mappings from $\mathcal{D}_{k_s}$ to $\mathcal{D}_{k_0}$. Similar to (7.13), we define

$$\oslash^{k_0}_{i_1,i_2,\cdots,i_{k_s}}(\delta^j_{k_s}) = \delta^{i_j}_{k_0}, \quad j = 1, \cdots, k_s. \tag{7.18}$$

Now the set

$$\mathcal{U}^{k_0}_{k_s} := \left\{ \oslash^{k_0}_{i_1,i_2,\cdots,i_{k_s}} \,\middle|\, 1 \le i_j \le k_0,\ j = 1, \cdots, k_s \right\}, \quad s = 1, \cdots, n,$$

form the set of all unary logical operators from $\mathcal{D}_{k_s}$ to $\mathcal{D}_{k_0}$.

Particularly, we consider the identifiers:

$$\nabla^{k_0}_{j,k_s}(x) := \begin{cases} \delta^1_{k_0}, & x = \delta^j_{k_s} \\ \delta^{k_0}_{k_0}, & \text{otherwise.} \end{cases} \tag{7.19}$$

In general form, we have

$$\nabla^{k_0}_{j,k_s} = \oslash^{k_0}_{\underbrace{k_0, \cdots, k_0}_{j-1}, 1, \underbrace{k_0, \cdots, k_0}_{k_s - j}}.$$

For notational compactness, when there is no possible confusion we simply denote

$$\nabla_j(\delta^t_{k_s}) := \begin{cases} \delta^1_{k_0}, & t = j \\ \delta^{k_0}_{k_0}, & \text{otherwise.} \end{cases} \tag{7.20}$$

In (7.20) the operator ∇_j can be considered as a general operator from Δ_{k_s} to Δ_{k_0}, where k_0 can be either the same as k_s or different from k_s.

Next, we deduce the (disjunctive) normal form for mix-valued logical functions. Assume $f(x_1, \cdots, x_n)$ is a mix-valued logical function as defined in Definition 7.3. Using truth table, it is easy to construct the structure matrix of f as

$$M_f \in \mathcal{L}_{k_0 \times k}, \quad \text{where } k = \prod_{j=1}^{n} k_j.$$

Split M_f into k_1 equal blocks as

$$M_f = [B_1\ B_2\ \cdots\ B_{k_1}].$$

Then we have the following result, which is parallel to Theorem 6.1 for standard logic, and Theorem 7.1 for k-valued logic.

Theorem 7.2. *Let $f(x_1, \cdots, x_n)$ be the mix-valued function defined in Definition 7.3. Then $f(x_1, \cdots, x_n)$ can be expressed as*

$$\begin{aligned} f(x_1, \cdots, x_n) = {} & (\nabla_1(x_1) \wedge f_1(x_2, \cdots, x_n)) \vee (\nabla_2(x_1) \wedge f_2(x_2, \cdots, x_n)) \vee \cdots \\ & \vee (\nabla_{k_0}(x_1) \wedge f_{k_0}(x_2, \cdots, x_n)), \end{aligned} \tag{7.21}$$

where $f_i(x_2, \cdots, x_n)$ has B_i as its structure matrix, $i = 1, \cdots, k_1$.

Using Theorem 7.2 repetitively, we finally can get the (disjunctive) normal form of mix-valued logical operators.

Corollary 7.3. *Let $f(x_1,\cdots,x_n)$ be the mix-valued logical function defined in Definition 7.3. Split M_f into k/k_n equal-size blocks*

$$M_f = [B_1\ B_2\ \cdots\ B_{k/k_n}],$$

denote

$$B_j = \delta_{k_0}[i_1^j, i_2^j, \cdots, i_{k_n}^j].$$

Then $f(x_1,\cdots,x_n)$ can be expressed as

$$\begin{aligned} f(x_1,\cdots,x_n) = \vee_{j_1=1}^{k_1} \vee_{j_2=1}^{k_2} \cdots \vee_{j_{n-1}=1}^{k_{n-1}} (\nabla_{j_1}(x_1) \wedge \nabla_{j_2}(x_2) \wedge \cdots \\ \wedge \nabla_{j_{n-1}}(x_{n-1}) \wedge \oslash^{k_0}_{i_1^s, i_2^s, \cdots, i_{k_n}^s}(x_n)\Big), \end{aligned} \tag{7.22}$$

where $s = \sum_{i=1}^{n-2}\Big((j_i - 1)\prod_{p=i+1}^{n-1} k_p\Big) + j_{n-1}$.

Remark 7.1.

(1) Both (6.18) (for classical logic) and (7.14) (for k-valued logic) may be considered as the special cases of (7.21).

(2) (7.22) is called the (disjunctive) normal form of mix-valued logical function $f(x_1,\cdots,x_n)$. (6.19) and (7.16) can be considered as its particular case.

(3) The set

$$\cup_{s=1}^{n}\left\{\mathcal{U}_{k_s}^{k_0}\right\} \cup \{\vee_{k_0}\} \cup \{\wedge_{k_0}\}$$

is adequate for the set of operators $\sigma : \mathcal{D}_{k_1} \times \cdots \times \mathcal{D}_{k_n} \to \mathcal{D}_{k_0}$.

Example 7.4. A mix-valued logical function $f : \mathcal{D}_2 \times \mathcal{D}_3 \times \mathcal{D}_2 \to \mathcal{D}_3$ has its structure matrix as

$$M_f = \delta_3[3\ 1\ 2\ 3\ 1\ 2\ 1\ 2\ 3\ 3\ 1\ 1].$$

Find its logical expression.

It is easy to calculate that

$$\begin{aligned} &f(x_1,x_2,x_3) \\ &= \big(\nabla_1(x_1) \wedge \nabla_1(x_2) \wedge \oslash^3_{3,1}(x_3)\big) \vee \big(\nabla_1(x_1) \wedge \nabla_2(x_2) \wedge \oslash^3_{2,3}(x_3)\big) \vee \\ &\quad \big(\nabla_1(x_1) \wedge \nabla_3(x_2) \wedge \oslash^3_{1,2}(x_3)\big) \vee \big(\nabla_2(x_1) \wedge \nabla_1(x_2) \wedge \oslash^3_{1,2}(x_3)\big) \vee \\ &\quad \big(\nabla_2(x_1) \wedge \nabla_2(x_2) \wedge \oslash^3_{3,3}(x_3)\big) \vee \big(\nabla_2(x_1) \wedge \nabla_3(x_2) \wedge \oslash^3_{1,1}(x_3)\big). \end{aligned}$$

7.3 General Logical Mappings

Let $x_i \in \mathcal{D}_{k_i}$, $i = 1, \cdots, n$ and $z_j \in \mathcal{D}_{s_j}$, $j = 1, \cdots, m$. Set $k = \prod_{i=1}^{n} k_i$ and $s = \prod_{j=1}^{m} s_j$. Assume a logical mapping

$$F : \prod_{i=1}^{n} \mathcal{D}_{k_i} \to \prod_{j=1}^{m} \mathcal{D}_{s_j} \tag{7.23}$$

is determined by

$$\begin{cases} z_1 = f_1(x_1, \cdots, x_n) \\ z_2 = f_2(x_1, \cdots, x_n) \\ \vdots \\ z_m = f_m(x_1, \cdots, x_n). \end{cases} \tag{7.24}$$

In vector form, we set $x = \ltimes_{i=1}^{n} x_i \in \Delta_k$ and $z = \ltimes_{j=1}^{m} z_j \in \Delta_s$, and denote by $M_j := M_{f_j} \in \mathcal{L}_{s_j \times k}$ the structure matrices of f_j.

As in Boolean case, to get the structure matrix of a logical function we may first need to add some fabricated arguments formally by using dummy matrix. Define a matrix, called the general dummy matrix, as

$$D_{p,q} := \mathbf{1}_p^T \otimes I_q, \tag{7.25}$$

and denote $D_{n,n}$ as D_n. Then, via a straightforward computation, we have

Proposition 7.2. *Let $x \in \Delta_p$ and $y \in \Delta_q$. Then*

$$\begin{aligned} D_{p,q} xy \quad &= y \\ D_{q,p} W_{[p,q]} xy &= x. \end{aligned} \tag{7.26}$$

Using (7.26), we can introduce some fabricated variables into a logical expression to complete all the arguments. Then the component-wise algebraic form of (7.24) can be obtained as

$$\begin{cases} z_1 = M_1 x \\ z_2 = M_2 x \\ \vdots \\ z_m = M_m x. \end{cases} \tag{7.27}$$

Next, we look for the structure matrices of F. Define a matrix, called the kth power-reducing matrix, as

$$M_{r,k} := \mathrm{diag}(\delta_k^1, \delta_k^2, \cdots, \delta_k^k). \tag{7.28}$$

Then it is easy to prove the following

Proposition 7.3. *Let $x \in \Delta_k$. Then*

$$x^2 = M_{r,k} x. \tag{7.29}$$

Using (7.29), we can reduce the power of a logical variable to 1. That is,

$$x^t = (M_{r,k})^{t-1} x. \tag{7.30}$$

Now we are ready to provide the structure matrix of F.

Theorem 7.3. *Consider a mapping, F, defined by (7.23) and (7.24). There is a unique matrix $M_F \in \mathcal{L}_{s\times k}$, called the structure matrix of F, such that*

$$z = M_F x. \tag{7.31}$$

Proof. Multiplying both sides of (7.27) yields

$$\begin{aligned} z &= M_1 x M_2 x \cdots M_m x \\ &= M_1 (I_k \otimes M_2) x^2 M_3 x \cdots M_m x \\ &= M_1 (I_k \otimes M_2) \cdots (I_{k^{m-1}} \otimes M_m) x^m \\ &= M_1 (I_k \otimes M_2) \cdots (I_{k^{m-1}} \otimes M_m) (M_{r,k})^{m-1} x. \end{aligned}$$

Hence

$$M_F = M_1 (I_k \otimes M_2) \cdots (I_{k^{m-1}} \otimes M_m) (M_{r,k})^{m-1}. \tag{7.32}$$

To see M_F is a logical matrix, it comes from the following claim, which is easily verifiable, that "the product of two logical matrices is a logical matrix".

To see M_F is unique, assume there is another such structure matrix, called M'_F, and the ith columns of these two matrices are different. That is, $\text{Col}_i(M_F) \neq \text{Col}_i(M'_F)$. Choose x_i, $i = 1, \cdots, n$ such that $x = \delta_k^i$. Then we have $z = F(x)$ equals to $\text{Col}_i(M_F)$ or $\text{Col}_i(M'_F)$ by using M_F or M'_F respectively. This is absurd. □

In fact, equation (7.32) may be considered as a formula to calculate the structure matrix of a mix-valued logical mapping. But the following property may provide a more convenient way for numerical calculation.

Let $x_i \in \mathcal{D}_{k_i}$, $i = 1, \cdots, n$, $y_p \in \mathcal{D}_{s_p}$, $p = 1, \cdots, m$, $z_q \in \mathcal{D}_{t_q}$, $q = 1, \cdots, r$. Set $k = \prod_{i=1}^n k_i$, $s = \prod_{p=1}^m s_p$, $t = \prod_{q=1}^r t_q$. Assume $F : \prod_{i=1}^n \mathcal{D}_{k_i} \to \prod_{p=1}^m \mathcal{D}_{s_p}$ and $G : \prod_{i=1}^n \mathcal{D}_{k_i} \to \prod_{q=1}^r \mathcal{D}_{t_q}$ have their structure matrices $M_F \in \mathcal{L}_{s\times k}$ and $M_G \in \mathcal{L}_{t\times k}$ respectively. The product mapping

$$\pi = F \times G : \prod_{i=1}^n \mathcal{D}_{k_i} \to \prod_{p=1}^m \mathcal{D}_{s_p} \prod_{q=1}^r \mathcal{D}_{t_q}$$

is defined as

$$\pi(x) = F(x) \ltimes G(x).$$

Assume $\pi(x) = F(x) \ltimes G(x)$ and the structure matrices of F and G are $M_F \in \mathcal{L}_{s\times k}$ and $M_G \in \mathcal{L}_{t\times k}$. Denote the structure matrix of π by $M_\pi \in \mathcal{L}_{st\times k}$. Using Proposition 6.11, we have

$$M_\pi = M_F * M_G. \tag{7.33}$$

Using this proposition to all the components of F, which is defined by (7.23) and (7.24), and assume the structure matrices of f_i, $i = 1, \cdots, m$ are M_i, then the structure matrix M_F of F can be calculated by

$$M_F = M_1 * M_2 * \cdots * M_m. \tag{7.34}$$

Example 7.5. Assume $x_1, x_3, z_1, z_2 \in \mathcal{D}$, $x_2, z_3 \in \mathcal{D}_3$, and the mapping $F : (x_1, x_2, x_3) \mapsto (z_1, z_2, z_3)$ is decided by

$$\begin{cases} z_1 = x_1 \wedge (\oslash^2_{2,1,2} x_2) \\ z_2 = (\oslash^2_{2,1,1} x_2) \vee x_3 \\ z_3 = \oslash^3_{1,3}(x_1 \leftrightarrow x_3). \end{cases} \tag{7.35}$$

(7.35) can be converted to its algebraic form as

$$\begin{cases} z_1 = \delta_2[1\ 2\ 2\ 2]x_1\delta_2[2\ 1\ 2]x_2 \\ z_2 = \delta_2[1\ 1\ 1\ 2]\delta_2[2, 1, 1]x_2x_3 \\ z_3 = \delta_3[1, 3]\delta_2[1\ 2\ 2\ 1]x_1x_3. \end{cases} \tag{7.36}$$

Consider z_1, we have

$$\begin{aligned} z_1 &= \delta_2[1\ 2\ 2\ 2]x_1\delta_2[2\ 1\ 2][I_3\ I_3]x_3x_2 \\ &= \delta_2[1\ 2\ 2\ 2]x_1\delta_2[2\ 1\ 2][I_3\ I_3]W_{[3,2]}x_2x_3 \\ &= \delta_2[1\ 2\ 2\ 2]x_1\delta_4[2\ 2\ 1\ 1\ 2\ 2\ 4\ 4\ 3\ 3\ 4\ 4]x_2x_3 \\ &= \delta_2[1\ 2\ 2\ 2]\left(I_2 \otimes \delta_4[2\ 2\ 1\ 1\ 2\ 2\ 4\ 4\ 3\ 3\ 4\ 4]\right)x_1x_2x_3 \\ &= \delta_2[2\ 2\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2]x_1x_2x_3. \end{aligned}$$

Similar calculation yields

$$\begin{cases} z_1 = \delta_2[2\ 2\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2]x_1x_2x_3 \\ z_2 = \delta_2[1\ 2\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 1\ 1]x_1x_2x_3 \\ z_3 = \delta_3[1\ 3\ 1\ 3\ 1\ 3\ 3\ 1\ 3\ 1\ 3\ 1]x_1x_2x_3. \end{cases}$$

Using (7.32), we can get

$$M_F = \delta_{12}[7\ 12\ 1\ 3\ 7\ 9\ 9\ 10\ 9\ 7\ 9\ 7].$$

Next, we consider how to retrieve the logical expressions of F from its structure matrix M_F. Assume F is defined by (7.23) with its structure matrix $M_F \in \mathcal{L}_{s\times k}$. Similar to Boolean case, we firstly define a set of logical matrices $S_j \in \mathcal{L}_{s_j\times s}$, called the retrievers as follows:

$$\begin{aligned}
S_1 &= \delta_{s_1}[\underbrace{1\ \cdots\ 1}_{s/s_1}\ \underbrace{2\ \cdots\ 2}_{s/s_1}\ \cdots\ \underbrace{s_1\ \cdots\ s_1}_{s/s_1}];\\
S_2 &= \delta_{s_2}[\underbrace{1\ \cdots\ 1}_{s/s_1s_2}\ \underbrace{2\ \cdots\ 2}_{s/s_1s_2}\ \cdots\ \underbrace{s_2\ \cdots\ s_2}_{s/s_1s_2}\\
&\qquad \cdots\ \underbrace{1\ \cdots\ 1}_{s/s_1s_2}\ \underbrace{2\ \cdots\ 2}_{s/s_1s_2}\ \cdots\ \underbrace{s_2\ \cdots\ s_2}_{s/s_1s_2}];\\
&\ \vdots\\
S_n &= \delta_{s_n}[1\ 2\ \cdots\ s_n\ \cdots\ 1\ 2\ \cdots\ s_n].
\end{aligned} \tag{7.37}$$

We have the following result.

Proposition 7.4. *Assume $z_j \in \mathcal{D}_{s_j}$, $j = 1, 2, \cdots, n$, denote by $s = \prod_{j=1}^n s_j$, let $z = \ltimes_{j=1}^n z_j$, then*

$$z_j = S_j z, \quad j = 1, 2, \cdots, n. \tag{7.38}$$

Proof. Since S_j has $\prod_{i=1}^j s_i$ equal-size blocks, if $z_j = \delta_{s_j}^p$, we have

$$\begin{aligned}
S_j z &= S_j z_1 \cdots z_{j-1} z_j z_{j+1} \cdots z_n\\
&= \delta_{s_j}[\underbrace{1\ \cdots\ 1}_{s/\prod_{i=1}^j s_i}\ \ \underbrace{2\ \cdots\ 2}_{s/\prod_{i=1}^j s_i}\ \ \cdots\ \underbrace{s_j\ \cdots\ s_j}_{s/\prod_{i=1}^j s_i}] z_j z_{j+1} \cdots z_n\\
&= \delta_{s_j}[\underbrace{p\ \cdots\ p}_{s/\prod_{i=1}^j s_i}] z_{j+1} \cdots z_n\\
&= \delta_{s_j}^p = z_j.
\end{aligned}$$

□

By Corollary 7.3 and Proposition 7.4, we have

Corollary 7.4. *The structure matrix M_j of f_j in (7.24) can be retrieved as follows:*

$$M_j = S_j M_F, \quad j = 1, 2, \cdots, n. \tag{7.39}$$

Using Corollary 7.4 we can get the structure matrix M_i which has $x_1, x_2, \cdots, x_n$ as its variables. But in general, there might be some fabricated variables, which do not affect the value of the f_i. To remove these variables, we need the following proposition.

Proposition 7.5. *Consider system (7.27). For arbitrary* $1 \le j \le n$, *split* $M_i W_{[k_j, \prod_{p=1}^{j-1} k_p]}$ *into* k_j *equal-size blocks as*

$$M_i W_{[k_j, \prod_{p=1}^{j-1} k_p]} = [B_1 \ B_2 \ \cdots \ B_{k_j}],$$

where

$$B_i = \mathrm{Blk}_i(M_i W_{[k_j, \prod_{p=1}^{j-1} k_p]}), \quad i = 1, 2, \cdots, k_j.$$

If all the blocks B_i *are the same, then* x_j *is a fabricated variable. Moreover, the equation of* z_i *can be replaced by*

$$z_i = M_i' x_1 \cdots x_{j-1} x_{j+1} \cdots x_n, \tag{7.40}$$

where

$$M_i' = \mathrm{Blk}_1(M_i W_{[k_j, \prod_{p=1}^{j-1} k_p]}) = M_i W_{[k_j, \prod_{p=1}^{j-1} k_p]} \delta_{k_j}^1.$$

Proof. Note that

$$\begin{aligned} z_i &= M_i x_1 \cdots x_{j-1} x_j x_{j+1} \cdots x_n \\ &= M_i W_{[k_j, \prod_{p=1}^{j-1} k_p]} x_j x_1 \cdots x_{j-1} x_{j+1} \cdots x_n, \end{aligned}$$

if x_j does not affect z_i, then z_i is invariant whatever the value of x_j is. Then we can simply set $x_j = \delta_{k_j}^1$ to simplify the expression, which yields (7.40) □

We give an example to depict this.

Example 7.6. Assume the structure matrix of a logical mapping with $x_1, x_3, z_1, z_2 \in \mathcal{D}$, and $x_2, z_3 \in \mathcal{D}_3$ is

$$M_F = \delta_{12}[7 \ 12 \ 1 \ 3 \ 7 \ 9 \ 9 \ 10 \ 9 \ 7 \ 9 \ 7].$$

Then using Corollary 7.4, we have

$$\begin{aligned} M_1 &= S_1 M_F = \delta_2[2 \ 2 \ 1 \ 1 \ 2 \ 2 \ 2 \ 2 \ 2 \ 2 \ 2 \ 2] \\ M_2 &= S_2 M_F = \delta_2[1 \ 2 \ 1 \ 1 \ 1 \ 1 \ 1 \ 2 \ 1 \ 1 \ 1 \ 1] \\ M_3 &= S_3 M_F = \delta_3[1 \ 3 \ 1 \ 3 \ 1 \ 3 \ 3 \ 1 \ 3 \ 1 \ 3 \ 1]. \end{aligned}$$

Next, consider M_1, it is easy to verify that

$$\begin{aligned} M_1 &= \delta_2[2\ 2\ 1\ 1\ 2\ 2 \mid 2\ 2\ 2\ 2\ 2\ 2] \\ M_1 W_{[2]} &= \delta_2[2\ 2\ 2\ 2 \mid 1\ 1\ 2\ 2 \mid 2\ 2\ 2\ 2] \\ M_1 W_{[3,4]} &= \delta_2[2\ 1\ 2\ 2\ 2\ 2 \mid 2\ 1\ 2\ 2\ 2\ 2]. \end{aligned}$$

We conclude that z_1 depends on x_1 and x_2 only. Then z_1 can be simplified as

$$z_1 = \delta_2[2\ 1\ 2\ 2\ 2\ 2]x_1x_2.$$

Similarly, we can remove the fabricated variables from other equations. We have

$$\begin{cases} z_1 = \delta_2[2\ 1\ 2\ 2\ 2\ 2]x_1x_2 \\ z_2 = \delta_2[1\ 2\ 1\ 1\ 1\ 1]x_2x_3 \\ z_3 = \delta_3[1\ 3\ 3\ 1]x_1x_3, \end{cases}$$

which is same to (7.36).

Using Theorem 7.2 we finally have

$$\begin{cases} z_1 = x_1 \wedge (\oslash^2_{2,1,2}x_2) \\ z_2 = \Delta^2_{1,3}(x_2) \wedge x_3 \vee \Delta^2_{2,3}(x_2) \vee \Delta^2_{3,3}(x_2) = (\oslash^2_{2,1,1}x_2) \vee x_3 \\ z_3 = \big(\Delta^3_{1,2}(x_1) \wedge \oslash^3_{1,3}(x_3)\big) \vee \big(\Delta^3_{2,2}(x_1) \wedge \oslash^3_{3,1}(x_3)\big) = \oslash^3_{1,3}(x_1 \leftrightarrow x_3). \end{cases}$$

7.4 Two Practical Examples

This section presents two application examples to show that the mix-valued logical functions have their practical backgrounds.

7.4.1 *Mix-Valued Logical Form of Rules in Fuzzy Control*

As one of the most successful intelligent control technologies, fuzzy control has attracted much attention from control community, and it has been used widely in industries. It is particularly suitable for a variety of control engineering problems, where the models are complicated, uncertain, or unclear. As pointed out in Passino and Yurkovich (1998) "In the fuzzy control design methodology, we ask this operator to write down a set of rules on how to control the process, then we incorporate these into a fuzzy controller that emulates the decision-making process of the human."

Rules play a key role in Fuzzy control. We describe this through the following example.

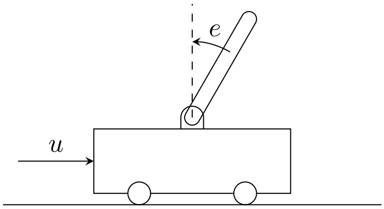


Fig. 7.1 An inverted pendulum

Example 7.7. Fig. 7.1 depicts an inverted pendulum. Denote by e the error, which is the angle leaving from the vertical position (with left side as positive value), $\dot{e}$ is the time derivative of e.

Quantizing the error, change-in-error and the force (control) into 5 levels as: positive-large (denoted by 2), positive-small (denoted by 1), zero (denoted by 0), negative-small (denoted by -1), and negative-large (denoted by -2). The control rules are presented as expert's linguistic description of how to perform the control. Say,

- **If** error is zero and change-in-error is zero **Then** force is zero
- **If** error is zero and change-in-error is positive-small **Then** force is negative small
- $\cdots$

Then the linguistic statements form a set of control rules, which are listed in Table 7.1 (Passino and Yurkovich, 1998).

Table 7.1 Rule table for the inverted pendulum

$e\backslash u\backslash \dot{e}$	-2	-1	0	1	2
-2	2	2	2	1	0
-1	2	2	1	0	-1
0	2	1	0	-1	-2
1	1	0	-1	-2	-2
2	0	-1	-2	-2	-2

Simply identifying $-2 \sim \delta_5^1$, $-1 \sim \delta_5^2$, $0 \sim \delta_5^3$, $1 \sim \delta_5^4$, and $2 \sim \delta_5^5$, we can see that $u(e, \dot{e}) : \mathcal{D}_5 \times \mathcal{D}_5 \to \mathcal{D}_5$ is a logical function. Its algebraic form is

$$u = M_u e(\dot{e}), \tag{7.41}$$

where the structure matrix of u can be easily calculated as

$$M_u = \delta_5[5\ 5\ 5\ 4\ 3\ 5\ 5\ 4\ 3\ 2\ 5\ 4\ 3\ 2\ 1\ 4\ 3\ 2\ 1\ 1\ 3\ 2\ 1\ 1\ 1]. \tag{7.42}$$

We leave to the reader to figure out the logical expression of $u = f(e, \dot{e})$.

In general, for a control system the controls may depend on several variables which can take different numbers of discrete numbers, hence when u is considered as logical functions, they are in general mix-valued logical functions. But to the authors' surprisal, most application examples in reference books the control functions are k-valued ones. A possible reason for this is the k-valued logical functions have corresponding circuit realization.

We consider another example.

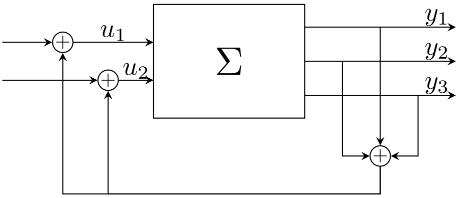


Fig. 7.2 Output-feedback control system

Example 7.8. Given a control system Σ as depicted in Fig. 7.2, which has two inputs u_1, u_2, and three outputs y_i, $i = 1, 2, 3$. Assume the controls can take values as:

$$\begin{aligned} &u_1 : 1,\ 0,\ -1; \\ &u_2 : 2,\ 1,\ 0,\ -1,\ -2; \end{aligned}$$

and the outputs are classified as

$$\begin{aligned} &y_1 : \text{high, low;} \\ &y_2 : \text{high, middle, low;} \\ &y_3 : \text{very high, high, middle, low, very low.} \end{aligned}$$

We simply let $u_1 \in \mathcal{D}_3$ and $u_2 \in \mathcal{D}_5$. Similarly, we have $y_1 \in \mathcal{D}_2$, $y_2 \in \mathcal{D}_3$, and $y_3 \in \mathcal{D}_5$. Consider the output feedback controls, which means the controls depend on outputs. Then the control becomes a mapping

$$\pi : \mathcal{D} \times \mathcal{D}_3 \times \mathcal{D}_5 \to \mathcal{D}_3 \times \mathcal{D}_5.$$

We need 30 statements to describe the control rules as, for instance,

"If y_1 is high and y_2 is high and y_3 is very high, then $u_1 = 1$, and $u_2 = 2$."

Denote by $u = u_1 \ltimes u_2$ and $y = \ltimes_{i=1}^{3} y_i$. Then u can be described as a mix-valued logical mapping, determined by its structure matrix M_u as

$$u = M_u y, \tag{7.43}$$

where $M_u \in \mathcal{L}_{15\times 30}$. As M_u is given, the mix-valued logical expressions for u_i, $i = 1, 2$ can also be constructed. For example, assume

$$M_u = \delta_{15}[6\ 7\ 8\ 9\ 9\ 6\ 6\ 6\ 6\ 6\ 11\ 12\ 12\ 12\ 12\ 1\ 2\ 3\ 4\ 4\ 6\ 6\ 6\ 6\ 6\ 11\ 12\ 12\ 12\ 12].$$

Then the logical expressions can be obtained as

$$\begin{cases} u_1 = \left(\oslash^3_{2,1} y_1\right) \wedge y_2 \\ u_2 = \left(\oslash^5_{4,1,2} y_2\right) \vee y_3. \end{cases}$$

Further discussion of fuzzy sets and fuzzy controls will be done in Chapters 8–10.

7.4.2 *Expression of Strategies of Dynamic Games*

As another application of mix-valued logic, we consider the strategies of infinitely repeated game, which is a kind of dynamic games. This problem will be discussed in detail later, here we give an example to depict it.

Example 7.9. In a game assume there are two players: P_1 and P_2. P_1 has 2 possible actions $S_1 = \{\alpha_1, \alpha_2\}$, and P_2 has 3 possible actions $S_2 = \{\beta_1, \beta_2, \beta_3\}$. Assume the game is infinitely repeated, and the strategies of each players at time $t+1$ depend on the strategies of the players at time t. Denote by $x(t)$ and $y(t)$ the strategies of P_1 and P_2 at time t respectively, then we have

$$\begin{cases} x(t+1) = f_1(x(t), y(t)) \\ y(t+1) = f_2(x(t), y(t)). \end{cases} \tag{7.44}$$

To use vector expression, we identify

$$\begin{aligned} &\alpha_1 \sim \delta_2^1, \quad \alpha_2 \sim \delta_2^2; \\ &\beta_1 \sim \delta_3^1, \quad \beta_2 \sim \delta_3^2, \quad \beta_3 = \delta_3^3. \end{aligned}$$

Then $x(t) \in \Delta_2$ and $y(t) \in \Delta_3$, and we can find the structure matrices $M_1 \in \mathcal{L}_{2\times 6}$ and $M_2 \in \mathcal{L}_{3\times 6}$ of f_1 and f_2 respectively, such that (7.44) can be expressed as

$$\begin{cases} x(t+1) = M_1 x(t) \ltimes y(t) \\ y(t+1) = M_2 x(t) \ltimes y(t). \end{cases} \tag{7.45}$$

Furthermore, setting $z(t) = x(t) \ltimes y(t)$, we have

$$z(t+1) = Lz(t), \tag{7.46}$$

where $L = M_1 * M_2 \in \mathcal{L}_{6\times 6}$.

To a numerical expression, we assume

$$L = \delta_6[1\ 3\ 5\ 2\ 4\ 6].$$

Then we have

$$M_1 = \delta_2[1\ 1\ 2\ 1\ 2\ 2], \quad M_2 = \delta_3[1\ 3\ 2\ 2\ 1\ 3].$$

It follows that

$$\begin{cases} f_1 = x_1 \wedge \oslash^2_{1,1,2} x_2 \vee (\neg x_1 \wedge \oslash^2_{1,2,2} x_2) \\ f_2 = \nabla^3_{1,2} x_1 \wedge \oslash^3_{1,3,2} x_2 \vee (\nabla^3_{2,2} x_1 \wedge \oslash^3_{2,1,3} x_2). \end{cases}$$

Exercises

7.1 For Boolean logic, use $\{\neg, \wedge, \vee\}$ to express the other binary operators.

7.2 For Boolean logic, use Proposition 7.1 to prove that $\{\neg, \rightarrow\}$ is an adequate set.

7.3 In fact, the general expression of logical operators is one-one correspondence to their structure matrices. Do the conversion for the following operations:

(i) Find the structure matrices of the following k-valued logical operators: (a) $\oslash_{*,*,2,*}$, (b) $\oslash_{3,3,*,*}$.

(ii) Find the structure matrices of the following mix-valued logical operators: (a) $\oslash^3_{2,1,2,3}$, (b) $\oslash^5_{2,4,1}$.

(iii) Find the following k-valued logical operators from their structure matrices: (a) $M_\sigma = \delta_4[2\ 3\ 3\ 4]$, (b) $M_\sigma = \delta_5[1\ 3\ 5\ 2\ 4]$.

(vi) Find the mix-valued logical operators from their structure matrices: (a) $M_\sigma = \delta_3[2\ 3\ 1\ 1]$, (b) $M_\sigma = \delta_5[4\ 3\ 1\ 2]$.

7.4 A 3-valued 3-ary logical function

$$f(x_1, x_2, x_3) = x_1 \leftrightarrow (x_2 \vee x_3),$$

where

$$\begin{aligned} p \rightarrow q &:= \neg p \vee q \\ p \leftrightarrow q &:= (p \rightarrow q) \wedge (q \rightarrow p). \end{aligned}$$

(i) Find the structure matrix of f.

(ii) Find the disjunctive normal form of f.

7.5 Prove De Morgan's Law for k-values logic.

7.6 State and prove the conjunctive normal form of k-valued r-ary logical expressions.

7.7 A mix-valued logical function $f : \mathcal{D}_2 \times \mathcal{D}_3 \times \mathcal{D}_2 \rightarrow \mathcal{D}_2$ is defined as

$$f(x_1, x_2, x_3) = \left[x_1 \wedge \oslash^2_{121}(x_2)\right] \leftrightarrow x_3.$$

(i) Find the structure matrix of f.

(ii) Find the disjunctive normal form of f.

7.8 A mix-valued logical function $f : \mathcal{D}_2 \times \mathcal{D}_3 \to \mathcal{D}_3$ has the structure matrix as

$$M_f = \delta_3[1\ 3\ 2\ 2\ 2\ 1].$$

Find its logical expression.

7.9 A mix-valued logical mapping F is defined by

$$\begin{cases} z_1 = f_1(x_1, x_2, x_3) \\ z_2 = f_2(x_1, x_2, x_3), \end{cases}$$

where $x_1, x_3, z_1 \in \mathcal{D}$ and $x_2, z_2 \in \mathcal{D}_3$. Let $z = z_1 z_2$ and $x = x_1 x_2 x_3$. Then the structure matrix of F is

$$M_F = \delta_6[1\ 3\ 5\ 2\ 4\ 6\ 2\ 4\ 6\ 1\ 3\ 5].$$

Find the logical expressions of f_1 and f_2.

7.10 Prove Proposition 7.2.

7.11 Prove Proposition 7.3.

7.12 Using the structure matrix (7.42) to figure out the logical expression of $u = f(e, \dot{e})$.

7.13 In a fuzzy control the control u_1 can take 2 values, and u_2 3 values. Let $u = u_1 \ltimes u_2$. Its relation with respect to the outputs y_1 and y_2 is shown in Table 7.2.

Table 7.2 Rule table for Exercise 7.13

$y_1\backslash u\backslash Y_2$	-2	-1	0	1	2
-2	-2	-1	-1	0	0
-1	-1	0	0	1	1
0	0	1	1	2	2
1	1	2	2	3	3
2	2	2	3	3	3

(i) Find the structure matrix M_F such that

$$u = M_F y,$$

where $u = u_1 u_2$, and $y = y_1 y_2$.

(ii) Figure out the logical expression of

$$\begin{cases} u_1 = f_1(y_1, y_2) \\ u_2 = f_2(y_1, y_2). \end{cases}$$

Chapter 8

Logical Matrix, Fuzzy Set and Fuzzy Logic

In 1965, L.A. Zadeh firstly proposed the fuzzy set theory to describe fuzzy nature in Zadeh (1965), which created a new area of fuzzy mathematics and applications.

In this chapter we first investigate the matrix expression of general fuzzy sets, their logical operators, etc. Then the fuzzy mappings and their expressions are studied. Finally, the fuzzy logic and the calculation of fuzzy logical functions via their matrix expressions are considered.

We refer to Kerre *et al.* (2004); Hu (2010) for some fundamental concepts and results about fuzzy logic, and to Dubois and Prade (2000) for some advanced discussions.

8.1 Matrices of General Logical Variables

Definition 8.1. Let $A = (a_{ij}) \in \mathcal{M}_{m\times n}$. A is called a k-valued matrix, if its entries $a_{ij} \in \mathcal{D}_k$. We allow $2 \le k \le \infty$. When $k = 2$, A is called a Boolean matrix. When $k = \infty$, $\mathcal{D}_\infty := [0, 1]$, and A is called a fuzzy matrix. The set of k-valued $m \times n$ matrices is denoted by $\mathcal{D}_k^{m\times n}$. When $k = 2$, the subscript k can be omitted, i.e., $\mathcal{D}^{m\times n} := \mathcal{D}_2^{m\times n}$. An alternative notation is: $\mathcal{B}_{m\times n} := \mathcal{D}^{m\times n}$.

When $m = 1$ ($n = 1$) it is called a row (column) k-valued vector. The set of n-dimensional k-valued vectors is denoted by $\mathcal{D}_k^m$.

Next, we define the logical operators on $\mathcal{D}_k^{m\times n}$.

Definition 8.2.

(1) Let $\alpha, \beta \in \mathcal{D}_k$. Then

$$\neg\alpha := 1 - \alpha. \tag{8.1}$$

$$\alpha \wedge \beta := \min\{\alpha, \beta\}. \tag{8.2}$$

$$\alpha \vee \beta := \max\{\alpha, \beta\}. \tag{8.3}$$

(2) Let $A = (a_{ij}), B = (b_{ij}) \in \mathcal{D}_k^{m\times n}$. Then

$$\neg A := (\neg a_{ij}). \tag{8.4}$$

$$A \wedge B := (a_{ij} \wedge b_{i,j}). \tag{8.5}$$

$$A \vee B := (a_{ij} \vee b_{i,j}). \tag{8.6}$$

We give some new notations:

(i) $\mathbf{1}_{m\times n}$: $\mathbf{1}_{m\times n} \in \mathcal{D}^{m\times n}$ with all entries equal to 1.
(ii) $\mathbf{1}_m := \mathbf{1}_{m\times 1}$. If m is obvious, it can be omitted.
(iii) $\mathbf{0}_{m\times n}$: $\mathbf{0}_{m\times n} \in \mathcal{D}^{m\times n}$ with all entries equal to 0.
(iv) $\mathbf{0}_m := \mathbf{0}_{m\times 1}$. If m is obvious, it can be omitted.
(v) Let $A = (a_{ij}), B = (b_{ij}) \in \mathcal{D}_k^{m\times n}$. Then

$$A \le B \;\Leftrightarrow\; a_{ij} \le b_{ij},\ \forall\, i, j.$$

(vi) Let $\alpha \in \mathcal{D}_k$ and $A = (a_{ij}) \in \mathcal{D}_k^{m\times n}$. Then

$$\alpha A = A\alpha := (\alpha \wedge a_{i,j}) \in \mathcal{D}_k^{m\times n}.$$

The following simple examples are used to depict the operators.

Example 8.1. Let

$$A = \begin{bmatrix} 0.2 & 0.5 \\ 1 & 0.7 \end{bmatrix}; \quad B = \begin{bmatrix} 0.4 & 0.6 \\ 0.8 & 0 \end{bmatrix}.$$

Then

$$\neg A = \begin{bmatrix} 0.8 & 0.5 \\ 0 & 0.3 \end{bmatrix}; \quad A \wedge B = \begin{bmatrix} 0.2 & 0.5 \\ 0.8 & 0 \end{bmatrix};$$

$$A \vee B = \begin{bmatrix} 0.4 & 0.6 \\ 1 & 0.7 \end{bmatrix}; \quad (0.5)A = \begin{bmatrix} 0.2 & 0.5 \\ 0.5 & 0.5 \end{bmatrix};$$

$$A \wedge \mathbf{1}_{2\times 2} = A = \begin{bmatrix} 0.2 & 0.5 \\ 1 & 0.7 \end{bmatrix}; \quad B \vee \mathbf{0}_{2\times 2} = B = \begin{bmatrix} 0.4 & 0.6 \\ 0.8 & 0 \end{bmatrix}.$$

The following properties are inherited from the corresponding properties for scalar logical variables.

Proposition 8.1. *Let $A, B, C, \mathbf{1}, \mathbf{0} \in \mathcal{D}_k^{m\times n}$. Then*

(i) (Idempotent Law)

$$A \wedge A = A; \quad A \vee A = A. \tag{8.7}$$

(ii)

$$A \wedge \mathbf{0} = \mathbf{0}; \quad A \vee \mathbf{1} = \mathbf{1}. \tag{8.8}$$

(iii)

$$A \vee \mathbf{0} = A; \quad A \wedge \mathbf{1} = A. \tag{8.9}$$

(iv) (Commutative Law)

$$A \wedge B = B \wedge A; \quad A \vee B = B \vee A. \tag{8.10}$$

(v) (Associative Law)

$$(A \wedge B) \wedge C = A \wedge (B \wedge C); \quad (A \vee B) \vee C = A \vee (B \vee C). \tag{8.11}$$

(vi) (Distributive Law)

$$\begin{aligned}(A \wedge B) \vee C = (A \vee C) \wedge (B \vee C);\\ (A \vee B) \wedge C = (A \wedge C) \vee (B \wedge C).\end{aligned} \tag{8.12}$$

(vii) (Absorptive Law)

$$(A \wedge B) \vee A = A; \quad (A \vee B) \wedge A = A. \tag{8.13}$$

(viii)

$$\neg(\neg A) = A. \tag{8.14}$$

(ix) (DeMorgan's Law)

$$\neg(A \wedge B) = (\neg A) \vee (\neg B); \quad \neg(A \vee B) = (\neg A) \wedge (\neg B). \tag{8.15}$$

(x) Assume $A \le B$. Then

$$A \wedge B = A; \quad A \vee B = B. \tag{8.16}$$

(xi) Assume $A \le B$ and $C \le D$. Then

$$A \wedge C \le B \wedge D; \quad A \vee C \le B \vee D. \tag{8.17}$$

(xii) Assume $A \le B$. Then

$$\neg A \ge \neg B. \tag{8.18}$$

8.2 Logical Operators for k-Valued Matrices

Definition 8.3. Consider $\mathcal{D}_k$. We define two logical operators as

(i) (Logical Addition)

$$\alpha\,(+)\,\beta := \alpha \vee \beta, \quad \alpha, \beta \in \mathcal{D}_k. \tag{8.19}$$

(ii) (Logical Product)

$$\alpha\,(\times)\,\beta := \alpha \wedge \beta, \quad \alpha, \beta \in \mathcal{D}_k. \tag{8.20}$$

Note that we also use the following notations for multi-addition and multi-product.

$$(+)_{i=1}^{n}\,\alpha_i = \alpha_1\,(+)\,\alpha_2\,(+)\cdots(+)\,\alpha_n;$$

$$(\times)_{i=1}^{n}\,\alpha_i = \alpha_1\,(\times)\,\alpha_2\,(\times)\cdots(\times)\,\alpha_n.$$

Using these, we can define the logical addition and product for matrices.

Definition 8.4.

(1) Let $\alpha \in \mathcal{D}_k$ and $A = (a_{i,j}) \in \mathcal{D}_k^{m\times n}$. Then

$$\alpha\,(\times)\,A = A\,(\times)\,\alpha := (\alpha \wedge a_{ij})\,. \tag{8.21}$$

(We simply use αA for $\alpha\,(\times)\,A$ and $A\alpha$ for $A\,(\times)\,\alpha$.)

(2) Let $A = (a_{i,j}) \in \mathcal{D}_k^{m\times n}$ and $B = (b_{i,j}) \in \mathcal{D}_k^{n\times p}$. Then

$$A\,(\times)\,B := C = (c_{ij}) \in \mathcal{D}_k^{m\times p}, \tag{8.22}$$

where

$$c_{ij} = (+)_{k=1}^{n}\,a_{ik}\,(\times)\,b_{kj}, \quad i = 1,\cdots,m;\ j = 1,\cdots p.$$

(3) Let $A = (a_{i,j}) \in \mathcal{D}_k^{m\times n}$ and $B = (b_{i,j}) \in \mathcal{D}_k^{p\times q}$. Then

$$A\,(\otimes)\,B := \begin{bmatrix} a_{11}\,(\times)\,B & a_{12}\,(\times)\,B & \cdots & a_{1n}\,(\times)\,B \\ a_{21}\,(\times)\,B & a_{22}\,(\times)\,B & \cdots & a_{2n}\,(\times)\,B \\ \vdots & & & \\ a_{m1}\,(\times)\,B & a_{m2}\,(\times)\,B & \cdots & a_{mn}\,(\times)\,B \end{bmatrix} \in \mathcal{D}_k^{mp\times nq}. \tag{8.23}$$

(4) Let $A \in \mathcal{D}_k^{n\times n}$. Then

$$A^{(m+1)} := A^{(m)}\,(\times)\,A, \quad m = 1,2,\cdots. \tag{8.24}$$

(5) Let $A = (a_{i,j}) \in \mathcal{D}_k^{m\times n}$, $B = (b_{i,j}) \in \mathcal{D}_k^{p\times q}$.

(i) If $n = pt$, then

$$A\,(\ltimes)\,B := A\,(\times)\,(B\,(\otimes)\,I_t)\,. \tag{8.25}$$

(ii) If $nt = p$, then

$$A\,(\ltimes)\,B := (A\,(\otimes)\,I_t)\,(\times)\,B. \tag{8.26}$$

(6) Let $A = \in \mathcal{D}_k^{m\times s}$, $B = \in \mathcal{D}_k^{n\times s}$. Then the Khatri-Rao-Boolean product

$$\begin{aligned} A\,(*)\,B\,:=\,&[\mathrm{Col}_1(A)\,(\ltimes)\,\mathrm{Col}_1(B)\ \ \mathrm{Col}_2(A)\,(\ltimes)\,\mathrm{Col}_2(B) \\ &\cdots\,\mathrm{Col}_s(A)\,(\ltimes)\,\mathrm{Col}_s(B)]\,. \end{aligned} \tag{8.27}$$

The following examples are used to depict these.

Example 8.2.

(1) Let

$$A = \begin{bmatrix} 0.2 & 0.4 \\ 0.6 & 0.8 \end{bmatrix}; \quad B = \begin{bmatrix} 0 & 0.2 & 0.4 \\ 0.6 & 0.8 & 1 \end{bmatrix}; \quad C = \begin{bmatrix} 0 & 0.8 \\ 0.2 & 1 \\ 0.4 & 0.7 \\ 0.6 & 0.3 \end{bmatrix}.$$

Then

$$A\,(\times)\,B = \begin{bmatrix} 0.4 & 0.4 & 0.4 \\ 0.6 & 0.8 & 0.8 \end{bmatrix};$$

$$A\,(\otimes)\,B = \begin{bmatrix} 0 & 0.2 & 0.2 & 0 & 0.2 & 0.4 \\ 0.2 & 0.2 & 0.2 & 0.4 & 0.4 & 0.4 \\ 0 & 0.2 & 0.4 & 0 & 0.2 & 0.4 \\ 0.6 & 0.6 & 0.6 & 0.6 & 0.8 & 0.8 \end{bmatrix};$$

$$A\,(\otimes)\,I_2 = \begin{bmatrix} 0.2 & 0 & 0.4 & 0 \\ 0 & 0.2 & 0 & 0.4 \\ 0.6 & 0 & 0.8 & 0 \\ 0 & 0.6 & 0 & 0.8 \end{bmatrix};$$

$$A\,(\ltimes)\,C = \begin{bmatrix} 0.4 & 0.4 \\ 0.4 & 0.3 \\ 0.4 & 0.7 \\ 0.6 & 0.6 \end{bmatrix}.$$

$$A^{(2)} = \begin{bmatrix} 0.4 & 0.4 \\ 0.6 & 0.8 \end{bmatrix}; \quad A^{(k)} = \begin{bmatrix} 0.4 & 0.4 \\ 0.6 & 0.8 \end{bmatrix}, \quad k \geq 3.$$

(2) Let

$$A = \begin{bmatrix} 0.2 & 0.3 \\ 0.7 & 0.9 \end{bmatrix}; \quad X = \begin{bmatrix} 0.3 \\ 0.5 \end{bmatrix}; \quad Y = \begin{bmatrix} 0.2 \\ 0.8 \end{bmatrix}; \quad \alpha = 0.4.$$

Then

$$\alpha A \,(\ltimes)(X \wedge Y) = 0.4 \begin{bmatrix} 0.2 & 0.3 \\ 0.7 & 0.9 \end{bmatrix} (\ltimes) \begin{bmatrix} 0.2 \\ 0.5 \end{bmatrix} = 0.4 \begin{bmatrix} 0.3 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.4 \end{bmatrix}.$$

(3) Let

$$A = \begin{bmatrix} 0.2 & 0.3 & 0 \\ 0.4 & 0.5 & 0.8 \end{bmatrix}; \quad B = \begin{bmatrix} 1 & 0.4 & 0.6 \\ 0.3 & 0.9 & 0.2 \\ 0 & 1 & 0.7 \end{bmatrix}.$$

Then

$$A\,(*)\,B = \begin{bmatrix} 0.2 & 0.3 & 0 \\ 0.2 & 0.3 & 0 \\ 0 & 0.3 & 0 \\ 0.4 & 0.4 & 0.6 \\ 0.3 & 0.5 & 0.2 \\ 0 & 0.5 & 0.7 \end{bmatrix}.$$

The following properties are easily verifiable.

Proposition 8.2. *Let $R, S, T \in \mathcal{D}_k^{n\times n}$ be k-valued matrices. Then*

(1)

$$R\,(\times)\,I = I\,(\times)\,R = R. \tag{8.28}$$

(2)

$$R\,(\times)\,\mathbf{0} = \mathbf{0}\,(\times)\,R = \mathbf{0}. \tag{8.29}$$

(3)

$$R^{(m+n)} = R^{(m)}\,(\times)\,R^{(n)}. \tag{8.30}$$

(4) Assume $S \le T$. Then

$$R\,(\times)\,S \le R\,(\times)\,T. \tag{8.31}$$

(5) (Associative Law)

$$(R\,(\times)\,S)\,(\times)\,T = R\,(\times)\,(S\,(\times)\,T). \tag{8.32}$$

(6) (Distributive Law)

(i)

$$R\,(\times)(S \vee T) = (R\,(\times)\,S) \vee (R\,(\times)\,T)\,. \tag{8.33}$$

(ii)

$$(S \vee T)\,(\times)\,R = (S\,(\times)\,R) \vee (T\,(\times)\,R)\,. \tag{8.34}$$

(iii)

$$R\,(\times)(S \wedge T) = (R\,(\times)\,S) \wedge (R\,(\times)\,T)\,. \tag{8.35}$$

(iv)

$$(S \wedge T)\,(\times)\,R = (S\,(\times)\,R) \wedge (T\,(\times)\,R)\,. \tag{8.36}$$

(7)

$$(R\,(\times)\,S)^T = S^T\,(\times)\,R^T. \tag{8.37}$$

8.3 Fuzzy Sets

Definition 8.5. Consider an objective set E, called a universe of discourse. A set A is called a fuzzy set over E if for each $e \in E$ there is a membership degree $\mu_A(e) = \alpha_e \in \mathcal{D}_\infty$. If $E = \{e_1, \cdots, e_n\}$ is a finite set and $\mu_A(e_i) = \alpha_i$, $i = 1, \cdots, n$, then A can be expressed as

$$A = \alpha_1/e_1 + \alpha_2/e_2 + \cdots + \alpha_n/e_n. \tag{8.38}$$

The set of fuzzy sets over the universe of discourse E is denoted by $\mathcal{F}(E)$.

Consider a set E, its power set, denoted by $\mathcal{P}(E)$, is the set of all subsets of E. That is,

$$\mathcal{P}(E) = \{S \mid S \subset E\}.$$

Consider $C \in \mathcal{P}(E)$. C can also be considered as a special fuzzy set, with

$$\mu_C(e) = \begin{cases} 1, & e \in C \\ 0, & \text{otherwise.} \end{cases}$$

Moreover, $p \in E$ can be considered as a special power set, which have

$$\mu_p(e) = \begin{cases} 1, & e = p \\ 0, & \text{otherwise.} \end{cases}$$

Under this unified description we have

$$E \subset \mathcal{P}(E) \subset \mathcal{F}(E). \tag{8.39}$$

To distinguish $C \in \mathcal{P}(E)$ with fuzzy sets, we call C a crisp set.

Definition 8.6. Let A and B be two fuzzy sets on E.

(1) $A = \varnothing$, if

$$\mu_A(e) = 0, \quad \forall\, e \in E.$$

(2) $A = E$, if

$$\mu_A(e) = 1, \quad \forall\, e \in E.$$

(3) $A \subset B$, if and only if

$$\mu_A(e) \le \mu_B(e), \quad \forall\, e \in E.$$

(4) $A \cap B$ is defined by

$$\mu_{A\cap B}(e) = \mu_A(e) \wedge \mu_B(e), \quad \forall\, e \in E.$$

(5) $A \cup B$ is defined by

$$\mu_{A\cup B}(e) = \mu_A(e) \vee \mu_B(e), \quad \forall\, e \in E.$$

(6) A^c is defined by

$$\mu_{A^c}(e) = \neg\mu_A(e), \quad \forall\, e \in E.$$

Definition 8.7. Assume the universe of discourse $|E| < \infty$, and a fuzzy set A on E is as in (8.38). Then the vector form of A, denoted by $\mathcal{V}_A$, is defined as

$$\mathcal{V}_A = (\alpha_1\ \alpha_2\ \cdots\ \alpha_n)^T \in \mathcal{D}_k^n. \tag{8.40}$$

Let $P = (p_{ij}), Q = (q_{ij}) \in M_{m\times n}$. Then $P \le Q$ means

$$p_{i,j} \le q_{i,j}, \quad i = 1, \cdots, m; j = 1, \cdots, n.$$

When $|E| < \infty$, according to Definition 8.6, we have that

Proposition 8.3. *Let A and B be two fuzzy sets on E.*

(1) $A = \varnothing$, if and only if

$$\mathcal{V}_A = \mathbf{0}.$$

(2) $A = E$, if and only if

$$\mathcal{V}_A = \mathbf{1}.$$

(3) $A \subset B$, if and only if

$$\mathcal{V}_A \le \mathcal{V}_B.$$

(4)

$$\mathcal{V}_{A\cap B} = \mathcal{V}_A \wedge \mathcal{V}_B. \tag{8.41}$$

(5)

$$\mathcal{V}_{A\cup B} = \mathcal{V}_A \vee \mathcal{V}_B. \tag{8.42}$$

(6)

$$\mathcal{V}_{A^c} = \neg\mathcal{V}_A. \tag{8.43}$$

Remark 8.1. Using Proposition 8.3 and the properties of logical operators, we can calculate the vector of any logical expressions. For instance,

$$\mathcal{V}_{(A\cap B)^c} = \neg\left(\mathcal{V}_A \wedge \mathcal{V}_B\right) = (\neg\mathcal{V}_A) \vee (\neg\mathcal{V}_B) = \mathcal{V}_{A^c} \vee \mathcal{V}_{B^c}. \tag{8.44}$$

$$\mathcal{V}_{(A\cup B)^c} = \neg\left(\mathcal{V}_A \vee \mathcal{V}_B\right) = (\neg\mathcal{V}_A) \wedge (\neg\mathcal{V}_B) = \mathcal{V}_{A^c} \wedge \mathcal{V}_{B^c}. \tag{8.45}$$

Example 8.3.

(1) Let $E = \{x_1, x_2, x_3, x_4, x_5\}$.

(i) $A = 0.1/x_1 + 0.3/x_2 + 1/x_5 \in \mathcal{F}(E)$. Then

$$\mathcal{V}_A = [0.1\ 0.3\ 0\ 0\ 1]^T.$$

(ii) $B = \{x_2, x_4, x_5\} \in \mathcal{P}(E)$. Then

$$\mathcal{V}_B = [0\ 1\ 0\ 1\ 1]^T.$$

(iii) $C = x_3 \in E$. Then

$$\mathcal{V}_C = [0\ 0\ 1\ 0\ 0]^T.$$

(2) Let $E = \{x_1, x_2, x_3\}$ and there are three fuzzy sets

$$A_1 = 0/x_1 + 0.25/x_2 + 0.75/x_3;$$
$$A_2 = 0.5/x_1 + 0.75/x_2 + 0/x_3;$$
$$A_3 = 1/x_1 + 0.5/x_2 + 0/x_3.$$

Then we can choose $k = 5$ and find their corresponding vector form as:

$$\mathcal{V}_1 = \mathcal{V}_{A_1} = \begin{bmatrix} 0 \\ 0.25 \\ 0.75 \end{bmatrix}; \quad \mathcal{V}_2 = \mathcal{V}_{A_2} = \begin{bmatrix} 0.5 \\ 0.75 \\ 0 \end{bmatrix}; \quad \mathcal{V}_3 = \mathcal{V}_{A_3} = \begin{bmatrix} 1 \\ 0.5 \\ 0 \end{bmatrix}.$$

Consider $S = A_1 \cup A_2 \cup A_3$. Then

$$\mathcal{V}_S = \mathcal{V}_1 \vee \mathcal{V}_2 \vee \mathcal{V}_3 = \begin{bmatrix} 1 \\ 0.75 \\ 0.75 \end{bmatrix}.$$

Equivalently, $S = 1/x_1 + 0.75/x_2 + 0.75/x_3$. Consider $T = A_1 \cap A_2 \cap A_3$. Then

$$\mathcal{V}_T = \mathcal{V}_1 \wedge \mathcal{V}_2 \wedge \mathcal{V}_3 = \begin{bmatrix} 0 \\ 0.25 \\ 0 \end{bmatrix}.$$

Equivalently, $T = 0/x_1 + 0.25/x_2 + 0/x_3$.

Throughout this chapter we assume

A1 The fuzzy sets, in which we are interested, is finite.
A2 The number of membership degrees for each fuzzy set is finite.

Theorem 8.1. *Under Assumptions* **A1** *and* **A2***, the universe of discourse can be equivalent to a finite partition.*

Proof. Assume the universe of discourse is E and the fuzzy sets concerned are $A_1, \cdots, A_s$, and the number of membership degrees for A_i is $k_i < \infty$, $i = 1, \cdots, s$. Define

$$E^i_j := \left\{ e \in E \;\middle|\; \mu_{A_i}(e) = \frac{j-1}{k_i - 1} \right\}, \quad j = 1, \cdots, k_i,\ i = 1, \cdots, s.$$

Using E^i_j, we define a set of subsets of E as

$$E_{j_1,\cdots,j_s} := E^1_{j_1} \cap E^2_{j_2} \cap \cdots \cap E^s_{j_s}, \quad j_i = 1, \cdots, k_i;\ i = 1, \cdots, s.$$

Then $\{ E_{j_1,\cdots,j_s} | \, j_i = 1, \cdots, k_i;\ i = 1, \cdots, s\}$ is a partition of E. That is

(1)

$$\cup_{j_1=1}^{k_1} \cup_{j_2=1}^{k_2} \cdots \cup_{j_s=1}^{k_s} E_{j_1,\cdots,j_s} = E;$$

(2)

$$E_{j_1,\cdots,j_s} \cap E_{j'_1,\cdots,j'_s} = \varnothing, \quad (j_1, \cdots, j_s) \neq (j'_1, \cdots, j'_s).$$

Now it is clear that when considering the fuzzy sets A_1, $\cdots$, A_s, we do not need to distinguish two points within a $E_{j_1,\cdots,j_s}$. Hence we can consider $E_{j_1,\cdots,j_s}$ as "one element" in the university of discourse E provided $E_{j_1,\cdots,j_s} \neq \varnothing$ (If $E_{j_1,\cdots,j_s} = \varnothing$, we can just ignore it). Then

$$E = \{ E_{j_1,\cdots,j_s} | \, E_{j_1,\cdots,j_s} \neq \varnothing;\ j_1, \cdots, j_s = 1, \cdots, k\}$$

can be treated as a finite set. □

We give an example to depict this.

Example 8.4. Let E be the set of human age. It could be $[0, \infty)$. Say, we are concerning two fuzzy sets: A: One is old; B: One is rational. Assume

$$\mu_A(x) = \begin{cases} 0, & x < 20 \\ \frac{1}{3}, & 20 \leq x < 40 \\ \frac{2}{3}, & 40 \leq x < 60 \\ 1, & x \geq 60, \end{cases}$$

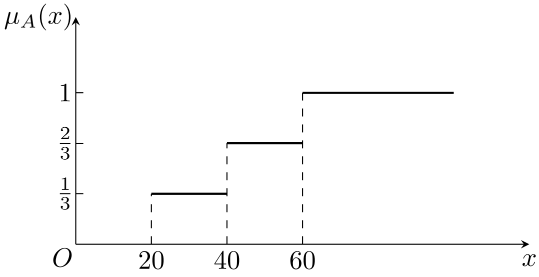


Fig. 8.1 Membership of A

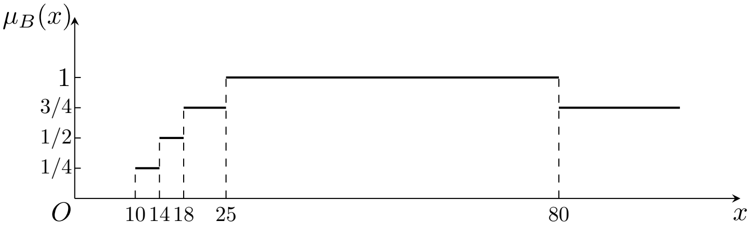


Fig. 8.2 Membership of B

$$\mu_B(x) = \begin{cases} 0, & x < 10 \\ \frac{1}{4}, & 10 \le x < 14 \\ \frac{1}{2}, & 14 \le x < 18 \\ \frac{3}{4}, & 18 \le x < 25, \text{ or } x \ge 80 \\ 1, & 25 \le x < 80. \end{cases}$$

Then the universe of discourse can be partitioned as

$$E = \{E_{11}, E_{12}, E_{13}, E_{14}, E_{24}, E_{25}, E_{35}, E_{45}, E_{44}\},$$

where

$$\begin{aligned} &E_{11} = [0, 10); \quad E_{12} = [10, 14); E_{13} = [14, 18); \\ &E_{14} = [18, 20); E_{24} = [20, 25); E_{25} = [20, 40); \\ &E_{35} = [40, 60); E_{45} = [60, 80); E_{44} = [80, \infty). \end{aligned}$$

We conclude that when only A, B and the fuzzy sets generated by them are concerned, E can be considered as a finite set with 9 elements.

Based on Theorem 8.1 we, hereafter, assume

A3 The universe of discourse of any fuzzy set is a finite set, i.e., $|E| < \infty$.

Then for each fuzzy set A we have $\mathcal{V}_A \in \mathcal{D}_k^n$, where $n = |E|$.

Note that when $k = 2$ we have a "complement rule" as

$$\mathcal{V}_A \vee \neg\mathcal{V}_A = \mathbf{1}; \quad \mathcal{V}_A \wedge \neg\mathcal{V}_A = \mathbf{0}. \tag{8.46}$$

When $k > 2$ (8.46) is not true.

8.4 Mappings over Fuzzy Sets

First, we consider the decomposition of a fuzzy set.

Definition 8.8. Let $\alpha \in \mathcal{D}_\infty$ and A be a fuzzy set. Then the α-truncated set of A is defined as

$$A_\alpha = \{e \,|\, \mu_A(e) \geq \alpha\}.$$

Note that the truncated set A_α is a crisp set. Let $\mathcal{V}_A$ be the vector expression of A. Then the components of $\mathcal{V}_{A_\alpha}$ can be determined as

$$[\mathcal{V}_{A_\alpha}]_i = \begin{cases} 0, & [\mathcal{V}_A]_i < \alpha \\ 1, & [\mathcal{V}_A]_i \geq \alpha. \end{cases}$$

The following decomposition theorem can be proved by a straightforward verification.

Theorem 8.2 (Decomposition Theorem). *Let A be a fuzzy set. Then*

$$\mathcal{V}_A = \vee_{\alpha \in \mathcal{D}_\infty} \alpha \mathcal{V}_{A_\alpha}. \tag{8.47}$$

Next, we consider how to extend a mapping $f : E \to F$ to the fuzzy sets $f : \mathcal{F}(E) \to \mathcal{F}(F)$. Recall the including relation (8.39). We first extend it to $f : \mathcal{P}(E) \to \mathcal{P}(F)$.

Definition 8.9. Let E and F be two arbitrary sets, and a mapping $f :$ $E \to F$ is given.

(1) f can naturally be extended to $f : \mathcal{P}(E) \to \mathcal{P}(F)$ as

$$f(S) = \{f(x)|x \in S\} \in \mathcal{P}(F), \quad S \in \mathcal{P}(E). \tag{8.48}$$

(2) $f^{-1} : \mathcal{P}(F) \to \mathcal{P}(E)$ is defined as

$$f^{-1}(T) = \{x \,|\, f(x) \in T\}, \quad T \in \mathcal{P}(F). \tag{8.49}$$

Next, we extend f further to $\mathcal{F}(E) \to \mathcal{F}(F)$. The following definition was proposed firstly by Zadeh (1972).

Definition 8.10. Assume $f : E \to F$ is given.

(1) Then f can be extended to $\mathcal{F}(E) \to \mathcal{F}(F)$ as follows:

$$\mu_{f(A)}(y) = \begin{cases} \vee_{x \in f^{-1}(y)} \mu_A(x), & A \in \mathcal{F}(E) \\ \varnothing, & f^{-1}(y) = \varnothing. \end{cases} \tag{8.50}$$

(2) The inverse $f^{-1} : \mathcal{F}(F) \to \mathcal{F}(E)$ is defined as

$$\mu_{f^{-1}(B)}(x) = \mu_B(f(x)). \tag{8.51}$$

Assume $E = \{e_1, e_2, \cdots, e_n\}$ and $F = \{d_1, d_2, \cdots, d_m\}$, and $f : E \to F$ is defined by

$$f(e_i) = d_{j_i}, \quad i = 1, \cdots, n;\ 1 \le j_i \le m.$$

Identifying e_i with its vector form as $e_i \sim \mathcal{V}_{e_i} = \delta_n^i$, $i = 1, \cdots, n$ and $d_j \sim \mathcal{V}_{d_j} = \delta_m^j$, $j = 1, \cdots, m$, we have, in vector form,

$$f(x) = M_f \mathcal{V}_x := \delta_m[j_1\ j_2\ \cdots\ j_n]\mathcal{V}_x, \tag{8.52}$$

where M_f is called the structure matrix of f.

Example 8.5. Let $E = \{1, 2, 3, 4, 5\}$, $F = \{0, 1, 2\}$. and $f : E \to F$ is defined by $f(x) = x^3 \pmod 3$. Then it is easy to figure out that

$$M_f = \delta_3[2, 3, 1, 2, 3]. \tag{8.53}$$

Using (8.52), let $x = 4 \sim \delta_5^4$. Then in vector form

$$f(x) = M_f \delta_5^4 = \delta_5^2.$$

Hence, $f(x) = 1$.

Theorem 8.3. *Assume $|E| = n$ and $|F| = m$ and $f : E \to F$ has its structure matrix $M_f \in \mathcal{L}_{m \times n}$. Then*

(1)

$$\mathcal{V}_{f(A)} = M_f\,(\times)\,\mathcal{V}_A, \quad \forall\, A \in \mathcal{F}(E). \tag{8.54}$$

(2)

$$\mathcal{V}_{f^{-1}(B)} = M_f^T\,(\times)\,\mathcal{V}_B, \quad \forall\, B \in \mathcal{F}(F). \tag{8.55}$$

Proof. (1) Assume $y = \delta_m$. If $f^{-1}(y) = \varnothing$, then $\mathrm{Row}_j\,(\times)(M_f) = 0$. It is not difficult to see that

$$\mathcal{V}_{f(A)} = \vee_{x\in E}\mathcal{V}_A(x).$$

Then (8.54) follows from the definition of the Boolean product of k-valued matrices.

(2) Let $x = \delta_n^i$. Then $f(x) = \mathrm{Row}_i(M_f)$. Say, $f(x) = \delta_m^j$. Then

$$\mu_B(f(x)) = \mathrm{Col}_j(\mathcal{V}_B) = f(x)^T\mathcal{V}_B.$$

Hence we have (8.55). □

Example 8.6. Consider the mapping defined in Example 8.5.

(1) Let $A = 0.3/1 + 0.8/2 + 1/4 + 0.5/4 \in \mathcal{P}(E)$. Then $\mathcal{V}_A = [0.3\ 0.8\ 0\ 1\ 0.5]^T$. Hence,

$$\begin{aligned}\mathcal{V}_{f(A)} &= M_f\,(\times)\,\mathcal{V}_A\\ &= \delta_3[2\ 3\ 1\ 2\ 3]\,(\times)\begin{bmatrix}0.3\\0.8\\0\\1\\0.5\end{bmatrix} = \begin{bmatrix}0\\1\\0.5\end{bmatrix}.\end{aligned}$$

It follows that $f(A) = 1/1 + 0.5/2$.

(2) Let $B = 0.2/0 + 0.8/1 + 0.4/2 \in \mathcal{P}(F)$. Then $\mathcal{V}_B = [0.2\ 0.8\ 0.4]^T$. Hence,

$$\begin{aligned}\mathcal{V}_{f^{-1}(B)} &= M_f^T\,(\times)\,\mathcal{V}_B\\ &= \begin{bmatrix}0\ 1\ 0\\0\ 0\ 0\\1\ 0\ 0\\0\ 1\ 0\\0\ 0\ 1\end{bmatrix}(\times)\begin{bmatrix}0.2\\0.8\\0.4\end{bmatrix} = \begin{bmatrix}0.8\\0\\0.2\\0.8\\0.4\end{bmatrix}.\end{aligned}$$

It follows that $f^{-1}(B) = 0.8/1 + 0.2/3 + 0.8/4 + 0.4/5$.

Corollary 8.1. *Let $f : E \to F$, where $|E| = n$ and $|F| = m$.*

(1) If f is one-to-one, then

$$\mathcal{V}_{f^{-1}(f(A))} = \mathcal{V}_A, \quad A \in \mathcal{F}(E). \tag{8.56}$$

(2) If f is one-to-one and onto, then

$$\mathcal{V}_{f(f^{-1}(B))} = \mathcal{V}_B, \quad B \in \mathcal{F}(F). \tag{8.57}$$

Proof. (1) Assume the structure matrix of f is $M_f = \delta_n[i_1\ i_2\ \cdots\ i_n]$. Since f is one-to-one, then when $p \neq q$ we have $i_p \neq i_q$. It follows that

$$\mathcal{V}_{f^{-1}(f(A))} = M_f^T M_f \mathcal{V}_A = I_n \mathcal{V}_A = \mathcal{V}_A.$$

(2) Since f is one-to-one and onto, it is easy to see that $M_f^T = M_f^{-1}$. Hence

$$\mathcal{V}_{f(f^{-1}(B))} = M_f M_f^T \mathcal{V}_B = \mathcal{V}_B.$$

□

The following properties are immediate consequence of the definitions.

Proposition 8.4.

(i)

$$f(A) = \varnothing \Leftrightarrow A = \varnothing. \tag{8.58}$$

(ii)

$$A \subset B \Rightarrow f(A) \subset f(B). \tag{8.59}$$

If f is one-to-one, the converse implement " $\Leftarrow$ " is also correct.

(iii)

$$f\left(\cup_{\lambda\in\Lambda} A_\lambda\right) = \cup_{\lambda\in\Lambda} f\left(A_\lambda\right). \tag{8.60}$$

(iv)

$$f\left(\cap_{\lambda\in\Lambda} A_\lambda\right) \subset \cap_{\lambda\in\Lambda} f\left(A_\lambda\right). \tag{8.61}$$

If f is one-to-one, then "$\subset$" can be replaced by "$=$".

(v)

$$f^{-1}(\varnothing) = \varnothing. \tag{8.62}$$

(vi) If f is onto, then

$$f^{-1}(B) = \varnothing \Rightarrow B = \varnothing. \tag{8.63}$$

(vii)

$$B_1 \subset B_2 \Rightarrow f^{-1}(B_1) \subset f^{-1}(B_2). \tag{8.64}$$

(viii)

$$f^{-1}\left(\cup_{\lambda\in\Lambda} B_\lambda\right) = \cup_{\lambda\in\Lambda} f^{-1}\left(B_\lambda\right). \tag{8.65}$$

(ix)

$$f^{-1}\left(\cap_{\lambda\in\Lambda} B_\lambda\right) = \cap_{\lambda\in\Lambda} f^{-1}\left(B_\lambda\right). \tag{8.66}$$

(x)

$$f^{-1}(B^c) = \left[f^{-1}(B)\right]^c. \tag{8.67}$$

8.5 Fuzzy Logic and Its Computation

Throughout this section we assume the universe of discourse $E = \{e_1, e_2, \cdots, e_n\}$ is unique for all fuzzy objects.

Definition 8.11.

(1) A fuzzy proposition a is a fuzzy set. Precisely, $a \in \mathcal{F}(E)$ is an element.
(2) A fuzzy logical variable x is a variable which takes values from $\mathcal{F}(E)$.
(3) A fuzzy logical function is an expression with some fuzzy propositions and fuzzy variables connected by (fuzzy) logical operators.

Remark 8.2.

(1) Traditionally, the logical operators allowed in a fuzzy logical function are $\{\neg,\ \wedge,\ \vee\}$. But we assume **A2** (refer to Section 7.3) holds, then all logical operators are allowed.
(2) In this section we consider only the k-valued (fuzzy) logic. The results obtained can easily be extended to mix-valued (fuzzy) logic.

Assume a, $x, x_1, \cdots, x_m$, and $f(x_1, \cdots, x_m)$ are fuzzy proposition, fuzzy logical variables, and fuzzy logical function respectively. Moreover, assume for any $\xi \in \mathcal{F}(E)$,

$$\mu_\xi(e) \in \mathcal{D}_k, \quad e \in E.$$

Then for a fixed $e_0 \in E$ the $\mu_a(e_0)$, $\mu_x(e_0)$ $\mu_{x_i}(e_0)$ $i = 1, \cdots, m$, and $\mu_f(e_0)$ are simply the k-valued proposition (or constant), k-valued logical variables, and k-valued logical functions. Therefore, as $|E| = n$ is assumed, they are n-dimensional k-valued proposition, n-dimensional k-valued logical variables, and n-dimensional k-valued logical function respectively.

Precisely speaking, a fuzzy logical variable x can be expressed in its vector form as

$$\mathcal{V}_x = (x^1, x^2, \cdots, x^n)^T \in \mathcal{D}_k^n.$$

Now since x^i is assumed to be a k-valued logical variable, that is the membership function can only take k different values, then

$$x^i \in \mathcal{D}_k, \quad i = 1, 2, \cdots, n.$$

Using the matrix expression of k-valued logic, the following result is obvious.

Proposition 8.5. *Let $x^1, \cdots, x^m$ be a set of fuzzy logical variables, and a fuzzy logical function $f(x^1, \cdots, x^m)$ has its structure matrix as $M_f \in$*

$\mathcal{L}_{k\times k^n}$. *Denote the vector form of the jth component of x^i by x^i_j, then*

$$x^i_j = [\mathcal{V}_{x^i}]_j \in \Delta_k.$$

Hence,

$$f_j = M_f x^1_j x^2_j \cdots x^n_j, \quad j = 1, \cdots, n. \tag{8.68}$$

We give an example to depict this.

Example 8.7. Assume the universe of discourse is $E = \{e_1, e_2, e_3, e_4\}$ and $k = 3$. $x, y, z \in \mathcal{F}(E)$. A fuzzy logical function f is defined as

$$f(x, y, z) = (x \wedge y) \leftrightarrow z. \tag{8.69}$$

Then we have the algebraic form of (8.69) as

$$f(x, y, z) = M_f xyz, \tag{8.70}$$

where the structure matrix of f is

$$M_f = M_{e,3}M_{c,3} = \delta_3[1\ 2\ 3\ 2\ 2\ 2\ 3\ 2\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 3\ 2\ 1\ 3\ 2\ 1\ 3\ 2\ 1\ 3\ 2\ 1].$$

Precisely speaking, (8.70) means

$$f(x, y, z)(e) = M_f x(e)y(e)z(e), \quad e \in E.$$

Now assume the vector form of x, y, and z are respectively as

$$\mathcal{V}_x = \begin{bmatrix} 0.5 \\ 0 \\ 0.5 \\ 1 \end{bmatrix} := \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}; \quad \mathcal{V}_y = \begin{bmatrix} 0 \\ 0.5 \\ 1 \\ 0.5 \end{bmatrix} := \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}; \quad \mathcal{V}_z = \begin{bmatrix} 0.5 \\ 1 \\ 1 \\ 0 \end{bmatrix} := \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix}.$$

Then

$$\begin{aligned} f(e_1) &= M_f x_1 y_1 z_1 = M_f \delta_3^2 \delta_3^3 \delta_3^2 = \delta_3^1; \\ f(e_2) &= M_f x_2 y_2 z_2 = M_f \delta_3^3 \delta_3^2 \delta_3^1 = \delta_3^3; \\ f(e_3) &= M_f x_3 y_3 z_3 = M_f \delta_3^2 \delta_3^1 \delta_3^1 = \delta_3^2; \\ f(e_4) &= M_f x_4 y_4 z_4 = M_f \delta_3^1 \delta_3^2 \delta_3^3 = \delta_3^2. \end{aligned}$$

We conclude that

$$f(x, y, z) = \begin{bmatrix} 1 \\ 0 \\ 0.5 \\ 0.5 \end{bmatrix}.$$

Exercises

8.1 Consider the 3-valued matrices

$$A = \begin{bmatrix} 0 & 0.5 & 0.5 & 1 \\ 1 & 0 & 0.5 & 0.5 \end{bmatrix}, \quad B = \begin{bmatrix} 0.5 & 0 & 0 & 1 \\ 0 & 1 & 0.5 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0.5 & 0.5 \\ 1 & 0 \\ 0.5 & 1 \end{bmatrix}.$$

Calculate the following:

(a) $$Y = 0.5A \ (+) \ B.$$

(b) $$Y = A(\ltimes)B.$$

(c) $$Y = A(*)B.$$

(d) $$Y = C(\otimes)I_2.$$

(e) $$Z_1 = (A(+)B)(\times)C.$$

(f) $$Z_2 = (A(\times)C)\,(+)\,(B(\times)C)\,.$$

Compare it with Z_1.

8.2 If $A \in \mathcal{D}_\infty^{m\times n}$ and $B \in \mathcal{D}_\infty^{p\times q}$ are two fuzzy matrices, show that $A \ltimes B$ may not be a fuzzy matrix but $A(\ltimes)B$ is a fuzzy matrix.

8.3 Let

$$A = \begin{bmatrix} 0.2 & 0.7 & 0.5 & 0.4 & 0.5 \\ 0 & 0.3 & 0.8 & 0.5 & 0.7 \end{bmatrix}, \quad B = \begin{bmatrix} 0.6 & 1 & 0.5 & 0.5 & 0.1 \\ 0.8 & 0.2 & 0.4 & 0 & 0.9 \end{bmatrix}.$$

Calculate (i) $0.25A(+)0.75B$, (ii) $A(\ltimes)B$, and (iii) $A(*)B$.

8.4 Prove Proposition 8.2.

8.5 Let $A \in \mathcal{D}_k^{n\times n}$ where $k < \infty$.

(i) Show that there exist two integer $r > 0$ and $p_0 \geq 0$ such that

$$A^{(p)} = A^{(p+r)}, \quad p \geq p_0. \tag{8.71}$$

(ii) Given any k-valued vector $X_0 \in \mathcal{D}_k^n$, show that there exists a $r_0 > 0$ such that

$$A^{(p)}(\times)X_0 = A^{(p+r_0)}(\times)X_0, \quad p \geq p_0. \tag{8.72}$$

(iii) Prove that r_0 is a factor of r.

8.6 (i) Let $a, b \in \mathcal{D}$. We define the addition $\langle + \rangle$ as

$$a \,\langle + \rangle\, b := a + b \ (\text{mod } 2) = a \bar{\vee} b.$$

Moreover, we define a product $\langle\times\rangle$ as

$$a\,\langle\times\rangle\,b := ab.$$

Show that $\mathcal{D}$ with $\langle+\rangle$ and $\langle\times\rangle$ form a vector space over $\mathcal{D}$.

(ii) Let $A, B \in \mathcal{B}_{m\times n}$ be two Boolean matrices. Define

$$A\,\langle+\rangle\,B := C \in \mathcal{B}_{m\times n},$$

where

$$c_{i,j} = a_{i,j}\,\langle+\rangle\,b_{i,j},\ i = 1, \cdots, m;\ j = 1, \cdots, n;$$

and

$$\alpha A := \left(\alpha\,\langle\times\rangle\,a_{i,j}\right).$$

Show that $\mathcal{B}_{m\times n}$ with $\langle+\rangle$ and $\langle\times\rangle$ form a vector space over $\mathcal{B}_{m\times n}$.

8.7 Let $A \in \mathcal{B}_{m\times n}$ and $B \in \mathcal{B}_{n\times p}$. Define the product $A\,\langle\times\rangle\,B$ as

$$A\,\langle\times\rangle\,B := C \in \mathcal{B}_{m\times p},$$

where

$$c_{i,j} = \langle+\rangle_{k=1}^{n}\, a_{i,k}\,\langle\times\rangle\,b_{k,j}, \quad i = 1, \cdots, m;\ j = 1, \cdots, p.$$

Show that this product satisfies

(i) (Distributive Law) Let $A, B \in \mathcal{B}_{m\times n}$, $C \in \mathcal{B}_{n\times p}$, and $D \in \mathcal{B}_{q\times m}$. Then

$$\begin{aligned}
&(\alpha A\,\langle+\rangle\,\beta B)\,\langle\times\rangle\,C = \alpha A\,\langle\times\rangle\,C\,\langle+\rangle\,\beta B\,\langle\times\rangle\,C;\\
&D\,\langle\times\rangle\,(\alpha A\,\langle+\rangle\,\beta B) = \alpha D\,\langle\times\rangle\,A\,\langle+\rangle\,\beta D\,\langle\times\rangle\,B;\\
&\qquad\qquad\qquad\qquad\qquad\qquad \alpha, \beta \in \mathcal{D}.
\end{aligned}$$

(ii) (Associative Law) Let $A \in \mathcal{B}_{m\times n}$, $B \in \mathcal{B}_{n\times p}$, and $C \in \mathcal{B}_{p\times q}$. Then

$$(A\,\langle\times\rangle\,B)\,\langle\times\rangle\,C = A\,\langle\times\rangle\,(B\,\langle\times\rangle\,C).$$

8.8 Let $X = [0, 100)$ be the universe of discourse. We are interested in three fuzzy sets A, B, and C as:

$$\mu_A(x) = \begin{cases} 0, & x < 30\\ \frac{1}{2}, & 30 \le x < 50\\ 1, & x \ge 50 \le x < 100, \end{cases}$$

$$\mu_B(x) = \begin{cases} 0, & x < 50\\ \frac{1}{2}, & 50 \le x < 80\\ 1, & 80 \le x < 100. \end{cases}$$

$$\mu_C(x) = \begin{cases} 0, & x < 70 \\ 1, & 70 \leq x < 100. \end{cases}$$

(i) Find a partition of X, such that the universe of discourse can be replaced by this finite set of partition.

(ii) Corresponding this new finite universe of discourse express A, B, and C into vector form. That is, find $\mathcal{V}_A$, $\mathcal{V}_B$, and $\mathcal{V}_C$.

(iii) Using vector form to calculate (a) $\mathcal{V}_{A\wedge B}$, (b) $\mathcal{V}_{(A\vee B)\wedge C}$.

(iv) Using (a) $\mathcal{V}_{A\wedge B}$, (b) $\mathcal{V}_{(A\vee B)\wedge C}$ obtained from above to find (a) $\mu_{A\wedge B}(x)$, (b) $\mu_{(A\vee B)\wedge C}(x)$ with respect to the original universe of discourse X.

8.9 Prove equation (8.67).

8.10 Assume a universe of discourse is $E = \{e_1, e_2, e_3\}$, and three fuzzy sets are

$$\begin{aligned} A &= 0.5/e_1 + 0.5/e_2 + 1/e_3, \\ B &= 0/e_1 + 0.5/e_2 + 0.5/e_3, \\ C &= 0.5/e_1 + 0/e_2 + 1/e_3. \end{aligned} \tag{8.73}$$

Find out $(A \cup B) \cap C$.

8.11 A universe E and fuzzy sets A, B, C are as given in (8.73). Assume another universe $F = \{f_1, f_2\}$, and a function $f : E \to F$ is defined as

$$f(e_1) = f(e_2) = f_1, \quad f(e_3) = f_2.$$

Find out the expression of $f^{-1}((A \cup B) \cap C)$ and $f^{-1}(((A \cup B) \cap C)^c)$.

8.12 E and A, B, C are given as in (8.73). There is a fuzzy logical function g such that

$$g(A, B, C) = 0/e_1 + 0.5/e^2 + 1/e^3.$$

Try to find out all the possible g.

8.13 Let $E = \{2, 3, 4, 5, 6, 7\}$ and $F = \{0, 1, 2, 3\}$. A mapping $\pi : E \to F$ is defined as $\pi(e) = e^2 + e \pmod 4$.

(i) Find the structure matrix of π.

(ii) Let $A, B \in \mathcal{F}(E)$ be

$$A = 0.2/3 + 0.5/5 + 0.9/7, \quad B = 0.3/2 + 0.6/4 + 0.5/6.$$

Calculate (a) $\pi(A)$, (b) $\pi(A \wedge B)$, and (c) $\pi(A \vee B)$.

(iii) Let $X = 0.4/0 + 0.8/2 + 0.3/3 \in \mathcal{F}(F)$. Calculate $\pi^{-1}(X)$.

8.14 Given the universe of discourse as $E = \{e_1, e_2, e_3, e_4, e_5\}$ and $k = 3$. $x, y \in \mathcal{F}(E)$. A fuzzy logical function f is defined as

$$f(x, y) = (x \wedge y) \leftrightarrow (x \vee y). \tag{8.74}$$

(i) Assume $\mathcal{V}_x = (1, 0.5, 0.5, 0, 1)^T$, and $\mathcal{V}_y = (0.5, 0, 1, 0, 1)^T$. Calculate $f(x, y)$.

(ii) Assume $\mathcal{V}_x = (0.5, 0, 0, 1, 1)^T$, and $\mathcal{V}_y = (0, 1, 0.5, 1, 1)^T$. Calculate $f(x, y)$.

This page intentionally left blank

Chapter 9

Fuzzy Relational Equation

Fuzzy relation plays a fundamental rule in the design of fuzzy controllers, fuzzy logical inferences, and the application of fuzzy control to engineering problems, the application of fuzzy inference to medical diagnosis etc. Y. Tsukamato et.al investigated the solvability of a class of lower-dimensional fuzzy relational equations (Tsukamoto, 1979). Then the problem of finding general fuzzy relations was considered by E. Sanchez, who proposed the so called fuzzy relational equation and provided some fundamental principles for solving it (Sanchez, 1996).

This chapter provides a method to find the complete set of solutions for a fuzzy relational equation by revealing and using the relationship between the parameter set solutions and the overall set of solutions

This chapter is based on Cheng *et al.* (2011a).

9.1 *k*-Valued Matrix and Fuzzy Relational Equations

First, we define a partial order $\geq$ on $\mathcal{D}_{\infty}^{m\times n}$ as follows.

Definition 9.1.

(1) Let $A=(a_{i,j})$, $B=(b_{i,j})$, and $A,B\in\mathcal{D}_{\infty}^{m\times n}$. We say $A\geq B$ if

$$a_{i,j}\geq b_{i,j},\quad i=1,\cdots,m;\ j=1,\cdots,n.$$

(2) If $A\geq B$ and $A\neq B$, then we say $A>B$.

(3) Let $\Theta\subset\mathcal{D}_{\infty}^{m\times n}$. $A\in\Theta$ is called a maximum (minimum) element, if there is no $B\in\Theta$ such that $B>A$ (correspondingly, $B<A$)

(4) $A\in\Theta$ is called the largest (smallest) element, if

$$A\geq B,\quad\forall\,B\in\Theta;\qquad(\text{correspondingly,}\quad A\leq B,\quad\forall\,B\in\Theta).$$

The following order-preserving property is an immediate consequence of the definitions of addition and product.

Proposition 9.1.

(1) Let $A,\ B,\ C,\ D \in \mathcal{D}_\infty^{m\times n}$. Assume $A \geq B$ and $C \geq D$. Then

$$A\,(+)\,C \geq B\,(+)\,D. \tag{9.1}$$

(2) Let $A,\ B \in \mathcal{D}_\infty^{m\times n}$ and $C,\ D \in \mathcal{D}_\infty^{n\times p}$. Assume $A \geq B$ and $C \geq D$. Then

$$A\,(\times)\,C \geq B\,(\times)\,D. \tag{9.2}$$

Throughout this chapter we consider only some finite universes of discourse. Particularly, we set

$$U = \{u_1, \cdots, u_m\}, \quad V = \{v_1, \cdots, v_n\}, \quad W = \{w_1, \cdots, w_s\}.$$

Then we have

Definition 9.2. Let $R \in \mathcal{F}(U \times V)$ be given. The relational matrix of R, denoted by M_R is defined as

$$M_R = \begin{bmatrix} \mu_R(u_1,v_1) & \mu_R(u_1,v_2) & \cdots & \mu_R(u_1,v_n) \\ \mu_R(u_2,v_1) & \mu_R(u_2,v_2) & \cdots & \mu_R(u_2,v_n) \\ \vdots & & & \\ \mu_R(u_m,v_1) & \mu_R(u_m,v_2) & \cdots & \mu_R(u_m,v_n) \end{bmatrix}, \tag{9.3}$$

where μ_R is the membership degree of R on $U \times V$.

For notational ease, R is also conventionally used for M_R. Hereafter, we use this convention.

Usually, two kinds of fuzzy relational equations (FRE) were investigated.

- Type 1: Let $A \in \mathcal{F}(U \times V)$, and $B \in \mathcal{F}(U \times W)$. We are looking for a fuzzy relation $X \in \mathcal{F}(V \times W)$, such that

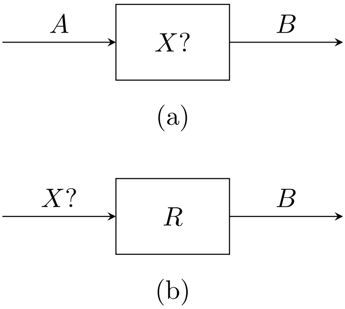
$$A\,(\times)\,X = B. \tag{9.4}$$

  We refer to Fig. 9.1 (a) for this type of FRE's, which are commonly used in the design of fuzzy controllers.

- Type 2: Let $R \in \mathcal{F}(V \times W)$, and $B \in \mathcal{F}(U \times W)$. We are looking for a fuzzy input $X \in \mathcal{F}(U \times V)$, such that

$$X\,(\times)\,R = B. \tag{9.5}$$

  We refer to Fig. 9.1 (b) for this type of FRE's, which may be used for a problem similar to the diagnosing diseases via symptoms, where the fuzzy relation is known (Sanchez, 1979).



(a)

(b)

Fig. 9.1 (a) Unknown fuzzy relation (b) Unknown fuzzy input

Taking a transpose on both sides of (9.5), we have

$$R^T\,(\times)\,X^T = B^T,$$

which has the same form as (9.4). We, therefore, consider the solvability of (9.4) only.

To state the existing result for solving (9.4), we need some preparations. The following statements are cited from Zhu (2005).

Definition 9.3 (Zhu, 2005). (i) Let $a,\ b \in [0,1]$. Then

$$a \oplus b = \begin{cases} b, & a > b \\ 1, & a \le b; \end{cases} \tag{9.6}$$

and

$$a \oslash b = \begin{cases} b, & a \ge b \\ 0, & a < b. \end{cases} \tag{9.7}$$

(ii) Let $A \in \mathcal{F}(U)$ and $B \in \mathcal{F}(V)$. Then $A \oplus B$ is defined by

$$\mu_{A\oplus B}(u,v) := \mu_A(u) \oplus \mu_B(v); \tag{9.8}$$

and $A^T \oslash B$ is defined by

$$\mu_{A^T\oslash B}(u,v) := \mu_A^T(u) \oslash \mu_B(v). \tag{9.9}$$

Then the following result is a commonly used method in nowadays fuzzy control design.

Theorem 9.1 (Zhu, 2005). *In equation (9.4) assume (i) $A \in \mathcal{F}(U)$ and $B \in \mathcal{F}(V)$; (ii) there exists at least one solution, then*

(i)

$$X^* = A^T \oplus B \tag{9.10}$$

is the largest solution;

(ii)

$$X_* = A^T \oslash B \tag{9.11}$$

is also a solution.

It is obvious that the result of Theorem 9.1 is very limited. It provides only some particular solutions under very strong constraints on A and B. The purpose of this chapter is to provide a general algorithm, which provides all the solutions of (9.4) without any restrictions on both A and B. This approach is based on the STP of matrices and the vector expression of multi-valued logic.

Recall that for an r-ary k-valued logical function f, there exists a unique logical matrix $M_f \in \mathcal{L}_{k\times k^r}$, called the structure matrix of f, such that in vector form

$$f(x_1, \cdots, x_r) = M_f \ltimes x_1 \ltimes \cdots \ltimes x_r. \tag{9.12}$$

Particularly, when σ is a unary operator, we can find its structure matrix M_σ such that

$$\sigma x = M_\sigma x, \quad x \in \Delta_k.$$

When σ is a binary operator, we can find its structure matrix M_σ such that

$$x\sigma y = M_\sigma \ltimes x \ltimes y, \quad x, y \in \Delta_k.$$

We refer to Chapter 6 for constructing the structure matrices of a k-valued logical functions.

9.2 Structure of the Set of Solutions

Consider equation (9.4). Let $A = (a_{i,j})$, $B = (b_{i,j})$, and $R = (x_{i,j})$, where $x_{i,j}$ are used to emphasize that R is the unknown matrix. We can further convert it into canonical linear algebraic equations as

$$A\,(\times)\,X_i = B_i, \quad i = 1, \cdots, s, \tag{9.13}$$

where $X_i = \mathrm{Col}_i(X)$ and $B_i = \mathrm{Col}_i(B)$.

Collecting different values of the entries of A and B as

$$S = \{a_{i,j}, b_{p,q} | i = 1, \cdots, m; j = 1, \cdots, n; p = 1, \cdots, m; q = 1, \cdots, s\},$$

and adding 1 and/or 0 when they are not in S, we construct an ordered set as

$$\Xi = \{\xi_i \,|\, i = 1, \cdots, r; \quad \text{and} \quad \xi_1 = 0 < \xi_2 < \cdots < \xi_{r-1} < \xi_r = 1\}.$$

Then we have $\Xi = S \cup \{0, 1\}$, which is called the parameter set.

Definition 9.4. Let $x \in [0, 1]$. Then we define

(i) $\pi_* : [0, 1] \to \Xi$ as

$$\pi_*(x) = \max_i \{\xi_i \in \Xi \,|\, \xi_i \le x\}\,; \tag{9.14}$$

(ii) $\pi^* : [0, 1] \to \Xi$ as

$$\pi^*(x) = \min_i \{\xi_i \in \Xi \,|\, \xi_i \ge x\}\,. \tag{9.15}$$

Note that if $x = \xi_i \in \Xi$, then

$$\pi_*(x) = \pi^*(x) = \xi_i.$$

Otherwise, there exists a unique i such that

$$\xi_i < x < \xi_{i+1}.$$

It follows that

$$\pi_*(x) = \xi_i; \quad \pi^*(x) = \xi_{i+1}.$$

For statement ease, we identify ξ_i with $\frac{i-1}{r-1}$, $i = 1, \cdots, r$. It means we identify Ξ with Δ_r. Then we have the following result, which allows us to search solutions from the finite set Ξ. A solution $x \in \Xi^n$ is called a parameter set solution (PSS).

Lemma 9.1. *Let $R = (x_{i,j}) \in \mathcal{D}_\infty^{n\times s}$ be a solution of (9.4). Then $\pi_*(R) := (\pi_*(x_{i,j}))$ is also a solution of (9.4).*

Proof. It suffices to prove it for each equation in (9.13), which is simply denoted by

$$A\,(\times)\,z = b. \tag{9.16}$$

Assume $z = (z_1, \cdots, z_n)^T$ is a solution of (9.16). Set $Z_0 = \{z_1, \cdots, z_n\}$ and define

$$Z^0 = Z_0 \backslash \Xi.$$

If $Z^0 = \varnothing$, we are done. Otherwise, we can find

$$z^0 = \max_{z_j}\{z_j \in Z^0\}.$$

Then there is an i_0 such that

$$\xi_{i_0} < z^0 < \xi_{i_0+1}. \tag{9.17}$$

Next, we replace all the elements in Z^0, which are greater than ξ_{i_0}, by ξ_{i_0}. Such a replacement converts Z_0 to a new set, called Z_1. We claim that Z_1 is also a solution of (9.16).

Consider a particular equation of (9.16), say, jth equation, which is

$$[a_{j,1} \wedge z_1] \vee [a_{j,2} \wedge z_2] \vee \cdots [a_{j,n} \wedge z_n] = b_j. \tag{9.18}$$

First, we assume $b_j \geq \xi_{i_0+1}$. Then there must be a term $a_{j,s} \wedge z_s$, which equals b_j. Then replacing any $\xi_{i_0} < z^0 < \xi_{i_0+1}$ by ξ_{i_0} will not affect the equality.

Next, we assume $b_j \leq \xi_{i_0}$. Then multiplying both sides of (9.18) by ξ_{i_0} (precisely, operating both sides by $\xi_{i_0}\wedge$). Then the right hand side is still b_j, and on the left hand side, since each term should be less than or equal to b_j, it changes nothing. But if there is a term, say, $a_{j,s} \wedge z_s$, which has z_s satisfying (9.17), then we can replace it by

$$\xi_{i_0} \wedge a_{j,s} \wedge z_s = a_{j,s} \wedge \xi_{i_0}.$$

We conclude that Z_1 is a solution of (9.16). Now for Z_1, we can do the same thing. That is, set

$$Z^1 = Z_1 \backslash \Xi$$

and define

$$z^1 = \max_{z_j}\{z_j \in Z^1\},$$

it follows that

$$z^1 < \xi_{i_0} < z^0.$$

Then we can find i_1 such that

$$\xi_{i_1} < z^1 < \xi_{i_1+1}. \tag{9.19}$$

Finally, in the solution $Z_1 = (z_1^1, \cdots, z_1^n)$ all z_1^j, satisfying (9.19), can be replaced by ξ_{i_1} to produce a new solution Z_2. Note that now $\xi_{i_1} < \xi_{i_0}$. Continuing this procedure, finally, we can have $Z^{k^*} = \varnothing$, where $k^* \leq r$. The conclusion follows. □

Similar to Lemma 9.1, we can prove the following result.

Lemma 9.2. *Let $R = (x_{i,j}) \in \mathcal{D}_\infty^{n\times s}$ be a solution of (9.4). Then $\pi^*(R) := (\pi^*(x_{i,j}))$ is also a solution of (9.4).*

Now we can prove the following result, which shows the structure of the set of solutions.

Theorem 9.2. *$R = (x_{i,j}) \in \mathcal{D}_{\infty}^{n\times s}$ is a solution of (9.4), if and only if both $\pi_*(R)$ and $\pi^*(R)$ are solutions of (9.4).*

Proof. The necessity comes from Lemmas 9.1 and 9.2. We prove the sufficiency. That is, if both $\pi_*(R)$ and $\pi^*(R)$ are solutions of (9.4), then so is R.

If $R = \pi_*(R)$ or $R = \pi^*(R)$, we are done. So we assume $R \neq \pi_*(R)$ and $R \neq \pi^*(R)$. We prove it by contradiction. Assume R is not a solution of (9.4). Since $R \geq \pi_*(R)$, according to Proposition 9.1 we have $A\,(\times)\,R \geq B$. But since R is not a solution, we have

$$A\,(\times)\,R > B.$$

Now since $\pi^*(R) \geq R$, we have

$$A\,(\times)\,\pi^*(R) \geq A\,(\times)\,R > B.$$

This is absurd. □

Theorem 9.2 gives a complete outline for the set of solutions. It has also clearly demonstrated that the set of PSS's is enough to describe the whole set of solutions.

We have the following useful proposition for the set of solutions. In fact, Theorem 9.1 shows the following result for the particular case, where A and B are restricted.

Proposition 9.2. *If there is a solution of (9.13) in Ξ^n. Then there is a largest solution in Ξ^n.*

Proof. If we can prove the maximum solution is unique, we are done. Now assume both z_1^* and z_2^* are two different maximum solutions. Because of the distributive property, it is easy to prove that $z_1^*\,(+)\,z_2^*$ is also a solution. But $z_1^*\,(+)\,z_2^* > z_1^*$ (or z_2^*), which is a contradiction. □

9.3 Solving Fuzzy Relational Equation

To solve the fuzzy relational equations we have only to solve equations in (9.13). That is, we have only to develop a method to solve (9.16).

Recall that in vector form we have the structure matrices such that all the logical expressions with operators can be expressed as a matrix product. Particularly, in this section we need the following expressions.

Using algebraic form, we can convert the left hand side (LHS) of equation (9.16) (as the jth equation of (9.13)), into the following form:

$$\begin{aligned} LHS &= (M_{d,r})^{n-1}(M_{c,r}a_{j,1}z_1)\cdots(M_{c,r}a_{j,n}z_n) \\ &= (M_{d,r})^{n-1}M_{c,r}a_{j,1}[I_r \otimes (M_{c,r}a_{j,2})][I_{r^2} \\ &\quad \otimes(M_{c,r}a_{j,3})]\cdots[I_{r^{n-1}} \otimes (M_{c,r}a_{j,n})] \ltimes_{i=1}^{n} z_i \\ &:= L_j z, \end{aligned} \tag{9.20}$$

where

$$\begin{aligned} L_j &= (M_{d,r})^{n-1}M_{c,r}a_{j,1}[I_r \otimes (M_{c,r}a_{j,2})][I_{r^2} \\ &\quad \otimes(M_{c,r}a_{j,3})]\cdots[I_{r^{n-1}} \otimes (M_{c,r}a_{j,n})] \in \mathcal{L}_{r\times r^n}; \\ z &= \ltimes_{i=1}^{n} z_i. \end{aligned}$$

Then (9.16) becomes

$$L_j z = b_j, \quad j = 1, \cdots, m. \tag{9.21}$$

In the following we briefly review the method for solving logical equations (9.21), which has been discussed in Chapter 5.

Multiplying both sides of m equations of (9.21), we can express (9.16) as

$$Lz = b, \tag{9.22}$$

where $L = L_1 * L_2 * \cdots * L_m \in \mathcal{L}_{r^m\times r^n}$, and $b = \ltimes_{i=1}^{m} b_i$. Here "$*$" is the Khatri-Rao product of matrices. (We refer to Chapter chap:1 for the definition.) Precisely,

$$\mathrm{Col}_t(L) = \mathrm{Col}_t(L_1) \ltimes \mathrm{Col}_t(L_2) \ltimes \cdots \ltimes \mathrm{Col}_t(L_s), \quad t = 1, \cdots, r^n.$$

Next, we show how to solve equation (9.22). Note that since L is a logical matrix, $b \in \Delta_{r^m}$ and $z \in \Delta_{r^n}$, the following result is obvious.

Theorem 9.3. *Equation (9.22) has solution, if and only if*

$$b \in \mathrm{Col}(L). \tag{9.23}$$

Now assume

$$\Lambda = \{\lambda \mid \mathrm{Col}_\lambda(L) = b\}.$$

Then the solution set is

$$\left\{ z_\lambda = \delta_{2^n}^{\lambda} \,\middle|\, \lambda \in \Lambda \right\}. \tag{9.24}$$

Finally, we have to convert z back to $(z_1, \cdots, z_n) \in \Xi^n$.

9.4 Numerical Examples

This section presents some examples to demonstrate the algorithm for solving the fuzzy relational equations. In fact, the method developed in previous section is applicable to general fuzzy logical equations. First example is a simple one, which is used to show the solving process.

Example 9.1. Consider the following logical equation

$$\begin{cases} x \wedge y = 0.32 \\ (\neg x) \vee y = 0.68. \end{cases} \tag{9.25}$$

First, one sees easily that the logical values can be divided into 4 levels. That is,

$$\Xi = \{1,\ 0.68,\ 0.32,\ 0\}.$$

Then we identify the values with their vector forms as

$$1 \sim \delta_4^1; \quad 0.68 \sim \delta_4^2; \quad 0.32 \sim \delta_4^3; \quad 0 \sim \delta_4^4.$$

Now (9.25) can be converted into its algebraic form as

$$\begin{cases} M_c^4 \ltimes x \ltimes y = \delta_4^3 \\ M_d^4 \ltimes (M_n^4 \ltimes x) \ltimes y = \delta_4^2. \end{cases} \tag{9.26}$$

Setting $z = x \ltimes y$, (9.26) can be converted as

$$\begin{cases} G_1 z = \delta_4^3 \\ G_2 z = \delta_4^2, \end{cases} \tag{9.27}$$

where

$$\begin{aligned} G_1 &= M_c^4 = \delta_4[1\ 2\ 3\ 4\ 2\ 2\ 3\ 4\ 3\ 3\ 3\ 4\ 4\ 4\ 4\ 4], \\ G_2 &= M_d^4 \ltimes M_n^4 = \delta_4[1\ 2\ 3\ 4\ 1\ 2\ 3\ 3\ 1\ 2\ 2\ 2\ 1\ 1\ 1\ 1]. \end{aligned}$$

Multiplying two equations together yields

$$L\,(\times)\,z = b, \tag{9.28}$$

where

$$L = G_1 * G_2 = \delta_{16}[1\ 6\ 11\ 16\ 5\ 6\ 11\ 15\ 9\ 10\ 10\ 14\ 13\ 13\ 13\ 13],$$

$$b = \delta_4^3 \ltimes \delta_4^2 = \delta_{16}^{10}.$$

Since

$$\mathrm{Col}_{10}(L) = \mathrm{Col}_{11}(L) = \delta_{16}^{10},$$

we have solutions

$$z_1 = \delta_{16}^{10}, \quad z_2 = \delta_{16}^{11}.$$

It follows that

$$\begin{cases} x_1 = \delta_4^3 \\ y_1 = \delta_4^2, \end{cases} \qquad \begin{cases} x_2 = \delta_4^3 \\ y_2 = \delta_4^3. \end{cases}$$

Back to the fuzzy values, we have

$$\begin{cases} x_1 = 0.32 \\ y_1 = 0.68, \end{cases} \qquad \begin{cases} x_2 = 0.32 \\ y_2 = 0.32. \end{cases}$$

The next example is from Sanchez (2002). It will be used to demonstrate the general structure of the solution set of fuzzy relational equations.

Example 9.2. Consider the following relational equation (Sanchez, 2002)

$$Q\,(\times)\,X = T, \tag{9.29}$$

where

$$Q = \begin{bmatrix} 0.2 & 0 & 0.8 & 1 \\ 0.4 & 0.3 & 0 & 0.7 \\ 0.5 & 0.9 & 0.2 & 0 \end{bmatrix}; \quad T = \begin{bmatrix} 0.7 & 0.3 & 1 \\ 0.6 & 0.4 & 0.7 \\ 0.8 & 0.9 & 0.2 \end{bmatrix}.$$

First, we figure out the levels of the membership degrees and identify them with their vector forms:

$$\begin{array}{l} 1 \sim \delta_{10}^1; \quad 0.9 \sim \delta_{10}^2; 0.8 \sim \delta_{10}^3; 0.7 \sim \delta_{10}^4; 0.6 \sim \delta_{10}^5; \\ 0.5 \sim \delta_{10}^6; 0.4 \sim \delta_{10}^7; 0.3 \sim \delta_{10}^8; 0.2 \sim \delta_{10}^9; 0 \sim \delta_{10}^{10}. \end{array}$$

We start by solving the first column of X. Let $X_1 = (x_{11}, x_{21}, x_{31}, x_{41})^T = \mathrm{Col}_1(X)$. Then the algebraic equation for X_1 becomes

$$\begin{cases} (\delta_{10}^9 \wedge x_{11}) \vee (\delta_{10}^{10} \wedge x_{21}) \vee (\delta_{10}^3 \wedge x_{31}) \vee (\delta_{10}^1 \wedge x_{41}) = \delta_{10}^4 \\ (\delta_{10}^7 \wedge x_{11}) \vee (\delta_{10}^8 \wedge x_{21}) \vee (\delta_{10}^{10} \wedge x_{31}) \vee (\delta_{10}^4 \wedge x_{41}) = \delta_{10}^5 \\ (\delta_{10}^6 \wedge x_{11}) \vee (\delta_{10}^2 \wedge x_{21}) \vee (\delta_{10}^9 \wedge x_{31}) \vee (\delta_{10}^{10} \wedge x_{41}) = \delta_{10}^3. \end{cases} \tag{9.30}$$

Let $x_1 = \ltimes_{i=1}^4 x_{i1}$. Then equation (9.30) can be converted into its algebraic form as

$$L\,(\times)\,x_1 = b_1, \tag{9.31}$$

where

$$\begin{aligned} L = \delta_{1000}[&32 \;\; 132 \;\; 232 \;\; 232 \;\; 242 \;\; 252 \;\; 262 \;\; 262 \;\; 262 \;\; 262 \cdots \\ &899 \;\; 40 \;\; 140 \;\; 240 \;\; 340 \;\; 450 \;\; 560 \;\; 670 \;\; 780 \;\; 890 \;\; 1000] \in \mathcal{L}_{1000\times 10000}, \end{aligned}$$

and

$$b_1 = \delta_{10}^4 \ltimes \delta_{10}^5 \ltimes \delta_{10}^3 = \delta_{1000}^{343}.$$

Using the Toolbox introduced in Appendix A, we can solve it out as

$$\begin{array}{lll}
X_1^1 = \delta_{10}[1\ 3\ 4\ 5]^T, & X_1^2 = \delta_{10}[2\ 3\ 4\ 5]^T, & X_1^3 = \delta_{10}[3\ 3\ 4\ 5]^T, \\
X_1^4 = \delta_{10}[4\ 3\ 4\ 5]^T, & X_1^5 = \delta_{10}[5\ 3\ 4\ 5]^T, & X_1^6 = \delta_{10}[6\ 3\ 4\ 5]^T, \\
X_1^7 = \delta_{10}[7\ 3\ 4\ 5]^T, & X_1^8 = \delta_{10}[8\ 3\ 4\ 5]^T, & X_1^9 = \delta_{10}[9\ 3\ 4\ 5]^T, \\
X_1^{10} = \delta_{10}[10\ 3\ 4\ 5]^T. & &
\end{array}$$

For the second column, we have

$$L\,(\times)\,x_2 = b_2, \tag{9.32}$$

where

$$b_2 = \delta_{10}^8 \ltimes \delta_{10}^7 \ltimes \delta_{10}^2 = \delta_{1000}^{762}.$$

Solving it, we have

$$\begin{array}{lll}
X_2^1 = \delta_{10}[1\ 1\ 8\ 8]^T, & X_2^2 = \delta_{10}[1\ 1\ 8\ 9]^T, & X_2^3 = \delta_{10}[1\ 1\ 8\ 10]^T, \\
X_2^4 = \delta_{10}[1\ 1\ 9\ 8]^T, & X_1^5 = \delta_{10}[1\ 1\ 10\ 8]^T, & X_2^6 = \delta_{10}[1\ 2\ 8\ 8]^T, \\
X_2^7 = \delta_{10}[1\ 2\ 8\ 9]^T, & X_2^8 = \delta_{10}[1\ 2\ 8\ 10]^T, & X_2^9 = \delta_{10}[1\ 2\ 9\ 8]^T, \\
X_2^{10} = \delta_{10}[1\ 2\ 10\ 8]^T, & X_2^{11} = \delta_{10}[2\ 1\ 8\ 8]^T, & X_2^{12} = \delta_{10}[2\ 1\ 8\ 9]^T, \\
X_2^{13} = \delta_{10}[2\ 1\ 8\ 10]^T, & X_2^{14} = \delta_{10}[2\ 1\ 9\ 8]^T, & X_2^{15} = \delta_{10}[2\ 1\ 10\ 8]^T, \\
X_2^{16} = \delta_{10}[2\ 2\ 8\ 8]^T, & X_2^{17} = \delta_{10}[2\ 2\ 8\ 9]^T, & X_2^{18} = \delta_{10}[2\ 2\ 8\ 10]^T, \\
X_2^{19} = \delta_{10}[2\ 2\ 9\ 8]^T, & X_2^{20} = \delta_{10}[2\ 2\ 10\ 8]^T, & X_2^{21} = \delta_{10}[3\ 1\ 8\ 8]^T, \\
X_2^{22} = \delta_{10}[3\ 1\ 8\ 9]^T, & X_2^{23} = \delta_{10}[3\ 1\ 8\ 10]^T, & X_2^{24} = \delta_{10}[3\ 1\ 9\ 8]^T, \\
X_2^{25} = \delta_{10}[3\ 1\ 10\ 8]^T, & X_2^{26} = \delta_{10}[3\ 2\ 8\ 8]^T, & X_2^{27} = \delta_{10}[3\ 2\ 8\ 9]^T, \\
X_2^{28} = \delta_{10}[3\ 2\ 8\ 10]^T, & X_2^{29} = \delta_{10}[3\ 2\ 9\ 8]^T, & X_2^{30} = \delta_{10}[3\ 2\ 10\ 8]^T, \\
X_2^{31} = \delta_{10}[4\ 1\ 8\ 8]^T, & X_2^{32} = \delta_{10}[4\ 1\ 8\ 9]^T, & X_2^{33} = \delta_{10}[4\ 1\ 8\ 10]^T, \\
X_2^{34} = \delta_{10}[4\ 1\ 9\ 8]^T, & X_2^{35} = \delta_{10}[4\ 1\ 10\ 8]^T, & X_2^{36} = \delta_{10}[4\ 2\ 8\ 8]^T, \\
X_2^{37} = \delta_{10}[4\ 2\ 8\ 9]^T, & X_2^{38} = \delta_{10}[4\ 2\ 8\ 10]^T, & X_2^{39} = \delta_{10}[4\ 2\ 9\ 8]^T, \\
X_2^{40} = \delta_{10}[4\ 2\ 10\ 8]^T, & X_2^{41} = \delta_{10}[5\ 1\ 8\ 8]^T, & X_2^{42} = \delta_{10}[5\ 1\ 8\ 9]^T, \\
X_2^{42} = \delta_{10}[5\ 1\ 8\ 10]^T, & X_2^{44} = \delta_{10}[5\ 1\ 9\ 8]^T, & X_2^{45} = \delta_{10}[5\ 1\ 10\ 8]^T, \\
X_2^{46} = \delta_{10}[5\ 2\ 8\ 8]^T, & X_2^{47} = \delta_{10}[5\ 2\ 8\ 9]^T, & X_2^{48} = \delta_{10}[5\ 2\ 8\ 10]^T, \\
X_2^{49} = \delta_{10}[5\ 2\ 9\ 8]^T, & X_2^{50} = \delta_{10}[5\ 2\ 10\ 8]^T, & X_2^{51} = \delta_{10}[6\ 1\ 8\ 8]^T, \\
X_2^{52} = \delta_{10}[6\ 1\ 8\ 9]^T, & X_2^{53} = \delta_{10}[6\ 1\ 8\ 10]^T, & X_2^{54} = \delta_{10}[6\ 1\ 9\ 8]^T, \\
X_2^{55} = \delta_{10}[6\ 1\ 10\ 8]^T, & X_2^{56} = \delta_{10}[6\ 2\ 8\ 8]^T, & X_2^{57} = \delta_{10}[6\ 2\ 8\ 9]^T, \\
X_2^{58} = \delta_{10}[6\ 2\ 8\ 10]^T, & X_2^{59} = \delta_{10}[6\ 2\ 9\ 8]^T, & X_2^{60} = \delta_{10}[6\ 2\ 10\ 8]^T, \\
X_2^{61} = \delta_{10}[7\ 1\ 8\ 8]^T, & X_2^{62} = \delta_{10}[7\ 1\ 8\ 9]^T, & X_2^{63} = \delta_{10}[7\ 1\ 8\ 10]^T, \\
X_2^{64} = \delta_{10}[7\ 1\ 9\ 8]^T, & X_2^{65} = \delta_{10}[7\ 1\ 10\ 8]^T, & X_2^{66} = \delta_{10}[7\ 2\ 8\ 8]^T, \\
X_2^{67} = \delta_{10}[7\ 2\ 8\ 9]^T, & X_2^{68} = \delta_{10}[7\ 2\ 8\ 10]^T, & X_2^{69} = \delta_{10}[7\ 2\ 9\ 8]^T, \\
X_2^{70} = \delta_{10}[7\ 2\ 10\ 8]^T. & &
\end{array}$$

Finally, for the last column, we have

$$L\left(\times\right)x_3 = b_3, \tag{9.33}$$

where

$$b_3 = \delta_{10}^1 \ltimes \delta_{10}^4 \ltimes \delta_{10}^9 = \delta_{1000}^{39}.$$

Solving it, we have

$$\begin{array}{lll}
X_3^1 = \delta_{10}[9\ 9\ 1\ 1]^T, & X_3^2 = \delta_{10}[9\ 9\ 2\ 1]^T, & X_3^3 = \delta_{10}[9\ 9\ 3\ 1]^T,\\
X_3^4 = \delta_{10}[9\ 9\ 4\ 1]^T, & X_3^5 = \delta_{10}[9\ 9\ 5\ 1]^T, & X_3^6 = \delta_{10}[9\ 9\ 6\ 1]^T,\\
X_3^7 = \delta_{10}[9\ 9\ 7\ 1]^T, & X_3^8 = \delta_{10}[9\ 9\ 8\ 1]^T, & X_3^9 = \delta_{10}[9\ 9\ 9\ 1]^T,\\
X_3^{10} = \delta_{10}[9\ 9\ 10\ 1]^T, & X_3^{11} = \delta_{10}[9\ 10\ 1\ 1]^T, & X_3^{12} = \delta_{10}[9\ 10\ 2\ 1]^T,\\
X_3^{13} = \delta_{10}[9\ 10\ 3\ 1]^T, & X_3^{14} = \delta_{10}[9\ 10\ 4\ 1]^T, & X_3^{15} = \delta_{10}[9\ 10\ 5\ 1]^T,\\
X_3^{16} = \delta_{10}[9\ 10\ 6\ 1]^T, & X_3^{17} = \delta_{10}[9\ 10\ 7\ 1]^T, & X_3^{18} = \delta_{10}[9\ 10\ 8\ 1]^T,\\
X_3^{19} = \delta_{10}[9\ 10\ 9\ 1]^T, & X_3^{20} = \delta_{10}[9\ 10\ 10\ 1]^T, & X_3^{21} = \delta_{10}[10\ 9\ 1\ 1]^T,\\
X_3^{22} = \delta_{10}[10\ 9\ 2\ 1]^T, & X_3^{23} = \delta_{10}[10\ 9\ 3\ 1]^T, & X_3^{24} = \delta_{10}[10\ 9\ 4\ 1]^T,\\
X_3^{25} = \delta_{10}[10\ 9\ 5\ 1]^T, & X_3^{26} = \delta_{10}[10\ 9\ 6\ 1]^T, & X_3^{27} = \delta_{10}[10\ 9\ 7\ 1]^T,\\
X_3^{28} = \delta_{10}[10\ 9\ 8\ 1]^T, & X_3^{29} = \delta_{10}[10\ 9\ 9\ 1]^T, & X_3^{30} = \delta_{10}[10\ 9\ 10\ 1]^T,\\
X_3^{31} = \delta_{10}[10\ 10\ 1\ 1]^T, & X_3^{32} = \delta_{10}[10\ 10\ 2\ 1]^T, & X_3^{33} = \delta_{10}[10\ 10\ 3\ 1]^T,\\
X_3^{34} = \delta_{10}[10\ 10\ 4\ 1]^T, & X_3^{35} = \delta_{10}[10\ 10\ 5\ 1]^T, & X_3^{36} = \delta_{10}[10\ 10\ 6\ 1]^T,\\
X_3^{37} = \delta_{10}[10\ 10\ 7\ 1]^T, & X_3^{38} = \delta_{10}[10\ 10\ 8\ 1]^T, & X_3^{39} = \delta_{10}[10\ 10\ 9\ 1]^T.
\end{array}$$

We conclude the following:

(1) We have totally 10 × 70 × 39 = 27300 solutions in Ξ^4.

(2) The largest solution, corresponding to the largest solutions of each column, is

$$X^* = [X_1^1\ \ X_2^1\ \ X_3^1] \sim \begin{bmatrix} 1 & 1 & 0.2\\ 0.8 & 1 & 0.2\\ 0.7 & 0.3 & 1\\ 0.6 & 0.3 & 1 \end{bmatrix}.$$

(3) There is no smallest solution, because for the second column there are two minimum solutions

$$X_2^{68} = \delta_{10}[7\ 2\ 8\ 10]^T \sim \begin{bmatrix} 0.4\\ 0.9\\ 0.3\\ 0 \end{bmatrix}; \quad X_2^{70} = \delta_{10}[7\ 2\ 10\ 8]^T \sim \begin{bmatrix} 0.4\\ 0.9\\ 0\\ 0.3 \end{bmatrix}.$$

Hence, we have also two minimum solutions for X as

$$X_*^1 = [X_1^{10}\ \ X_2^{68}\ \ X_3^{39}] \sim \begin{bmatrix} 0 & 0.4 & 0\\ 0.8 & 0.9 & 0\\ 0.7 & 0.3 & 0.2\\ 0.6 & 0 & 1 \end{bmatrix},$$

and

$$X_*^2 = [X_1^{10} \quad X_2^{70} \quad X_3^{39}] \sim \begin{bmatrix} 0 & 0.4 & 0 \\ 0.8 & 0.9 & 0 \\ 0.7 & 0 & 0.2 \\ 0.6 & 0.3 & 1 \end{bmatrix}.$$

(4) The solution provided in Sanchez (2002) is

$$R = \begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0.8 & 1 & 0 \\ 0.7 & 0 & 0.5 \\ 0.6 & 0.3 & 1 \end{bmatrix},$$

which corresponds to

$$X = [X_1^8 \ X_2^{55} \ X_3^{16}].$$

(5) Finally, we consider the set of all solutions. Note that for PSS's we have

$$X_1 = \delta_{10}[a \ 3 \ 4 \ 5]^T, \quad 1 \le a \le 10.$$

It follows from Theorem 9.2 that

$$X_1 = \begin{bmatrix} \alpha \\ 0.8 \\ 0.7 \\ 0.6 \end{bmatrix}, \quad \text{where } 0 \le \alpha \le 1.$$

Similarly, we can calculate that X_2 is either

$$X_2^1 = \begin{bmatrix} \alpha \\ \beta \\ 0.3 \\ \eta \end{bmatrix}, \quad \text{where } 0.4 \le \alpha \le 1, \ 0.9 \le \beta \le 1, \ 0 \le \eta \le 0.3;$$

or

$$X_2^2 = \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ 0.3 \end{bmatrix}, \quad \text{where } 0.4 \le \alpha \le 1, \ 0.9 \le \beta \le 1, \ 0 \le \gamma \le 0.3.$$

X_3 can be expressed as

$$X_3^1 = \begin{bmatrix} 0.2 \\ \beta \\ \gamma \\ 1 \end{bmatrix}, \quad 0 \le \beta \le 0.2; \ 0 \le \gamma \le 1;$$

or

$$X_3^2 = \begin{bmatrix} \alpha \\ 0.2 \\ \gamma \\ 1 \end{bmatrix}, \quad 0 \le \alpha \le 0.2;\ 0 \le \gamma \le 1;$$

or

$$X_3^3 = \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ 1 \end{bmatrix}, \quad 0 \le \alpha \le 0.2;\ 0 \le \beta \le 0.2;\ 0.2 \le \gamma \le 1.$$

Summarizing them, we have 6 groups of solutions (with possible overlaps), which are expressed as

$$R_1 = \begin{bmatrix} 0 \le r_{11} \le 1 & 0.4 \le r_{12} \le 1 & 0.2 \\ 0.8 & 0.9 \le r_{22} \le 1 & 0 \le r_{23} \le 0.2 \\ 0.7 & 0.3 & 0 \le r_{33} \le 1 \\ 0.6 & 0 \le r_{42} \le 0.3 & 1 \end{bmatrix};$$

or

$$R_2 = \begin{bmatrix} 0 \le r_{11} \le 1 & 0.4 \le r_{12} \le 1 & 0 \le r_{13} \le 0.2 \\ 0.8 & 0.9 \le r_{22} \le 1 & 0.2 \\ 0.7 & 0.3 & 0 \le r_{33} \le 1 \\ 0.6 & 0 \le r_{42} \le 0.3 & 1 \end{bmatrix};$$

or

$$R_3 = \begin{bmatrix} 0 \le r_{11} \le 1 & 0.4 \le r_{12} \le 1 & 0 \le r_{13} \le 0.2 \\ 0.8 & 0.9 \le r_{22} \le 1 & 0 \le r_{23} \le 0.2 \\ 0.7 & 0.3 & 0.2 \le r_{33} \le 1 \\ 0.6 & 0 \le r_{42} \le 0.3 & 1 \end{bmatrix};$$

or

$$R_4 = \begin{bmatrix} 0 \le r_{11} \le 1 & 0.4 \le r_{12} \le 1 & 0.2 \\ 0.8 & 0.9 \le r_{22} \le 1 & 0 \le r_{23} \le 0.2 \\ 0.7 & 0 \le r_{32} \le 0.3 & 0 \le r_{33} \le 1 \\ 0.6 & 0.3 & 1 \end{bmatrix};$$

or

$$R_5 = \begin{bmatrix} 0 \le r_{11} \le 1 & 0.4 \le r_{12} \le 1 & 0 \le r_{13} \le 0.2 \\ 0.8 & 0.9 \le r_{22} \le 1 & 0.2 \\ 0.7 & 0 \le r_{32} \le 0.3 & 0 \le r_{33} \le 1 \\ 0.6 & 0.3 & 1 \end{bmatrix};$$

or

$$R_6 = \begin{bmatrix} 0 \le r_{11} \le 1 & 0.4 \le r_{12} \le 1 & 0 \le r_{13} \le 0.2 \\ 0.8 & 0.9 \le r_{22} \le 1 & 0 \le r_{23} \le 0.2 \\ 0.7 & 0 \le r_{32} \le 0.3 & 0.2 \le r_{33} \le 1 \\ 0.6 & 0.3 & 1 \end{bmatrix}.$$

Remark 9.1. Searching all the solutions and providing the overall picture of the set of solutions are significant in applications. For instance, in designing fuzzy controllers it provides a knowledge for finding "best" solutions, or to diagnose where is the problem if there is no solution. Then further improvement can be directed.

Exercises

9.1 Let $\Theta \subset \mathcal{D}_{\infty}^{m\times n}$.

(i) Does a maximum (minimum) element of Θ imply that it is a largest (smallest) element? If "yes", prove it; if "no", give a counterexample.

(ii) Does a largest (smallest) element of Θ imply that it is a maximum (minimum) element? If "yes", prove it; if "no", give a counterexample.

9.2 Consider $\mathcal{B}_{m\times n}$ with the matrix addition $\langle + \rangle$ and product $\langle \times \rangle$. (We refer to Exercise 8.6 of Chapter 8 for the definition.) Check whether the order-preserving property, i.e., Proposition 9.1, remains true?

9.3 Assume $U = \{-2,-1,0,1,2\}$, $V = \{-3,-2,-1,0,1\}$, $f(u,v) = (u+v)^2$, and $m^* = \max_{(u,v)\in U\times V} f(u,v)$. Define

$$\mu_R(u,v) := \frac{f(u,v)}{m^*}, \quad (u,v) \in U \times V.$$

Calculate the relational matrix of R.

9.4 Prove the first part of Theorem 9.1. That is, $A^T \oplus B$ is the largest solution.

9.5 Prove Lemma 9.2.

9.6 Consider the following FRE:

$$\begin{bmatrix} 1 \; 0.5 \; 0.5 \end{bmatrix} (\times) \begin{bmatrix} x_1 & x_4 \\ x_2 & x_5 \\ x_3 & x_6 \end{bmatrix} = \begin{bmatrix} 0 \; 0.5 \end{bmatrix}.$$

(i) Find the largest solution using Theorem 9.1.

(ii) Find all the solutions.

9.7 Solve the following FREs:

(i)

$$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 0 \end{bmatrix} (\times) \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} = \begin{bmatrix} 0.5 & 0.4 \\ 0.6 & 0.5 \end{bmatrix}.$$

(ii)

$$\begin{bmatrix} x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \end{bmatrix} (\times) \begin{bmatrix} 0.8 & 0.3 \\ 0 & 0.4 \\ 0.3 & 1 \end{bmatrix} = \begin{bmatrix} 0.4 & 0 \\ 1 & 0.5 \end{bmatrix}.$$

9.8 (i) Consider equation (9.4). One way to solve it is converting it into s equations as in (9.13), and then solve R column by column. Prove that we can also convert it into

$$(I_p(\otimes)A)\,(\times)V_c(R) = V_c(B),$$

and then solve whole R simultaneously.

(ii) Carefully check that the technique developed in Chapter 3 is applicable to solving fuzzy relational equations. That is, when the matrix addition and production have been replaced by $(+)$ and $(\times)$ respectively, the computation rules used in Chapter 3 remain correct.

9.9 Solving the following fuzzy relational equations:

(i)

$$\begin{bmatrix} 0 & 0.3 \\ 0.5 & 0 \end{bmatrix} (\times) \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} = \begin{bmatrix} 0.3 & 1 \\ 0 & 0.5 \end{bmatrix} (\times) \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} (+) \begin{bmatrix} 0 & 0.3 \\ 1 & 0.5 \end{bmatrix}.$$

(ii)

$$\begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0 \end{bmatrix} (\times) \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} (+) \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix}^T (\times) \begin{bmatrix} 0.2 & 1 \\ 1 & 0.5 \end{bmatrix} = \begin{bmatrix} 0.2 & 0.8 \\ 1 & 0.5 \end{bmatrix}.$$

9.10 For fuzzy relations $A \in \mathcal{F}(U \times V)$, and $B \in \mathcal{F}(V \times W)$, we can define another composition of fuzzy relations $A \odot B \in \mathcal{F}(U \times W)$ as

$$\mu_{A\odot B}(u_i, w_j) = \bigwedge_k (\mu_A(u_i, v_k) \vee \mu_B(v_k, w_j)).$$

(i) Prove

$$A^C \odot B^C = (A\,(\times)\,B)^C; \quad (A \odot B)^C = A^C\,(\times)\,B^C.$$

(ii) Solve

$$\begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} \odot \begin{bmatrix} 1 & 0.5 \\ 0.5 & 0 \end{bmatrix} = \begin{bmatrix} 0.5 & 0.4 \\ 0.6 & 0.5 \end{bmatrix}.$$

Chapter 10

Fuzzy Control with Coupled Fuzzy Relations

The fuzzy control theory is a combination of the fuzzy logic theory and the automatic control technology, and it plays more and more important roles in industrial controls (Lewis *et al.*, 2002; Li *et al.*, 2010; Mamdani, 1974; Zhang *et al.*, 2009). There are many nice references on fuzzy controls, e.g., Lee (2005); Zhu (2005). Fuzzy control for multiple-input multiple-output (MIMO) systems is theoretically more challenging and practically more useful. A traditional way to design multiple fuzzy controls is to design each controls separately (Lee, 2005; Zhu, 2005). This method can only be used to systems with decomposable multiple fuzzy relations. In this chapter we consider the fuzzy control of systems with coupled multiple fuzzy relations. First, operations of multiple fuzzy relations are considered. Then the synthesis of fuzzy control systems with coupled multiple fuzzy relations is investigated. Finally, some numerical examples are presented to illustrate the design technique.

This chapter is based on Feng *et al.* (2011).

10.1 Multiple Fuzzy Relations

10.1.1 *Matrix Expression*

This subsection considers the matrix expression of multiple fuzzy relations. In most literature the fuzzy relations are defined over two universes of discourse. In practice the fuzzy relations over multiple universes of discourse need to be treated. We give the following definition first.

Definition 10.1. Let E_i, $i = 1, \cdots, k$ be k universes of discourse. A multiple fuzzy relation R over E_i, $i = 1, \cdots, k$ is a fuzzy set on the product space $\prod_{i=1}^{k} E_i = E_1 \times \cdots \times E_k$. Precisely, for each point $(e_1, \cdots, e_k) \in$

$\prod_{i=1}^{k} E_i$ there is a membership degree $\mu_R(e_1, \cdots, e_k) \in [0,1]$.

Assume $E_i = \{e^i_1, \cdots, e^i_{n_i}\}$, $i = 1, \cdots, k$. Denote by

$$r_{j_1,\cdots,j_k} = \mu_R(e^1_{j_1}, \cdots, e^k_{j_k}), \quad j_t = 1, \cdots, n_t,\ t = 1, \cdots, k.$$

They are called the relational parameters.

Let $\{\alpha_1, \cdots, \alpha_p\}$ and $\{\beta_1, \cdots, \beta_q\}$ $(p + q = k)$ be a partition of $\{1, 2, \cdots, k\}$. We can arrange the data

$$\{\, r_{j_1,\cdots,j_k} \,|\, j_t = 1, \cdots, n_t,\ t = 1, \cdots, k\}$$

into a matrix $M_R \in \mathcal{M}_{(n_{\alpha_1}\times\cdots\times n_{\alpha_p})\times(n_{\beta_1}\times\cdots\times n_{\beta_q})}$ in the order of $\mathrm{id}(j_{\alpha_1}, \cdots, j_{\alpha_p}; n_{\alpha_1}, \cdots, n_{\alpha_p}) \times \mathrm{id}(j_{\beta_1}, \cdots, j_{\beta_q}; n_{\beta_1}, \cdots, n_{\beta_q})$. The matrix M_R is called the relational matrix of R.

Since the relational parameters can be arranged into a matrix according to different orders, the relational matrix of a relation R is not unique. We give an example to depict this.

Example 10.1. Let $E = \{e_1, e_2, e_3, e_4\}$, $F = \{f_1, f_2, f_3\}$, and $G = \{g_1, g_2\}$ be three universes of discourse. R is a fuzzy set on $E \times F \times G$, and

$$r_{j_1,j_2,j_3} = \mu_R(e_{j_1}, f_{j_2}, g_{j_3}), \quad j_1 = 1, 2, 3, 4;\ j_2 = 1, 2, 3;\ j_3 = 1, 2.$$

(1) Assume we arrange M_R in the order of $\mathrm{id}(j_1; 4) \times \mathrm{id}(j_2, j_3; 3, 2)$, then we have

$$M_R = \begin{bmatrix} r_{1,1,1} & r_{1,1,2} & r_{1,2,1} & r_{1,2,2} & r_{1,3,1} & r_{1,3,2} \\ r_{2,1,1} & r_{2,1,2} & r_{2,2,1} & r_{2,2,2} & r_{2,3,1} & r_{2,3,2} \\ r_{3,1,1} & r_{3,1,2} & r_{3,2,1} & r_{3,2,2} & r_{3,3,1} & r_{3,3,2} \\ r_{4,1,1} & r_{4,1,2} & r_{4,2,1} & r_{4,2,2} & r_{4,3,1} & r_{4,3,2} \end{bmatrix}.$$

(2) Assume we arrange M_R in the order of $\mathrm{id}(j_1, j_3; 4, 2) \times \mathrm{id}(j_2; 3)$, then we have

$$M_R = \begin{bmatrix} r_{1,1,1} & r_{1,2,1} & r_{1,3,1} \\ r_{1,1,2} & r_{1,2,2} & r_{1,3,2} \\ r_{2,1,1} & r_{2,2,1} & r_{2,3,1} \\ r_{2,1,2} & r_{2,2,2} & r_{2,3,2} \\ r_{3,1,1} & r_{3,2,1} & r_{3,3,1} \\ r_{3,1,2} & r_{3,2,2} & r_{3,3,2} \\ r_{4,1,1} & r_{4,2,1} & r_{4,3,1} \\ r_{4,1,2} & r_{4,2,2} & r_{4,3,2} \end{bmatrix}.$$

(3) Assume we arrange M_R into a row in the order of $\mathrm{id}(j_1, j_2, j_3; 4, 3, 2)$, then we have

$$\begin{aligned}M_R = [\, & r_{1,1,1}\ r_{1,1,2}\ r_{1,2,1}\ r_{1,2,2}\ r_{1,3,1}\ r_{1,3,2}\ r_{2,1,1}\ r_{2,1,2} \\ & r_{2,2,1}\ r_{2,2,2}\ r_{2,3,1}\ r_{2,3,2}\ r_{3,1,1}\ r_{3,1,2}\ r_{3,2,1}\ r_{3,2,2} \\ & r_{3,3,1}\ r_{3,3,2}\ r_{4,1,1}\ r_{4,1,2}\ r_{4,2,1}\ r_{4,2,2}\ r_{4,3,1}\ r_{4,3,2}].\end{aligned}$$

Different forms of relational matrices of a relation are used to emphasize the relations corresponding to different splits, and they can then be used for different purposes. This fact will be demonstrated in next section.

10.1.2 *Multiple Fuzzy Inference*

Assume we have a fuzzy relation $R \in \mathcal{F}(P \times Q)$, where P and Q are two finite sets. The fuzzy inference means given a fuzzy set $A \in \mathcal{F}(Q)$, using R we can get a fuzzy set $B \in \mathcal{F}(P)$. In fact,

$$\mathcal{V}_B = M_R \,(\times)\, \mathcal{V}_A. \tag{10.1}$$

Note that taking transpose, we can have an $A \in \mathcal{F}(Q)$ from $B \in \mathcal{F}(P)$ as

$$\mathcal{V}_A = M_R^T \,(\times)\, \mathcal{V}_B. \tag{10.2}$$

This is called the fuzzy inference (Hu, 2010).

Next, we consider multiple fuzzy inference.

Definition 10.2. Assume a multiple relation is given on the product space as $R \in \mathcal{F}\left(\prod_{i=1}^{k} E_i\right)$. Let $\alpha = \{\alpha_1, \cdots, \alpha_p\}$ and $\beta = \{\beta_1, \cdots, \beta_q\}$ be a partition of $\{1, 2, \cdots, k\}$. Given a set of fuzzy sets A_{α_i}, $i = 1, \cdots, p$, a multiple fuzzy inference means we can find, using A_{α_i}, a fuzzy relation

$$R_\beta \in \mathcal{F}\left(\prod_{j=1}^{q} E_{\beta_j}\right).$$

Precisely,

$$M_{R_\beta} = M_R \,(\times)\, \ltimes_{j=1}^{p} \mathcal{V}_{A_{\alpha_j}}. \tag{10.3}$$

Particularly, as $|\beta| = 1$, M_{R_β} becomes $\mathcal{V}_{A_\beta}$, where $A_\beta \in \mathcal{F}(E_\beta)$.

We give an example to show how to do this.

Example 10.2. Assume we have universes of discourse E, F, and G as shown in Example 10.1. Moreover, a relation $R \in \mathcal{F}(E \times F \times G)$ is given, i.e.,

$$r_{i,j,k} = \mu_R(e_i, f_j, g_k), \quad i = 1, 2, 3, 4;\ j = 1, 2, 3;\ k = 1, 2$$

are known.

Now assume we have $A \in \mathcal{F}(E)$ and $B \in \mathcal{F}(F)$. Using R, (10.3) can provide a fuzzy inference $C \in \mathcal{F}(G)$. The following is a numerical example.

Assume M_R is arranged as $\mathrm{id}(k;2) \times \mathrm{id}(i,j;4,3)$ as

$$M_R = \begin{bmatrix} 0\ 0.3\ 0.7\ \ 1\ \ 0.5\ 0.9\ 0.4\ 0.1\ 0\ 0.3\ 0.1\ 0 \\ 1\ 0.2\ 0.3\ 0.5\ 0.3\ 0.2\ 0.7\ 0.8\ 1\ 0.4\ 0.6\ 1 \end{bmatrix},$$

and the vector forms of A and B are

$$\mathcal{V}_A = [0.1\ 0.5\ 1\ 0.4]^T;$$
$$\mathcal{V}_B = [0.8\ 0.7\ 0.5]^T.$$

Then we have C as

$$\mathcal{V}_C = M_R\,(\ltimes)\,\mathcal{V}_A\,(\ltimes)\,\mathcal{V}_B = [0.5\ 0.7]^T.$$

That is, $C = 0.5/g_1 + 0.7/g_2$.

When only a fuzzy set $A \in \mathcal{F}(E)$ is given. Then the fuzzy inference provides a relation $R' \in \mathcal{F}(F,G)$, which has its relational matrix as

$$M_{R'} = M_R\,(\ltimes)\,\mathcal{V}_A.$$

If we use previous R and A, then we have

$$M_{R'} = \begin{bmatrix} 0.5\ 0.5\ 0.5 \\ 0.7\ 0.8\ \ 1 \end{bmatrix}.$$

Remark 10.1. When a multiple fuzzy relation is considered, we must be aware of the order of the arrangement. Proper order should be chosen for corresponding fuzzy inference.

10.1.3 *Compounded Multiple Fuzzy Relations*

This subsection considers the composition of multiple fuzzy relations. We first consider the classical case (Hu, 2010; Zhu, 2005).

Definition 10.3. Let E, F, G be three sets, R and S be two relations over $E \times F$ and $F \times G$ respectively. That is, $R \in \mathcal{F}(E \times F)$, $S \in \mathcal{F}(F \times G)$. Then the compounded relation $R \circ S \in \mathcal{F}(E \times G)$ is a relation on $E \times G$, defined as

$$\mu_{R\circ S}(e,g) = \vee_{d\in F}\left[\mu_R(e,d) \wedge \mu_S(d,g)\right], \quad e \in E,\ g \in G.$$

The following result is an immediate consequence of the definition.

Proposition 10.1. *Let E, F, and G be three finite sets, and $R \in \mathcal{F}(E\times F)$, $S \in \mathcal{F}(F \times G)$. Assume R and S have their relational matrices as M_R and M_S respectively. Then*

$$M_{R\circ S} = M_R\,(\times)\,M_S. \tag{10.4}$$

Consider the relations on same universe of discourse (Zhu, 2005).

Definition 10.4.

(1) $R \in \mathcal{F}(E \times E)$ is called an identity relation, if

$$\mu_R(x,y) = \begin{cases} 1, & x = y, \\ 0, & \text{otherwise.} \end{cases}$$

Note that if $|E| = n$, then identity relation R has its matrix form $M_R = I_n$.

(2) $R \in \mathcal{F}(E \times E)$ is said to be self-related, if

$$\mu_R(x,x) = 1, \quad \forall x \in E.$$

It is said to be self-unrelated, if

$$\mu_R(x,x) = 0, \quad \forall x \in E.$$

(3) $R \in \mathcal{F}(E \times E)$ is said to be symmetric, if

$$\mu_R(x,y) = \mu_R(y,x), \quad \forall x, y \in E.$$

(4) $R \in \mathcal{F}(E \times E)$ is said to be transitive, if

$$M_R \,(\times)\, M_R = (M_R)^{(2)} \le M_R.$$

For the sake of compactness, in the following definition we consider only relation on three sets. More than three cases can be treated in exactly the same way.

Definition 10.5. Let X, Y and Z be three sets. A relation among them is a fuzzy set $R \in \mathcal{F}(X \times Y \times Z)$.

Assume $X = \{x_1, x_2, \cdots, x_m\}$, $Y = \{y_1, y_2, \cdots, y_n\}$, and $Z = \{z_1, z_2, \cdots, z_r\}$. Then we can arrange $\{\mu_R(x_i, y_j, z_k) \,|\, i = 1, \cdots, m; j = 1, \cdots, n; k = 1, \cdots, r\}$ into a relational matrix. Using the order of $\mathrm{id}(i; m) \times \mathrm{id}(j, k; n, r)$, then we have

$$M_{R(X \times YZ)} = \begin{bmatrix} \mu_A(x_1,y_1,z_1) & \cdots & \mu_A(x_1,y_1,z_r) & \cdots & \mu_A(x_1,y_n,z_1) & \cdots & \mu_A(x_1,y_n,z_r) \\ \mu_A(x_2,y_1,z_1) & \cdots & \mu_A(x_2,y_1,z_r) & \cdots & \mu_A(x_2,y_n,z_1) & \cdots & \mu_A(x_2,y_n,z_r) \\ \vdots & & & & & & \\ \mu_A(x_m,y_1,z_1) & \cdots & \mu_A(x_m,y_1,z_r) & \cdots & \mu_A(x_m,y_n,z_1) & \cdots & \mu_A(x_m,y_n,z_r) \end{bmatrix}. \tag{10.5}$$

If we use the order of $\mathrm{id}(j;n) \times \mathrm{id}(i,k;m,r)$, then we have

$$M_{R(Y\times XZ)} = \begin{bmatrix} \mu_A(x_1,y_1,z_1) & \cdots & \mu_A(x_1,y_1,z_r) & \cdots & \mu_A(x_m,y_1,z_1) & \cdots & \mu_A(x_m,y_1,z_r) \\ \mu_A(x_1,y_2,z_1) & \cdots & \mu_A(x_1,y_2,z_r) & \cdots & \mu_A(x_m,y_2,z_1) & \cdots & \mu_A(x_m,y_2,z_r) \\ \vdots & & & & & & \\ \mu_A(x_1,y_n,z_1) & \cdots & \mu_A(x_1,y_n,z_r) & \cdots & \mu_A(x_m,y_n,z_1) & \cdots & \mu_A(x_m,y_n,z_r) \end{bmatrix}. \tag{10.6}$$

For multiple fuzzy relations we define two kinds of compositions. First, we extend Definition 10.3 to multiple fuzzy relations. Let $\{E_1,\cdots,E_\alpha; F_1,\cdots,F_\beta; G_1,\cdots,G_\gamma\}$ be a set of three groups of universes of discourse. Let

$$\begin{aligned} E_i &= \left\{e_1^i,\cdots,e_{n_i^\alpha}^i\right\}, \quad i=1,\cdots,\alpha; \\ F_i &= \left\{f_1^i,\cdots,f_{n_i^\beta}^i\right\}, \quad i=1,\cdots,\beta; \\ G_i &= \left\{g_1^i,\cdots,g_{n_i^\gamma}^i\right\}, \quad i=1,\cdots,\gamma. \end{aligned}$$

Assume we have two multiple fuzzy relations

$$\begin{aligned} R &\in \mathcal{F}(E_1\times\cdots\times E_\alpha\times F_1\times\cdots\times F_\beta), \\ S &\in \mathcal{F}(F_1\times\cdots\times F_\beta\times G_1\times\cdots\times G_\gamma). \end{aligned} \tag{10.7}$$

Then the relational matrix of R, denoted by $M_{R(E_1\cdots E_\alpha\times F_1\cdots F_\beta)}$ is arranging

$$\left\{\mu_{e^1_{\xi_1}\cdots e^\alpha_{\xi_\alpha} f^1_{\eta_1}\cdots f^\beta_{\eta_\beta}} \,\middle|\, 1\le \xi_i\le\alpha; 1\le\eta_i\le\beta\right\}$$

in the order of

$$\mathrm{id}\left(\xi_1,\cdots,\xi_\alpha; n_1^\alpha,\cdots,n_\alpha^\alpha\right)\times \mathrm{id}\left(\eta_1,\cdots,\eta_\beta; n_1^\beta,\cdots,n_\beta^\beta\right).$$

Similarly, the structure matrix of S, denoted by $M_{S(F_1\cdots F_\beta\times G_1\cdots G_\gamma)}$ is arranging

$$\left\{\mu_{f^1_{\eta_1}\cdots f^\beta_{\eta_\beta} g^1_{\zeta_1}\cdots g^\gamma_{\zeta_\gamma}} \,\middle|\, 1\le \eta_i\le\beta; 1\le\zeta_i\le\gamma\right\}$$

in the order of

$$\mathrm{id}\left(\eta_1,\cdots,\eta_\beta; n_1^\beta,\cdots,n_\beta^\beta\right)\times \mathrm{id}\left(\zeta_1,\cdots,\zeta_\gamma; n_1^\gamma,\cdots,n_\gamma^\gamma\right).$$

Using these notations, we can give the following definition for transitive composition.

Definition 10.6. Let R and S be as in (10.7). The transitive composition of R and S, denoted as $R\circ S$, is a relation of $R\circ S\in\mathcal{F}(E_1\cdots E_\alpha\times G_1\cdots G_\gamma))$, with its relational matrix as

$$M_{R\circ S} = M_{R(E_1\cdots E_\alpha\times F_1\cdots F_\beta)}\,(\times)\,M_{S(F_1\cdots F_\beta\times G_1\cdots G_\gamma)}, \tag{10.8}$$

where $M_{R(E_1\cdots E_\alpha\times F_1\cdots F_\beta)}$ and $M_{S(G_1\cdots G_\gamma\times F_1\cdots F_\beta)}$ are the relational matrices of R and S respectively.

Next, we define another composition called the jointed composition.

Definition 10.7. Assume R and S are given as in (10.7) and their relational matrices are given as

$$M_{R(E_1\cdots E_\alpha\times F_1\cdots F_\beta)};\quad M_{S(G_1\cdots G_\gamma\times F_1\cdots F_\beta)}. \tag{10.9}$$

The jointed composition of R and S, denoted by $R*S \in \mathcal{F}(E_1\cdots E_\alpha G_1\cdots G_\gamma \times F_1\cdots F_\beta)$, is defined by its relational matrix as

$$\begin{aligned}&M_{R*S}(E_1\cdots E_\alpha G_1\cdots G_\gamma\times F_1\cdots F_\beta)\\ =&M_{R(E_1\cdots E_\alpha\times F_1\cdots F_\beta)}\,(*)\,M_{S(G_1\cdots G_\gamma\times F_1\cdots F_\beta)}.\end{aligned} \tag{10.10}$$

Both transitive and jointed compositions can be used to deduce required multiple fuzzy relations as we wish. We give an example to show how to use both to manipulate the multiple fuzzy relations. They will be used in the following fuzzy control design.

Example 10.3. Let $X=\{x_1,x_2\}$, $Y=\{y_1,y_2,y_3\}$, $Z=\{z_1,z_2\}$, and $W=\{w_1,w_2,w_3\}$. We have $R\in\mathcal{F}(X\times Y\times Z)$, $S\in\mathcal{F}(Y\times W)$, and $T\in\mathcal{F}(Z\times W)$. Their relational matrices are

$$M_{R(X\times YZ)}=\begin{bmatrix}0.2&0&0.1&0.5&0.9&1\\0.4&0.3&0.7&0.8&0&0\end{bmatrix};$$

$$M_S=\begin{bmatrix}0&0.5&0.6\\0.1&0.3&0.8\\0.2&0.7&1\end{bmatrix};\quad M_T=\begin{bmatrix}0&0.1&0.9\\0.5&1&0.3\end{bmatrix}.$$

Intuitively, W has relation with X through both Y and Z. Therefore, we intend to find the relation of X and W. First, we calculate product M_S with M_T to get a relation in $Y\times Z\times W$ as

$$M_{S*T(YZ\times W)}=M_S\,(*)\,M_T=\begin{bmatrix}0&0.1&0.6\\0&0.5&0.3\\0&0.1&0.8\\0.1&0.3&0.3\\0&0.1&0.9\\0.2&0.7&0.3\end{bmatrix}.$$

Then $\Psi=R\circ(S*T)\in\mathcal{F}(X,W)$ has its relational matrix as

$$M_\Psi=M_{R(X\times YZ)}\,(\times)\,M_{S*T(YZ\times W)}=\begin{bmatrix}0.2&0.7&0.9\\0.1&0.3&0.7\end{bmatrix}.$$

Finally, it is worth noting that a general method for solving fuzzy relational equations was proposed in previous chapter, which is based on Cheng *et al.* (2011a) and is useful in fuzzy control design.

10.2 Fuzzy Control of Coupled Multiple Fuzzy Relations

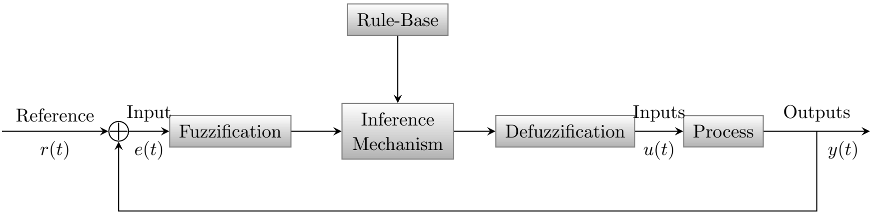


Fig. 10.1 A fuzzy control system

Fig. 10.1 (Passino and Yurkovich, 2002) shows the structure of a fuzzy control system. In this section, a new framework, based on matrix approach, is investigated.

10.2.1 *Fuzzification via Dual Fuzzy Structure*

This section considers only the fuzzification. We first introduce a dual fuzzy structure.

Definition 10.8.

(1) Let E be a universe of discourse, and $\mathcal{A} = \{A_1, \cdots, A_k\}$ be a set of fuzzy sets on E. Then $(E, \mathcal{A})$ is called a fuzzy structure. The support of A_i is defined as

$$\mathrm{Supp}(A_i) = \{e \in E \,|\, \mu_{A_i}(e) \neq 0\} \subset E.$$

(2) Assume E is a well ordered set. $\mathcal{A} = \{A_1, \cdots, A_k\}$ is called a set of degree-based fuzzy sets, if

$$\begin{aligned} \sup(\mathrm{Supp}(A_i)) &< \sup(\mathrm{Supp}(A_{i+1})), \\ \text{and } \inf(\mathrm{Supp}(A_i)) &< \inf(\mathrm{Supp}(A_{i+1})), \\ i &= 1, \cdots, k-1. \end{aligned} \tag{10.11}$$

(3) Given a fuzzy structure $(E, \mathcal{A})$ as in item (1). Assume E is a well ordered set and $\mathcal{A}$ is a set of degree-based fuzzy sets, i.e., (10.11) is satisfied. Then we may consider $(\mathcal{A}, E)$ as a fuzzy structure, where $\mathcal{A} = \{A_1, \cdots, A_k\}$ is considered as a universe of discourse, each $e \in E$ is a fuzzy set, with

$$\mu_e(A_i) := \mu_{A_i}(e), \quad i = 1, \cdots, k. \tag{10.12}$$

This fuzzy structure is called the dual structure of $(E, \mathcal{A})$. When $\mathcal{A}$ is a finite set, the dual structure has a finite universe of discourse.

In fact, the process of fuzzification is basically finding the dual structure. We use an example to describe this.

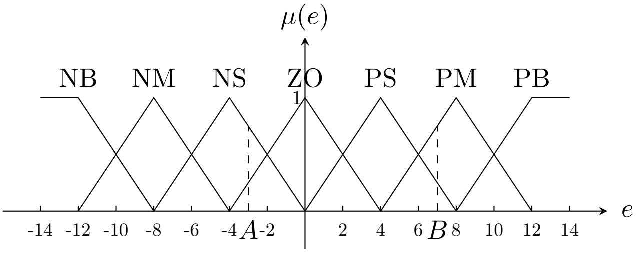


Fig. 10.2 The membership functions of fuzzy set

Example 10.4. Consider a measurement error $e \in [-14, 14]$. We consider 7 fuzzy sets, which have linguistic statements respectively as: NB (negative big), NM (negative medium), NS (negative small), ZO (zero), PS (positive small), PM (positive medium), and PB (positive big). The membership functions of the fuzzy sets are depicted in Fig. 10.2. Now for each point e in the universe of discourse $E = [-14, 14]$, we know its membership degrees for each fuzzy set. For instance, for point A, we have

$$\mu_{NS}(A) = 0.75;\ \mu_{ZO}(A) = 0.25;\ \mu_Y(A) = 0,\ Y = NB, NM, PS, PM, PB. \tag{10.13}$$

Similarly, for point B we have

$$\mu_{PS}(B) = 0.25;\ \mu_{PM}(B) = 0.75;\ \mu_Y(B) = 0,\ Y = NB, NM, NS, ZO, PB. \tag{10.14}$$

Checking the fuzzification in fuzzy control, one sees easily that what we need to do is to convert a point $e \in E$ to a fuzzy set. To do this, people usually interchange the universe of discourse with the set of degree-based fuzzy sets. Precisely, as in Example 10.4 we consider

$$D := \{d_1 = NB,\ d_2 = NM,\ d_3 = NS,\ d_4 = ZO,$$
$$d_5 = PS,\ d_6 = PM,\ d_7 = PB\}$$

as the universe of discourse and consider each $e \in E$ as a fuzzy set. In this consideration we can express (10.13) and (10.14) respectively as

$$\mu_A(d_3) = 0.75;\ \mu_A(d_4) = 0.25;\ \mu_A(d_i) = 0,\ i = 1, 2, 5, 6, 7. \tag{10.15}$$

$$\mu_B(d_5) = 0.25;\ \mu_B(d_6) = 0.75;\ \mu_B(d_i) = 0,\ i = 1, 2, 3, 4, 7. \tag{10.16}$$

In vector form we have

$$\begin{aligned}\mathcal{V}_A &= [0\ 0\ 0.75\ 0.25\ 0\ 0\ 0]^T;\\ \mathcal{V}_B &= [0\ 0\ 0\ 0\ 0.25\ 0.75\ 0]^T.\end{aligned} \tag{10.17}$$

10.2.2 *Design of Fuzzy Controller*

Roughly speaking, a fuzzy controller is a fuzzy inference mechanism. Assume the system in Fig. 10.1 has m inputs and p outputs, then the fuzzy controller has the form as

$$\Sigma \in \mathcal{F}(Y_1 \times \cdots \times Y_p \times U_1 \times \cdots \times U_m), \tag{10.18}$$

where Y_i, $i = 1, \cdots, p$, and U_j, $j = 1, \cdots, m$ have been fuzzificated, and the fuzzification process has been described in the previous subsection. Note that for the controller $\{Y_i\}$ becomes the input set and $\{U_i\}$ the output set. We give an example to depict this.

Example 10.5 (Zhu, 2005). *Assume a system has a single control U, which depends on A and B. Both A and B have 7 levels as $\{NB, NM, NS, ZO, PS, PM, PB\}$, and U has 13 levels as $\{NVB, NB, NMB, NMS, NS, NVS, ZO, PVS, PS, PMS, PMB, PB, PVB\}$. Using above listed natural orders and dual fuzzy structure, we denote by*

$$\begin{aligned}E_A &= \{a_1, \cdots, a_7\},\\ E_B &= \{b_1, \cdots, b_7\},\\ E_U &= \{u_1, \cdots, u_{13}\},\end{aligned}$$

the universes of discourse for A, B, and U respectively.

(1) In general, we can arrange

$$\{\, \mu_\Sigma(a_i, b_j, u_k) \,|\, i = 1, \cdots, 7;\ j = 1, \cdots, 7;\ k = 1, \cdots, 13\}$$

into a matrix M_Σ in the order of $\mathrm{id}(k; 13) \times \mathrm{id}(i, j; 7, 7)$. Then

$$M_\Sigma \in \mathcal{D}_\infty^{13\times 7^2}. \tag{10.19}$$

(2) Particularly, we may use the "If $A = \times$, and $B = \times$, then $U = \times$" rules, a rule table can be obtained. For instance, we have Rule Table 10.1.

To use vector expression we identify

$$\begin{aligned}&NB \sim \delta_7^1;\ NM \sim \delta_7^2;\ \cdots;\ PB \sim \delta_7^7;\\ &-1 \sim \delta_{13}^1;\ -0.8 \sim \delta_{13}^2;\ \cdots;\ 1 \sim \delta_{13}^{13}.\end{aligned} \tag{10.20}$$

Table 10.1 Rule table

$A\backslash U\backslash B$	NB	NM	NS	ZO	PS	PM	PB
NB	-1	-0.8	-0.6	-0.4	-0.2	-0.1	0
NM	-0.8	-0.6	-0.4	-0.2	-0.1	0	0.1
NS	-0.6	-0.4	-0.2	-0.1	0	0.1	0.2
ZO	-0.4	-0.2	-0.1	0	0.1	0.2	0.4
PS	-0.2	-0.1	0	0.1	0.2	0.4	0.6
PM	-0.1	0	0.1	0.2	0.4	0.6	0.8
PB	0	0.1	0.2	0.4	0.6	0.8	1

Then M_Σ can be expressed as

$$\begin{aligned} M_\Sigma = \delta_{13}[\, & 1\ \ 2\ \ 3\ \ 4\ \ 5\ \ 6\ \ 7\ \ 2\ \ 3\ \ 4\ \ 5\ \ 6\ \ 7\ \ 8 \\ & 3\ \ 4\ \ 5\ \ 6\ \ 7\ \ 8\ \ 9\ \ 4\ \ 5\ \ 6\ \ 7\ \ 8\ \ 9\ \ 10 \\ & 5\ \ 6\ \ 7\ \ 8\ \ 9\ \ 10\ \ 11\ \ 6\ \ 7\ \ 8\ \ 9\ \ 10\ \ 11\ \ 12 \\ & 7\ \ 8\ \ 9\ \ 10\ \ 11\ \ 12\ \ 13] \in \mathcal{B}_2^{13\times 49}. \end{aligned} \tag{10.21}$$

It is easy to see that (10.21), which is obtained by if-then rules, is a particular case of (10.19).

In general case a fuzzy controller is mathematically equivalent to a fuzzy relation (10.18). We describe this as follows. First, we specify degree-based fuzzy sets of Y_i and U_j as

$$\begin{aligned} E_{Y_i} &= \left\{y^i_1, \cdots, y^i_{\alpha_i}\right\}, \quad i = 1, \cdots, p; \\ E_{U_j} &= \left\{u^j_1, \cdots, u^j_{\beta_j}\right\}, \quad j = 1, \cdots, m. \end{aligned} \tag{10.22}$$

Note that y^i_k, $k = 1, \cdots, \alpha_i$, correspond to "negative big", "negative middle", ..., which are what we mean the degree-based fuzzy sets.

We use dual fuzzy structure. That is, consider E_{Y_i} as the universe of discourse for Y_i, and E_{U_j} as the universe of discourse for U_j; and meanwhile, consider each true value $y_i \in Y_i$ as a fuzzy set over E_{Y_i}, and $u_j \in U_j$ as a fuzzy set over E_{Y_i}.

A fuzzy controller is a fuzzy relation among the set of outputs and controls

$$\{Y_1,\ \cdots,\ Y_p;\ U_1,\ \cdots,\ U_m\}.$$

This fuzzy relation comes from experience etc. Now assume it is known as Σ. That is, Σ is a fuzzy relation on $\prod_{i=1}^p Y_i \times \prod_{j=1}^m U_j$. Then for each element in this product space we have its membership degree as

$$\begin{aligned} &\mu_\Sigma\left(y^1_{\xi_1}, \cdots, y^p_{\xi_p}, u^1_{\eta_1}, \cdots, u^m_{\eta_m}\right) := \gamma^{\xi_1\cdots\xi_p}_{\eta_1\cdots\eta_m}, \\ &\quad \xi_i = 1, \cdots, \alpha_i,\ i = 1, \cdots, p;\ \eta_j = 1, \cdots, \beta_j,\ j = 1, \cdots, m. \end{aligned} \tag{10.23}$$

Arranging

$$\left\{ \gamma_{\eta_1\cdots\eta_m}^{\xi_1\cdots\xi_p} \,\middle|\, \xi_i = 1,\cdots,\alpha_i,\ i=1,\cdots,p;\ \eta_j = 1,\cdots,\beta_j,\ j=1,\cdots,m \right\}$$

into a relational matrix in the order of $\mathrm{id}(\eta_1,\cdots,\eta_m;\beta_1,\cdots,\beta_m) \times \mathrm{id}(\xi_1,\cdots,\xi_p;\,\alpha_1,\cdots,\alpha_p)$, we have

$$M_\Sigma = \begin{bmatrix} \gamma_{1\cdots11}^{1\cdots11} & \gamma_{1\cdots11}^{1\cdots12} & \cdots & \gamma_{1\cdots11}^{1\cdots1\alpha_p} & \cdots & \gamma_{1\cdots11}^{\alpha_1\cdots\alpha_p} \\ \gamma_{1\cdots12}^{1\cdots11} & \gamma_{1\cdots12}^{1\cdots12} & \cdots & \gamma_{1\cdots12}^{1\cdots1\alpha_p} & \cdots & \gamma_{1\cdots12}^{\alpha_1\cdots\alpha_p} \\ \vdots & & & & & \\ \gamma_{\beta_1\cdots\beta_m}^{1\cdots11} & \gamma_{\beta_1\cdots\beta_m}^{1\cdots12} & \cdots & \gamma_{\beta_1\cdots\beta_m}^{1\cdots1\alpha_p} & \cdots & \gamma_{\beta_1\cdots\beta_m}^{\alpha_1\cdots\alpha_p} \end{bmatrix}. \tag{10.24}$$

As aforementioned that a fuzzy controller is essentially a fuzzy relation among the outputs and the controls. As the fuzzy relation is expressed by a matrix, we are convinced that a fuzzy controller is essentially a relational matrix, which is expressed as (10.24).

Now given a set of $y_i \in Y_i$, $i = 1,\cdots,p$. Recall that in dual fuzzy structure they are fuzzy sets and fuzzufication provides their vector forms as $\mathcal{V}_{y_i}$, $i = 1,\cdots,p$. Then the fuzzy controller produces the corresponding controls u_j, $j = 1,\cdots,m$ as

$$\ltimes_{j=1}^{m} \mathcal{V}_{u_j} = M_\Sigma \ltimes_{i=1}^{p} \mathcal{V}_{y_i}. \tag{10.25}$$

We give the following example to depict it.

Example 10.6. Recall Example 10.5. Assume the fuzzification of both outputs A and B is ruled by Fig. 10.2. Moreover, assume $A = -3$ and $B = 7$. Recall (10.17), we have

$$\mathcal{V}_A = \begin{bmatrix} 0 \\ 0 \\ 0.75 \\ 0.25 \\ 0 \\ 0 \\ 0 \end{bmatrix}; \quad \mathcal{V}_B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0.25 \\ 0.75 \\ 0 \end{bmatrix}.$$

Using (10.3), then the corresponding control is

$$\begin{aligned} \mathcal{V}_u &= M_\Sigma \,(\times)\, \mathcal{V}_A \,(\times)\, \mathcal{V}_B \\ &= 0.25\,(\times)\,\mathrm{Col}_{19}(M_\Sigma)\,(+)\,0.75\,(\times)\,\mathrm{Col}_{20}(M_\Sigma)\,(+) \\ &\quad\ 0.25\,(\times)\,\mathrm{Col}_{26}(M_\Sigma)\,(+)\,0.25\,(\times)\,\mathrm{Col}_{27}(M_\Sigma) \\ &= 0.25\,(\times)\,\delta_{13}^{7}\,(+)\,0.75\,(\times)\,\delta_{13}^{8}\,(+)\,0.25\,(\times)\,\delta_{13}^{8}\,(+)\,0.25\,(\times)\,\delta_{13}^{9} \\ &= [0,0,0,0,0,0,0.25,0.75,0.25,0,0,0,0]^T. \end{aligned} \tag{10.26}$$

10.2.3 *Defuzzification*

From last subsection one sees that the engine of a fuzzy controller is a fuzzy relation, which provides a fuzzy mapping from outputs to controls. So we need two auxiliary works: (i) fuzzification, which converts the outputs into fuzzy sets; (ii) defuzzification, which converts the result obtained from the fuzzy mapping — which are fuzzy sets — back to control values.

Defuzzification is one of the key issues in fuzzy control. Many efforts have been put on it and various techniques have been proposed (Runkler, 1996; Saade and Diab, 2000; Soriano *et al.*, 2005).

There are several standard methods to do the defuzzification. In the following, our method is based on "Weighted Average" method (Zhu, 2005), but our modification makes it possible to convert the product fuzzy set (equivalently, a fuzzy relation over multiple u_i) back to separated u_i, $i = 1, \cdots, m$.

Since the interesting case is $m > 1$, we first show what is the modified Weighted Average we proposed.

What we obtained from the fuzzy mapping is

$$\mathcal{V}_u := \mathcal{V}_{u_1} \, (\ltimes) \cdots (\ltimes) \, \mathcal{V}_{u_m}. \tag{10.27}$$

The purpose of defuzzification is to provide controls $(u_1, \cdots, u_m)$ from a fuzzy set $u \in \mathcal{F}(U_1 \times \cdots \times U_m)$. We propose two methods to deal with defuzzification of multiple control case.

Method 1: Joined Defuzzification (JD)

This method defuzzificates $u = (u_1, \cdots, u_m)$ simultaneously. We first use a simple example to depict it.

Example 10.7. Assume there are two controls u_1 and u_2 with $u_1 \in [-4, 4]$, and $u_2 \in [-6, 6]$. Moveover, their degree-based fuzzy sets are depicted in Fig. 10.3.

For u_1 we identify

$$\begin{aligned} &V_1 := NB \sim \delta_5^1;\ V_2 := NS \sim \delta_5^2;\ V_3 := ZO \sim \delta_5^3;\\ &V_4 := PS \sim \delta_5^4;\ V_5 := PB \sim \delta_5^5. \end{aligned}$$

For u_2 we identify

$$\begin{aligned} &W_1 = NB \sim \delta_7^1; W_2 := NM \sim \delta_7^2; W_3 := NS \sim \delta_7^3; W_4 := ZO \sim \delta_7^4;\\ &W_5 := PS \sim \delta_7^5; W_6 := PM \sim \delta_7^6;\ W_7 := PB \sim \delta_7^7. \end{aligned}$$

Then we have

$$\mu_{V_i \times W_j}(u_1, u_2) = \mu_{V_i}(u_1) \wedge \mu_{W_j}(u_2), \quad i = 1, \cdots, 5;\ j = 1, \cdots, 7. \tag{10.28}$$

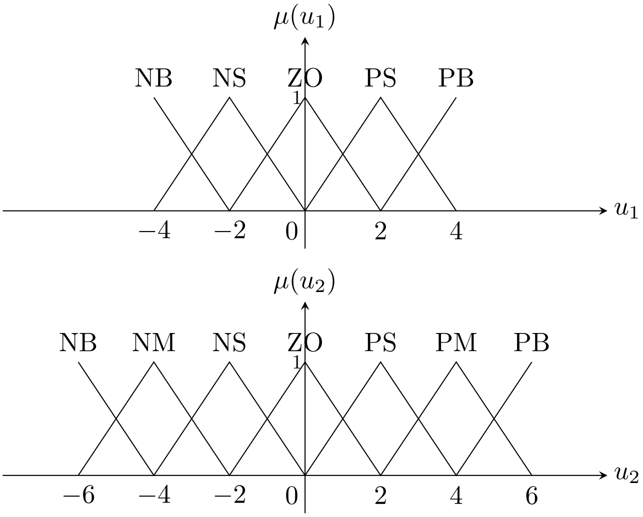


Fig. 10.3 Degree-based fuzzy sets of u_1 and u_2

Note that in this example we have unique

$$\mu^{-1}_{V_i\times W_j}(1), \quad i = 1, \cdots, 5;\ j = 1, \cdots, 7. \tag{10.29}$$

For instance,

$$\mu^{-1}_{V_1\times W_1}(1) = (-4,-6),\ \mu^{-1}_{V_1\times W_2}(1) = (-4,-4),\ \cdots,\ \mu^{-1}_{V_5\times W_7}(1) = (4,6). \tag{10.30}$$

Then we may choose $\mu^{-1}_{V_i\times W_j}(1)$ as the defuzzificated value of $\delta_5^i \ltimes \delta_7^j$.

Denote the fuzzy values obtained from (10.27) as

$$\mathcal{V}_u = [b_{11}\ \cdots\ b_{17}\ b_{21}\ \cdots\ b_{27}\ \cdots\ b_{51}\ \cdots\ b_{57}]^T. \tag{10.31}$$

Then we take the weighted values as the defuzzificated controls. That is,

$$(u_1,u_2) = \sum_{i=1}^{5}\sum_{j=1}^{7}\left(\frac{b_{i,j}}{\sum\limits_{i=1}^{5}\sum\limits_{j=1}^{7} b_{ij}}\right)\mu^{-1}_{V_i\times W_j}(1). \tag{10.32}$$

We can extend the procedure proposed in Example 10.7 to general case. Assume the degree-based fuzzy sets for controls are of the forms as isosceles triangle or isosceles trapezoid. Then

$$\mu^{-1}_{U^1_{i_1}\times\cdots\times U^m_{i_m}}(1), \quad i_s = 1, \cdots, \beta_s;\ s = 1, \cdots, m \tag{10.33}$$

are either a point or a segment. Then we can take the average $\overline{\mu^{-1}_{U^1_{i_1}\times\cdots\times U^m_{i_m}}(1)}$ as the defuzzification of $\delta^{i_1}_{\beta_1}\times\delta^{i_2}_{\beta_2}\times\cdots\times\delta^{i_m}_{\beta_m}$. (Note that in general, we may replace 1 by a properly chosen $0 \ll \varepsilon < 1$ to avoid possible empty set or improve the approximation, etc.)

Assume the fuzzy values obtained from (10.27) are

$$\mathcal{V}_u = [b_{1\cdots 11}\ \cdots b_{1\cdots 1\beta_m}\cdots b_{1\beta_2\cdots\beta_m}\ \cdots b_{\beta_1\beta_2\cdots\beta_m}]^T. \tag{10.34}$$

Using this modified weighted average, the defuzzificated controls are chosen as

$$(u_1,\cdots,u_m) = \sum_{j_1=1}^{\beta_1}\cdots\sum_{j_m=1}^{\beta_m}\left(\frac{b_{j_1\cdots j_m}}{\sum\limits_{i_1=1}^{\beta_1}\cdots\sum\limits_{i_m=1}^{\beta_m} b_{i_1\cdots i_m}}\right)\overline{\mu^{-1}_{U^1_{j_1}\times\cdots\times U^m_{j_m}}(1)}. \tag{10.35}$$

Method 2: Separated-Defuzzification (SD)

In this method we first calculate the fuzzy values $\mathcal{V}_{u_i}$, $i=1,\cdots,m$ from $\mathcal{V}_u$, and then defuzzificate each u_i from its $\mathcal{V}_{u_i}$.

We still use Example 10.7 to depict it. Assume the fuzzy values of $\mathcal{V}_u$ is obtained as in (10.31). Then we calculate

$$b^1_i = (+)^7_{j=1}\, b_{ij}, \quad i=1,\cdots,5. \tag{10.36}$$

Then we have

$$\mathcal{V}_{u_1} = (b^1_1,\ b^1_2,\ b^1_3,\ b^1_4,\ b^1_5)^T. \tag{10.37}$$

Using it, we can defuzzificate u_1.

Similarly, we have

$$b^2_j = (+)^5_{i=1}\, b_{ij}, \quad j=1,\cdots,7. \tag{10.38}$$

Then we have

$$\mathcal{V}_{u_2} = (b^2_1,\ b^2_2,\ \cdots,\ b^2_7)^T. \tag{10.39}$$

Using it, we can defuzzificate u_2.

In general assume we have fuzzy values as (10.34). Then we calculate

$$b^\alpha_k = (+)^{\beta_1}_{i_1=1}\cdots(+)^{\beta_{\alpha-1}}_{i_{\alpha-1}=1}(+)^{\beta_{\alpha+1}}_{i_{\alpha+1}=1}\cdots(+)^{\beta_m}_{i_m=1}\, b_{i_1\cdots i_{\alpha-1}ki_{\alpha+1}\cdots i_m}, \qquad k=1,\cdots,\beta_\alpha. \tag{10.40}$$

It follows that

$$\mathcal{V}_{u_\alpha} = (b_1^\alpha,\ b_2^\alpha,\ \cdots,\ b_{\beta_\alpha}^\alpha)^T, \quad \alpha = 1, \cdots, m. \tag{10.41}$$

Finally, we have

$$u_\alpha = \sum_{j=1}^{\beta_\alpha} \left(\frac{b_j^\alpha}{\sum\limits_{j=1}^{\beta_\alpha} b_j^\alpha} \right) \mu_{U_j^\alpha}^{-1}(1), \quad \alpha = 1, \cdots, m. \tag{10.42}$$

Note that when $m = 1$ Methods 1 and 2 provide the same solution.

Example 10.8. Consider the control system discussed in Examples 10.5 and 10.6. Using (10.20), we have $\delta_{13}^7 \sim 0$, $\delta_{13}^8 \sim 0.1$, $\delta_{13}^9 \sim 0.2$. Assume the degree-based fuzzy sets for control have the form as isosceles triangle. Then we have

$$\mu_{u_7}^{-1}(1) = 0, \quad \mu_{u_8}^{-1}(1) = 0.1, \quad \mu_{u_9}^{-1}(1) = 0.2.$$

Using defuzzification formula (10.32), we have the control

$$u = \frac{0.25 \times 0 + 0.75 \times 0.1 + 0.25 \times 0.2}{0.25 + 0.75 + 0.25} = 0.1.$$

Alternatively, using formula (10.42), same result can be obtained.

Method 1 (JD) is particularly suitable for the case when the controls are coupled, i.e., interacted to each other. While Method 2 (SD) is particularly suitable for the case when the controls are independent to each other. Experience is necessary for judging whether the controls are coupled.

10.3 Numerical Solution for Fuzzy Control Design

In this section we use two numerical examples to show how to design fuzzy control for systems with multiple fuzzy relations.

Example 10.9. Recall Example 10.5. Assume in addition to control u, we have control v with its rule table in Table 10.2.

CD Method: Using the traditional design (CD) method (Lee, 2005), u and v are treated independently. Hence, similar to (10.21), we have the fuzzy relation for y_1, y_2, v as

Table 10.2 Rule table for v

$A\backslash U\backslash B$	NB	NM	NS	ZO	PS	PM	PB
NB	-1	-1	-1	-1	-1	-1	-1
NM	-1	-1	-1	-1	-1	-1	1
NS	-1	-1	-1	-1	-1	1	1
ZO	-1	-1	-1	-1	1	1	1
PS	-1	-1	-1	1	1	1	1
PM	-1	-1	1	1	1	1	1
PB	-1	1	1	1	1	1	1

$$\begin{aligned} M_{\Sigma'} = \delta_2[\,&1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 2\\ &1\ 1\ 1\ 1\ 1\ 2\ 2\ 1\ 1\ 1\ 1\ 2\ 2\ 2\\ &1\ 1\ 1\ 2\ 2\ 2\ 2\ 1\ 1\ 2\ 2\ 2\ 2\ 2\\ &1\ 2\ 2\ 2\ 2\ 2\ 2] \in \mathcal{B}_2^{2\times 49}. \end{aligned} \tag{10.43}$$

Taking $y_1 = A = -3, y_2 = B = 7$, similar to Example 10.6, we have

$$\mathcal{V}_v = M_{\Sigma'}\,(\times)\,\mathcal{V}_A\,(\times)\,\mathcal{V}_B = [0.25, 0.75]^T. \tag{10.44}$$

After defuzzification, the control v is obtained as

$$v = \frac{0.25 \times (-1) + 0.75 \times 1}{0.25 + 0.75} = 0.5.$$

Combining it with the result obtained in Example 10.5, we have $u = 0.1$ and $v = 0.5$.

JD Method: Assume u and v are strongly coupled. Then we use JD method.

According to Definition 10.7, the fuzzy relation for y_1, y_2, u, v is obtained as

$$\begin{aligned} M &= M_\Sigma *_{\mathcal{B}} M_{\Sigma'}\\ &= \delta_{26}[1\ 3\ 5\ 7\ 9\ 11\ 13\ 3\ 5\ 7\ 9\ 11\ 13\ 16\\ &\qquad 5\ 7\ 9\ 11\ 13\ 16\ 18\ 7\ 9\ 11\ 13\ 16\ 18\ 20\\ &\qquad 9\ 11\ 13\ 16\ 18\ 20\ 22\ 11\ 13\ 16\ 18\ 20\ 22\ 24\\ &\qquad 13\ 16\ 18\ 20\ 22\ 24\ 26] \in \mathcal{B}_2^{26\times 49}. \end{aligned} \tag{10.45}$$

Using JD method and assume $y_1 = A = -3, y_2 = B = 7$, we can derive that

$$\begin{aligned} \mathcal{V}_u\,(\times)\,\mathcal{V}_v &= M\,(\times)\,\mathcal{V}_A\,(\times)\,\mathcal{V}_B\\ &= 0.25\,(\times)\,\mathrm{Col}_{19}(M)\,(+)\,0.75\,(\times)\,\mathrm{Col}_{20}(M)\,(+)\\ &\quad\ 0.25\,(\times)\,\mathrm{Col}_{26}(M)\,(+)\,0.25\,(\times)\,\mathrm{Col}_{27}(M)\\ &= 0.25\,(\times)\,\delta_{26}^{13}\,(+)\,0.75\,(\times)\,\delta_{26}^{16}\,(+)\,0.25\,(\times)\,\delta_{26}^{16}\,(+)\,0.25\,(\times)\,\delta_{26}^{18}\\ &= [\underbrace{0, 0, \cdots, 0}_{12}, 0.25, 0, 0, 0.75, 0, 0.25, \underbrace{0, 0, \cdots, 0}_{8}]^T. \end{aligned} \tag{10.46}$$

Using defuzzification formula (10.32), the control (u, v) is obtained as

$$(u,v)=\frac{0.25\times(0,-1)+0.75\times(0.1,1)+0.25\times(0.2,1)}{0.25+0.75+0.25}=(0.1,0.6).$$

SD Method: Assume u and v are likely independent. Then we use S-D method. That is, decompose $\mathcal{V}_u$ and $\mathcal{V}_v$ before defuzzification. From (10.45), we have

$$\begin{aligned}\mathcal{V}_u\,(\times)\,\mathcal{V}_v &= 0.25\,(\times)\,\delta_{26}^{13}\,(+)\,0.75\,(\times)\,\delta_{26}^{16}\,(+)\,0.25\,(\times)\,\delta_{26}^{18}\\ &= 0.25\,(\times)(\delta_{13}^{7}\,(\times)\,\delta_2^1)\,(+)\,0.75\,(\times)(\delta_{13}^{8}\,(\times)\,\delta_2^2)\\ &\quad(+)\,0.25\,(\times)(\delta_{13}^{9}\,(\times)\,\delta_2^2).\end{aligned}\tag{10.47}$$

Thus we get

$$\begin{aligned}\mathcal{V}_u &= 0.25\,(\times)\,\delta_{13}^{7}\,(+)\,0.75\,(\times)\,\delta_{13}^{8}\,(+)\,0.25\,(\times)\,\delta_{13}^{9}\\ &= [0,0,0,0,0,0,0.25,0.75,0.25,0,0,0,0]^T,\end{aligned}$$

and

$$\mathcal{V}_v = 0.25\,(\times)\,\delta_2^1\,(+)\,0.75\,(\times)\,\delta_2^2\,(+)\,0.25\,(\times)\,\delta_2^2=[0.25,0.75]^T,$$

which are same as those in (10.26) and (10.44). After defuzzification, the control $(u, v) = (0.1, 0.5)$ is obtained, which is just same as what we have from CD Method.

In previous example, as well as in traditional fuzzy control design, it is assumed that the experience provides some rules as (Lee, 2005)

$$\text{If } y_1=*,\cdots,y_p=*,\ \text{then } u_1=*,\cdots,u_m=*.\tag{10.48}$$

We should say that this is a special case, where the interactions among u_i, $i = 1, \cdots, m$ are ignored. General rules must be

$$\text{If } y_1=*,\cdots,y_p=*,\ \text{then } g(u_1,\cdots,u_m)=a,\tag{10.49}$$

where g is a logical function. When such rules are considered, the traditional design method can not be applicable. It can only be solved by the methods proposed in previous section. We give an example to demonstrate it.

Example 10.10. Recall Examples 10.5 and 10.9.

Assume the fuzzy relation for y_1, y_2, u, v is $\tilde{M} = M + M_1 + M_2$ where M takes same vales as in Example 10.9, and

$$\begin{aligned}M_1 = 0.8\,(\times)\,\delta_{26}[&2\ \ 4\ \ 6\ \ 8\ \ 10\ \ 12\ \ 14\ \ 4\ \ 6\ \ 8\ \ 10\ \ 12\ \ 14\ \ 17\\ &6\ \ 8\ \ 10\ \ 12\ \ 14\ \ 17\ \ 19\ \ 8\ \ 10\ \ 12\ \ 14\ \ 17\ \ 19\ \ 21\\ &10\ \ 12\ \ 14\ \ 17\ \ 19\ \ 21\ \ 23\ \ 12\ \ 14\ \ 17\ \ 19\ \ 21\ \ 23\ \ 25\\ &14\ \ 17\ \ 19\ \ 21\ \ 23\ \ 25\ \ 26]\in\mathcal{D}_2^{26\times 49},\end{aligned}\tag{10.50}$$

and

$$\begin{aligned} M_2 = 0.5\,(\times)\,\delta_{26}[&3\ \ 5\ \ 7\ \ 9\ \ 11\ \ 13\ \ 15\ \ 5\ \ 7\ \ 9\ \ 11\ \ 13\ \ 15\ \ 18 \\ &7\ \ 9\ \ 11\ \ 13\ \ 15\ \ 18\ \ 20\ \ 9\ \ 11\ \ 13\ \ 15\ \ 18\ \ 20\ \ 22 \\ &11\ \ 13\ \ 15\ \ 18\ \ 20\ \ 22\ \ 24\ \ 13\ \ 15\ \ 18\ \ 20\ \ 22\ \ 24\ \ 26 \\ &15\ \ 18\ \ 20\ \ 22\ \ 24\ \ 26\ \ 26] \in \mathcal{D}_2^{26\times 49}. \end{aligned} \tag{10.51}$$

It is easy to check that there do not exist matrices $M_3 \in \mathcal{M}_{13\times 49}$ and $M_4 \in \mathcal{M}_{2\times 49}$ such that $\tilde{M} = M_3 * M_4$. That is, the control rules cannot be separated with respect to each controls. Hence, the CD method is not applicable.

Still taking $y_1 = A = -3, y_2 = B = 7$, we use two methods to defuzzificate the controls.

JD Method: Using the method developed in previous section, we have

$$\begin{aligned} &\mathcal{V}_u\,(\times)\,\mathcal{V}_v = \tilde{M}\,(\times)\,\mathcal{V}_A\,(\times)\,\mathcal{V}_B \\ &\quad = 0.25\,(\times)\,\mathrm{Col}_{19}(\tilde{M})\,(+)\,0.75\,(\times)\,\mathrm{Col}_{20}(\tilde{M}) \\ &\quad\quad (+)\,0.25\,(\times)\,\mathrm{Col}_{26}(\tilde{M})\,(+)\,0.25\,(\times)\,\mathrm{Col}_{27}(\tilde{M}) \\ &\quad = 0.25\,(\times)\,\delta_{26}^{13}\,(+)\,0.25\,(\times)\,0.8\,(\times)\,\delta_{26}^{14}\,(+)\,0.25\,(\times)\,0.5\,(\times)\,\delta_{26}^{15} \\ &\quad\quad (+)\,0.75\,(\times)\,\delta_{26}^{16}\,(+)\,0.75\,(\times)\,0.8\,(\times)\,\delta_{26}^{17}\,(+)\,0.75\,(\times)\,0.5\,(\times)\,\delta_{26}^{18} \\ &\quad\quad (+)\,0.25\,(\times)\,\delta_{26}^{16}\,(+)\,0.25\,(\times)\,0.8\,(\times)\,\delta_{26}^{17}\,(+)\,0.25\,(\times)\,0.5\,(\times)\,\delta_{26}^{18} \\ &\quad\quad (+)\,0.25\,(\times)\,\delta_{26}^{18}\,(+)\,0.25\,(\times)\,0.8\,(\times)\,\delta_{26}^{19}\,(+)\,0.25\,(\times)\,0.5\,(\times)\,\delta_{26}^{20} \\ &\quad = [\underbrace{0, 0, \cdots, 0}_{12}, 0.25, 0.25, 0.25, 0.75, 0.75, 0.5, 0.25, 0.25, \underbrace{0, 0, \cdots, 0}_{6}]^T. \end{aligned}$$

Via defuzzification formula (10.32), the control (u, v) is obtained that is

$$\begin{aligned} (u, v) = &[0.25 \times (0, -1) + 0.25 \times (0, 1) + 0.25 \times (0.1, -1) + \\ &0.75 \times (0.1, 1) + 0.75 \times (0.2, -1) + 0.5 \times (0.2, 1) + \\ &0.25 \times (0.3, -1) + 0.25 \times (0.3, 1)]/ \\ &(0.25 + 0.25 + 0.25 + 0.75 + 0.75 + 0.5 + 0.25 + 0.25) \\ = &(\tfrac{2}{13}, \tfrac{1}{13}). \end{aligned}$$

SD Method: First, we decompose $\mathcal{V}_u$ and $\mathcal{V}_v$ before defuzzification. Similar to Example 10.9, we get

$$\begin{aligned} \mathcal{V}_u &= 0.25\,(\times)\,\delta_{13}^{7}\,(+)\,0.75\,(\times)\,\delta_{13}^{8}\,(+)\,0.75\,(\times)\,\delta_{13}^{9}\,(+)\,0.25\,(\times)\,\delta_{13}^{10} \\ &= [\underbrace{0, \cdots, 0}_{6}, 0.25, 0.75, 0.75, 0.25, 0, 0, 0]^T, \end{aligned}$$

and

$$\mathcal{V}_v = 0.75\,(\times)\,\delta_2^1\,(+)\,0.75\,(\times)\,\delta_2^2 = [0.75, 0.75]^T,$$

Using defuzzification formula (10.34), we have the control (u, v), where

$$u = \frac{0.25 \times 0 + 0.75 \times 0.1 + 0.75 \times 0.2 + +0.25 \times 0.3}{0.25 + 0.75 + 0.75 + 0.25} = 0.225,$$

and

$$v = \frac{0.75 \times (-1) + 0.75 \times 1}{0.75 + 0.75} = 0.$$

Exercises

10.1 Consider a relation R over E_i, $i = 1, \cdots, k$. How many different ways to construct the relational matrix of R, if

(i) the indexes of the rows and columns are inherited from the index of E_i;

(ii) the index orders of the rows and columns can be arbitrary assigned.

10.2 Assume $U = \{u_1, u_2, u_3\}$, $V = \{v_1, v_2\}$, $W = \{w_1, w_2, w_3\}$ are three universes of discourse, and

$$R = \begin{bmatrix} 0.5 & 0.9 & 0 \\ 1 & 0.6 & 0.3 \\ 0.7 & 0.4 & 0.5 \end{bmatrix} \in \mathcal{F}(U \times W), \quad S = \begin{bmatrix} 0.4 & 0.8 & 0.2 \\ 0.8 & 0 & 1 \end{bmatrix} \in \mathcal{F}(V \times W).$$

(i) Using R and S to construct a fuzzy relation P between UV and W. (Hint: $P = R(*)S$. Explain why?)

(ii) Using R and S to construct a fuzzy relation Q between U and V. (Hint: $Q = R(\times)S^T$. Explain why?)

10.3 Let $E = \{e_1, e_2, e_3, e_4\}$ and $R \in \mathcal{F}(\mathcal{E} \times \mathcal{E})$. Check if R is (a) identity; (b) self-related; (c) self-unrelated; (d) symmetric; (e) transitive, when the relational matrix of R is as follows.

(i)

$$M_R = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

(ii)

$$M_R = \begin{bmatrix} 1 & 0.5 & 0 & 0 \\ 0 & 1 & 0.5 & 0 \\ 0 & 0 & 1 & 0.5 \\ 0.5 & 0 & 0 & 1 \end{bmatrix}.$$

(iii)

$$M_R = \begin{bmatrix} 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

(iv)

$$M_R = \begin{bmatrix} 0.2 & 0.2 & 0 & 0 \\ 0.2 & 0.2 & 0 & 0 \\ 0 & 0 & 0.3 & 0.3 \\ 0 & 0 & 0.3 & 0.3 \end{bmatrix}.$$

10.4 Let $U = \{u_1, u_2, u_3\}$, $V = \{v_1, v_2, v_3\}$, $W = \{w_1, w_2\}$, and a fuzzy relation $R \in \mathcal{F}(U \times V \times W)$ be

$$R_{U\times VW} = \begin{bmatrix} 0.2 & 0.8 & 0.5 & 0.75 & 0 & 1 \\ 0 & 0.5 & 0.4 & 0.7 & 0.3 & 0 \\ 0.5 & 0.2 & 0.7 & 0.2 & 0.1 & 0.9 \end{bmatrix}.$$

Moreover, assume $A \in \mathcal{F}(V)$ with $\mathcal{V}_A = (0.5, 0.3, 0.8)^T$, and $B \in \mathcal{F}(W)$ with $\mathcal{V}_B = (0.2, 1)^T$ are given.

(i) Using A and B to find a $C \in \mathcal{F}(U)$.

(ii) Using B to get a fuzzy relation $S_{U\times V} \in \mathcal{F}(U \times V)$.

(iii) Using A and $S_{U\times V}$ to get a $C' \in \mathcal{F}(U)$. Check that $C' = C$. Why?

10.5 Assume $U = \{u_1, u_2\}$, $V = \{v_1, v_2, v_3\}$, $W = \{w_1, w_2, w_3\}$, and $H = \{h_1, h_2\}$ are four universes of discourse, and a fuzzy relation $R \in \mathcal{F}(U \times V \times W \times H)$ is expressed as

$$R_{UV\times WH} = \begin{bmatrix} 0.8 & 0.2 & 0.5 & 0.3 & 1 & 1 \\ 0 & 0 & 0.45 & 0.6 & 0.3 & 0.7 \\ 0.4 & 0.4 & 0.8 & 0.7 & 0.9 & 0 \\ 1 & 0 & 1 & 0.7 & 0.8 & 0 \\ 0.4 & 0.5 & 0.3 & 0.6 & 0.7 & 1 \\ 0.4 & 0.2 & 0.3 & 0.2 & 0.1 & 0 \end{bmatrix}.$$

(i) Rewrite R as (a) $R_{U\times VWH}$; (b) $R_{UW\times VH}$, (c) $R_{UWH\times V}$.

(ii) Using $B \in \mathcal{F}(V)$ with $\mathcal{V}_B = (0.3, 0.5, 0.8)^T$, $C \in \mathcal{F}(W)$ with $\mathcal{V}_C = (0.8, 0.4, 0.2)^T$, and $D \in \mathcal{F}(H)$ with $\mathcal{V}_D = (0.6, 0.3)^T$ to find a fuzzy set $A \in \mathcal{F}(U)$ using R.

(iii) Using $E \in \mathcal{F}(V)$ with $\mathcal{V}_E = (0.5, 0.2, 0.9)^T$, and $F \in \mathcal{F}(W)$ with $\mathcal{V}_C = (0.1, 0.8, 0.1)^T$ to find a fuzzy relation $S \in \mathcal{F}(U \times H)$, using R.

10.6 Consider Example 10.4. Express the following e as a vector $\mathcal{V}_e$ with respect to the dual fuzzy structure. (i) $e = -5$; (ii) $e = 1$; (iii) $e = 3.25$.

10.7 For fuzzy relations $S \in \mathcal{F}(X \times Z)$, $T \in \mathcal{F}(Y \times Z)$, $S * T$ is defined in (10.10).

(i) Prove that

$$\mu_{T*S}(x, y, z) = \mu_T(x, z) \wedge \mu_S(y, z).$$

(ii) We can also define another jointed composition $T \star S$ as

$$\mu_{T\star S}(x, y, z) = \mu_T(x, z) \vee \mu_S(y, z).$$

Prove that

$$(\mathbf{1} - M_T) * (\mathbf{1} - M_S) = (\mathbf{1} - M_S \star M_T),$$

where **1** is the matrix with all entries being 1.

10.8 Assume the fuzzy matrices of the fuzzy relations $A \in \mathcal{F}(Y \times X)$, $B \in \mathcal{F}(Z \times X)$, $C \in \mathcal{F}(Z \times W)$ are, respectively,

$$M_A = \begin{bmatrix} 1 & 0 \\ 0.6 & 1 \\ 0 & 0.4 \end{bmatrix}, \quad M_B = \begin{bmatrix} 0.5 & 1 \\ 0.5 & 0 \end{bmatrix}, \quad M_C = \begin{bmatrix} 0 & 0.6 \\ 0.4 & 1 \end{bmatrix}.$$

Find a fuzzy relation $X \in \mathcal{F}(Y \times W)$ satisfying

$$M_{(A*B)\circ(X*D)} = \begin{bmatrix} 0.5 & 0.4 \\ 1 & 0 \end{bmatrix}.$$

10.9 Assume the degree-based fuzzy sets of variables e_1, e_2, u are described as Fig 10.4, and the rule table for fuzzificated E_1, E_2, and U is as Table 10.3.

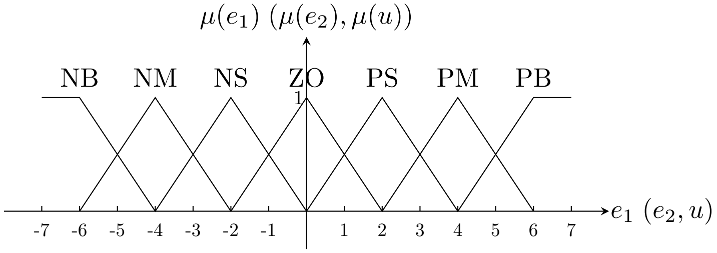


Fig. 10.4 The membership functions of fuzzy set

Table 10.3 Rule table for Exercise 10.9

$E_1\backslash U\backslash E_2$	NB	NM	NS	ZO	PS	PM	PB
NB	PB	PB	PB	PB	PM	PS	ZO
NM	PB	PB	PM	PM	PS	ZO	ZO
NS	PB	PM	PM	PS	ZO	ZO	NS
ZO	PM	PS	PS	ZO	NS	NS	NM
PS	PS	ZO	ZO	NS	NM	NM	NB
PM	ZO	ZO	NS	NM	NM	NB	NB
PB	ZO	NS	NM	NB	NB	NB	NB

(i) For $e_1 = 5.2$, $e_2 = -1.3$, find a control u.

(ii) Add a new control $v \in [-1, 1]$ whose fuzzy set can be described as

$$\mu_N(v) = 0.5 + 0.5v, \quad \mu_P(v) = 0.5 - 0.5v.$$

Set $N \sim \delta_2^1, P \sim \delta_2^2$, and assume the fuzzy matrix of fuzzy control rule for v is

$$\begin{array}{rl} M_{\Sigma'} = 0.5 \times \delta_2[& 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 2 \\ & 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 2\ \ 2\ \ 1\ \ 1\ \ 1\ \ 1\ \ 2\ \ 2\ \ 2 \\ & 1\ \ 1\ \ 1\ \ 2\ \ 2\ \ 2\ \ 2\ \ 1\ \ 1\ \ 2\ \ 2\ \ 2\ \ 2\ \ 2 \\ & 1\ \ 2\ \ 2\ \ 2\ \ 2\ \ 2\ \ 2]. \end{array}$$

Denoting the fuzzy matrix of fuzzy control u by M_Σ, then we construct a control rule as

$$M = (M_\Sigma * M_{\Sigma'})\,(\times)\,M_1,$$

where $M_1 = (m_{i,j}) \in \mathcal{D}_2^{49\times 49}$, and

$$m_{i,j} = \begin{cases} 1, & j > i \\ 0.5, & j = i \\ 0, & j < i. \end{cases}$$

Also for $e_1 = 5.2$, $e_2 = -1.3$, find controls u and v using JD and SD method respectively.

This page intentionally left blank

Chapter 11

Representation of Boolean Functions

Logical functions (or Boolean functions) and the Boolean field are very important in many applications, including computer science (Wolfram, 1986), cryptography (Carlet, 2010), and many other fields. In this chapter we first introduce different expressions of Boolean functions, including polynomial form, structure matrix expression, Walsh transformation. Then the conversions among different forms are considered. In addition to expressions, some fundamental properties, such as linearity and nonlinearity and symmetry, are investigated. This chapter is based on Wen *et al.* (2000) and Zhao *et al.* (2010a).

11.1 Boolean Functions in Galois Field $\mathbb{Z}_2$

Let p be a prime number and denote by

$$\mathbb{Z}_p = \{0, 1, \cdots, p-1\}.$$

Define the addition $\langle + \rangle$ and the multiplication $\langle \times \rangle$ over $\mathbb{Z}_p$ as

$$\begin{cases} a \langle + \rangle b := a + b \pmod p \\ a \langle \times \rangle b := ab \pmod p. \end{cases} \tag{11.1}$$

Then $(\mathbb{Z}_p, \langle + \rangle, \langle \times \rangle)$ becomes a field, which is called a Galois field. Note that when $p = 2$ we have $\mathbb{Z}_2 = \mathcal{D}$. For statement ease, we simply call $\mathbb{Z}_p$ the Galois field. Throughout this chapter we assume $p = 2$ we call $\mathbb{Z}_2$ a Boolean field. In this case, instead of $\mathcal{D}$, we use $\mathbb{Z}_2$, which means the operators on $\mathcal{D}$ are $\langle + \rangle$ and $\langle \times \rangle$.

It is obvious that in $\mathbb{Z}_2$ $\langle + \rangle$ and $\langle \times \rangle$ are two logical operators. In fact, we have

$$\langle + \rangle = \bar{\vee}, \quad \langle \times \rangle = \wedge.$$

Now a natural question is: Is $\{\langle + \rangle, \langle \times \rangle\}$ an adequate set of logical operators? The answer is: "Yes". Because

$$\neg x = 1 \langle + \rangle x,$$

and it is well known that $\{\neg, \wedge\}$ is an adequate set (refer to Proposition 7.1). It follows that any Boolean function can be expressed via $\langle + \rangle$ and $\langle \times \rangle$. Throughout this chapter a logical function is called a Boolean function. For the sake of compactness, we simply denote

$$\begin{cases} a \langle + \rangle b = a + b \\ a \langle \times \rangle b = ab, \quad a, b \in \mathbb{Z}_2. \end{cases}$$

Consider an element in $x = \{x_1, \cdots, x_n\} \in \mathbb{Z}_2^n$. We propose the following three ways to express the ordered form of x.

(i) Vector Form:

$$V_x = (x_1, x_2, \cdots, x_n), \quad x_i \in \mathbb{Z}_2, \; i = 1, \cdots, n. \tag{11.2}$$

(ii) Scalar Form: Consider $x = \{x_1, x_2, \cdots, x_n\}$ as a binary number $x_1 x_2 \cdots x_n$. Converting it into a decimal form, we have a number as

$$\chi_x = x_1 2^{n-1} + x_2 2^{n-2} + \cdots + x_n, \tag{11.3}$$

where $0 \leq \chi_x \leq 2^n - 1$.

(iii) STP Form: Identify $1 \sim \delta_2^1$ and $0 \sim \delta_2^2$, then $x_i \in \Delta$ and we set

$$x := \ltimes_{i=1}^n x_i \in \Delta_{2^n}. \tag{11.4}$$

It is obvious that these three expressions are equivalent. To convert one form to another, we need the following formulas. We leave the proof to the reader.

Proposition 11.1. *Let χ_x be a scalar form of $x \in \Delta_{2^n}$. Then*

$$x = \delta_{2^n}^{2^n - \chi_x}. \tag{11.5}$$

Equivalently, let $x = \delta_{2^n}^t$. Then

$$\chi_x = 2^n - t. \tag{11.6}$$

Using the definitions and Proposition 11.1, it is easy to convert an element in $\mathbb{Z}_2^n$ from one form to another. We give an example to show this.

Example 11.1. Let $n = 8$. Then

(i) Assume $\chi_x = 51$. Then

$$V_x = (0, 0, 1, 1, 0, 0, 1, 1); \quad x = \delta_{2^8}^{205}.$$

(ii) Assume $V_x = (1, 1, 0, 0, 1, 0, 1, 0)$. Then

$$\chi_x = 2^7 + 2^6 + 2^3 + 2^1 = 202; \quad x = \delta_{2^8}^{54}.$$

(iii) Assume $x = \delta_{2^8}^{120}$. Then

$$\chi_x = 2^8 - 120 = 136; \quad V_x = (1, 0, 0, 0, 1, 0, 0, 0).$$

Let $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$ be a logical mapping. It is well known that there exists a matrix $M_f \in \mathcal{L}_{2\times 2^n}$, called the structure matrix of f, such that in vector form f can be expressed as

$$y := f(x_1, \cdots, x_n) = M_f \ltimes_{i=1}^n x_i, \quad x_i \in \Delta. \tag{11.7}$$

When $(x_1, \cdots, x_n)$ are expressed into its scalar form, the mapping can be expressed into its vector form as

$$V_f := (f(0), f(1), \cdots, f(2^n - 1)). \tag{11.8}$$

Then it follows from the definition that

Proposition 11.2. *Denote the first row of M_f as m^f, (i.e., $m^f = \mathrm{Row}_1(M_f)$ is the truth vector of f). Then*

$$m_i^f = (V_f)_{2^n+1-i}, \quad i = 1, \cdots, 2^n. \tag{11.9}$$

Example 11.2.

(1) Let $M_f = \delta_2[1\ 1\ 2\ 1\ 2\ 1\ 1\ 2]$. Then

$$V_f = (0\ 1\ 1\ 0\ 1\ 0\ 1\ 1).$$

(2) Let $V_f = (1\ 1\ 0\ 1\ 1\ 0\ 1\ 0)$. Then

$$M_f = \delta_2[2\ 1\ 2\ 1\ 1\ 2\ 1\ 1].$$

Finally, it is worth noting that using different notations for corresponding different forms is not convenient in use. Hence, as a convention in cryptography, in the following we may not distinct the three forms of $x \in \mathbb{Z}_2^n$. That is,

$$\begin{aligned} &V_x = (x_1, \cdots, x_n)^T, \quad x_i \in \mathcal{D} \Leftrightarrow \\ &x = \ltimes_{i=1}^n x_i, \quad x_i = \delta_{2^n}^{2^n - \chi_x} \in \Delta_2 \Leftrightarrow \\ &\chi_x = x_1 2^{n-1} + x_2 2^{n-2} + \cdots + x_n = 2^n - t, \text{ while } x = \delta_{2^n}^t. \end{aligned}$$

We will only use x for its different forms. To be more specific, when the vector forms are used, we may also use $\mathbf{x}$ for V_x, and $\mathbf{f}$ for V_f.

11.2 Polynomial Form of Boolean Functions

To get polynomial expression of Boolean functions we need some new notations.

Definition 11.1.

(1) Let $x, c \in \mathbb{Z}_2^n$ be $x = \{x_1, \cdots, x_n\}$ and $c = \{c_1, \cdots, c_n\}$. Define (as a notation)

$$x_i^1 := x_i, \quad x_i^0 = \neg x_i. \tag{11.10}$$

Then

$$x_i^{c_i} = \begin{cases} 1, & x_i = c_i \\ 0, & x_i \neq c_i. \end{cases} \tag{11.11}$$

(2)

$$\mathbf{x}^{\mathbf{c}} := \prod_{i=1}^{n} x_i^{c_i} = \begin{cases} 1, & \mathbf{x} = \mathbf{c} \\ 0, & \mathbf{x} \neq \mathbf{c}. \end{cases} \tag{11.12}$$

Precisely speaking, (11.10) is a definition, (11.11) and (11.12) follow from (11.10). According to the definition, we have the following proposition.

Proposition 11.3. *Let $f : \mathbb{Z}_2^n \to \mathbb{Z}^2$. $x = \{x_1, \cdots, x_n\} \in \mathbb{Z}_2^n$. Then we have the following two expressions.*

(1) (power form)

$$f(x) = \sum_{i=0}^{2^n-1} f(i)\mathbf{x}^{\mathbf{i}}. \tag{11.13}$$

(2) (polynomial form)

$$\begin{aligned} f(x) &= a_0 + a_1x_1 + \cdots a_nx_n + a_{12}x_1x_2 + \cdots \\ &\quad + a_{n-1\ n}x_{n-1}x_n + \cdots + a_{12\cdots n}x_1x_2\cdots x_n \\ &= a_0 + \sum_{k=1}^{n} \sum_{1\leq j_1<\cdots<j_k\leq n} a_{j_1\cdots j_k}x_{j_1}\cdots x_{j_k}. \end{aligned} \tag{11.14}$$

Proof. Note that $\mathbf{x}^{\mathbf{x}} = 1$, and $\mathbf{x}^{\mathbf{i}} = 0$ for $\mathbf{i} \neq \mathbf{x}$. Then (11.13)follows from (11.12). By definition,

$$x_i^0 = \begin{cases} 1, & x_i = 0 \\ 0, & x_i = 1. \end{cases}$$

Hence

$$x_i^0 = x_i + 1.$$

Then (11.14) comes from (11.13) by replacing x_i^0 by $x_i + 1$ and multiplying all factors out. □

(11.14) is called the polynomial form of $f(x)$. $\deg(f(x))$ is defined as the degree of the polynomial form of $f(x)$. When $\deg(f(x)) = 1$ it is called an affine function. An affine function with $a_0 = 0$ is called a linear function. In literature an affine function is sometimes also called a linear function.

Example 11.3. Consider

$$f(x_1, x_2, x_3) = (x_1 \wedge x_2) \leftrightarrow x_3.$$

Then we have

$$\begin{aligned}
&f(0) = f(0,0,0) = 1,\ f(1) = f(0,0,1) = 0,\ f(2) = f(0,1,0) = 1,\\
&f(3) = f(0,1,1) = 0,\ f(4) = f(1,0,0) = 1,\ f(5) = f(1,0,1) = 0,\\
&f(6) = f(1,1,0) = 0,\ f(7) = f(1,1,1) = 1.
\end{aligned}$$

Hence the vector form of f is

$$\mathbf{f} = (1\ 0\ 1\ 0\ 1\ 0\ 0\ 1).$$

The polynomial form of f is:

$$\begin{aligned}
f(x) &= x_1^0x_2^0x_3^0 + x_1^0x_2^1x_3^0 + x_1^1x_2^0x_3^0 + x_1^1x_2^1x_3^1\\
&= (1+x_1)(1+x_2)(1+x_3) + (1+x_1)x_2(1+x_3)\\
&\quad + x_1(1+x_2)(1+x_3) + x_1x_2x_3\\
&= 1 + x_3 + x_1x_2.
\end{aligned}$$

Denote by $\mathcal{B}_n^{\mathcal{F}}$ the set of logical functions $\mathbb{Z}_2^n \to \mathbb{Z}_2$, which is a vector space over $\mathbb{Z}_2$. Denote by $\mathcal{B}_n^{\mathcal{A}}$ its affine subspace and $\mathcal{B}_n^{\mathcal{L}}$ its linear subspace.

Next, we consider the conversions between the polynomial form and the structure matrix of a Boolean function f. Since f can be expressed as

$$\begin{aligned}
f(x) &= M_f x\\
&= m_f \begin{bmatrix} x_1^1 \\ x_1^0 \end{bmatrix} \begin{bmatrix} x_2^1 \\ x_2^0 \end{bmatrix} \cdots \begin{bmatrix} x_n^1 \\ x_n^0 \end{bmatrix}\\
&= m_f \begin{bmatrix} x_1 \\ x_1 + 1 \end{bmatrix} \begin{bmatrix} x_2 \\ x_2 + 1 \end{bmatrix} \cdots \begin{bmatrix} x_n \\ x_n + 1 \end{bmatrix}\\
&= m_f \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \end{bmatrix} \cdots \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ x_n \end{bmatrix} \qquad (11.15)\\
&= m_f \left(\underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}}_{n} \right) \begin{bmatrix} 1 \\ x_1 \end{bmatrix} \begin{bmatrix} 1 \\ x_2 \end{bmatrix} \cdots \begin{bmatrix} 1 \\ x_n \end{bmatrix}\\
&:= m_f P_n \xi_n := \alpha \xi_n,
\end{aligned}$$

where the last second equality of (11.15) comes from (2.20), and

$$\alpha = m_f P_n, \tag{11.16}$$

$$P_n = \left(\underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \otimes \cdots \otimes \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}}_{n} \right), \tag{11.17}$$

$$\xi_n = \begin{bmatrix} 1 \\ x_1 \end{bmatrix} \begin{bmatrix} 1 \\ x_2 \end{bmatrix} \cdots \begin{bmatrix} 1 \\ x_n \end{bmatrix} \tag{11.18}$$

is a basis of the polynomials on $\mathbb{Z}_2^n$, we have already converted M_f to its polynomial form $\alpha\xi_n$.

Alternatively, we can express f into a natural alphabetic and power increasing form as

$$f(x) = \beta\eta_n, \tag{11.19}$$

where η_n is an alphabetic and power increasing basis as

$$\eta_n = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_n \\ x_1x_2 \\ \vdots \\ x_{n-1}x_n \\ x_1x_2x_3 \\ \vdots \\ x_{n-1}x_{n-1}x_n \\ \vdots \\ x_1x_2\cdots x_n \end{bmatrix}. \tag{11.20}$$

We would like to find the relationship between η_n and ξ_n. To this end, we may consider the position which $x_{i_1}x_{i_2}\cdots x_{i_n}$ appears in ξ_n

To be specific, let $\mu_{i_1,i_2,\cdots,i_r}$, $i_1 < i_2 \cdots < i_r$ be the position where $x_{i_1}x_{i_2}\cdots x_{i_r}$ appears in ξ_n, consider

Case 1:

$$\begin{bmatrix} 1 \\ x_{n-1} \end{bmatrix} \begin{bmatrix} 1 \\ x_n \end{bmatrix} = \begin{bmatrix} 1 \\ x_n \\ x_{n-1} \\ x_{n-1}x_n \end{bmatrix}. \tag{11.21}$$

We have $\mu_n = 2^0 + 1$ $\mu_{n-1} = 2^1 + 1$ $\mu_{n-1,n} = 2^1 + 2^0 + 1$.
Case 2:

$$\begin{bmatrix} 1 \\ x_{n-2} \end{bmatrix} \begin{bmatrix} 1 \\ x_{n-1} \end{bmatrix} \begin{bmatrix} 1 \\ x_n \end{bmatrix} = \begin{bmatrix} 1 \\ x_{n-2} \end{bmatrix} \begin{bmatrix} 1 \\ x_n \\ x_{n-1} \\ x_{n-1}x_n \end{bmatrix} = \begin{bmatrix} 1 \\ x_n \\ x_{n-1} \\ x_{n-1}x_n \\ x_{n-2} \\ x_{n-2}x_n \\ x_{n-2}x_{n-1} \\ x_{n-2}x_{n-1}x_n \end{bmatrix}. \tag{11.22}$$

$\vdots$

Case 2^n:

$$\begin{bmatrix} 1 \\ x_1 \end{bmatrix} \begin{bmatrix} 1 \\ x_2 \end{bmatrix} \cdots \begin{bmatrix} 1 \\ x_n \end{bmatrix} = \xi_n. \tag{11.23}$$

Then we have

Theorem 11.1.

$$\mu_{i_1,i_2,\cdots,i_r} = \sum_{j=1}^{r} 2^{n-i_j} + 1. \tag{11.24}$$

Proof. For any j, the position where x_{n-j} appears for the first time is $\mu_{n-j}{=}2^j + 1$

Then for any $x_{n-i_1}x_{n-i_2}\cdots x_{n-i_r}$, we can arrange x_{n-i_1}, $x_{n-i_1}x_{n-i_2}$, $x_{n-i_1}x_{n-i_2}\cdots x_{n-i_r}$ in a sequence, and in that way the conclusion follows.

□

Using Theorem 11.1, we construct Φ_n as follows

$$\Phi_n = \delta_{2^n}[1, \phi_1, \phi_2, \cdots, \phi_n], \tag{11.25}$$

where

$$\phi_r = (\mu_{1,2,\cdots,r}, \mu_{1,2,\cdots,r+1}, \cdots, \mu_{n-r+1,n-r+2,\cdots,n}), \quad r = 1, 2, \cdots, n.$$

Then it is ready to check that

$$\Phi_n^T \xi_n = \eta_n. \tag{11.26}$$

Finally, observing $f = \alpha\xi_n = \beta\eta_n$, we have

$$\begin{cases} \alpha = m_f P_n \\ \beta = \alpha\Phi_n = m_f P_n \Phi_n. \end{cases} \tag{11.27}$$

We give an example to depict the conversions.

Example 11.4. Consider a Boolean function

$$f = x_1 \wedge \neg(x_2 \to x_3).$$

The truth table of f is

$$m_f = [0\ 1\ 0\ 0\ 0\ 0\ 0\ 0].$$

By straightforward computation we have

$$P_3 = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \otimes \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \otimes \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&1&0&1 \\ 0&0&0&0&1&1&1&1 \\ 0&0&0&1&0&0&0&1 \\ 0&0&1&1&0&0&1&1 \\ 0&1&0&1&0&1&0&1 \\ 1&1&1&1&1&1&1&1 \end{bmatrix}.$$

$$\Phi_3 = \delta_8[1, \phi_1, \phi_2, \phi_3],$$

where

$$\begin{aligned} \phi_1 &= (\mu_1, \mu_2, \mu_3), \\ \phi_2 &= (\mu_{1,2}, \mu_{1,3}, \mu_{2,3}), \\ \phi_3 &= (\mu_{1,2,3}). \end{aligned}$$

By Theorem 11.1, we have

$$\Phi_3 = \delta_8[1\ 5\ 3\ 2\ 7\ 6\ 4\ 8].$$

Then using (11.27),

$$\beta = m_f P_3 \Phi_3 = [0\ 0\ 0\ 0\ 1\ 0\ 0\ 1].$$

Thus

$$f(x) = x_1 x_2 + x_1 x_2 x_3.$$

11.3 Walsh Transformation

Definition 11.2. Let $x = \{x_1, \cdots, x_n\}$, $\omega = \{\omega_1, \cdots, \omega_n\} \in \mathbb{Z}_2^n$.

(1) The inner product of x and ω is defined as

$$\mathbf{x} \cdot \boldsymbol{\omega} = x_1\omega_1 + \cdots + x_n\omega_n \in \mathbb{Z}_2. \tag{11.28}$$

(2) Given an $\boldsymbol{\omega} \in \mathbb{Z}_2^n$. Define a function $\mathbb{Z}_2^n \to \mathbb{Z}_2$ as

$$Q_\omega(x) = (-1)^{\boldsymbol{\omega}\cdot\mathbf{x}}, \tag{11.29}$$

and in vector form:

$$\mathbf{Q}_\omega = (Q_\omega(0), Q_\omega(1), \cdots, Q_\omega(2^n - 1)).$$

Lemma 11.1. *Assume* $\omega \neq 0$. *Then*

$$\sum_{x=0}^{2^n-1} (-1)^{\boldsymbol{\omega}\cdot\mathbf{x}} = 0. \tag{11.30}$$

Proof. Assume $\omega_i \neq 0$. For each $\mathbf{x} = (x_1, \cdots, x_n)$ satisfying $\boldsymbol{\omega} \cdot \mathbf{x} = 0$, we construct an $\mathbf{x}^* = (x_1, \cdots, \neg x_i, \cdots, x_n)$, which satisfies $\boldsymbol{\omega} \cdot \mathbf{x} = 1$. Since $\mathbf{x} \leftrightarrow \mathbf{x}^*$ is a one-to-one correspondence, it follows that

$$\big|\{\mathbf{x} \,|\, \boldsymbol{\omega} \cdot \mathbf{x} = 0\}\big| = \big|\{\mathbf{x} \,|\, \boldsymbol{\omega} \cdot \mathbf{x} = 1\}\big|.$$

Then (11.30) is obvious. □

Proposition 11.4. $\{\mathbf{Q}_\omega \,|\, \omega = 0, 1, \cdots, 2^n - 1\}$ *is a set of orthogonal functions. Precisely,*

$$\mathbf{Q}_\alpha \cdot \mathbf{Q}_\beta = \begin{cases} 2^n, & \boldsymbol{\alpha} = \boldsymbol{\beta} \\ 0, & \boldsymbol{\alpha} \neq \boldsymbol{\beta}. \end{cases} \tag{11.31}$$

Proof. Assume $\boldsymbol{\alpha} = \boldsymbol{\beta}$. Then

$$\begin{aligned} \mathbf{Q}_\alpha \cdot \mathbf{Q}_\beta &= \sum_{x=0}^{2^n-1} (-1)^{\boldsymbol{\alpha}\cdot\mathbf{x}}(-1)^{\boldsymbol{\beta}\cdot\mathbf{x}} \\ &= \sum_{x=0}^{2^n-1} (-1)^{2\boldsymbol{\alpha}\cdot\mathbf{x}} \\ &= \sum_{x=0}^{2^n-1} (-1)^0 = 2^n. \end{aligned}$$

Assume $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$. Since $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$, $\boldsymbol{\alpha} + \boldsymbol{\beta} \neq \mathbf{0}$. Using Lemma 11.1, we have

$$\mathbf{Q}_\alpha \cdot \mathbf{Q}_\beta = \sum_{x=0}^{2^n-1} (-1)^{(\boldsymbol{\alpha}+\boldsymbol{\beta})\cdot\mathbf{x}} = 0.$$

□

Since $\{\mathbf{Q}_\omega \,|\, \omega = 0, 1, \cdots, 2^n - 1\}$ is a set of orthogonal functions, then for any $f \in \mathcal{B}_n^{\mathcal{F}}$ its vector form can be expressed as

$$\mathbf{f} = \sum_{\omega=0}^{2^n-1} S_f(\omega)\mathbf{Q}_\omega. \tag{11.32}$$

Then $\{S_f(\omega)\}$ is called the first Walsh transformation of f.

Proposition 11.5. *The first Walsh transformation is calculated as*

$$S_f(\omega) = 2^{-n} \sum_{x=0}^{2^n-1} f(x)Q_x(\omega). \tag{11.33}$$

Proof. For any fixed $\omega_0 \in \mathbb{Z}_2^n$, we have

$$\begin{aligned} & \mathbf{f} \cdot \mathbf{Q}_{\omega_0} \\ & = \left(\sum_{\omega=0}^{2^n-1} S_f(\omega)\mathbf{Q}_\omega(x) \right) \cdot \mathbf{Q}_{\omega_0}(x) \\ & = S_f(\omega_0)\mathbf{Q}_{\omega_0}(x) \cdot \mathbf{Q}_{\omega_0}(x) \\ & = 2^n S_f(\omega_0). \end{aligned} \tag{11.34}$$

On the other hand, we have

$$\mathbf{f} \cdot \mathbf{Q}_{\omega_0} = \sum_{x=0}^{2^n-1} Q_{\omega_0}(x)f(x).$$

Hence,

$$S_f(\omega_0) = 2^{-n} \sum_{x=0}^{2^n-1} Q_{\omega_0}(x)f(x).$$

□

Next, we consider another Walsh transformation. Define

$$g(x) := 1 - 2f(x). \tag{11.35}$$

We have the expression of $\mathbf{g}$ over the basis $\{\mathbf{Q}_\omega | \omega = 0, 1, \cdots, 2^n - 1\}$ as

$$\mathbf{g} = \sum_{\omega=0}^{2^n-1} S_{(f)}(\omega)\mathbf{Q}_\omega(x). \tag{11.36}$$

Again, for a fixed $\boldsymbol{\omega}_0 \in \mathbb{Z}_2^n$, we have

$$\mathbf{g} \cdot \mathbf{Q}_{\omega_0} = \sum_{x=0}^{2^n-1} (1 - 2f(x))Q_{\omega_0}(x). \tag{11.37}$$

It is easy to check that

$$(-1)^{f(x)} = 1 - 2f(x). \tag{11.38}$$

Hence,

$$\mathbf{g} \cdot \mathbf{Q}_{\omega_0} = \sum_{x=0}^{2^n-1} (-1)^{f(x)} Q_{\omega_0}(x). \tag{11.39}$$

On the other hand, we have

$$\mathbf{g} \cdot \mathbf{Q}_{\omega_0} = \left(\sum_{\omega=0}^{2^n-1} S_{(f)}(\omega) Q_\omega(x) \right) \cdot Q_{\omega_0}(x) = 2^n \times S_{(f)}(\omega_0).$$

We conclude that

$$S_{(f)}(\omega) = \frac{1}{2^n} \sum_{x=0}^{2^n-1} (-1)^{f(x)} Q_\omega(x).$$

Definition 11.3. For a Boolean function $f(x)$,

$$S_{(f)}(\omega) = \frac{1}{2^n} \sum_{x=0}^{2^n-1} (-1)^{f(x)} Q_\omega(x) \tag{11.40}$$

is called the second Walsh transformation of f.

From (11.35) and (11.36) we have

$$f(x) = \frac{1}{2} - \frac{1}{2} \sum_{\omega=0}^{2^n-1} S_{(f)}(\omega) Q_\omega(x). \tag{11.41}$$

Next, we consider the relationship between the two Walsh Transformations. We have

Proposition 11.6. *$S_{(f)}(\omega)$ and $S_f(\omega)$ have the following relationships*

$$S_{(f)}(\omega) = \begin{cases} -2S_f(\omega), & \boldsymbol{\omega} \neq \mathbf{0} \\ 1 - 2S_f(\omega), & \boldsymbol{\omega} = \mathbf{0}. \end{cases} \tag{11.42}$$

Proof. Using (11.38), we have

$$\begin{aligned} S_{(f)}(\omega) &= \tfrac{1}{2^n} \sum_{x=0}^{2^n-1} (-1)^{f(x)} Q_\omega(x) \\ &= \tfrac{1}{2^n} \sum_{x=0}^{2^n-1} (1 - 2f(x)) Q_\omega(x) \\ &= \tfrac{1}{2^n} \sum_{x=0}^{2^n-1} Q_\omega(x) - \tfrac{2}{2^n} \sum_{x=0}^{2^n-1} f(x) Q_\omega(x). \end{aligned} \tag{11.43}$$

Using Lemma 11.1, we have (11.42). □

In the following we consider some interesting properties of Walsh transformations.

Proposition 11.7. *Let $S_f(\omega)$ be the Walsh transformation of $f(x)$, then for any $a \in \mathbb{Z}_2^n$, the Walsh transformation of $f(x+a)$ is*

$$S_{f(x+a)}(\omega) = Q_a(\omega) S_f(\omega). \tag{11.44}$$

Proof. By definition, the Walsh transformation of $f(x+a)$ is

$$\begin{aligned} S_{f(x+a)}(\omega) &= 2^{-n} \sum_{x=0}^{2^n-1} (-1)^{\boldsymbol{\omega}\cdot\mathbf{x}} f(x+a) \\ &= (-1)^{2\boldsymbol{\omega}\cdot\mathbf{a}} \cdot 2^{-n} \sum_{x=0}^{2^n-1} (-1)^{\boldsymbol{\omega}\cdot\mathbf{x}} f(x+a) \\ &= (-1)^{\boldsymbol{\omega}\cdot\mathbf{a}} \cdot 2^{-n} \sum_{x=0}^{2^n-1} (-1)^{\boldsymbol{\omega}\cdot(\mathbf{x}+\mathbf{a})} f(x+a) \\ &= (-1)^{\boldsymbol{\omega}\cdot\mathbf{a}} \cdot 2^{-n} \sum_{x=0}^{2^n-1} (-1)^{\boldsymbol{\omega}\cdot\mathbf{x}} f(x) \\ &= Q_a(\omega) S_f(\omega). \end{aligned}$$

□

Proposition 11.8. *Let $S_f(\omega)$ be the Walsh transformation of $f(x)$, and $S_g(\omega)$ the Walsh transformation of $g(x)$. Then the Walsh transformation of $af(x) + bg(x)$ is*

$$S_{af+bg}(\omega) = aS_f(\omega) + bS_g(\omega). \tag{11.45}$$

Proof. Starting from its definition, the Walsh transformation of $af(x) + bg(x)$ is

$$\begin{aligned} S_{af+bg}(\omega) &= 2^{-n} \sum_{x=0}^{2^n-1} (-1)^{\boldsymbol{\omega}\cdot\mathbf{x}} (af(x) + bg(x)) \\ &= a \cdot 2^{-n} \sum_{x=0}^{2^n-1} (-1)^{\boldsymbol{\omega}\cdot\mathbf{x}} f(x) + b \cdot 2^{-n} \sum_{x=0}^{2^n-1} (-1)^{\boldsymbol{\omega}\cdot\mathbf{x}} g(x) \\ &= aS_f(\omega) + bS_g(\omega). \end{aligned}$$

□

Proposition 11.9 (Plancheral Equation).

$$\sum_{\omega=0}^{2^n-1} S_f^2(\omega) = S_f(0). \tag{11.46}$$

Proof. Since

$$S_f(\omega) = 2^{-n} \sum_{x=0}^{2^n-1} Q_\omega(x) f(x),$$

we have

$$\begin{aligned}
\sum_{\omega=0}^{2^n-1} S_f^2(\omega) &= 2^{-2n} \sum_{\omega=0}^{2^n-1} dsum_{x=0}^{2^n-1} Q_\omega(x) f(x) S_f(\omega) \\
&= 2^{-2n} dsum_{x=0}^{2^n-1} f(x) \left(\sum_{\omega=0}^{2^n-1} Q_\omega(x) S_f(\omega) \right) \\
&= 2^{-2n} dsum_{x=0}^{2^n-1} f^2(x) \\
&= 2^{-2n} dsum_{x=0}^{2^n-1} f(x).
\end{aligned}$$

In another hand

$$S_f(0) = 2^{-2n} \sum_{\omega=0}^{2^n-1} f(x) Q_x(0) = 2^{-2n} \sum_{\omega=0}^{2^n-1} f(x).$$

Hence

$$\sum_{\omega=0}^{2^n-1} S_f^2(\omega) = S_f(0).$$

□

Then, it is easy to get the following proposition

Proposition 11.10 (Parseval Equation). *For the second Walsh transformation we have*

$$\sum_{\omega=0}^{2^n-1} S_{(f)}^2(\omega) = 1. \tag{11.47}$$

Next, we investigate the matrix converting form between $\mathbf{f}$ and its Walsh transformation $\mathbf{S}_f$. Because of the symmetry, we denote

$$Q(\omega, x) := Q_\omega(x).$$

From (11.32), we know that

$$\begin{aligned}&(f(0), f(1), \cdots, f(2^n-1))\\&= (S_f(0), S_f(1), \cdots, S_f(2^n-1))\times\\&\begin{bmatrix} Q(0,0) & Q(0,1) & \cdots & Q(0,2^n-1)\\ Q(1,0) & Q(1,1) & \cdots & Q(1,2^n-1)\\ \\ Q(2^n-1,0) & Q(2^n-1,1) & \cdots & Q(2^n-1,2^n-1)\end{bmatrix} \qquad (11.48)\\&:= (S_f(0), S_f(1), \cdots, S_f(2^n-1))H_n.\end{aligned}$$

The above equation can also be expressed briefly as

$$\mathbf{f} = \mathbf{S}_f H_n. \tag{11.49}$$

Definition 11.4. Let $A=(a_{ij})\in\mathcal{M}_{s\times s}$. A is called a Hadamard matrix, if it satisfies

(i)

$$a_{ij}=\pm1, \quad 1\le i,j\le s;$$

(ii)

$$A^TA=AA^T=sI_s.$$

Proposition 11.11. *The transfer matrix H_n, defined in (11.48), satisfies*

(i)

$$H_{n+1}=H_1\otimes H_n; \tag{11.50}$$

(ii)

$$H_nH_n=2^nI_{2^n}; \tag{11.51}$$

(iii) H_n is a Hadamard Matrix.

Proof. Item (ii) follows from Proposition 11.4 immediately. Then (iii) is obvious. We prove (i) only. Consider H_{n+1}. It can be expressed as

$$H_{n+1}=\begin{bmatrix}H_n^1 & H_n^2\\ H_n^2 & H_n^3\end{bmatrix},$$

where

$$H_n^1=\begin{bmatrix} Q(0,0) & \cdots & Q(0,2^n-1)\\ \vdots & & \\ Q(2^n-1,0) & \cdots & Q(2^n-1,2^n-1)\end{bmatrix};$$

$$H_n^2 = \begin{bmatrix} Q(0,2^n) & \cdots & Q(0,2^{n+1}-1) \\ \vdots & & \\ Q(2^n-1,2^n) & \cdots & Q(2^n-1,2^{n+1}-1) \end{bmatrix};$$

$$H_n^3 = \begin{bmatrix} Q(2^n,2^n) & \cdots & Q(2^n,2^{n+1}-1) \\ \vdots & & \\ Q(2^{n+1}-1,2^n) & \cdots & Q(2^{n+1}-1,2^{n+1}-1) \end{bmatrix}.$$

Consider H_n^1, and denote it as

$$H_n^1 = \left(Q^1(\omega,x)\right),$$

where

$$\mathbf{x} = (0,x_1,x_2,\cdots,x_{2^n}), \quad \boldsymbol{\omega} = (0,\omega_1,\omega_2,\cdots,\omega_{2^n}),$$

and both $(x_1,x_2,\cdots,x_{2^n})$ and $(\omega_1,\omega_2,\cdots,\omega_{2^n})$ run from $(0,\cdots,0)$ to $(1,\cdots,1)$. Then it is obvious that $H_n^1 = H_n$.

Next, consider H_n^2, and denote it as

$$H_n^2 = \left(Q^2(\omega,x)\right),$$

where

$$\mathbf{x} = (1,x_1,x_2,\cdots,x_{2^n}), \quad \boldsymbol{\omega} = (0,\omega_1,\omega_2,\cdots,\omega_{2^n}),$$

and both $(x_1,x_2,\cdots,x_{2^n})$ and $(\omega_1,\omega_2,\cdots,\omega_{2^n})$ run from $(0,\cdots,0)$ to $(1,\cdots,1)$. Then it is obvious that $H_n^2 = H_n$.

Finally, we consider H_n^3, and denote it as

$$H_n^3 = \left(Q^3(\omega,x)\right),$$

where

$$\mathbf{x} = (1,x_1,x_2,\cdots,x_{2^n}), \quad \boldsymbol{\omega} = (1,\omega_1,\omega_2,\cdots,\omega_{2^n}).$$

Similar argument shows that $H_n^3 = -H_n$.

Note that

$$H(1) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

then it is clear that

$$H_{n+1} = H_1 \otimes H_n.$$

□

We consider an example.

Example 11.5. Recall the Boolean function in Example 11.4,

$$f = x_1 \wedge \neg(x_2 \to x_3).$$

Thus

$$\mathbf{f} = (f(0)\ f(1)\ \cdots\ f(7)) = (0\ 0\ 0\ 0\ 0\ 0\ 1\ 0).$$

Using (11.49), we have

$$\mathbf{f} = \mathbf{S}_f H_3.$$

From Proposition 11.11, we know that $H_3 H_3 = 8I_3$, thus

$$\mathbf{S}_f = \frac{1}{8}\mathbf{f} H_3.$$

By (11.50),

$$H_3 = H_1 \otimes H_1 \otimes H_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix}.$$

Thus,

$$\mathbf{S}_f = (\frac{1}{8}, \frac{1}{8}, -\frac{1}{8}, -\frac{1}{8}, -\frac{1}{8}, -\frac{1}{8}, \frac{1}{8}, \frac{1}{8}).$$

By Proposition 11.6,

$$\mathbf{S}_{(f)} = (\frac{3}{4}, -\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}).$$

Moreover, it is easy to check that Plancheral Equation (11.46) and Parseval Equation (11.47) are satisfied.

11.4 Linear Structure

Definition 11.5. Let $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$ ba a Boolean function.

(1) $a \in \mathbb{Z}_2^n$ is called an invariant linear structure (ILS) of f, if

$$f(x + a) + f(x) = 0. \tag{11.52}$$

(2) $a \in \mathbb{Z}_2^n$ is called a variant linear structure (VLS) of f, if

$$f(x+a)+f(x)=1. \tag{11.53}$$

(3) Denote by

$$\begin{array}{l} E_0 := \{a \in \mathbb{Z}_2^n \mid f(x+a)+f(a)=0\} \\ E_1 := \{a \in \mathbb{Z}_2^n \mid f(x+a)+f(a)=1\} \\ E \ := E_0 \cup E_1. \end{array} \tag{11.54}$$

Then E is called the linear structure subspace of f.

Proposition 11.12.

(1) $E_0 \cap E_1 = \varnothing$.
(2) E is a vector space and $E_0 \subset E$ is a vector subspace.
(3) Assume $E_1 \neq \varnothing$. Then $E_1 = E_0 + a$, where $a \in E_1$.

Proof. (1) It follows from definitions immediately.

(2)
- Let $a, b \in E_0$. Then

$$\begin{array}{l} f(x+a+b)+f(x) \\ = f(x+a+b)+f(x+a)+f(x+a)+f(x) \\ = 0+0=0. \end{array}$$

That is, $a+b \in E_0 \subset E$.

- Let $a \in E_0$ and $b \in E_1$. Then

$$\begin{array}{l} f(x+a+b)+f(x) \\ = f(x+a+b)+f(x+a)+f(x+a)+f(x) \\ = 1+0=1. \end{array}$$

That is, $a+b \in E_1 \subset E$.

- Let $a, b \in E_1$. Then

$$\begin{array}{l} f(x+a+b)+f(x) \\ = f(x+a+b)+f(x+a)+f(x+a)+f(x) \\ = 1+1=0. \end{array}$$

That is, $a+b \in E_0 \subset E$.

Hence E is a vector space. Moreover, case 1 proved that E_0 is a vector subspace.

(3) It was proved that if $a \in E_0$, $b \in E_1$, then $a+b \in E_1$. That is, $E_0 + b \subset E_1$. Now assume $\xi \in E_1$, then $\xi + b \in E_0$, and

$$\xi = (\xi + b) + b \in E_0 + b,$$

which means $E_1 \subset E_0 + b$. The conclusion follows.

□

We have the following corollary.

Corollary 11.1.

(1)

$$|E_0| = 2^r, \tag{11.55}$$

where r is the dimension of E_0.

(2) Either $E_1 = \varnothing$, or

$$|E_1| = |E_0|. \tag{11.56}$$

***Proof*.** (1) As a vector subspace, (11.55) is trivial.

(2) Assume $E_1 \neq \varnothing$, and let $b \in E_1$. Define $\pi_b : E_0 \to E_1$ as $x \mapsto x + b$, then it is easy to check that π is one-to-one and onto.

□

Definition 11.6. Given a Boolean function f.

(i) Let $|E| = 2^q$. Then q is called the dimension of linear structure of f. When $q > 0$, f is called a logical function with linear structure (LFLS).

(ii) For an LFLS f, if $E_0 \neq \{0\}$, then f is said to be of type I, if $E_0 = \{0\}$, it is said to be of type II.

Next, we consider how to calculate E_0 and E_1. Let $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$ be a logical function with its structure matrix $M_f \in \mathcal{L}_{2\times 2^n}$. Denote by $\alpha = \ltimes_{i=1}^n a_i$, $x = \ltimes_{i=1}^n x_i$. Then it is easy to see that $(a_1, \cdots, a_n) \in E_1$, if and only if

$$M_f M_p a_1 x_1 M_p a_2 x_2 \cdots M_p a_n x_n = M_f x_1 x_2 \cdots x_n. \tag{11.57}$$

A straightforward computation shows that (11.57) is equivalent to

$$M_f M_p \ltimes_{i=1}^{n-1} (I_{2^{2i}} \otimes M_p) \ltimes_{i=1}^{n-1} \left(I_{2^i} \otimes W_{[2,2^i]}\right) \alpha x = M_f x. \tag{11.58}$$

Define

$$\Psi_f := M_f M_p \ltimes_{i=1}^{n-1} (I_{2^{2i}} \otimes M_p) \ltimes_{i=1}^{n-1} \left(I_{2^i} \otimes W_{[2,2^i]}\right),$$

and split it into 2^n blocks as

$$\Psi_f = [\psi_1 \ \psi_2 \ \cdots \ \psi_{2^n}],$$

where $\psi_k = \mathrm{Blk}_k(\Psi_f)$, $k = 1, 2, \cdots, 2^n$. Then the following result can be verified by a straightforward computation.

Theorem 11.2. *Let $\alpha = \ltimes_{i=1}^n a_i := \delta_{2^n}^i$. Then $(a_1, \cdots, a_n) \in E_0$, if and only if $\psi_i = M_f$. $(a_1, \cdots, a_n) \in E_1$, if and only if $\psi_i = M_n M_f$.*

Example 11.6. (1) Assume the structure matrix of f is

$$M_f = \delta_2[2\ 2\ 2\ 1\ 2\ 1\ 1\ 1].$$

Then

$$M_n M_f = \delta_2[1\ 1\ 1\ 2\ 1\ 2\ 2\ 2].$$

It is easy to calculate that

$$\begin{array}{rl}\Psi_f = [1 & 1\ 1\ 2\ 1\ 2\ 2\ 2\ 1\ 1\ 2\ 1\ 2\ 1\ 2\ 2 \\ 1 & 2\ 1\ 1\ 2\ 2\ 1\ 2\ 2\ 1\ 1\ 1\ 2\ 2\ 2\ 1 \\ 1 & 2\ 2\ 2\ 1\ 1\ 1\ 2\ 2\ 1\ 2\ 2\ 1\ 1\ 2\ 1 \\ 2 & 2\ 1\ 2\ 1\ 2\ 1\ 1\ 2\ 2\ 2\ 1\ 2\ 1\ 1\ 1].\end{array}$$

Since $\psi_8 = M_f$ and $\psi_1 = M_n M_f$, we have that

$$(0,0,0,0,0,0,0,0) \in E_0,$$

and

$$(1,1,1,1,1,1,1,1) \in E_1.$$

Now $|E_0| = |E_1| = 1$, hence $|E| = 2$, $\dim(E) = 1$, and f is an LFLS.

(2) Assume the structure matrix of f is

$$M_f = \delta_2[2\ 2\ 2\ 2\ 2\ 2\ 2\ 1].$$

Then

$$M_n M_f = \delta_2[1\ 1\ 1\ 1\ 1\ 1\ 1\ 2].$$

It is easy to calculate that

$$\begin{array}{rl}\Psi_f = [1 & 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 2\ 2\ 2\ 2\ 2\ 2 \\ 2 & 2\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 2\ 2\ 2\ 2 \\ 2 & 2\ 2\ 2\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 2\ 2 \\ 2 & 2\ 2\ 2\ 2\ 2\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1].\end{array}$$

Since only $\psi_8 = M_f$, we have that

$$E_0 = \{(0,0,0,0,0,0,0,0)\},$$

and

$$E_1 = \varnothing.$$

Now $|E_0| = 1$ and $|E_1| = 0$, hence $|E| = 1$ f is not an LFLS.

11.5 Nonlinearity

Definition 11.7.

(1) Let $\mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{Z}_2^n$. The Hamming weight of $\mathbf{x}$ is defined as

$$w_H(\mathbf{x}) = \big|\,\{i \,|\, x_i \neq 0\}\,\big|. \tag{11.59}$$

(2) Let $f, g \in \mathcal{B}_n^{\mathcal{F}}$. The Hamming distance of f and g is defined as

$$d_H(f, g) := w_H(\mathbf{f} + \mathbf{g}). \tag{11.60}$$

Next, we consider the nonlinearity of a Boolean function.

Definition 11.8. Let $f \in \mathcal{B}_n^{\mathcal{F}}$.

(1) The nonlinearity of f, denoted by N_f, is defined as

$$N_f := \min_{\ell \in \mathcal{B}_n^{\mathcal{A}}} d_H(\mathbf{f}, \boldsymbol{\ell}). \tag{11.61}$$

(2) The linearity of f, denoted by C_f, is defined as

$$C_f := \max_{\ell \in \mathcal{B}_n^{\mathcal{A}}} d_H(\mathbf{f}, \boldsymbol{\ell}). \tag{11.62}$$

Definition 11.9. Assume $\ell(x) \in \mathcal{B}_n^{\mathcal{A}}$ satisfies

$$d_H(\boldsymbol{\ell}, \mathbf{f}) = N_f. \tag{11.63}$$

Then $\ell(x)$ is called the best linear approximation of $f(x)$.

To calculate the nonlinearity of an $f \in \mathcal{B}_n^{\mathcal{F}}$, we consider the linear equivalence of f. The following theory shows the probability of linear equivalence of a Boolean function via Walsh transformation (Wen *et al.*, 2000).

Theorem 11.3. *Let $\boldsymbol{\omega} = (\omega_1, \omega_2 \cdots \omega_n)$, $\mathbf{x} = (x_1, x_2 \cdots x_n) \in \mathbb{Z}_2^n$, and $x \in \mathbb{Z}_2^n$ be identically distributed. Then we have*

$$P\left(\{x | f(x) = \boldsymbol{\omega} \cdot \mathbf{x}\}\right) = \frac{1 + S_{(f)}(\omega)}{2}, \tag{11.64}$$

$$P\left(\{x | f(x) \neq \boldsymbol{\omega} \cdot \mathbf{x}\}\right) = \frac{1 - S_{(f)}(\omega)}{2}. \tag{11.65}$$

***Proof*.** Since

$$\begin{aligned}
S_{(f)}(x) &= 2^{-1}\sum_{x=0}^{2^n-1}(-1)^{f(x)}(-1)^{\boldsymbol{\omega}\cdot\mathbf{x}}\\
&= 2^{-1}\left(\sum_{\{x|f(x)=\boldsymbol{\omega}\cdot\mathbf{x}\}}(-1)^{f(x)+\boldsymbol{\omega}\cdot\mathbf{x}}+\sum_{\{x|f(x)\neq\boldsymbol{\omega}\cdot\mathbf{x}\}}(-1)^{f(x)+\boldsymbol{\omega}\cdot\mathbf{x}}\right)\\
&= 2^{-1}\left(\sum_{\{x|f(x)=\boldsymbol{\omega}\cdot\mathbf{x}\}}1+\sum_{\{x|f(x)\neq\boldsymbol{\omega}\cdot\mathbf{x}\}}(-1)\right)\\
&= P\left(\{x|f(x)=\boldsymbol{\omega}\cdot\mathbf{x}\}\right)-P\left(\{x|f(x)\neq\boldsymbol{\omega}\cdot\mathbf{x}\}\right),
\end{aligned}$$

and $P\left(\{x|f(x)=\boldsymbol{\omega}\cdot\mathbf{x}\}\right)+P\left(\{x|f(x)\neq\boldsymbol{\omega}\cdot\mathbf{x}\}\right)=1$, (11.64) and (11.65) can be obtained. □

Theorem 11.4. *Let $f\in\mathcal{B}_n^{\mathcal{F}}$ and denote*

$$a=\max_{0\leq\omega\leq 2^n-1}\left|S_{(f)}(\omega)\right|.$$

Then

$$N_f=2^n\left(\frac{1-a}{2}\right); \tag{11.66}$$

and

$$C_f=2^n\left(\frac{1+a}{2}\right). \tag{11.67}$$

***Proof*.** For any affine function $\ell(x)=\boldsymbol{\omega}\cdot\mathbf{x}+\omega_0$ we have

$$\begin{aligned}
&P\left(\{x|f(x)=\ell(x)\}\right)=P\left(\{x|f(x)=\boldsymbol{\omega}\cdot\mathbf{x}+\omega_0\}\right)\\
&\quad=\begin{cases}P\left(\{x|f(x)=\boldsymbol{\omega}\cdot\mathbf{x}\}\right), & \omega_0=0\\ P\left(\{x|f(x)=\boldsymbol{\omega}\cdot\mathbf{x}+1\}\right)=P\left(\{x|f(x)\neq\boldsymbol{\omega}\cdot\mathbf{x}\}\right), & \omega_0=1.\end{cases}
\end{aligned}$$

Hence

$$P\left(\{x|f(x)=\ell(x)\}\right)=\begin{cases}P\left(\{x|f(x)=\boldsymbol{\omega}\cdot\mathbf{x}\}\right), & \ell(x)=\boldsymbol{\omega}\cdot\mathbf{x}\\ P\left(\{x|f(x)\neq\boldsymbol{\omega}\cdot\mathbf{x}\}\right), & \ell(x)=\boldsymbol{\omega}\cdot\mathbf{x}+1.\end{cases}$$

According to Theorem 11.3, we have

$$P\left(\{x|f(x)=\ell(x)\}\right)=\begin{cases}\frac{1+S_{(f)}(\omega)}{2}, & \ell(x)=\boldsymbol{\omega}\cdot\mathbf{x}\\ \frac{1-S_{(f)}(\omega)}{2}, & \ell(x)=\boldsymbol{\omega}\cdot\mathbf{x}+1.\end{cases}$$

It follows that

$$\max_{\ell \in \mathcal{B}_n^{\mathcal{A}}} P\left(\{x | f(x) = \ell(x)\}\right) = \frac{1+a}{2}; \tag{11.68}$$

and

$$\min_{\ell \in \mathcal{B}_n^{\mathcal{A}}} P\left(\{x | f(x) = \ell(x)\}\right) = \frac{1-a}{2}. \tag{11.69}$$

Using (11.68), we have

$$\begin{aligned} N_f &= \min\nolimits_{\ell \in \mathcal{B}_n^{\mathcal{A}}} w_H(\mathbf{f} + \boldsymbol{\ell}) \\ &= \min\nolimits_{\ell \in \mathcal{B}_n^{\mathcal{A}}} 2^n P\left(\{x | f(x) \neq \ell(x)\}\right) \\ &= \min\nolimits_{\ell \in \mathcal{B}_n^{\mathcal{A}}} 2^n \left(1 - P\left(\{x | f(x) = \ell(x)\}\right)\right) \\ &= 2^n \left(1 - \max\nolimits_{\ell \in \mathcal{B}_n^{\mathcal{A}}} P\left(\{x | f(x) = \ell(x)\}\right)\right) \\ &= 2^n \left(\tfrac{1-a}{2}\right). \end{aligned}$$

Using (11.69), a similar argument shows that

$$C_f = 2^n \left(\frac{1+a}{2}\right).$$

□

An immediate consequence is

Corollary 11.2. *For any* $f \in \mathcal{B}_n^{\mathcal{F}}$,

$$N_f + C_f = 2^n. \tag{11.70}$$

From Theorem 11.4 one sees that when a reaches its smallest value the corresponding N_f becomes the largest one. Since

$$\sum_{\omega=0}^{2^n-1} S_{(f)}^2 = 1,$$

when $\left|S_{(f)}\right| = \text{const}$, a reaches the smallest. In this case we have

$$\left|S_{(f)}\right| = 2^{-\frac{n}{2}}. \tag{11.71}$$

Hence we have

$$N_f = 2^n \left(\frac{1 - 2^{-\frac{n}{2}}}{2}\right) = 2^{n-1} - 2^{\frac{n}{2}-1}. \tag{11.72}$$

Therefore, we have the following result.

Proposition 11.13.

$$N_f \leq 2^{n-1} - 2^{\frac{n}{2}-1}. \tag{11.73}$$

When (11.72) holds, f has the highest nonlinear degree. Such a Boolean function is called a bent function or a complete nonlinear function, which is very important in cryptography (Carlet, 2010).

11.6 Symmetry of Boolean Function

Recall that $\mathbf{S}_n$ is the nth order symmetric group. Denote by $\mathbf{H}_n < \mathbf{S}_n$ a subgroup of $\mathbf{S}_n$.

Recall Definition 6.6, a Boolean function $f(x) \in \mathcal{B}_n^{\mathcal{F}}$ is said to be symmetric with respect to $\mathbf{H}_n$, if

$$f(x_{\sigma(1)}, \cdots, x_{\sigma(n)}) = f(x_1, \cdots, x_n), \quad \forall\, \sigma \in \mathbf{H}_n. \tag{11.74}$$

Let $\pi_i(f)$ be the ith degree homogeneous part of $f(x)$.

Theorem 11.5. *f is symmetric with respect to $\mathbf{H}_n$, if and only if $\pi_i(f)$, $i = 0, 1, \cdots, n$ are symmetric with respect to $\mathbf{H}_n$.*

Proof. Sufficiency is obvious. We prove the necessity. Assume f is symmetric with respect to $\mathbf{H}_n$. We prove it by contradiction. Assume there exists at least one i, such that $\pi_i(f)$ is not symmetric with respect to $\mathbf{H}_n$. Assume $i > 0$ be the smallest such i. We express $\pi_i(f)$ as

$$\pi_i(f) = \sum_{1 \le j_1 < j_2 < \cdots < j_i \le n} c_{j_1 \cdots j_i} x_{j_1} \cdots x_{j_i}. \tag{11.75}$$

Note that not all $c_{j_1 \cdots j_i} = 0$. Otherwise, $\pi_i(f) = 0$ and hence it is symmetric. For any $c_{j_1 \cdots j_i} = 1$ if $c_{\sigma(j_1) \cdots \sigma(j_i)} = 1$ for all $\sigma \in \mathcal{H}$, we are done. So we assume there exists $\sigma \in \mathbf{H}_n$ such that $c_{\sigma(j_1) \cdots \sigma(j_i)} = 0$. Let $x_0 = (x_1, \cdots, x_n)$ be determined by

$$x_j = \begin{cases} 1, & j \in \{j_1, \cdots, j_i\} \\ 0, & \text{otherwise.} \end{cases}$$

Then it is easy to see that

$$\begin{aligned} \pi_i(f)(x_1, \cdots, x_n) \quad &= 1 \\ \pi_i(f)(x_{\sigma(1)}, \cdots, x_{\sigma(n)}) &= 0. \end{aligned}$$

Note that

$$\pi_k(x_0) = 0, \quad k > i.$$

Hence

$$f(x_1, \cdots, x_n) \neq f(x_{\sigma(1)}, \cdots, x_{\sigma(n)}).$$

This is a contradiction. □

The following corollaries are obvious.

Corollary 11.3. *f is symmetric with respect to $\mathbf{S}_n$, if and only if for each $1 \leq i \leq n-1$, the coefficients of ith homogeneous terms are the same. Precisely,*

$$c_{j_1 \cdots j_i}, \quad 1 \leq j_1 < j_2 < \cdots < j_i \leq n$$

are identically 1 or 0.

Symmetry with respect to $\mathbf{S}_n$ is also called the complete symmetry.

Corollary 11.4. *There are 2^{n+1} completely symmetric Boolean functions in $\mathcal{B}_n^{\mathcal{F}}$.*

Proof. According to Corollary 11.3, if f is symmetric with respect to $\mathbf{S}_n$ we have either $\pi_i(f) \equiv 0$ or $\pi_i(f) \not\equiv 0$. Precisely, we have $\pi_i(f) = \mathbf{0}_{d_i}$ or $\pi_i(f) = \mathbf{1}_{d_i}$. The conclusion follows. □

We may name the 2^{n+1} completely symmetric Boolean functions as $f_0, f_1, \cdots, f_{2^n-1}$, where f_i is decided in the following way: Converting i into binary form as

$$i \sim i_n i_{n-1} \cdots i_0.$$

Then

$$f_i(x) = \sum_{j=0}^{n} i_j P_j(x), \quad i = 0, 1, \cdots, 2^{n+1} - 1, \tag{11.76}$$

where

$$P_j(x) = \sum_{1 \leq k_1 < \cdots < k_j \leq n} \prod_{s=1}^{j} x_{k_s}.$$

Then we have the following result.

Proposition 11.14. *Let $\{f_i \,|\, i = 0, 1, \cdots, 2^{n+1}-1\}$ be the set of completely symmetric Boolean functions, with the indices being determined as in the above. Then the Hamming weight of f_i is*

$$w_H(\mathbf{f}_i) = \sum_{j=0}^{n} i_j \binom{n}{j}, \quad i = 0, 1, \cdots, 2^{n+1} - 1. \tag{11.77}$$

Proof. Since

$$w_H(\mathbf{P}_j) = \binom{n}{j}, \quad j = 0, 1, \cdots, n,$$

the conclusion follows from (11.76) immediately. □

To consider the symmetry with respect to $\mathbf{H}_n$, we need some additional concepts.

Definition 11.10. Let G be a group, and S a nonempty set. A mapping $G \times S \to S$ is called the group action of G on S, if it satisfies

(i)

$$e(s) = s, \quad \forall\, s \in S,$$

where e is the identity of G.

(ii)

$$g_1(g_2(s)) = (g_1 g_2)(s), \quad \forall\, s \in S.$$

Definition 11.11. Assume G acts on S and $s \in S$.

(1) The trajectory of s under the action of G is defined as

$$Gs := \{gs \mid g \in G\}. \tag{11.78}$$

(2) The stability subgroup of s is defined as

$$G_s := \{g \in G \mid g(s) = s\}. \tag{11.79}$$

It is obvious that the trajectories $\{Gs \mid s \in S\}$ form a partition of S. Because for $s_1, s_2 \in S$, either $Gs_1 = Gs_2$ or $Gs_1 \cap Gs_2 = \varnothing$.

For group action we have the following properties (Dixon and Mortimer, 1996).

Proposition 11.15. *The length of trajectory Gs is*

$$|Gs| = \frac{|G|}{|G_s|}. \tag{11.80}$$

The number of the trajectories can be obtained via the following theorem.

Theorem 11.6 (Burnside Lemma). *Assume G, acting on S, forms m trajectories. Then*

$$m|G| = \sum_{g \in G} |Fix(g)|. \tag{11.81}$$

Here

$$Fix(g) = \{s \mid g(s) = s\}.$$

Construct a sequence of sets as:

$$S_i = \big\{ \{j_1, \cdots, j_i\} \subset \mathbb{Z} \mid 1 \le j_t \le n, t = 1, \cdots, i; j_p \ne j_q, p \ne q \big\}, \\ i = 1, \cdots, n-1.$$

Let $\mathbf{H}_n < \mathbf{S}_n$. The action of $\mathbf{H}_n$ on S_i is defined as:

$$\sigma(\{j_1, \cdots, j_i\}) := \{\sigma(j_1), \cdots, \sigma(j_i)\}. \tag{11.82}$$

Using Theorem 11.5 and arguing as for Corollary 11.3, we can prove the following

Theorem 11.7. *Assume the number of trajectories of $\mathbf{H}_n$ acting on S_i is m_i, $i = 1, \cdots, n-1$. Then the number of Boolean functions symmetric with respect to $\mathbf{H}_n$ is*

$$m = 2^{2^{2+\sum_{i=1}^{n-1} m_i}}. \tag{11.83}$$

Assume a subgroup $\mathbf{C}_n < \mathbf{S}_n$ is generated by $(1, 2, \cdots, n)$, that is, $\mathbf{C}_n = \langle(1, 2, \cdots, n)\rangle$, which is called a cyclic subgroup. A Boolean function $f \in \mathcal{B}_n^{\mathcal{F}}$ is said to be cyclically symmetric, if it is symmetric with respect to the cyclic subgroup.

Example 11.7. Let $n = 4$, and $\mathbf{C}_4 = \langle(1, 2, 3, 4)\rangle$ be the cyclic subgroup generated by $(1, 2, 3, 4)$. Then

$$\begin{aligned} S_1 &= \{1, 2, 3, 4\}; \\ S_2 &= \big\{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}\big\}; \\ S_3 &= \big\{\{1, 2, 3\}, \{1, 2, 4\}, \{1, 3, 4\}, \{2, 3, 4\}\big\}. \end{aligned}$$

It is easy to check that each S_1 or S_3 has only one trajectory. S_2 has two trajectories, which are

$$\begin{aligned} &\{1, 2\} \to \{2, 3\} \to \{3, 4\} \to \{4, 1\} \to; \\ &\{1, 3\} \to \{2, 4\} \to . \end{aligned}$$

We conclude that all the $f \in \mathcal{B}_4^{\mathcal{F}}$, which are cyclically symmetric with respect to $\mathbf{C}_4 = G\{(1, 2, 3, 4)\}$, can be expressed as

$$\begin{aligned} f(x) = a_0 &+ a_1(x_1 + x_2 + x_3 + x_4) + a_2(x_1x_2 + x_2x_3 + x_3x_4 + x_4x_1) \\ &+ a_3(x_1x_3 + x_2x_4) + a_4(x_1x_2x_3 + x_1x_2x_4 + x_1x_3x_4 + x_2x_3x_4), \end{aligned} \tag{11.84}$$

where $a_i = 0$ or 1, $i = 0, 1, 2, 3, 4$.

Exercises

11.1 Assume p is a prime number. Show that $(\mathbb{Z}_p, \langle + \rangle, \langle \times \rangle)$ is a field, but $(\mathcal{D}, (+), (\times))$ is not.

11.2 Prove Proposition 11.1.

11.3 Let $\mathbf{x} = (1,1,0,1,0,1,0,1)$. Find its scalar form χ_x.

11.4 Prove Proposition 11.2.

11.5 Prove Proposition 11.10.

11.6 Let $f(x) = (x_1 \vee x_2) \to x_3$. Calculate M_f, $\mathbf{S}_f$, $\mathbf{S}_{(f)}$, and the polynomial form of f.

11.7 Check whether the f given in above exercise is (a) affine? (b) linear?

11.8 Prove Theorem 11.2.

11.9 Prove that the Hamming distance defined in Definition 11.7 is a distance.

(Hint: A distance should satisfy: (i) $d(x,y) \geq 0$, and $d(x,y) = 0$ if and only if $x = y$; (ii) $d(x,y) = d(y,x)$; (iii) $d(x,y) + d(y,z) \geq d(x,z)$.)

11.10 Calculate N_f and C_f for $f(x) = (x_1 \vee x_2) \to x_3$.

11.11 Show that when $\omega^* = \arg\max_\omega |S_{(f)}(\omega)|$, $\boldsymbol{\omega} \cdot \mathbf{x}$ or $\boldsymbol{\omega} \cdot \mathbf{x} + 1$ is the best linear approximation of $f(x)$.

11.12 Find the best linear approximation of $f(x) = (x_1 \vee x_2) \to x_3$.

11.13 Show that

$$S_f(0) = 2^{-n} w_H(f). \tag{11.85}$$

11.14 Assume G acts on S and $s \in S$. Prove that $G_s < G$, i.e., G_s is a subgroup of G.

11.15 Prove that the action of $\mathbf{H}_n$ on S_i, defined by (11.82), is a group action.

11.16 Find out all the $f \in \mathcal{B}_5^{\mathcal{F}}$, which are symmetric with respect to $\mathbf{C}_5 = G\{(1,2,3,4,5)\}$. Give a general form of this set of Boolean functions.

This page intentionally left blank

Chapter 12

Decomposition of Logical Functions

Decomposition is one of the most efficient ways to realize Boolean networks via circuits economically. If the decomposition exists, it can significantly reduce the area, delay, and power for logical synthesis. This chapter first considers the bi-decomposition of Boolean functions. Both disjoint and non-disjoint decompositions are considered. Then the results are extended to k-valued logical functions and to mix-valued logical case. Necessary and sufficient conditions are obtained for each case. This chapter is based on Cheng and Xu (2011).

12.1 Disjoint Bi-Decomposition

Let $f : \mathcal{D}^n \to \mathcal{D}$ be a Boolean function, which is assumed to be realized by a logical circuit (also called a network) (see Fig. 12.1 (a)).

(1) Let $\{X_1, X_2\}$ be a partition of $X = \{x_1, x_2, \cdots, x_n\}$. If f can be expressed as

$$f(X) = F(\phi(X_1), \psi(X_2)), \tag{12.1}$$

then f can be realized by a disjoint bi-decomposed circuit as in Fig. 12.1 (b).

(2) Let $\{X_1, X_2, X_3\}$ be a partition of X. If f can be expressed as

$$f(X) = F(\phi(X_1, X_2), \psi(X_2, X_3)), \tag{12.2}$$

then f can be realized by a non-disjoint bi-decomposed circuit as in Fig. 12.1 (c).

Here, f, F, ϕ, ψ are all Boolean functions.

Decomposition is one of the most efficient way to realize networks via circuits economically. If the decomposition exists, it can significantly reduce

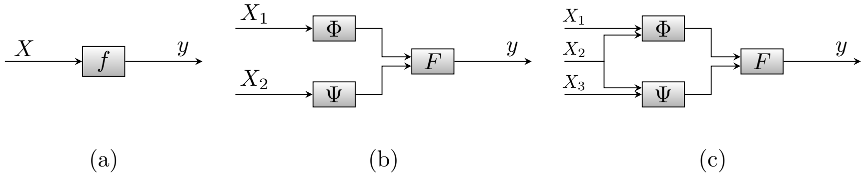


Fig. 12.1 (a) Boolean function (b) Disjoint decomposition (c) Non-disjoint decomposition

the area, delay, and power for logical synthesis (Hachtel and Somenzi, 2000). Therefore, it becomes a long standing research topic since 1950s. There are some interesting and useful results. For instance, when the number of inputs of a switching circuit is small, the Quine-McCluskey procedure is widely used for designing two-stage networks (Davio *et al.*, 1978). When a large number of inputs are involved, decomposition chart method to multi-level minimization was proposed by Ashenhurst (1957), and was later further discussed and developed by Curtis (1962); Roth and Karp (1962). There are many articles devoted to developing efficient algorithms for decomposition of switching functions, as well as multi-valued logic functions (Choudhury and Mohanram, 2010; Sasao and Fujita, 1996; Sasao and Butler, 1997; Sasao, 1999; Posthoff and Steinbach, 2004; Mishchenko *et al.*, 2001; Brayton and Khatri, 1999).

We start with disjoint decomposition.

Definition 12.1. Let $f : \mathcal{D}^n \to \mathcal{D}$ be a Boolean function, and $\Gamma \cup \Lambda$ a partition of $\{1, 2, \cdots, n\}$. f is said to be bi-decomposable with respect to Γ and Λ if there exist three Boolean functions $F : \mathcal{D}^2 \to \mathcal{D}$, $\phi : \{x_\gamma | \gamma \in \Gamma\} \to \mathcal{D}$, and $\psi : \{x_\lambda | \lambda \in \Lambda\} \to \mathcal{D}$, such that

$$f(x_1, \cdots, x_n) = F(\phi(x_\gamma | \gamma \in \Gamma), \psi(x_\lambda | \lambda \in \Lambda)). \tag{12.3}$$

First, we assume

$$\Gamma = \{1, 2, \cdots, k\}, \quad \text{and } \Lambda = \{k+1, k+2, \cdots, n\}. \tag{12.4}$$

Definition 12.2. Let $M = \delta_2[i_1 \; i_2 \; \cdots \; i_{2^k}] \in \mathcal{L}_{2 \times 2^k}$.

(1) M is called a constant function matrix, if

$$i_1 = i_2 = \cdots = i_{2^k}.$$

That is, it is the structure matrix of a constant function.

(2)

$$\neg M := \delta_2[1-i_1\ 1-i_2\ \cdots\ 1-i_{2^k}] \in \mathcal{L}_{2\times 2^k}$$

is called the compliment of M. M and $\neg M$ are called the two complimented matrices.

Theorem 12.1. *Let $f : \mathcal{D}^n \to \mathcal{D}$ be a Boolean function with its structure matrix M_f, being split into 2^k equal blocks as*

$$M_f = [M_1\ M_2\ \cdots\ M_{2^k}], \tag{12.5}$$

where each $M_i \in \mathcal{L}_{2\times 2^{n-k}}$.

f is bi-decomposable with respect to the partition in (12.4), if and only if, the set $\{M_i \,|\, i = 1, \cdots, 2^k\}$ consists of one of the following four possible cases:

(i) two constant matrices; or
(ii) one constant matrix and one non-constant matrix; or
(iii) one non-constant matrix; or
(iv) two complemented non-constant matrices.

Proof. (Necessity) Assume there are three functions F, ϕ, and ψ, such that (12.1) holds. Denote the structure matrix of f by $M_f \in \mathcal{L}_{2\times 2^n}$, and it is split as in (12.5). Assume the structure matrix of F is

$$M_F = \delta_2[i_1\ i_2\ i_3\ i_4];$$

the structure matrix of ϕ is

$$M_\phi = \delta_2[j_1\ j_2\ \cdots\ j_{2^k}];$$

and the structure matrix of ψ is $M_\psi \in \mathcal{L}_{2\times 2^{n-k}}$. Then we have

$$M_f x = M_F M_\phi x^1 M_\psi x^2 = M_F M_\phi \left(I_{2^k} \otimes M_\psi\right) x, \tag{12.6}$$

where $x = \ltimes_{i=1}^n x_i$, $x^1 = \ltimes_{i=1}^k x_i$, and $x^2 = \ltimes_{i=k+1}^n x_i$.

Since x is arbitrary, we have

$$M_f = M_F M_\phi \left(I_{2^k} \otimes M_\psi\right). \tag{12.7}$$

We first calculate $M_F M_\phi$, which is denoted by

$$M_F M_\phi := [N_1\ N_2\ \cdots\ N_{2^k}].$$

Then a straightforward computation shows that

$$N_s = \begin{cases} \delta_2[i_1\ i_2], & j_s = 1 \\ \delta_2[i_3\ i_4], & j_s = 2, \end{cases} \qquad s = 1, 2, \cdots, 2^k.$$

It follows that if we denote

$$M_F M_\phi \left(I_{2^k} \otimes M_\psi\right) = \left[W_1\ W_2\ \cdots\ W_{2^k}\right].$$

Then

$$W_s = \begin{cases} \delta_2[i_1\ i_2]M_\psi, & j_s = 1 \\ \delta_2[i_3\ i_4]M_\psi, & j_s = 2, \end{cases} \qquad s = 1, 2, \cdots, 2^k.$$

For (12.7) to be true, we need

$$M_i = W_i, \quad i = 1, \cdots, 2^k. \tag{12.8}$$

Now if $i_1 = i_2$ and $i_3 = i_4$, we have case (i); if either $i_1 = i_2$ or $i_3 = i_4$ but not both, we have case (ii); if $i_1 = i_3$ and $i_2 = i_4$, but $i_1 \neq i_2$, we have case (iii); and if $i_1 \neq i_2$, $i_3 \neq i_4$, and $i_1 \neq i_3$, then we have either $\delta_2[i_1\ i_2] = I_2$ and $\delta_2[i_1\ i_2] = \neg I_2$, or $\delta_2[i_1\ i_2] = \neg I_2$ and $\delta_2[i_1\ i_2] = I_2$, then we have case (iv).

(sufficiency) Since i_1, i_2, i_3, i_4 and $j_1, \cdots, j_{2^k}$ are completely free, if the structure matrix M_f of f belongs to one of the above four cases, we can first choose i_1, i_2, i_3, i_4, according to the type of M_f. (Please refer to the proof of the necessity to see how to choose this.) Then we can choose j_s, $s = 1, \cdots, 2^s$, according to the type of M_s. Finally, M_ψ can be determined automatically.

□

Remark 12.1.

(1) We can state Theorem 12.1 alternatively as: f is bi-decomposed with respect to Γ and Λ (as aforementioned), if and only if the structure matrix of f can be expressed as

$$M_f = \left[\mu_1 M_\psi\ \mu_2 M_\psi\ \cdots\ \mu_{2^k} M_\psi\right], \tag{12.9}$$

where

$$M_\psi \in \mathcal{L}_{2\times 2^{n-k}};$$

$\mu_i \in S$, $\forall i$, where S can be one of the following types:

- **Type 1**:

$$S = S_1 = \{\delta_2[1\ 1],\ \delta_2[2\ 2]\};$$

- **Type 2**:

$$S = S_2 = \{\delta_2[1\ 1],\ \delta_2[1\ 2]\} \text{ or } \{\delta_2[2\ 2],\ \delta_2[1\ 2]\};$$

- **Type 3**:

$$S = S_3 = \{\delta_2[1\ 2]\} \text{ or } \{\delta_2[2\ 1]\}\,;$$

- **Type 4**:

$$S = S_4 = \{\delta_2[1\ 2],\ \delta_2[2\ 1]\}\,.$$

(2) We ignore the case when the type consists of only one constant matrix, say $\delta_2[1\ 1]$ (or $\delta_2[2\ 2]$). Because in this case f is a constant mapping, i.e., $f \equiv 1$ or $f \equiv 0$.

(3) Actually, Type 2 may have two other cases $S_2 = \{\delta_2[1\ 1],\ \delta_2[2\ 1]\}$ or $S_2 = \{\delta_2[2\ 2],\ \delta_2[2\ 1]\}$. But they can be realized by using the above S and replacing ψ by $\neg\psi$.

(4) Type 3 ($S = S_3$) is a trivial case, because it means f is independent of ϕ. So we may ignore this trivial case.

(5) This result is basically the same as the one in Sasao and Butler (1997).

Note that from the structure matrix M_f of f it is easy to figure out the set $\{\mu_1, \cdots, \mu_{2^k}\}$ provided the conditions of Theorem 12.1 are satisfied. Because we can first find constant function matrices (CFM). If there are two CFMs, we are done. If there is only one CFM, then there is only one non-constant function matrix, and we can choose another μ as $\mu = \delta_2[1\ 2]$. If there is no CFM, we should have $\mu_1 = \delta_2[1\ 2]$ and $\mu_2 = \delta_2[2\ 1]$.

The following corollary gives the way to construct the decomposition.

Corollary 12.1. *Assume the structure matrix M_f of f, as in (12.9), satisfies the conditions of Theorem 12.1. Then the structure matrices of F, ϕ, and ψ can be figured out by the following process:*

(1) If the set $\{\mu_1, \cdots, \mu_{2^k}\}$ contains only one element $\delta_2[p, q]$, then

$$M_F = [\delta_2[p, q]\ \delta_2[p, q]]\,; \tag{12.10}$$

otherwise the set contains two elements $\delta_2[p_1, q_1]$, $\delta_2[p_2, q_2]$, then

$$M_F = [\delta_2[p_1, q_1]\ \delta_2[p_2, q_2]]\,. \tag{12.11}$$

(2) Say,

$$\mu_i = M_F \delta_2^{t_i}, \quad i = 1, \cdots, 2^k,$$

then

$$M_\phi = \delta_2\left[t_1\ t_2\ \cdots\ t_{2^k}\right]. \tag{12.12}$$

(3) M_ψ can be constructed by (12.9).

Using these M_F, M_ϕ, and M_ψ, we can construct the decomposition.

Next, we discuss the general case where (12.4) is not true. That is, $\{\Gamma, \Lambda\}$ is an arbitrary partition of $\{1, 2, \cdots, n\}$ (with $\Gamma \neq \varnothing$ and $\Lambda \neq \varnothing$). First, the order of ϕ and ψ does not matter, because, say,

$$f(X) = F(\phi(X_1), \psi(X_2))$$

has its algebraic form as $f(x) = M_f x$, then

$$M_f x = M_F M_\phi x^1 M_\psi x^2 = M_F W_{[2,2]} M_\psi x^2 M_\phi x^1 = \tilde{F}(\phi(X_2), \psi(X_1)),$$

where $\tilde{F}$ has its structure matrix as $M_{\tilde{F}} := M_F W_{[2,2]}$. Now since we consider all possible M_F, and $M_{\tilde{F}} := M_F W_{[2,2]}$ is another possible M_F, with this $M_{\tilde{F}}$, the order of ϕ and ψ has been reversed. Based on this consideration we can choose $k \leq n/2$ variables as the arguments of the second function ψ. We conclude that

Proposition 12.1. *Let $n_0 = \left[\frac{n}{2}\right]$, where $[r]$ denotes the largest integer $s \leq r$. Then there are*

$$\binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n_0} \tag{12.13}$$

possible Λ, which contain $1, 2, \cdots, n_0$ possible arguments of ψ, corresponding to each term in (12.13).

Let $\Lambda = \{j_1, j_2, \cdots, j_s\}$ be the selected variables, where $s \leq n_0$ and $j_1 < j_2 < \cdots j_s$, and $\Gamma = \Lambda^c$. Denote

$$x^1 := \ltimes_{i\in\Gamma} x_i; \quad \text{and} \quad x^2 = \ltimes_{i=1}^{s} x_{j_i}.$$

Then we have

$$\begin{aligned} f(x_1, \cdots, x_n) &= M_f \ltimes_{i=1}^{n} x_i \\ &= M_f W_{[2,2^{j_s-1}]} x_{j_s} \ltimes_{i\neq j_s} x_i \\ &= M_f W_{[2,2^{j_s-1}]} W_{[2,2^{j_{s-1}}]} x_{j_{s-1}} x_{j_s} \ltimes_{i\notin\{j_s, j_{s-1}\}} x_i \\ &= \cdots \\ &= M_f W_{[2,2^{j_s-1}]} W_{[2,2^{j_{s-1}}]} \cdots W_{[2,2^{j_1+(s-2)}]} \ltimes_{i=1}^{s} x_{j_i} \ltimes_{i\in\Gamma} x_i \\ &= M_f W_{[2,2^{j_s-1}]} W_{[2,2^{j_{s-1}}]} \cdots W_{[2,2^{j_1+(s-2)}]} W_{[2^{n-s},2^s]} x^1 x^2. \end{aligned}$$

From the above argument we have the following result, which tells us when the arguments for ψ are chosen, how to use Theorem 12.1 (or Remark 12.1) to check possible decomposition.

Theorem 12.2. *Using above notations, when $\{x_{j_1}, \cdots, x_{j_s}\}$ are chosen as the arguments for possible ψ, the structure matrix, corresponding to $x^1 = \ltimes_{i\in\Gamma} x_i$ and $x^2 = \ltimes_{i\in\Lambda} x_i$, is*

$$\tilde{M}_f = M_f \Psi, \tag{12.14}$$

where $\Psi \in \mathcal{L}_{2^n \times 2^n}$, *called the rearrangement matrix, is*

$$\Psi = \ltimes_{k=s}^{1} W_{[2,2^{j_k+(s-1)-k}]} \ltimes W_{[2^{n-s},2^s]}.$$

Example 12.1.

(1) Assume

$$f(x_1, x_2, x_3, x_4) = (x_1 \leftrightarrow x_2) \vee (x_3 \wedge x_4).$$

Here f is already in bi-decomposed form. We check the conditions of Theorem 12.1. First, we have

$$M_f = \delta_2[1\ 1\ 1\ 1\ \ 1\ 2\ 2\ 2\ \ 1\ 2\ 2\ 2\ \ 1\ 1\ 1\ 1]. \qquad (12.15)$$

It is of the type of $S = S_2$. Choose $\delta_2[i_1\ i_2] = \delta_2[1\ 1]$ and $\delta_2[i_3\ i_4] = \delta_2[1\ 2]$, then $M_F = \delta_2[1\ 1\ 1\ 2]$. Let $M_\phi = \delta_2[j_1\ j_2\ j_3\ j_4]$, it is clear that $j_1 = j_4 = 1$ and $j_2 = j_3 = 2$; and for M_ψ we have $M_\psi = [1\ 2\ 2\ 2]$.
Note that we can also choose $\delta_2[i_1\ i_2] = \delta_2[1\ 2]$ and $\delta_2[i_3\ i_4] = \delta_2[1\ 1]$. Then new $\tilde{M}_\phi = \neg M_\phi$.

(2) Assume a Boolean function $f(x_1, x_2, x_3, x_4)$ has its structure matrix as

$$M_f = \delta_2[1\ 2\ 2\ 2\ \ 2\ 1\ 1\ 1\ \ 1\ 2\ 2\ 2\ \ 1\ 2\ 2\ 2]. \qquad (12.16)$$

We consider the possible decomposition. Obviously, it is of type $S = S_4$. Choosing $\delta_2[i_1\ \ i_2] = \delta_2[1\ \ 2]$ and $\delta_2[i_3\ \ i_4] = \delta_2[2\ \ 1]$, then $M_\phi = \delta_2[1\ 2\ 1\ 1]$ and $M_\psi = [1\ 2\ 2\ 2]$. It follows that f can be decomposed as

$$f(x_1, x_2, x_3, x_4) = [(x_1 \wedge x_2) \vee (\neg x_1)] \leftrightarrow (x_3 \wedge x_4).$$

(3) Assume a Boolean function $f(x_1, x_2, x_3, x_4, x_5)$ has its structure matrix as

$$M_f = \delta_2[1\ 1\ 2\ 2\ \ 1\ 2\ 2\ 1\ \ 1\ 1\ 2\ 2\ \ 1\ 1\ 2\ 2\ \ 2\ 2\ 1\ 1\ \ 2\ 1\ 1\ 2\ \ 2\ 2\ 1\ 1\ \ 2\ 2\ 1\ 1]. \qquad (12.17)$$

It is obvious that M_f does not satisfy the requirements of Theorem 12.1.
Next, we try to choose proper variable(s) for the argument(s) of ψ. By trial-and-error, we choose $\{x_1, x_4\}$. Using (12.14), we have

$$\begin{aligned}\tilde{M}_f &= M_f W_{[2,2^3]} W_{[2,2]} W_{[2^3,2^2]}\\ &= \delta_2[1\ 2\ 2\ 1\ \ 1\ 2\ 2\ 1\ \ 1\ 2\ 2\ 1\ \ 2\ 1\ 1\ 2\ \ 1\ 2\ 2\ 1\ \ 1\ 2\ 2\ 1\ \ 1\ 2\ 2\ 1\ \ 1\ 2\ 2\ 1].\end{aligned}$$

It is clear that $\tilde{M}_f$ is of Type 4, and it can be easily constructed as

$$\begin{aligned}f(x) &= \tilde{M}_f x_2 x_3 x_5 x_1 x_4 = M_f M_\phi x_2 x_3 x_5 M_\psi x_1 x_4\\ &= \delta_2[1\ 2\ 2\ 1]\delta_2[1\ 1\ 1\ 2\ 1\ 1\ 1\ 1] x_2 x_3 x_5 \delta_2[1\ 2\ 2\ 1] x_1 x_4.\end{aligned}$$

Since $M_F = \delta_2[1\ 2\ 2\ 1]$, we have

$$f(x) = \phi(x_2, x_3, x_5) \leftrightarrow \psi(x_1, x_4).$$

Since $M_\phi = \delta_2[1\ 1\ 1\ 2\ 1\ 1\ 1\ 1]$,

$$\phi(x_2, x_3, x_5) = [x_2 \wedge (x_3 \vee x_5)] \vee \neg x_2.$$

Since $M_\psi = \delta_2[1\ 2\ 2\ 1]$,

$$\psi(x_1, x_4) = x_1 \leftrightarrow x_4.$$

Finally, $f(x)$ has the decomposed form as

$$f(x) = \{[x_2 \wedge (x_3 \vee x_5)] \vee \neg x_2\} \leftrightarrow \{x_1 \leftrightarrow x_4\}.$$

12.2 Non-Disjoint Bi-Decomposition

Definition 12.3. Let $f : \mathcal{D}^n \to \mathcal{D}$ be a Boolean function, $\Gamma \cup \Theta \cup \Lambda$ be a partition of $\{1, 2, \cdots, n\}$. f is said to be bi-decomposed with respect to $\Gamma \cup \Theta$ and $\Lambda \cup \Theta$ if there exist three Boolean functions $F : \mathcal{D}^2 \to \mathcal{D}$, $\phi : \{x_\gamma \,|\, \gamma \in \Gamma \cup \Theta\} \to \mathcal{D}$, and $\psi : \{x_\lambda \,|\, \lambda \in \Theta \cup \Lambda\} \to \mathcal{D}$, such that

$$f(x_1, \cdots, x_n) = F(\phi(x_\gamma \,|\, \gamma \in \Gamma \cup \Theta), \psi(x_\lambda \,|\, \lambda \in \Theta \cup \Lambda). \tag{12.18}$$

For statement ease, let

$$\begin{aligned} X^1 &= \{x^1_1, \cdots, x^1_{k_1}\} = \{x_i | i \in \Gamma\}; \\ X^2 &= \{x^2_1, \cdots, x^2_{k_2}\} = \{x_i | i \in \Theta\}; \\ X^3 &= \{x^3_1, \cdots, x^3_{k_3}\} = \{x_i | i \in \Lambda\}. \end{aligned} \tag{12.19}$$

Theorem 12.3. *Let $f : \mathcal{D}^n \to \mathcal{D}$ be a Boolean function with its structure matrix M_f. f can be decomposed as in (12.18), if and only if its structure matrix can be expressed as*

$$\begin{aligned} M_f = \Big[&\mu_{1,1} M^1_\psi \ \ \mu_{1,2} M^2_\psi \ \ \cdots \ \ \mu_{1,2^{k_2}} M^{2^{k_2}}_\psi \\ &\mu_{2,1} M^1_\psi \ \ \mu_{2,2} M^2_\psi \ \ \cdots \ \ \mu_{2,2^{k_2}} M^{2^{k_2}}_\psi \\ &\vdots \\ &\mu_{2^{k_1},1} M^1_\psi \ \ \mu_{2^{k_1},2} M^2_\psi \ \ \cdots \ \ \mu_{2^{k_1},2^{k_2}} M^{2^{k_2}}_\psi \Big], \end{aligned} \tag{12.20}$$

where each

$$M^s_\psi \in \mathcal{L}_{2 \times 2^{k_3}}, \quad s = 1, \cdots, 2^{k_2};$$

$$\mu_{i,j} \in S, \quad i = 1, \cdots, 2^{k_1},\ j = 1, \cdots, 2^{k_2},$$

S equals to one of the S_1, S_2, S_3, or S_4, which are defined in Remark 12.1.

Proof. (Necessity) Assume there are three functions F, ϕ, and ψ, such that (12.18) holds.

Assume the structure matrix of F is

$$M_F = \delta_2[i_1\ i_2\ i_3\ i_4];$$

the structure matrix of ϕ is

$$M_\phi = \delta_2[j_1\ j_2\ \cdots\ j_{2^{k_1+k_2}}];$$

and the structure matrix of ψ is expressed as

$$M_\psi = \left[M_\psi^1\ M_\psi^2\ \cdots\ M_\psi^{2^{k_2}}\right] \in \mathcal{L}_{2\times 2^{k_2+k_3}},$$

where

$$M_\psi^i \in \mathcal{L}_{2\times 2^{k_3}}, \quad i = 1,\cdots,2^{k_2}.$$

Then we have

$$M_f x = M_F M_\phi x^1 x^2 M_\psi x^2 x^3, \tag{12.21}$$

where $x = \ltimes_{i=1}^n x_i$, $x^j = \ltimes_{i=1}^{k_j} x_i^j$, $j = 1,2,3$.

Converting the RHS of (12.21) into normal form, we have that

$$M_f = M_F M_\phi \left(I_{2^{k_1+k_2}} \otimes M_\psi\right)\left(I_{2^{k_1}} \otimes M_r^{2^{k_2}}\right). \tag{12.22}$$

We first calculate $M_F M_\phi$, which is denoted as

$$M_F M_\phi := [N_1\ N_2\ \cdots\ N_{2^{k_1+k_2}}]. \tag{12.23}$$

Similar to the disjoint case, we have

$$N_s = \begin{cases} \delta_2[i_1\ i_2], & j_s = 1 \\ \delta_2[i_3\ i_4], & j_s = 2, \end{cases} \qquad s = 1,2,\cdots,2^{k_1+k_2}. \tag{12.24}$$

Next, we calculate $\left(I_{2^{k_1+k_2}} \otimes M_\psi\right)\left(I_{2^{k_1}} \otimes M_r^{2^{k_2}}\right)$:

$$\left(I_{2^{k_1+k_2}} \otimes M_\psi\right)\left(I_{2^{k_1}} \otimes M_r^{2^{k_2}}\right) = I_{2^{k_1}} \otimes \left[\left(I_{2^{k_2}} \otimes M_\psi\right) M_r^{2^{k_2}}\right]. \tag{12.25}$$

We simplify $\left(I_{2^{k_2}} \otimes M_\psi\right) M_r^{2^{k_2}}$ first. Note that $\left(I_{2^{k_2}} \otimes M_\psi\right) \in \mathcal{L}_{2^{k_2+1}\times 2^{2k_2+k_3}}$ and $M_r^{2^{k_2}} \in \mathcal{L}_{2^{2k_2}\times 2^{k_2}}$, converting them back to conventional matrix product we have

$$\left(I_{2^{k_2}} \otimes M_\psi\right) M_r^{2^{k_2}} = \left(I_{2^{k_2}} \otimes M_\psi\right)\left(M_r^{2^{k_2}} \otimes I_{2^{k_3}}\right), \tag{12.26}$$

and

$$I_{2^{k_2}} \otimes M_\psi = \left.\begin{bmatrix} M_\psi & 0 & \cdots & 0 \\ 0 & M_\psi & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & M_\psi \end{bmatrix}\right\} 2^{k_2};$$

$$M_r^{2^{k_2}} \otimes I_{2^{k_3}} = \begin{bmatrix} \left.\begin{bmatrix} I_{2^{k_3}} \\ 0 \\ \vdots \\ 0 \end{bmatrix}\right\} 2^{k_2} & 0 & \cdots & 0 \\ 0 & \left.\begin{bmatrix} 0 \\ I_{2^{k_3}} \\ \vdots \\ 0 \end{bmatrix}\right\} 2^{k_2} & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \left.\begin{bmatrix} 0 \\ 0 \\ \vdots \\ I_{2^{k_3}} \end{bmatrix}\right\} 2^{k_2} \end{bmatrix}.$$

It follows that

$$\left(I_{2^{k_2}} \otimes M_\psi\right) M_r^{2^{k_2}} = \begin{bmatrix} M_\psi^1 & 0 & \cdots & 0 \\ 0 & M_\psi^2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & M_\psi^{2^{k_2}} \end{bmatrix}. \tag{12.27}$$

Putting (12.25), (12.26), and (12.27) together, (12.20) follows immediately.

(sufficiency) Using

$$M_\psi = \left[M_\psi^1 \; M_\psi^2 \; \cdots \; M_\psi^{2^{k_2}}\right]$$

as the structure matrix of ψ yields ψ. Denote

$$M_\phi = \left[M_\phi^{1,1} \; \cdots \; M_\phi^{1,k_2} \; \cdots \; M_\phi^{2^{k_1},1} \; \cdots \; M_\phi^{2^{k_1},2^{k_2}}\right].$$

According to $\mu_{\alpha,\beta}$ we can uniquely determine $M_\phi^{\alpha,\beta}$. Precisely, we set

$$M_\phi^{\alpha,\beta} = \begin{cases} \delta_2^1, & \mu_{\alpha,\beta} = \delta_2[i_1, i_2] \\ \delta_2^2, & \mu_{\alpha,\beta} = \delta_2[i_3, i_4]. \end{cases}$$

Using this pair of $\{\phi, \psi\}$, it is easy to check that the factorization (12.18) holds. □

Remark 12.2. The explicit expression in Theorem 12.3 is an improvement of the known implicit form in Sasao and Butler (1997), because it is not only straightforward verifiable but also easily used to construct the decomposition.

The following corollary provides a procedure to construct the decomposition.

Corollary 12.2. *Assume the structure matrix M_f of f, as in (12.20), satisfies the conditions of Theorem 12.3. Then we have the following structures for F, ψ, and ψ respectively.*

(1) If the set $\{\mu_{1,1}, \cdots, \mu_{2^{k_1},2^{k_2}}\}$ contains only one element $\delta_2[p,q]$, then

$$M_F = [\delta_2[p,q] \; \delta_2[p,q]]\,; \tag{12.28}$$

otherwise assume the set contains two elements $\delta_2[p_1,q_1]$, $\delta_2[p_2,q_2]$, then

$$M_F = [\delta_2[p_1,q_1] \; \delta_2[p_2,q_2]]\,. \tag{12.29}$$

(2) Consider $\mu_{i,j}$. If

$$\mu_{i,j} = M_F \delta_2^{t_{i,j}}, \quad i = 1, \cdots, 2^{k_1};\; j = 1, \cdots, 2^{k_2},$$

then

$$M_\phi = \delta_2 \left[t_{1,1} \;\cdots\; t_{1,2^{k_2}} \;\;\cdots\;\; t_{2^{k_1},1} \;\cdots\; t_{2^{k_1},2^{k_2}} \right]. \tag{12.30}$$

(3)

$$M_\psi = \left[M_\psi^1 \; M_\psi^2 \;\cdots\; M_\psi^{2^{k_2}} \right]. \tag{12.31}$$

Using these M_F, M_ϕ, and M_ψ, we can construct the decomposition.

Remark 12.3. As for arbitrary order variables, the basic idea of Theorem 12.2 remains applicable. When

$$\Theta = \{x^2\} = \{x_{j_1}, \cdots, x_{j_{k_2}}\}; \quad \text{and} \quad \Lambda = \{x_{j_{k_2+1}}, \cdots, x_{j_{k_3}}\}$$

are chosen as the arguments x^2 and x^3 of possible ψ, the structure matrix corresponding to $x^1 = \ltimes_{i\in\Gamma} x_i$, $x^2 = \ltimes_{i\in\Theta} x_i$ and $x^3 = \ltimes_{i\in\Lambda} x_i$, is

$$\begin{aligned} \tilde{M}_f = M_f \ltimes_{i=k_3}^{1} W_{[2,2^{j_i+(k_3-1)-i}]} \ltimes W_{[2^{k_1+k_2},2^{k_3}]} \\ \ltimes_{i=k_2}^{1} W_{[2,2^{j_i+(k_2-1)-i}]} \ltimes W_{[2^{k_1},2^{k_2}]}. \end{aligned} \tag{12.32}$$

We give two examples to depict Theorem 12.3 and Remark 12.3.

Example 12.2.

(1) Consider a Boolean function $f(x_1, x_2, x_3, x_4, x_5, x_6)$ with its structure matrix as

$$M_f = \delta_2[\ 2\ 1\ 1\ 2\ \ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ \ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 2\ \ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ \ 1\ 1\ 1\ 1 \\ 2\ 1\ 1\ 2\ \ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ \ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 2\ \ 1\ 2\ 1\ 2\ 2\ 2\ 1\ 1\ \ 1\ 1\ 1\ 1].$$

Obviously, it is of type $S = S_2$. Choose $\delta_2[i_1\ i_2] = \delta_2[1\ 2]$ and $\delta_2[i_3\ i_4] = \delta_2[2\ 2]$, then $M_F = \delta_2[1\ 2\ 2\ 2]$. It can be easily seen that

$$M_\phi = \delta_2[1\ 2\ 1\ 1\ 1\ 2\ 1\ 1\ 1\ 2\ 1\ 1\ 1\ 1\ 1\ 1],$$
$$M_\psi = \delta_2[2\ 1\ 1\ 2\ 1\ 2\ 1\ 2\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 1].$$

Now since $M_F = \delta_2[1\ 2\ 2\ 2]$, we have

$$f(x_1, x_2, x_3, x_4, x_5, x_6) = \phi(x_1, x_2, x_3, x_4) \wedge \psi(x_3, x_4, x_5, x_6).$$

The functions ϕ and ψ can be constructed via their structure matrices via standard procedure. Finally, f can be decomposed as

$$f(x_1, x_2, x_3, x_4, x_5, x_6) \\ = [(x_1 \vee x_2) \wedge x_3 \to x_4] \wedge [(x_4 \wedge x_5) \leftrightarrow \neg(x_3 \to x_6)].$$

(2) Assume a Boolean function $f(x_1, x_2, x_3, x_4, x_5)$ has its structure matrix as

$$M_f = \delta_2[1\ 2\ 1\ 2\ \ 2\ 2\ 2\ 2\ \ 1\ 1\ 2\ 1\ \ 2\ 1\ 2\ 1\ \ 1\ 1\ 1\ 1\ \ 1\ 1\ 1\ 1\ \ 1\ 2\ 2\ 2\ \ 1\ 2\ 1\ 2].$$

It is obvious that M_f does not satisfy the requirements of Theorem 12.3.

Hence, we try to choose proper variable(s) for the argument(s) of possible ψ. Say, we choose $\Theta = \{3, 5\}, \Lambda = \{1\}$. Using (12.32), we have

$$\tilde{M}_f = M_f W_{[2,2^4]} W_{[2,2^3]} W_{[2,2^4]} W_{[2,2^4]} = \\ \delta_2[1\ 1\ 2\ 1\ \ 2\ 1\ 2\ 1\ \ 1\ 1\ 2\ 1\ \ 2\ 1\ 2\ 1\ \ 1\ 1\ 1\ 2\ \ 2\ 1\ 1\ 2\ \ 2\ 2\ 1\ 2\ \ 2\ 1\ 1\ 2].$$

It is clear that $\tilde{M}_f$ is of Type 4: $\delta_2[i_1\ i_2] = \delta_2[1\ 2]$ and $\delta_2[i_3\ i_4] = \delta_2[2\ 1]$. So $M_F = \delta_2[1\ 2\ 2\ 1]$, and we have

$$f(x) = \phi(x_2, x_3, x_4, x_5) \leftrightarrow \psi(x_1, x_3, x_5).$$

Following the procedure in Corollary 12.2, we have $M_\phi = \delta_2[1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 2\ 2\ 2\ 1\ 2]$ and $M_\psi = \delta_2[1\ 1\ 2\ 1\ 2\ 1\ 2\ 1]$. From M_ϕ and M_ψ we can get

$$\phi(x_2, x_3, x_4, x_5) = x_2 \vee [(x_3 \to x_4) \wedge x_5].$$

and

$$\psi(x_1, x_3, x_5) = \neg x_1 \vee (x_3 \wedge x_5).$$

Thus, finally, $f(x)$ has the decomposed form as

$$f(x) = \{x_2 \vee [(x_3 \to x_4) \wedge x_5]\} \leftrightarrow [\neg x_1 \vee (x_3 \wedge x_5)].$$

12.3 Decomposition of Multi-Valued Logical Functions

This section considers the decomposition of multi-valued logical functions. We look for a general formula as in Boolean case.

Let $f(x_1, x_2, \cdots, x_n)$ be an r-valued logical function. Identifying

$$\delta_r^i \sim \frac{r-i}{r-1}, \quad i = 1, 2, \cdots, r,$$

we can express $f : \Delta_{r^n} \to \Delta_r$, which is called the vector form of f. Similar to Boolean case, we know that in vector form there exists a unique $M_f \in \mathcal{L}_{r \times r^n}$, called the structure matrix of f, such that

$$f(x_1, \cdots, x_n) = M_f x, \tag{12.33}$$

where $x = \ltimes_{i=1}^n x_i$.

Definition 12.4. Let $f : \mathcal{D}_r^n \to \mathcal{D}_r$ be an r-valued logical function.

(1) Assume $\Gamma \cup \Lambda$ is a partition of $\{1, 2, \cdots, n\}$. f is said to be bi-decomposable with respect to Γ and Λ if there exist three r-valued logical functions $F : \mathcal{D}_r^2 \to \mathcal{D}_r$, $\phi : \{x_\gamma | \gamma \in \Gamma\} \to \mathcal{D}_r$, and $\psi : \{x_\lambda | \lambda \in \Lambda\} \to \mathcal{D}_r$, such that

$$f(x_1, \cdots, x_n) = F\left(\phi(x_\gamma | \gamma \in \Gamma), \psi(x_\lambda | \lambda \in \Lambda)\right). \tag{12.34}$$

(2) Assume $\Gamma \cup \Theta \cup \Lambda$ is a partition of $\{1, 2, \cdots, n\}$. f is said to be bi-decomposable with respect to $\Gamma \cup \Theta$ and $\Lambda \cup \Theta$ if there exist three Boolean functions $F : \mathcal{D}_r^2 \to \mathcal{D}_r$, $\phi : \{x_\gamma | \gamma \in \Gamma \cup \Theta\} \to \mathcal{D}_r$, and $\psi : \{x_\lambda | \lambda \in \Theta \cup \Lambda\} \to \mathcal{D}_r$, such that

$$f(x_1, \cdots, x_n) = F\left(\phi(x_\gamma | \gamma \in \Gamma \cup \Theta), \psi(x_\lambda | \lambda \in \Theta \cup \Lambda\right). \tag{12.35}$$

First we consider the disjoint case. It is clear that there are r^r mappings from $\mathcal{D}_r \to \mathcal{D}_r$. In vector form they can be expressed as

$$b_i = T_i x, \quad i = 1, 2, \cdots, r^r,$$

where $x, b_i \in \Delta_r$ and $T_i \in \mathcal{L}_{r \times r}$.

We use $\{T_i\}$ to describe F. Choosing r elements from $\mathcal{L}_{r \times r}$, say,

$$\mathcal{T} = \{T_1, T_2, \cdots, T_r\} \subset \mathcal{L}_{r \times r},$$

then we say that F has Type $\mathcal{T}$, if the structure matrix of F is

$$M_F = [T_1 \ T_2 \ \cdots \ T_r].$$

As we see in Boolean case, the order of $\{T_i \,|\, i = 1, \cdots, r^2\}$ does not affect the decomposition.

Similar to (12.5), we first split M_f into r^k equal blocks as

$$M_f = [M_1 \; M_2 \; \cdots \; M_{r^k}]. \tag{12.36}$$

Then we can prove the following:

Theorem 12.4. *Let $f : \mathcal{D}^n \to \mathcal{D}$ be an r-valued logical function with its structure matrix M_f, being split into r^k equal blocks as in (12.36). Assume Γ and Λ form a partition as in (12.4). f is decomposable with respect to the partition in (12.4), if and only if, there exist*

(i) a type $\mathcal{T} = \{T_1, T_2, \cdots, T_r\} \subset \mathcal{L}_{r\times r}$,
(ii) a logical matrix $M_\psi \in \mathcal{L}_{r\times r^{n-k}}$,

such that

$$M_i = T_{s_i} M_\psi, \quad \textit{where } T_{s_i} \in \mathcal{T}, \;\; i = 1, \cdots, r^k. \tag{12.37}$$

Remark 12.4.

(1) The number of types for r-valued logical functions is

$$N_r = \binom{r^r}{r} = \frac{r^r!}{r!(r^e - r)!},$$

which is a large number. For instance, when $r = 3$ the $N_3 = 2925$, when $r = 4$ the $N_4 = 174792640$ etc. It is very difficult to verify all such types. For practical circuit design, we may only be interested in some particular types. For instance, the most commonly used F is either $\vee$ or $\wedge$. It is easy to figure out that their corresponding types are

- $r = 3$

$$\begin{aligned} \mathcal{T}_\vee &= \{\delta_3[1\;1\;1], \delta_3[1\;2\;2], \delta_3[1\;2\;3]\}; \\ \mathcal{T}_\wedge &= \{\delta_3[1\;2\;3], \delta_3[2\;2\;3], \delta_3[3\;3\;3]\}. \end{aligned} \tag{12.38}$$

- $r = 4$

$$\begin{aligned} \mathcal{T}_\vee &= \{\delta_4[1\;1\;1\;1], \delta_4[1\;2\;2\;2], \delta_4[1\;2\;3\;3], \delta_4[1\;2\;3\;4]\}; \\ \mathcal{T}_\wedge &= \{\delta_4[1\;2\;3\;4], \delta_4[2\;2\;3\;4], \delta_4[3\;3\;3\;4], \delta_4[4\;4\;4\;4]\}. \end{aligned} \tag{12.39}$$

(2) If the partition is in arbitrary order the re-ordering result, Theorem 12.4, remains applicable via replacing (12.5) by the following equation (12.40).

$$\tilde{M}_f = M_f \ltimes_{k=s}^{1} W_{[r, r^{j_k+(s-1)-k}]} \ltimes W_{[r^{n-s}, r^s]}. \tag{12.40}$$

Example 12.3. Let $f(x_1, x_2, x_3, x_4) : \mathcal{D}_3^4 \to \mathcal{D}_3$ be a 3-valued logical function with its structure matrix as

$$\begin{aligned} M_f = \delta_3[\,&1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ \ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ \ 1\ 2\ 3\ 2\ 2\ 2\ 3\ 2\ 1\\ &1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ \ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ \ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\\ &1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ \ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ \ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\,]. \end{aligned} \tag{12.41}$$

We try $\mathcal{T} = \mathcal{T}_\vee$ as in (12.38), that is,

$$\mathcal{T} = \{T_1 = \delta_3[1\ 1\ 1],\ T_2 = \delta_3[1\ 2\ 2],\ T_3 = \delta_3[1\ 2\ 3]\};$$

and choose

$$M_\psi = \delta_3[1\ 2\ 3\ 2\ 2\ 2\ 3\ 2\ 1].$$

It is easy to check that

$$\begin{aligned} T_1 M_\psi &= [1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1],\\ T_2 M_\psi &= [1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1],\\ T_3 M_\psi &= [1\ 2\ 3\ 2\ 2\ 2\ 3\ 2\ 1]. \end{aligned}$$

Comparing each block of M_f with above product forms, it follows immediately that

$$M_\phi = \delta_3[1\ 2\ 3\ 1\ 2\ 2\ 1\ 1\ 1].$$

If we define the 3-valued logical operators → and ↔ by using the corresponding formulas of Boolean functions as

$$\begin{aligned} A \to B &:= (A \wedge B) \vee \neg A,\\ A \leftrightarrow B &:= (A \to B) \wedge (B \to A), \end{aligned}$$

then it is easy to verify that

$$M_\to = M_\phi, \quad \text{and } \ M_\leftrightarrow = M_\psi.$$

Eventually, we have the decomposed f as

$$f(x_1, x_2, x_3, x_4) = (x_1 \to x_2) \vee (x_3 \leftrightarrow x_4)\,.$$

Next, we consider non-disjoint case.

Definition 12.5. Let $f : \mathcal{D}_r^n \to \mathcal{D}_r$ be an r-valued logical function, $\Gamma\cup\Theta\cup\Lambda$ be a partition of $\{1, 2, \cdots, n\}$. f is decomposable with respect to $\Gamma \cup \Theta$ and $\Lambda \cup \Theta$ (as in (12.18)), if there exist three r-valued logical functions $F : \mathcal{D}_r^2 \to \mathcal{D}_r$, $\phi : \{x_\gamma | \gamma \in \Gamma \cup \Theta\} \to \mathcal{D}_r$, and $\psi : \{x_\lambda | \lambda \in \Theta \cup \Lambda\} \to \mathcal{D}_r$, such that

$$f(x_1, \cdots, x_n) = F(\phi(x_\gamma | \gamma \in \Gamma \cup \Theta), \psi(x_\lambda | \lambda \in \Theta \cup \Lambda). \tag{12.42}$$

Theorem 12.5. *Let $f : \mathcal{D}_r^n \to \mathcal{D}_r$ be an r-valued logical function with its structure matrix M_f. f can be decomposed as in (12.42) with respect to the partition as in (12.19), if and only if*

(i) there exists a type $\mathcal{T} \subset \mathcal{L}_{r\times r}$,
(ii) there exist

$$M_\psi^i \in \mathcal{L}_{r\times r^{k_3}}, \quad i = 1, \cdots, r^{k_2}, \tag{12.43}$$

such that the structure matrix of f can be expressed as

$$\begin{aligned} M_f = \Big[&\mu_{1,1}M_\psi^1 \;\; \mu_{1,2}M_\psi^2 \;\; \cdots \;\; \mu_{1,r^{k_2}}M_\psi^{r^{k_2}} \\ &\mu_{2,1}M_\psi^1 \;\; \mu_{2,2}M_\psi^2 \;\; \cdots \;\; \mu_{2,r^{k_2}}M_\psi^{r^{k_2}} \\ &\vdots \\ &\mu_{r^{k_1},1}M_\psi^1 \;\; \mu_{r^{k_1},2}M_\psi^2 \;\; \cdots \;\; \mu_{r^{k_1},r^{k_2}}M_\psi^{r^{k_2}}\Big], \end{aligned} \tag{12.44}$$

where

$$\mu_{i,j} \in \mathcal{T}, \quad i = 1, \cdots, r^{k_1}, \; j = 1, \cdots, r^{k_2}.$$

Remark 12.5. Similar to Corollary 12.1 for disjoint case (Corollary 12.2 for non-disjoint case), when the conditions in part 1 (part 2) of Theorem 12.5 are satisfied the corresponding decomposition can be easily constructed by using the structure matrices M_F, M_ϕ, and M_ψ.

Example 12.4. Let $f(x_1, x_2, x_3, x_4) : \mathcal{D}_3^4 \to \mathcal{D}_3$ be a 3-valued logical function with its structure matrix as

$$\begin{aligned} M_f = \delta_3[&1\ 1\ 1\ 2\ 2\ 2\ 3\ 3\ 3\ \ 2\ 2\ 1\ 2\ 2\ 2\ 3\ 3\ 3\ \ 3\ 2\ 1\ 3\ 2\ 2\ 3\ 3\ 3 \\ &1\ 1\ 1\ 2\ 2\ 2\ 3\ 3\ 3\ \ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ \ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2 \\ &1\ 1\ 1\ 2\ 2\ 2\ 3\ 3\ 3\ \ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ \ 1\ 2\ 3\ 1\ 2\ 2\ 1\ 1\ 1]. \end{aligned}$$

If we define the 3-valued logical operators $\leftrightarrow$ as in Example 12.3 above, and we try $\mathcal{T} = \mathcal{T}_\leftrightarrow$ as in (12.38), that is,

$$\mathcal{T} = \{T_1 = \delta_3[1\ 2\ 3],\ T_2 = \delta_3[2\ 2\ 2],\ T_3 = \delta_3[3\ 2\ 1]\}.$$

Choosing

$$M_\psi^1 = \delta_3[1\ 1\ 1\ 2\ 2\ 2\ 3\ 3\ 3],$$

it is easy to check that

$$\begin{aligned} T_1M_\psi^1 &= [1\ 1\ 1\ 2\ 2\ 2\ 3\ 3\ 3], \\ T_2M_\psi^1 &= [2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2], \\ T_3M_\psi^1 &= [3\ 3\ 3\ 2\ 2\ 2\ 1\ 1\ 1]. \end{aligned}$$

Similarly, choosing

$$M_\psi^2 = \delta_3[2\ 2\ 1\ 2\ 2\ 2\ 3\ 3\ 3]$$

yields

$$\begin{aligned} T_1 M_\psi^2 &= [2\ 2\ 1\ 2\ 2\ 2\ 3\ 3\ 3],\\ T_2 M_\psi^2 &= [2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2],\\ T_3 M_\psi^2 &= [2\ 2\ 3\ 2\ 2\ 2\ 1\ 1\ 1]; \end{aligned}$$

and choosing

$$M_\psi^3 = \delta_3[3\ 2\ 1\ 3\ 2\ 2\ 3\ 3\ 3],$$

yields

$$\begin{aligned} T_1 M_\psi^3 &= [3\ 2\ 1\ 3\ 2\ 2\ 3\ 3\ 3],\\ T_2 M_\psi^3 &= [2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2],\\ T_3 M_\psi^3 &= [1\ 2\ 3\ 1\ 2\ 2\ 1\ 1\ 1]. \end{aligned}$$

Comparing each block of M_f with the above product forms, it is easy to see that (12.44) is satisfied. It follows immediately that

$$\begin{aligned} M_\psi &= [M_\psi^1\ M_\psi^2\ M_\psi^3]\\ &= \delta_3[1\ 1\ 1\ 2\ 2\ 2\ 3\ 3\ 3\quad 2\ 2\ 1\ 2\ 2\ 2\ 3\ 3\ 3\quad 3\ 2\ 1\ 3\ 2\ 2\ 3\ 3\ 3], \end{aligned}$$

and

$$M_\phi = \delta_3[1\ 1\ 1\ 1\ 2\ 2\ 1\ 2\ 3].$$

Thus,

$$f(x_1, x_2, x_3, x_4) = M_f x_1 x_2 x_3 x_4 = M_\leftrightarrow M_\phi x_1 x_2 M_\psi x_2 x_3 x_4.$$

Back to logical form, we have the decomposed f as

$$f(x_1, x_2, x_3, x_4) = (x_1 \vee x_2) \leftrightarrow [(x_2 \vee \neg x_4) \wedge x_3].$$

12.4 Decomposition of Mix-Valued Logical Functions

Definition 12.6.

(1) Let $f : \mathcal{D}_{r_1} \times \mathcal{D}_{r_2} \to \mathcal{D}_{r_0}$ be a mix-valued logical function. f is said to be decomposable with respect to $\mathcal{D}_{r_1}$ and $\mathcal{D}_{r_2}$, if there exist $F : \mathcal{D}_{r_0} \to \mathcal{D}_{r_0}$, $\phi : \mathcal{D}_{r_1} \to \mathcal{D}_{r_0}$, and $\psi : \mathcal{D}_{r_2} \to \mathcal{D}_{r_0}$, such that

$$f(x_1, x_2) = F(\phi(x_1), \psi(x_2)), \quad x_1 \in \mathcal{D}_{r_1},\ x_2 \in \mathcal{D}_{r_2}. \tag{12.45}$$

(2) Let $f : \mathcal{D}_{r_1} \times \mathcal{D}_{r_2} \times \mathcal{D}_{r_3} \to \mathcal{D}_{r_0}$ be a mix-valued logical function. f is said to be decomposable with respect to $\mathcal{D}_{r_1} \times \mathcal{D}_{r_2}$ and $\mathcal{D}_{r_2} \times \mathcal{D}_{r_3}$, if there exist $F : \mathcal{D}_{r_0} \to \mathcal{D}_{r_0}$, $\phi : \mathcal{D}_{r_1} \times \mathcal{D}_{r_2} \to \mathcal{D}_{r_0}$, and $\psi : \mathcal{D}_{r_2} \times \mathcal{D}_{r_3} \to \mathcal{D}_{r_0}$, such that

$$f(x_1, x_2, x_3) = F(\phi(x_1, x_2), \psi(x_2, x_3)), \quad x_1 \in \mathcal{D}_{r_1},\ x_2 \in \mathcal{D}_{r_2},\ x_3 \in \mathcal{D}_{r_3}. \tag{12.46}$$

Remark 12.6. In fact, the decompositions defined in Definition 12.6 are the most general ones. To see this, we consider (in vector form) $x^1 = \ltimes_{i=1}^{k} x_i$, $x^2 = \ltimes_{i=1}^{n-k} x_i$.

(i) Let $x_i \in \Delta_2$, $\forall i$ (i.e., $r_1 = 2^k$, $r_2 = 2^{n-k}$), and choose $r_0 = 2$. Then we have the bi-decomposition of Boolean functions.

(ii) Let $x_i \in \Delta_r$, $\forall i$ (i.e., $r_1 = r^k$, $r_2 = r^{n-k}$), and choose $r_0 = r$. Then we have the bi-decomposition of r-valued logical functions.

(iii) Let x_i as in the above case 1 (case 2), and choose $r_0 = 2^s$ ($r_0 = r^s$). Then we have the bi-decomposition of Boolean (r-valued) multi-input multi-output (MIMO) mappings.

Using the argument for Boolean or multi-valued case, we can have the following general result immediately.

Theorem 12.6.

(1) Let $f : \mathcal{D}_{r_1} \times \mathcal{D}_{r_2} \to \mathcal{D}_{r_0}$ be a mix-valued logical function with its structure matrix as

$$M_f = [M_1\ M_2\ \cdots\ M_{r_1}], \tag{12.47}$$

where $M_i \in \mathcal{L}_{r_0 \times r_2}$. f has a decomposed form with respect to $\mathcal{D}_{r_1}$ and $\mathcal{D}_{r_2}$, if and only if there exist

(i) a type $\mathcal{T} = \{T_1, T_2, \cdots, T_{r_0}\} \subset \mathcal{L}_{r_0 \times r_0}$,

(ii) a logical matrix $M_\psi \in \mathcal{L}_{r_0 \times r_2}$,

such that

$$M_i = T_{s_i} M_\psi, \quad \text{where } T_{s_i} \in \mathcal{T},\ \ i = 1, \cdots, r_1. \tag{12.48}$$

(2) Let $f : \mathcal{D}_{r_1} \times \mathcal{D}_{r_2} \times \mathcal{D}_{r_3} \to \mathcal{D}_{r_0}$ be a mix-valued logical function. f is decomposable with respect to $\mathcal{D}_{r_1} \times \mathcal{D}_{r_2}$ and $\mathcal{D}_{r_2} \times \mathcal{D}_{r_3}$, if and only if

(i) there exists a type $\mathcal{T} \subset \mathcal{L}_{r_0 \times r_0}$,

(ii) there exist

$$M_\psi^i \in \mathcal{L}_{r_0 \times r_3}, \quad i = 1, \cdots, r_2, \tag{12.49}$$

such that the structure matrix of f can be expressed as

$$\begin{aligned}M_f = \Big[&\mu_{1,1}M_\psi^1\ \ \mu_{1,2}M_\psi^2\ \ \cdots\ \ \mu_{1,r_2}M_\psi^{r_2}\\ &\mu_{2,1}M_\psi^1\ \ \mu_{2,2}M_\psi^2\ \ \cdots\ \ \mu_{2,r_2}M_\psi^{r_2}\\ &\vdots\\ &\mu_{r_1,1}M_\psi^1\ \ \mu_{r_1,2}M_\psi^2\ \ \cdots\ \ \mu_{r_1,r_2}M_\psi^{r_2}\Big],\end{aligned} \tag{12.50}$$

where

$$\mu_{i,j} \in \mathcal{T}, \quad i = 1, \cdots, r_1,\ j = 1, \cdots, r_2.$$

Example 12.5. Consider a mix-valued logical function $f(x_1, x_2, x_3) : \mathcal{D}_3 \times \mathcal{D} \times \mathcal{D} \to \mathcal{D}$, which structure matrix is

$$M_f = \delta_2[1\ 2\ 1\ 1\ \ 1\ 1\ 1\ 1\ \ 1\ 2\ 1\ 1].$$

If we choose

$$\mathcal{T}_\vee = \{T_1 = \delta_2[1\ 1], T_2 = \delta_2[1\ 2]\},$$

and

$$M_\psi = \delta_2[1\ 2\ 1\ 1],$$

it is easy to check that

$$\begin{aligned}T_1 M_\psi &= \delta_2[1\ 1\ 1\ 1],\\ T_2 M_\psi &= \delta_2[1\ 2\ 1\ 1].\end{aligned}$$

Thus,

$$M_\phi = \delta_2[2\ 1\ 2],$$

and

$$f(x_1, x_2, x_3) = M_\vee M_\phi x_1 M_\psi x_2 x_3.$$

In logical form, we have

$$f(x_1, x_2, x_3) = \nabla_{2,3}^2(x_1) \vee (x_2 \to x_3).$$

Decomposition of a logical function is not only useful in circuit design but also useful in many other problems. For instance, in Chapter 15 the decomposition of mix-valued function will be used to get the normal form of dynamic-algebraic Boolean networks.

Exercises

12.1 A logical function $f(x_1, \cdots, x_n)$ has its structure matrix $M_f \in \mathcal{L}_{2\times 2^n}$. Show that

$$M_{\neg f} = \neg M_f.$$

12.2 Recall Remark 12.1. Assume f has a structure matrix as

$$M_f = [\mu_1 M_\psi \; \mu_2 M_\psi \; \cdots \; \mu_{2^k} M_\psi].$$

(i) Show that if

$$\mu_i \in S_2 = \{\delta_2[1\ 1],\ \delta_2[1\ 2]\}, \quad i = 1, \cdots, 2^k,$$

then f has same type of M_f with

$$\mu_i \in S_2' = \{\delta_2[2\ 2],\ \delta_2[1\ 2]\}, \quad i = 1, \cdots, 2^k,$$

and vice versa.

(ii) Show that if

$$\mu_i \in S_3 = \{\delta_2[1\ 2]\}, \quad i = 1, \cdots, 2^k,$$

then f has same type of M_f with

$$\mu_i \in \{\delta_2[2\ 1]\}, \quad i = 1, \cdots, 2^k,$$

and vice versa.

12.3 Consider the following bi-decomposed Boolean functions. Figure out M_ψ and μ_i's.

(i)

$$f(x_1, x_2, x_3, x_4) = (x_1 \vee x_2) \leftrightarrow (x_3 \wedge x_4).$$

(ii)

$$f(x_1, x_2, x_3, x_4, x_5) = (x_1 \to (x_2 \wedge x_3)) \vee (\neg x_4 \bar{\vee} x_5).$$

12.4 Consider the following bi-decomposed Boolean functions. Figure out M_ψ, μ_i's and the rearrangement matrix Ψ.

(i)

$$f(x_1, x_2, x_3) = x_3 \to [x_1 \wedge x_2).$$

(ii)

$$f(x_1, x_2, x_3, x_4, x_5) = (x_1 \to (x_3 \vee x_5)) \wedge (\neg x_2 \to x_4).$$

12.5 The structure matrix of a Boolean function f is given as follows. Find the bi-decomposition of f.

(i)

$$M_f = \delta_2[1\ 1\ 1\ 1\ 1\ 2\ 2\ 1\ 1\ 2\ 2\ 1\ 1\ 2\ 2\ 1].$$

(Hint: Try $\Gamma = \{1, 2\}, \Lambda = \{3, 4\}$.)

(ii)

$$M_f = \delta_2[1\ 1\ 1\ 1\ 1\ 2\ 1\ 1].$$

(Hint: Try all possible partitions.)

12.6 Assume $f(x_1, x_2, x_3, x_4, x_5)$ has the following non-disjoint decomposition

$$f(x_1, x_2, x_3, x_4, x_5) = F\left(\phi(x_1, x_2, x_3), \psi(x_3, x_4, x_5)\right).$$

Assume

$$M^1_\psi = \delta_2[1\ 2\ 2\ 1], \quad M^2_\psi = \delta_2[1\ 1\ 1\ 2];$$

$$\begin{aligned}
&\mu_{11} = \delta_2[1\ 1],\ \mu_{12} = \delta_2[1\ 2],\\
&\mu_{21} = \delta_2[1\ 2],\ \mu_{22} = \delta_2[1\ 2],\\
&\mu_{31} = \delta_2[1\ 2],\ \mu_{32} = \delta_2[1\ 1],\\
&\mu_{41} = \delta_2[1\ 2],\ \mu_{42} = \delta_2[1\ 2].
\end{aligned}$$

Find F, ϕ, and ψ, and express f into a decomposed form.

12.7 For k-valued logic, define

(1) $$x \to y := \neg x \vee y. \tag{12.51}$$

(2) $$x \leftrightarrow y := (x \to y) \wedge (y \to x). \tag{12.52}$$

(3) $$x \bar{\vee} y = \neg(x \leftrightarrow y). \tag{12.53}$$

(i) When $r = 3$, use (12.51)–(12.53) to express (a) $\mathcal{T}_\to$; (b) $\mathcal{T}_\leftrightarrow$; (c) $\mathcal{T}_{\bar{\vee}}$ into the form of (12.38).

(ii) When $r = 4$, use (12.51)–(12.53) to express (a) $\mathcal{T}_\to$; (b) $\mathcal{T}_\leftrightarrow$; (c) $\mathcal{T}_{\bar{\vee}}$ into the form of (12.39).

12.8 Assume $f : \mathcal{D}_3^3 \to \mathcal{D}_3$ has structure matrix

$$M_f = \delta_3[1, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 2, 3].$$

Find its possible bi-decomposition.

12.9 Assume $f : \mathcal{D}^3 \times \mathcal{D}_3^2 \to \mathcal{D}$. We decompose it into

$$f(x_1, x_2, x_3, x_4, x_5) = F(\phi(x_1, x_2, x_3), \psi(x_4, x_5)),$$

where $\phi : \mathcal{D}^3 \to \mathcal{D}$, $\psi : \mathcal{D}_3^2 \to \mathcal{D}$.

(i) How many possible F's need to be considered?

(ii) How many possible ϕ's and ψ's need to be considered?

12.10 Assume $f(x_1, x_2, x_3, x_4, x_5) : \mathcal{D}_3 \times \mathcal{D} \times \mathcal{D} \times \mathcal{D} \times \mathcal{D}_3 \to \mathcal{D}$,

$$\begin{aligned} M_f = \delta[&1\ 1\ 1\ 1\ 2\ 1\ 1\ 2\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 1\ 1 \\ &1\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 2\ 1\ 1\ 1 \\ &1\ 1\ 2\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 1 \\ &2\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 1\ 1\ 1]. \end{aligned}$$

Find its possible bi-decomposition.

(Hint: Try the bi-decomposition as $F[\phi(x_2, x_3, x_4), \psi(x_1, x_5)]$, where $\phi : \mathcal{D}^3 \to \mathcal{D}$ and $\psi : \mathcal{D}_3^2 \to \mathcal{D}$. For this purpose, you first need to find $\tilde{M}_f$ such that $f(x_1, x_2, x_3, x_4, x_5) = M_f x_1 x_2 x_3 x_4 x_5 = \tilde{M}_f x_2 x_3 x_4 x_1 x_5$.)

12.11 Recall the structure matrix of (12.16). Consider it as the structure matrix of a mapping $f : \mathcal{D}_4 \times \mathcal{D}_4 \to \mathcal{D}_2$, and use the general result in Theorem 12.6 to solve the decomposition problem.

12.12 Recall the structure matrix of (12.17). Consider it as the structure matrix of a mapping $f : \mathcal{D}_8 \times \mathcal{D}_4 \to \mathcal{D}_2$, and solve the decomposition problem.

12.13 Recall the structure matrix of (12.41). Consider it as the structure matrix of a mapping $f : \mathcal{D}_9 \times \mathcal{D}_9 \to \mathcal{D}_3$, and solve the decomposition problem.

12.14 Consider a logical function $f : \mathcal{D}_k \to \mathcal{D}$, and consider a partition

$$\{1, 2, \cdots, k\} = \Gamma \cup \Theta \cup \Lambda, \tag{12.54}$$

where $\Gamma = \{1, 2, \cdots, k_1\}$, $\Theta = \{k_1+1, k_1+2, \cdots, k_2\}$, and $\Lambda = \{k_2+1, k_2+2, \cdots, k\}$.

(i) Find the necessary and sufficient condition for f to be compounded bi-decomposition of $\phi(x_\gamma|\gamma \in \Gamma)$, $\psi(x_\theta|\theta \in \Theta)$, $\zeta(x_\lambda|\lambda \in \Lambda)$ as follows:

$$f(x_1, \cdots, x_n) = F_1\left(F_2(\phi(x), \psi(x)), \zeta(x)\right),$$

where $F_i : \mathcal{D}^2 \to \mathcal{D}$, $i = 1, 2$.

(ii) Consider an alternative compounded form as

$$f(x_1, \cdots, x_n) = F_1\left(\phi(x), F_2(\psi(x), \zeta(x))\right).$$

(iii) When (12.54) is an arbitrary partition, (i.e., the elements in three subsets are not in given order), reconsider the above compounded bi-decomposition problems.

Chapter 13

Boolean Calculus

This chapter considers the Boolean calculus, including Boolean derivatives and Boolean integrals. Formulas are obtained to calculate Boolean derivatives. Its applications to solving Boolean algebraic/differential equations and to fault detection of combinational circuits etc. are investigated. Then the Boolean integrals are defined as the inverse of the Boolean derivative in certain sense. Three kinds of integrals are proposed. The inverse of a partial derivative with respect to x_i is called the ith primitive function. The inverse of a differential form is called the indefinite integral. A necessary and sufficient condition for the existence of the indefinite integral is proved. Using the unique indefinite integral (up to complement equivalence), definite integral is also defined. Easily computable formulas are provided for solving each kind of integrals. This chapter is mainly based on Cheng *et al.* (2011d).

13.1 Boolean Derivatives

The first version of Boolean differential calculus was proposed by Daniell (1917). Some forty years later after Shannon proposed the switching algebra in the evaluation of switching circuit designing, it was discovered that the partial derivatives of Boolean functions are particularly useful in switching theory (Reed, 1954; Akers, 1959). Since then, the Boolean derivative has been developed quickly, both in view of applications and for its own algebraic interest (Posthoff and Steinbach, 2004; Brown, 2003; Yanushkevich, 1998; Schneeweiss, 1989; Thayse, 1984; Davio *et al.*, 1978).

The Boolean derivative used in this book is defined as follows (Bochmann, 1978), which is the commonly adopted one.

Definition 13.1. Let $f(x_1, \cdots, x_n) : \mathcal{D}^n \to \mathcal{D}$ be a Boolean function.

(1) The Boolean derivative of f with respect to x_i is defined as

$$\frac{\partial f}{\partial x_i} = f(x_1, \cdots, x_i, \cdots, x_n) \, \langle + \rangle \, f(x_1, \cdots, \neg x_i, \cdots, x_n). \tag{13.1}$$

(2) The higher-order derivative of f with respect to x_{i_1}, $\cdots$, x_{i_k} is defined recursively as

$$\frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}} = \frac{\partial}{\partial x_{i_1}} \left(\frac{\partial}{\partial x_{i_2}} \left(\cdots \left(\frac{\partial f}{\partial x_{i_k}} \right) \right) \right). \tag{13.2}$$

Note that as in Chapter 11 throughout this chapter $\langle + \rangle := \bar{\vee}$, that is, $+$ mod 2.

We cite some basic properties in the following. According to the definition, they can be proved by straightforward computations.

Proposition 13.1 (Vichniac, 1990). *(1) For a constant $c \in \mathcal{D}$,*

$$\frac{\partial c}{\partial x_i} = 0. \tag{13.3}$$

(2) $\frac{\partial f}{\partial x_i}$ is independent of x_i, and hence

$$\frac{\partial^2 f}{\partial^2 x_i} = 0. \tag{13.4}$$

(3)

$$\frac{\partial x_i}{\partial x_i} = 1. \tag{13.5}$$

(4)

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}. \tag{13.6}$$

(5)

$$\frac{\partial (f_1 \, \langle + \rangle \, f_2)}{\partial x_i} = \frac{\partial f_1}{\partial x_i} \, \langle + \rangle \, \frac{\partial f_2}{\partial x_i}. \tag{13.7}$$

(6)

$$\frac{\partial (f_1 f_2)}{\partial x_i} = \frac{\partial f_1}{\partial x_i} f_2 \, \langle + \rangle \, f_1 \frac{\partial f_2}{\partial x_i} \, \langle + \rangle \, \frac{\partial f_1}{\partial x_i} \frac{\partial f_2}{\partial x_i}. \tag{13.8}$$

(7)

$$\frac{\partial(x_i x_j)}{\partial x_i} = x_j. \tag{13.9}$$

(8) Denote $\bar{f} := \neg f$*, and* $\bar{x} := \neg x$*, then*

$$\frac{\partial \bar{f}}{\partial x_i} = \frac{\partial f}{\partial x_i}, \quad \textit{and} \quad \frac{\partial f}{\partial \bar{x}_i} = \frac{\partial f}{\partial x_i}. \tag{13.10}$$

Our first goal is to get the formula, i.e., the structure matrix of $\frac{\partial}{\partial x_i}$, which is denoted by $M_{\partial_i f}$. Let M_f be the structure matrix of f and $x := \ltimes_{i=1}^n x_i$. Using (13.1), we have

$$\frac{\partial f}{\partial x_i} = M_{\partial_i f} x = M_f x \,\langle + \rangle\, M_f x_1 \cdots \neg x_i \cdots x_n. \tag{13.11}$$

Using vector form, a standard procedure can convert the right hand side of (13.11) into a canonical form, which provides the structure matrix of $\frac{\partial}{\partial x_i}$ as (Li and Wang, 2010)

$$M_{\partial_i f} = M_{\triangledown} M_f \left[I_{2^n} \otimes M_f \left(I_{2^{i-1}} \otimes M_n \right) \right] M_{r,2^n}, \quad 1 \le i \le n. \tag{13.12}$$

Then the higher-order derivatives can also be calculated recursively.

In the following we shall give an explicit form of the structure matrices of the derivatives. Consider the structure matrix of $g(x_1, \cdots, x_n) := f(x_1, \cdots, \neg x_i, \cdots, x_n)$. Assume the structure matrices of f and g are M_f and M_g respectively. A straightforward computation shows that

$$\begin{aligned} M_g x &= M_f x_1 \cdots M_\neg x_i \cdots x_n \\ &= M_f \left(I_{2^{i-1}} \otimes M_n \right) x. \end{aligned}$$

That is,

$$M_g = M_f \left(I_{2^{i-1}} \otimes M_n \right). \tag{13.13}$$

Let $m_f = \mathrm{Row}_1(M_f)$ and $m_g = \mathrm{Row}_1(M_g)$. In fact, $m_f = T_f^T$, where T_f is the truth vector. But for convenience we also call m_f the truth vector of f.

The following proposition is easily verifiable. We leave the verification to the reader.

Proposition 13.2. *Assume* $f(x_1, \cdots, x_n)$ *and* $g(x_1, \cdots, x_n)$ *have their truth vectors as* $m_f, m_g \in \mathcal{B}_{2^n}$ *respectively, and* σ *is a binary logical operator. Then*

$$m_{f \sigma g} = m_f \sigma m_g. \tag{13.14}$$

Note that $m_f \sigma m_g = ((m_f)_1 \sigma (m_g)_1, \cdots, (m_f)_{2^n} \sigma (m_g)_{2^n})$.

Using Proposition 13.2, we have

$$m_{\partial_i f} = m_f \left\langle + \right\rangle m_f (I_{2^{i-1}} \otimes M_n). \tag{13.15}$$

Using the distributive law of STP, we can calculate that

$$\begin{aligned} m_{\partial_i f} &= m_f \left\langle + \right\rangle m_f (I_{2^{i-1}} \otimes M_n) \\ &= m_f \, (\ltimes) \, I_{2^i} \left\langle + \right\rangle m_f \, (\ltimes) (I_{2^{i-1}} \otimes M_n) \\ &= m_f \, (\ltimes) \left(I_{2^i} \left\langle + \right\rangle (I_{2^{i-1}} \otimes M_n) \right) \\ &= m_f \, (\ltimes) \left(I_{2^{i-1}} \otimes (I_2 \left\langle + \right\rangle M_n) \right) \\ &= m_f \, (\ltimes) \left(I_{2^{i-1}} \otimes \mathbf{1}_{2\times 2} \right). \end{aligned}$$

Summarizing the above argument, we have

Theorem 13.1. *Let $f(x_1, \cdots, x_n)$ be a Boolean function with structure matrix M_f. Then the structure matrix of $\frac{\partial f}{\partial x_i}$, denoted by $M_{\partial_i f}$, is*

$$M_{\partial_i f} = \begin{bmatrix} m_f \, (\ltimes) \, \Xi_n^i \\ \neg m_f \, (\ltimes) \, \Xi_n^i \end{bmatrix} \tag{13.16}$$

where

$$\Xi_n^i = I_{2^{i-1}} \otimes \mathbf{1}_{2\times 2}.$$

Hence, in vector form,

$$\frac{\partial f}{\partial x_i} = M_{\partial_i f} x, \tag{13.17}$$

where $x = \ltimes_{i=1}^n x_i$. Moreover,

$$m_{\partial_i f} = m_f \Xi_n^i. \tag{13.18}$$

As we know that $\frac{\partial f}{\partial x_i}$ is independent of x_i, so one may be interested in an alternative expression as

$$\frac{\partial f}{\partial x_i} = M_{\partial_{[i]} f} x_1 \cdots x_{i-1} \hat{x}_i x_{i+1} \cdots x_n, \tag{13.19}$$

where notation "$\hat{x}_i$" means x_i is omitted.

To calculate $M_{\partial_{[i]} f}$, we divide $M_{\partial_i f}$ into 2^i equal-size blocks as

$$M_{\partial_i f} = [C_1 \; C_2 \; \cdots \; C_{2^i}].$$

One sees easily that to get $M_{\partial_{[i]} f}$ from $M_{\partial_i f}$, we need only to pick out all odd (or even) blocks. It can be done by right-multiplying

$$\left(I_{2^{i-1}} \otimes \begin{bmatrix} I_{2^{n-i}} \\ \mathbf{0}_{2^{n-i} \times 2^{n-i}} \end{bmatrix} \right).$$

That is,

$$M_{\partial_{[i]}f} = \begin{bmatrix} m_f \,(\ltimes)[\Psi_n^i]^T \\ \neg m_f \,(\ltimes)[\Psi_n^i]^T \end{bmatrix} \tag{13.20}$$

where

$$\begin{aligned} \Psi_n^i &= \left(I_{2^{i-1}} \otimes \left[I_{2^{n-i}} \ \mathbf{0}_{2^{n-i}\times 2^{n-i}}\right]\right)(\ltimes)\left(I_{2^{i-1}} \otimes \mathbf{1}_{2\times 2}\right) \\ &= I_{2^{i-1}} \otimes \mathbf{1}_2^T \otimes I_{2^{n-i}}. \end{aligned}$$

Then we have the following

Corollary 13.1. *Assume the truth vector of a logical function $f(x_1, \cdots, x_n)$ is m_f. Then the truth vector of $\frac{\partial f}{\partial x_i}$, in condensed form, is*

$$m_{\partial_{[i]}f}^T = \Psi_n^i m_f^T. \tag{13.21}$$

It is easy to check that Corollary 13.1 coincides with the result in Agaian *et al.* (1995, 2010).

The following Corollaries 13.2 and 13.3 are convenient in numerical computation. We leave the proves to the reader.

Corollary 13.2. *Divide m_f into 2^i blocks*

$$m_f = (c_{1,1}\ c_{1,2}\ c_{2,1}\ c_{2,2}\ \cdots\ c_{2^{i-1},1}\ c_{2^{i-1},2}).$$

Then $m_{\partial_{[i]}f}$ can be calculated directly by

$$m_{\partial_{[i]}f} = (c_{1,1}\,\langle+\rangle\,c_{1,2}\ c_{2,1}\,\langle+\rangle\,c_{2,2}\ \cdots\ c_{2^{i-1},1}\,\langle+\rangle\,c_{2^{i-1},2}). \tag{13.22}$$

Corollary 13.3. *The truth table of $\frac{\partial^k f}{\partial x_{i_1}\cdots\partial x_{i_k}}$ (in condensed form) is (assume $i_1 > i_2 > \cdots > i_k$):*

$$m_{\partial_{[i_1,\cdots,i_k]}f}^T = \Psi_{n-k+1}^{i_k}\Psi_{n-k+2}^{i_{k-1}}\cdots\Psi_n^{i_1} m_f^T. \tag{13.23}$$

Note that in (13.23) we require $i_1 > i_2 > \cdots > i_k$ because otherwise, the later positions need to be adjusted. For instance, say, $i_1 < i_2$, then (refer to Section 12.3) after integral with respect to $d[i_1]$ the position for i_2 becomes $i_2 - 1$. So we need this order. Because of (13.6), we can assume this without loss of generality.

We give an example to show how to calculate the derivatives. To this end, we introduce the MacLaurin expansion of a Boolean function.

Theorem 13.2 (Akers, 1959). *A Boolean function $f(x_1, \cdots, x_n)$ has its MacLaurin expansion as*

$$\begin{aligned}
&f(x_1, \cdots, x_n) = \\
&f(\mathbf{0}) \langle + \rangle \langle + \rangle_{i=1}^{n} \left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{0}} \wedge x_i \langle + \rangle \langle + \rangle_{1 \le i_1 < i_2 \le n} \left.\frac{\partial^2 f}{\partial x_{i_1} \partial x_{i_2}}\right|_{\mathbf{0}} \wedge x_{i_1} \wedge x_{i_2} \\
&\langle + \rangle \langle + \rangle_{1 \le i_1 < i_2 < i_3 \le n} \left.\frac{\partial^3 f}{\partial x_{i_1} \partial x_{i_2} \partial x_{i_3}}\right|_{\mathbf{0}} \wedge x_{i_1} \wedge x_{i_2} \wedge x_{i_3} \langle + \rangle \cdots \\
&\langle + \rangle \left.\frac{\partial^n f}{\partial x_1 \partial x_2 \cdots \partial x_n}\right|_{\mathbf{0}} \wedge x_1 \wedge x_2 \wedge \cdots \wedge x_n.
\end{aligned} \tag{13.24}$$

Example 13.1. Assume $f(x_1, x_2, x_3, x_4) = (x_1 \bar{\vee} x_2) \to (x_3 \vee x_4)$. It is easy to calculate its truth table as

$$m_f = [1\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 1\ 1\ 1\ 0\ 1\ 1\ 1\ 1].$$

Using (13.21),

$$\Psi_4^1 = \begin{bmatrix} I_8 & I_8 \end{bmatrix},$$

the truth table of $\frac{\partial f}{\partial x_1}$ is

$$m_{\partial_{[1]} f} = m_f \,(\ltimes)\, \left[\Psi_4^1\right]^T = [0\ 0\ 0\ 1\ 0\ 0\ 0\ 1].$$

Similarly,

$$\Psi_4^2 = \begin{bmatrix} I_4 & I_4 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_4 & I_4 \end{bmatrix}.$$

$$m_{\partial_{[2]} f} = m_f \,(\ltimes)\, \left[\Psi_4^2\right]^T = [0\ 0\ 0\ 1\ 0\ 0\ 0\ 1].$$

$$\Psi_4^3 = \begin{bmatrix} I_2 & I_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_2 & I_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & I_2 & I_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & I_2 & I_2 \end{bmatrix}.$$

$$m_{\partial_{[3]} f} = m_f \,(\ltimes)\, \left[\Psi_4^3\right]^T = [0\ 0\ 0\ 1\ 0\ 1\ 0\ 0].$$

$$\Psi_4^4 = \begin{bmatrix} 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ \vdots \\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1 \end{bmatrix}.$$

$$m_{\partial_{[4]}f} = m_f\,(\ltimes)\left[\Psi_4^4\right]^T = [0\;0\;0\;1\;0\;1\;0\;0].$$

Using (13.23), we can also easily calculate that

$$\begin{array}{l}
m_{\partial_{[1,2]}f} = [0\;0\;0\;0];\; m_{\partial_{[1,3]}f} = [0\;1\;0\;1];\; m_{\partial_{[1,4]}f} = [0\;1\;0\;1];\\
m_{\partial_{[2,3]}f} = [0\;1\;0\;1];\; m_{\partial_{[2,4]}f} = [0\;1\;0\;1];\; m_{\partial_{[3,4]}f} = [0\;1\;1\;0];\\
m_{\partial_{[1,2,3]}f} = [0\;0]; \quad m_{\partial_{[1,2,4]}f} = [0\;0]; \quad m_{\partial_{[1,3,4]}f} = [1\;1];\\
m_{\partial_{[2,3,4]}f} = [1\;1]; \quad m_{\partial_{[1,2,3,4]f}} = 0.
\end{array}$$

Note that evaluating the Boolean derivatives at $\mathbf{0}$ is equivalent to taking last element of its corresponding true table. Hence we have:

$$\begin{array}{llll}
f(\mathbf{0}) = 1; & \frac{\partial f}{\partial x_1}|_{\mathbf{0}} = 1, & \frac{\partial f}{\partial x_2}|_{\mathbf{0}} = 1, & \frac{\partial f}{\partial x_3}|_{\mathbf{0}} = 0\\
\frac{\partial f}{\partial x_4}|_{\mathbf{0}} = 0; & \frac{\partial^2 f}{\partial x_1 \partial x_2}|_{\mathbf{0}} = 0, & \frac{\partial^2 f}{\partial x_1 \partial x_3}|_{\mathbf{0}} = 1, & \frac{\partial^2 f}{\partial x_1 \partial x_4}|_{\mathbf{0}} = 1,\\
\frac{\partial^2 f}{\partial x_2 \partial x_3}|_{\mathbf{0}} = 1, & \frac{\partial^2 f}{\partial x_2 \partial x_4}|_{\mathbf{0}} = 1, & \frac{\partial^2 f}{\partial x_3 \partial x_4}|_{\mathbf{0}} = 0; & \frac{\partial^3 f}{\partial x_1 \partial x_2 \partial x_3}|_{\mathbf{0}} = 0,\\
\frac{\partial^3 f}{\partial x_1 \partial x_2 \partial x_4}|_{\mathbf{0}} = 0, & \frac{\partial^3 f}{\partial x_1 \partial x_3 \partial x_4}|_{\mathbf{0}} = 1, & \frac{\partial^3 f}{\partial x_2 \partial x_3 \partial x_4}|_{\mathbf{0}} = 1; & \frac{\partial^4 f}{\partial x_1 \partial x_2 \partial x_3 \partial x_4}|_{\mathbf{0}} = 0.
\end{array}$$

Then we have the MacLaurin expansion of $f(x)$ as

$$\begin{aligned}
f(x_1, x_2, x_3, x_4) = {} & 1 \langle + \rangle x_1 \langle + \rangle x_2 \langle + \rangle x_1 \wedge x_3 \langle + \rangle x_1 \wedge x_4 \langle + \rangle x_2 \wedge x_3\\
& \langle + \rangle x_2 \wedge x_4 \langle + \rangle x_1 \wedge x_3 \wedge x_4 \langle + \rangle x_2 \wedge x_3 \wedge x_4.
\end{aligned} \tag{13.25}$$

13.2 Boolean Differential Equations

Firstly, we consider the solution of Boolean equations, which involves a known Boolean function $f(x_1, \cdots, x_n)$ and its Boolean derivatives, as

$$\begin{cases}
G_1\left(x_i, f, \frac{\partial f}{\partial x_i}, \cdots, \frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}}\right) = c_1\\
G_2\left(x_i, f, \frac{\partial f}{\partial x_i}, \cdots, \frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}}\right) = c_2\\
\vdots\\
G_s\left(x_i, f, \frac{\partial f}{\partial x_i}, \cdots, \frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}}\right) = c_s.
\end{cases} \tag{13.26}$$

Using (13.23), solving the equations (13.26) is standard. We describe it via the following algorithm.

Algorithm 1.

- *Step 1:* Convert each logical equation into its algebraic form as
$$M_i x = c_i, \quad i = 1, \cdots, s, \tag{13.27}$$
where $M_i \in \mathcal{L}_{2\times 2^n}$.
- *Step 2:* Multiply all equations in (13.27) together to build a system as
$$Mx = c, \tag{13.28}$$
where $x = \ltimes_{i=1}^n x_i$, $c = \ltimes_{i=1}^s c_i$, and $M \in \mathcal{L}_{2^s\times 2^n}$ is constructed as
$$M = M_1 * M_2 * \cdots * M_s, \tag{13.29}$$
where $*$ is the Khatri-Rao product. Precisely,
$$\mathrm{Col}_i(M) = \ltimes_{j=1}^s \mathrm{Col}_i(M_j), \quad i = 1, \cdots, 2^n.$$
- *Step 3:* Find all the solutions $\delta_{2^n}^j$, which satisfies $\mathrm{Col}_j(M) = c$.

The fault detection of combinational circuits (Keren, 2008), Li and Wang (2010) is a typical example of this problem. Let $f(x_1, \cdots, x_n)$ be a Boolean function describing a combinational circuit. The test vector set for double stuck-at faults $x_i(s-a-\alpha)$, $x_j(s-a-\beta)$ is the set of solutions of

$$\bar{x}_i^\alpha x_j^\beta \frac{\partial f}{\partial x_i} \,\langle + \rangle\, x_i^\alpha \bar{x}_j^\beta \frac{\partial f}{\partial x_j} \,\langle + \rangle\, \bar{x}_i^\alpha \bar{x}_j^\beta \frac{\partial^2 f}{\partial x_i \partial x_j} = 1, \tag{13.30}$$

where $\alpha, \beta \in \mathcal{D}$, and $x^1 := x$, $x^0 := \bar{x}$.

We give an example to depict it.

Example 13.2. Assume a combinational circuit is described as (Li and Wang, 2010)

$$\begin{aligned} f(x_1, \cdots, x_5) = \neg\,\{\neg[x_2 \vee (\neg x_1 \wedge \neg x_3)\} \vee \neg(x_1 \vee x_5) \\ \vee \neg(x_4 \vee x_5) \vee \neg[\neg x_3 \vee (\neg x_2 \wedge \neg x_4)]. \end{aligned} \tag{13.31}$$

We look for the test vector set for the double stuck at $x_3(s-a-1)$, and $x_4(s-a-0)$.

That is, to solve the equation

$$\bar{x}_3 \bar{x}_4 \frac{\partial f}{\partial x_3} \,\langle + \rangle\, x_3 x_4 \frac{\partial f}{\partial x_4} \,\langle + \rangle\, \bar{x}_3 x_4 \frac{\partial^2 f}{\partial x_3 \partial x_4} = 1. \tag{13.32}$$

The structure matrix of f is

$$M_f = \delta_2[1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 2\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 2].$$

Then, using Corollary 13.2, it is easy to obtain that

$$\begin{aligned} M_{f_{[3]}} &= \delta_2[1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 1\ 2\ 1\ 2\ 2\ 2\ 1\ 2], \\ M_{f_{[4]}} &= \delta_2[2\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 2], \\ M_{f_{[3,4]}} &= \delta_2[2\ 1\ 2\ 2\ 2\ 2\ 1\ 2]. \end{aligned}$$

Using a standard computing process, the algebraic form of (13.31) is obtained as

$$Mx = 1,$$

where $x = \ltimes_{i=1}^{5} x_i$, and

$$M = \delta_2[2\ 1\ 2\ 2\ 2\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 2\ 2\ 2\ 2\ 2\ 1\ 2\ 1\ 2].$$

Thus, the solution is

$$\left\{x = \delta_{32}^{i} | i = 2, 6, 7, 23, 29, 31\right\},$$

or, in scalar form

$$\begin{aligned} \{&(1,1,1,1,0),\ (1,1,0,0,1),\ \ (1,1,0,0,0), \\ &(0,1,0,0,1),\ \ (0,0,0,1,1),\ (0,0,0,0,1)\}. \end{aligned}$$

Next, we consider the case that in equations (13.26), the Boolean function $f(x_1, \cdots, x_n)$ is unknown and with a set of the boundary conditions $f(\mathbf{0})$ and some Boolean derivatives of f at $\mathbf{0}$. We call this kind of equations the Boolean differential equations (BDE). If a Boolean function $g(x_1, \cdots, x_n)$ satisfies (13.26) and the boundary conditions, it is called a solution of the BDE with boundary conditions. We use an example to depict how to solve BDE.

Example 13.3. Consider the following Boolean differential equation with initial condition $F(0) = 0$:

$$\begin{cases} \frac{\partial F}{\partial x_3} = \neg x_1 \wedge \neg x_4 \\ \frac{\partial^2 F}{\partial x_1 \partial x_4} = \neg(x_2 \vee x_3) \vee (x_2 \wedge x_3) \\ \frac{\partial^2 F}{\partial x_2 \partial x_4} = \neg x_1 \\ \frac{\partial^2 F}{\partial x_1 \partial x_3} \vee \frac{\partial^2 F}{\partial x_1 \partial x_2} = 1. \end{cases} \tag{13.33}$$

In vector form we have

$$\begin{cases} \partial_{[3]}(M_F) = \delta_2[2\ 2\ 2\ 2\ 2\ 1\ 2\ 1] \\ \partial_{[1,4]}(M_F) = \delta_2[1\ 2\ 2\ 1] \\ \partial_{[2,4]}(M_F) = \delta_2[2\ 2\ 1\ 1] \\ \partial_{[1,3]}(M_F) \vee \partial_{[1,2]}(M_F) = \delta_2[1\ 1\ 1\ 1\ 1\ 1\ 1\ 1]. \end{cases} \tag{13.34}$$

Assume the first row of M_F is

$$m_F = [a_1\ a_2\ \cdots\ a_{16}].$$

Then from (13.34) we know that

$$\begin{array}{ll}
a_1 \langle + \rangle a_3 = 0 & a_2 \langle + \rangle a_4 = 0 \\
a_5 \langle + \rangle a_7 = 0 & a_6 \langle + \rangle a_8 = 0 \\
a_9 \langle + \rangle a_{11} = 0 & a_{10} \langle + \rangle a_{12} = 1 \\
a_{13} \langle + \rangle a_{15} = 0 & a_{14} \langle + \rangle a_{16} = 1 \\
a_1 \langle + \rangle a_2 \langle + \rangle a_9 \langle + \rangle a_{10} = 1 & a_3 \langle + \rangle a_4 \langle + \rangle a_{11} \langle + \rangle a_{12} = 0 \\
a_5 \langle + \rangle a_6 \langle + \rangle a_{13} \langle + \rangle a_{14} = 0 & a_7 \langle + \rangle a_8 \langle + \rangle a_{15} \langle + \rangle a_{16} = 1 \\
a_1 \langle + \rangle a_2 \langle + \rangle a_5 \langle + \rangle a_6 = 0 & a_3 \langle + \rangle a_4 \langle + \rangle a_7 \langle + \rangle a_8 = 0 \\
a_9 \langle + \rangle a_{10} \langle + \rangle a_{13} \langle + \rangle a_{14} = 1 & a_{11} \langle + \rangle a_{12} \langle + \rangle a_{15} \langle + \rangle a_{16} = 1 \\
a_3 \langle + \rangle a_7 \langle + \rangle a_{11} \langle + \rangle a_{15} = 1 & a_4 \langle + \rangle a_8 \langle + \rangle a_{12} \langle + \rangle a_{16} = 0.
\end{array}$$

Since $F(\mathbf{0}) = 0$, we know that $a_{16} = 0$, then the solution is

$$\begin{aligned}
m_F &= \mathrm{Row}_1(M_F) \\
&= [\begin{array}{llll}
a & b & a & b \\
c & a \langle + \rangle \neg b \langle + \rangle \neg c & c & a \langle + \rangle \neg b \langle + \rangle \neg c \\
\neg b \langle + \rangle \neg c & a \langle + \rangle \neg c & \neg b \langle + \rangle \neg c & a \langle + \rangle c \\
a \langle + \rangle \neg b & 1 & a \langle + \rangle \neg b & 0
\end{array}]
\end{aligned}$$

where a, b and c can be arbitrary Boolean numbers. For instance,

(i) Let $a = 1$, $b = 0$, and $c = 1$. Then we have

$$\begin{aligned}
F(x_1, x_2, x_3, x_4) &= (x_1 \wedge x_4) \vee (\neg x_1 \wedge x_2 \wedge x_3) \\
&\quad \vee (\neg x_1 \wedge x_2 \wedge x_4) \vee (\neg x_1 \wedge \neg x_2 \wedge x_3 \wedge \neg x_4).
\end{aligned}$$

(ii) Let $a = 0$, $b = 0$, and $c = 1$. Then we have

$$F(x_1, x_2, x_3, x_4) = [x_1 \wedge (x_2 \bar{\vee} x_4)] \vee [\neg x_1 \wedge (\neg x_2 \bar{\vee} x_3) \wedge \neg x_4].$$

13.3 Boolean Integral

There is no commonly used definition for Boolean integral. Tucker *et al.* (1988) provides a framework for Boolean integral. Unfortunately, the Boolean derivative used in Tucker *et al.* (1988) is different from the standard one, and hence the integral is in-consistent with the aforementioned Boolean derivative. Moreover, the computation problem has not been solved yet there.

In the following we define the Boolean integrals in the sense that they are precisely the inverse of the Boolean derivatives.

13.3.1 *Primitive Function*

First, we define the primitive function of a given Boolean function.

Definition 13.2. Given a Boolean function $f(x_1, \cdots, x_n)$. $F(x_1, \cdots, x_{i-1}, z, x_i, \cdots, x_n)$ is called the ith primitive function of $f(x)$ (or the ith partial integral of $f(x)$), denoted by

$$\int f(x_1, \cdots, x_n) d[i] = F(x_1, \cdots, x_{i-1}, z, x_i, \cdots, x_n), \tag{13.35}$$

if

$$\frac{\partial F}{\partial z} = f(x_1, \cdots, x_n). \tag{13.36}$$

In the light of Corollary 13.1, the problem becomes solving the equation

$$m_F \left[\Psi^i_{n+1}\right]^T = m_f. \tag{13.37}$$

To express the result in a condensed form we propose a notation as follows: Let $a, b \in \mathcal{B}_{2^n}$ be two Boolean vectors, and $1 \leq i \leq n+1$. Split a and b into 2^{i-1} equal-size blocks as

$$a = [a_1 \; a_2 \; \cdots \; a_{2^{i-1}}]; \quad b = [b_1 \; b_2 \; \cdots \; b_{2^{i-1}}],$$

and define

$$a \dashv_{[i]} b = \left[b_1 \; a_1 \langle + \rangle b_1 \; b_2 \; a_2 \langle + \rangle b_2 \; \cdots \; b_{2^{i-1}} \; a_{2^{i-1}} \langle + \rangle b_{2^{i-1}}\right].$$

Note that $\dashv_{[i]}$ is not commutative. That is, in general

$$a \dashv_{[i]} b \neq b \dashv_{[i]} a.$$

Using this notation, we have

Theorem 13.3. *Assume $f(x_1, \cdots, x_n)$ is a Boolean function with its truth table m_f. Then F is the ith primitive function ($1 \leq i \leq n+1$), if and only if, there is a constant $c \in \mathcal{B}_{2^n}$, such that*

$$m_F = m_f \dashv_{[i]} c. \tag{13.38}$$

Proof. Splitting m_F into 2^i equal-size blocks, and label them as

$$m_F = \left[c_{1,1} \; c_{1,2} \; c_{2,1} \; c_{2,2} \; \cdots \; c_{2^{i-1},1} \; c_{2^{i-1},2}\right].$$

Denote by

$$m_f = [\alpha_1 \; \alpha_2 \; \cdots \; \alpha_{2^{i-1}}].$$

Then we know that F is an ith primitive function of f, if and only if

$$c_{j,1} \langle + \rangle c_{j,2} = \alpha_j, \quad j = 1, \cdots, 2^{i-1}. \tag{13.39}$$

Consider equation (13.39) component-wise, one sees easily that if $c_{j,1}$ is fixed, there exists unique $c_{j,2}$, satisfying (13.39). Taking this into consideration, we can choose $c_{j,1}$ arbitrarily, and set $c_{j,2} = \alpha_j \langle + \rangle c_{j,1}$, which is the only solution of (13.39). The conclusion follows. □

We give some examples to demonstrate this.

Example 13.4. Assume $f(x_1, x_2, x_3) = x_3 \wedge (x_1 \vee (x_2 \leftrightarrow x_3))$. Find

$$\int f(x_1, x_2, x_3)d[2].$$

It is easy to calculate that

$$m_f = [1\ 0\ 1\ 0\ 1\ 0\ 0\ 0].$$

Assume

$$F(x_1, z, x_2, x_3) = \int f(x_1, x_2, x_3)d[2],$$

with its truth table as

$$m_F = [\alpha_1\ \alpha_2\ \alpha_3\ \alpha_4\ \alpha_5\ \alpha_6\ \alpha_7\ \alpha_8\ \alpha_9\ \alpha_{10}\ \alpha_{11}\ \alpha_{12}\ \alpha_{13}\ \alpha_{14}\ \alpha_{15}\ \alpha_{16}]^T.$$

Setting

$$\alpha_5 = c_1\ \alpha_6 = c_2\ \alpha_7 = c_3\ \alpha_8 = c_4\ \alpha_{13} = c_5\ \alpha_{14} = c_6\ \alpha_{15} = c_7\ \alpha_{16} = c_8.$$

We have $\alpha_1 = c_1 \langle + \rangle (m_f)_1 = \neg c_1$. A similar calculation shows

$$m_F = [\neg c_1\ c_2\ \neg c_3\ c_4\ c_1\ c_2\ c_3\ c_4\ \neg c_5\ c_6\ c_7\ c_8\ c_5\ c_6\ c_7\ c_8]^T.$$

Given any set of parameters, we can find a corresponding primitive function. Hence, we have totally 2^8 primitive functions. To give a particular primitive function, we choose a set of parameters as, say, $c_1 = c_2 = c_3 = c_4 = 0$, $c_5 = c_6 = c_7 = c_8 = 1$. Then we have

$$m_F = [1\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 1]^T.$$

If follows that

$$\begin{aligned} F(x_1, z, x_2, x_3) = {} & (x_1 \wedge z \wedge x_3) \vee ((\neg x_1) \wedge z \\ & \wedge ((x_2 \wedge (\neg x_3)) \vee (\neg x_2))) \vee ((\neg x_1) \wedge (\neg z)). \end{aligned}$$

Consider multiple integral case, we define

$$\begin{aligned} & \int \cdots \int f(x_1, \cdots, x_n)d[i_1] \cdots d[i_k] \\ & := \int \left(\cdots \int \left(\int f(x_1, \cdots, x_n)d[i_1] \right) d[i_2] \cdots \right) d[i_k]. \end{aligned} \tag{13.40}$$

Note that to make the index meaningful we require that

$$1 \le i_j \le n + j, \quad j = 1, \cdots, k.$$

Example 13.5. Assume $f(x_1, x_2, x_3)$ is the same as in Example 13.4. Find

$$\iint f(x_1, x_2, x_3)d[3]d[5].$$

Similar to the technique used in Example 13.4, we calculate

$$G(x_1, x_2, z_1, x_3) = \int f(x_1, x_2, x_3) d[3].$$

It is ready to see that

$$m_G = [c_1 \langle + \rangle 1 \; c_2 \; c_1 \; c_2 \; c_3 \langle + \rangle 1 \; c_4 \; c_3 \; c_4 \; c_5 \langle + \rangle 1 \; c_6 \; c_5 \; c_6 \; c_7 \; c_8 \; c_7 \; c_8].$$

Then we calculate

$$F(x_1, x_2, z_1, x_3, z_2) = \int G(x_1, x_2, z_1, x_3) d[5].$$

Then we have

$$\begin{aligned} m_F = [& d_1 \langle + \rangle c_1 \langle + \rangle 1 \; d_1 \quad d_2 \langle + \rangle c_2 \quad d_2 \quad d_3 \langle + \rangle c_1 \quad d_3 \quad d_4 \langle + \rangle c_2 \quad d_4 \\ & d_5 \langle + \rangle c_3 \langle + \rangle 1 \; d_5 \quad d_6 \langle + \rangle c_4 \quad d_6 \quad d_7 \langle + \rangle c_3 \quad d_7 \quad d_8 \langle + \rangle c_4 \quad d_8 \\ & d_9 \langle + \rangle c_5 \langle + \rangle 1 \; d_9 \quad d_{10} \langle + \rangle c_6 \; d_{10} \; d_{11} \langle + \rangle c_5 \; d_{11} \; d_{12} \langle + \rangle c_6 \; d_{12} \\ & d_{13} \langle + \rangle c_7 \qquad d_{13} \; d_{14} \langle + \rangle c_8 \; d_{14} \; d_{15} \langle + \rangle c_7 \; d_{15} \; d_{16} \langle + \rangle c_8 \; d_{16}], \end{aligned}$$

where c_i, $i = 1, \cdots, 8$, d_j, $j = 1, \cdots, 16$ are arbitrary Boolean numbers.

For instance, we may have some special F as follows.

(i) Assume $c_i = 0$, $\forall i$, and $d_j = 0$, $\forall j$. Then

$$F(x_1, \cdots, x_5) = (x_1 \vee x_2) \wedge x_3 \wedge x_4 \wedge x_5.$$

(ii) Assume $c_i = 0$, $\forall i$, $d_1 = d_5 = d_9 = d_{16} = 1$, and $d_j = 0$, $j \neq 1, 5, 9, 16$. Then

$$F(x_1, \cdots, x_5) = [(x_1 \vee x_2) \wedge x_3 \wedge x_4 \wedge \neg x_5] \vee [\neg(x_1 \vee x_2 \vee x_3 \vee x_4)].$$

13.3.2 *Indefinite Integral*

Definition 13.3. Given a logical function $F(x_1, \cdots, x_n)$. Its differential form, denoted by dF, is defined as

$$dF := \frac{\partial F}{\partial x_1} dx_1 + \cdots + \frac{\partial F}{\partial x_n} dx_n. \tag{13.41}$$

Note that in (13.41) the symbol " + " is considered as only an adjacent notation, but not an operator.

Definition 13.4. Given a set of functions

$$f_i(x_1, \cdots, x_{i-1}, \hat{x}_i, x_{i+1} \cdots, x_n), \quad i = 1, \cdots, n.$$

A function $F(x_1, \cdots, x_n)$ is called the indefinite integral of the differential form

$$dh = f_1 dx_1 + f_2 dx_2 + \cdots + f_n dx_n$$

(or simply, integral of $\{f_1, \cdots, f_n\}$), if

$$\frac{\partial F}{\partial x_i} = f_i, \quad i = 1, \cdots, n. \tag{13.42}$$

Note that according to equation (13.10) one sees that if F is an indefinite integral of dh, then so is $\bar{F}$.

Next, we consider when the indefinite integral exists.

Theorem 13.4. *Consider a differential form*

$$\begin{aligned} dh = {} & h_1(\hat{x}_1, x_2, \cdots, x_n)dx_1 + h_2(x_1, \hat{x}_2, \cdots, x_n)dx_2 \\ & + \cdots + h_n(x_1, x_2 \cdots \hat{x}_n)dx_n. \end{aligned}$$

There exists at least a pair of complemented indefinite integrals, if and only if

$$\frac{\partial h_i}{\partial x_j} = \frac{\partial h_j}{\partial x_i}, \quad 1 \le i < j \le n. \tag{13.43}$$

Proof. Necessity is trivial. We prove the sufficiency. Using $\{h_i | i = 1, \cdots, n\}$, we can calculate

$$\frac{\partial h_i}{\partial x_j} = \frac{\partial h_j}{\partial x_i}, \quad 1 \le i < j \le n.$$

Similarly, we have third-order cross derivatives as

$$\frac{\partial^2 h_i}{\partial x_j \partial x_k} = \frac{\partial^2 h_j}{\partial x_i \partial x_k} = \frac{\partial^2 h_k}{\partial x_i \partial x_j},$$

and even higher order cross derivatives. Using the obtained partial derivatives and following the form of MacLaurin expansion we can construct

$$\begin{aligned} F(x_1, \cdots, x_n) = {} & c \langle + \rangle \langle + \rangle_{i=1}^{n} h_i|_{\mathbf{0}} \wedge x_i \langle + \rangle \langle + \rangle_{1 \le i_1 < i_2 \le n} \left. \frac{\partial h_{i_1}}{\partial x_{i_2}} \right|_{\mathbf{0}} \wedge x_{i_1} \wedge x_{i_2} \\ & \langle + \rangle \langle + \rangle_{1 \le i_1 < i_2 < i_3 \le n} \left. \frac{\partial^2 h_{i_1}}{\partial x_{i_2} \partial x_{i_3}} \right|_{\mathbf{0}} \wedge x_{i_1} \wedge x_{i_2} \wedge x_{i_3} \langle + \rangle \cdots \\ & \langle + \rangle \left. \frac{\partial^{n-1} h_1}{\partial x_2 \cdots \partial x_n} \right|_{\mathbf{0}} \wedge x_1 \wedge x_2 \wedge \cdots \wedge x_n. \end{aligned} \tag{13.44}$$

Then it is ready to verify that

$$F(x) = \int dh.$$

□

The following result comes from the constructive proof of Theorem 13.4.

Corollary 13.4. *If $\int dh$ exists, then it is unique (up to a complement equivalence).*

In the following when we consider the integral of a differential form, we assume

A1 Integrable condition (13.43) holds.

Hence as long as dh is integrable, we can write an indefinite integral as

$$\int dh = F(x) \langle + \rangle c,$$

where $c \in \mathcal{D}$.

In later use, we would like to specify F. So we also use the following notation:

$$\int dh = F(x), \quad F(0) = 0;$$

$$\int \bar{d}h = \bar{F}(x), \quad \bar{F}(0) = 1.$$

Next, we consider how to calculate the indefinite integral. In fact, the constructive proof already provides a method to find the integral. We are looking another simple proof.

The following result is an immediate consequence of Corollary 13.1.

Theorem 13.5. *Each indefinite integral of a differential form $dh = f_1 dx_1 + \cdots + f_n dx_n$ has a solution z of the following linear Galois algebraic system as its truth table.*

$$\Psi_n \, (\ltimes) \, z = b, \tag{13.45}$$

where

$$\Psi_n = \begin{bmatrix} \Psi_n^1 \\ \Psi_n^2 \\ \vdots \\ \Psi_n^n \end{bmatrix} \in \mathcal{B}_{n2^{n-1} \times 2^n}; \quad \textit{and} \quad b = \begin{bmatrix} m_{f_1}^T \\ m_{f_2}^T \\ \vdots \\ m_{f_n}^T \end{bmatrix} \in \mathcal{B}_{n2^{n-1}}.$$

It is worth noting that to get the default solution, we need to set the last component of z, $z_{2^n} = 0$, which corresponds to $F(0) = 0$.

Example 13.6.

(1) Assume $n = 2$. Then we have

$$\Psi_2 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

Case 1: Assume $f_1 = x_2$, $f_2 = \neg x_1$. Then we have $m_{f_1} = [1\ 0]$, $m_{f_2} = [0\ 1]$. Equation (13.45) becomes

$$\begin{bmatrix} 1\,0\,1\,0 \\ 0\,1\,0\,1 \\ 1\,1\,0\,0 \\ 0\,0\,1\,1 \end{bmatrix} (\ltimes) \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Equivalently, we have

$$\begin{cases} z_1 \langle + \rangle z_3 = 1 \\ z_2 \langle + \rangle z_4 = 0 \\ z_1 \langle + \rangle z_2 = 0 \\ z_3 \langle + \rangle z_4 = 1. \end{cases}$$

Setting $z_4 = 0$, which corresponds to $F(\mathbf{0}) = 0$, we have

$$\begin{cases} z_1 = 0 \\ z_2 = 0 \\ z_3 = 1 \\ z_4 = 0. \end{cases}$$

That is, $m_F = [0\ 0\ 1\ 0]$. Hence, $F = (\neg x_1) \wedge x_2$. Writing it into integral form, we have

$$\int x_2 dx_1 + \neg x_1 dx_2 = (\neg x_1) \wedge x_2. \tag{13.46}$$

We also have

$$\int x_2 \bar{d} x_1 + \neg x_1 \bar{d} x_2 = (\neg x_1) \wedge x_2 \langle + \rangle 1 = x_1 \vee (\neg x_2). \tag{13.47}$$

Case 2: Assume $f_1 = x_2$, $f_2 = 1$. Then we have $m_{f_1} = [1\ 0]$ and $m_{f_2} = [1\ 1]$. It is easy to check that there is no solution. Hence the integral does not exist.

(2) Assume $n = 3$. Then it is easy to calculate that

$$\Psi_3 = \begin{bmatrix} 1&0&0&0&1&0&0&0\\ 0&1&0&0&0&1&0&0\\ 0&0&1&0&0&0&1&0\\ 0&0&0&1&0&0&0&1\\ 1&0&1&0&0&0&0&0\\ 0&1&0&1&0&0&0&0\\ 0&0&0&0&1&0&1&0\\ 0&0&0&0&0&1&0&1\\ 1&1&0&0&0&0&0&0\\ 0&0&1&1&0&0&0&0\\ 0&0&0&0&1&1&0&0\\ 0&0&0&0&0&0&1&1 \end{bmatrix}.$$

Consider the indefinite integral of

$$dh = (x_3 \rightarrow x_2)dx_1 + \neg(x_1 \wedge x_3)dx_2 + (x_1 \wedge (\neg x_2)dx_3.$$

It is easy to calculate that

$$m_{f_1} = [1\ 1\ 0\ 1]; \quad m_{f_2} = [0\ 1\ 1\ 1]; \quad m_{f_3} = [0\ 1\ 0\ 0].$$

Consider

$$\Psi_3 z = b,$$

where

$$b = [1\ 1\ 0\ 1\ 0\ 1\ 1\ 1\ 0\ 1\ 0\ 0]^T.$$

Equivalently, we have

$$\begin{cases} z_1 \langle + \rangle z_5 = z_2 \langle + \rangle z_6 = z_4 \langle + \rangle z_8 = z_2 \langle + \rangle z_4 \\ \quad = z_5 \langle + \rangle z_7 = z_6 \langle + \rangle z_8 = z_3 \langle + \rangle z_4 = 1 \\ z_3 \langle + \rangle z_7 = z_1 \langle + \rangle z_3 = z_1 \langle + \rangle z_2 = z_5 \langle + \rangle z_6 = z_7 \langle + \rangle z_8 = 0. \end{cases}$$

Setting $z_8 = 0$, we have the following solution

$$m_F = [0\ 0\ 0\ 1\ 1\ 1\ 0\ 0].$$

It is easy to calculate that the corresponding function, which leads to

$$\begin{aligned} &\int (x_3 \rightarrow x_2)dx_1 + \neg(x_1 \wedge x_3)dx_2 + (x_1 \wedge (\neg x_2))dx_3 \\ &\quad = \neg(x_1 \leftrightarrow x_2) \wedge (\neg x_1 \vee \neg x_3), \end{aligned} \tag{13.48}$$

and

$$\begin{aligned} &\int (x_3 \rightarrow x_2)\bar{d}x_1 + \neg(x_1 \wedge x_3)\bar{d}x_2 + (x_1 \wedge (\neg x_2))\bar{d}x_3 \\ &\quad = (x_1 \leftrightarrow x_2) \vee (x_1 \wedge x_3). \end{aligned} \tag{13.49}$$

13.3.3 *Definite Integral*

When the indefinite integral of $\int dh$ exists, it is a pair $(F, \bar{F})$. Then we define the definite integral as follows.

Definition 13.5. Assume there is a differential form dh as

$$\begin{aligned} dh = & f_1(\hat{x}_1, \cdots, x_n)dx_1 + f_2(x_1, \hat{x}_2, \cdots, x_n)dx_2 \\ & + \cdots + f_n(x_1, \cdots, x_{n-1}, \hat{x}_n)dx_n, \end{aligned}$$

a subset $S \subset \mathcal{D}^n$, and a logical function $g(x)$.

Assume $\int dh = F(x)$ (with $F(0) = 0$). Then we define

$$\int_S g(x)dh = \sum_{x \in S} g(x) \wedge F(x). \tag{13.50}$$

(Note that here the summation $\sum$ is the conventional plus on $\mathbb{R}$.)

We also define the integral with respect to $\bar{F}$ as

$$\int_S g(x)\bar{d}h = \sum_{x \in S} g(x) \wedge \bar{F}(x). \tag{13.51}$$

S is called the integral domain and $g(x)$ the integrand.

It is easy to show that the definite integral, defined in Definition 13.5, satisfies some basic properties of the definite integral. For instance, we have

(1) If $f(x) \le g(x)$, then

$$\int_S f(x)dh \le \int_S g(x)dh. \tag{13.52}$$

(2) If $S_1 \subseteq S_2$, then

$$\int_{S_1} g(x)dh \le \int_{S_2} g(x)dh. \tag{13.53}$$

(3)

$$\int_{S_1 \cup S_2} g(x)dh = \int_{S_1} g(x)dh + \int_{S_2} g(x)dh - \int_{S_1 \cap S_2} g(x)dh. \tag{13.54}$$

Define

$$\text{supp}(f) = \{x | f(x) \neq 0\}.$$

Let $F = \int dh$. Then we have

$$\int_S g(x)dh = |\text{supp}(F) \cap \text{supp}(g) \cap S|.$$

Example 13.7. Recall Example 13.6. We consider the definite integrals using the corresponding indefinite integrals.

(1) $n = 2$.

Case 1: Assume $f_1 = x_2$, $f_2 = \neg x_1$. Using (13.44), the default indefinite integral is $F = x_1 \wedge (\neg x_2)$. Assume $S = \{(x_1, x_2) | x_1 \to x_2 = 1\}$. Then

$$S = \{(1,1), (0,1), (0,0)\}.$$

Let the integrand be $g(x_1, x_2) = x_1 \leftrightarrow x_2$. Then it is easy to calculate that

$$\int_S (x_1 \leftrightarrow x_2) x_2 dx_1 + \neg x_1 dx_2 = 0;$$

and

$$\int_S (x_1 \leftrightarrow x_2) x_2 \bar{d}x_1 + \neg x_1 \bar{d}x_2 = 2.$$

Case 2: Assume $f_1 = x_2$, $f_2 = 1$. Then

$$\int_S x_2 dx_1 + dx_2$$

does not exist

(2) Consider $n = 3$ and

$$dh = (x_3 \to x_2)dx_1 + \neg(x_1 \wedge x_3)dx_2 + (x_1 \wedge (\neg x_2)dx_3.$$

Assume

$$S = \left\{ x \in \mathcal{D}^3 \middle| \, (x_1 \vee x_2) \wedge x_3 = 0 \right\}.$$

Then

$$S = \{(1,1,0),\ (1,0,0),\ (0,1,0),\ (0,0,0),\ (0,0,1)\}.$$

Let $g(x) = x_1 \vee (x_3 \to x_2)$. Using (13.48), a straightforward computation shows that

$$\int [x_1 \vee (x_3 \to x_2)](x_3 \to x_2)dx_1 + \neg(x_1 \wedge x_3)dx_2 + (x_1 \wedge (\neg x_2) = 2.$$

Exercises

13.1 Given a Boolean function

$$f(x_1, x_2, x_3, x_4) = (\neg x_1 \wedge x_2) \leftrightarrow [x_3 \vee (x_4 \to x_2)].$$

(i) Calculate

$$\frac{\partial f}{\partial x_i}, \quad i = 1, 2, 3, 4.$$

(ii) Calculate

$$\frac{\partial^2 f}{\partial x_2^2}, \quad \frac{\partial^2 f}{\partial x_2 x_3}, \quad \frac{\partial^2 f}{\partial x_1 x_4}.$$

(iii) Calculate

$$\frac{\partial^2 f}{\partial x_1 x_2 x_3}, \quad \frac{\partial^3 f}{\partial x_2 x_3 x_4}.$$

13.2 Prove equations (13.3)–(13.10) in Proposition 13.1.
13.3 Prove equation (13.14) in Proposition 13.2.
13.4 Prove equation (13.22) in Corollary 13.2.
13.5 Prove equations (13.22) and (13.23) in Corollary 13.3.
13.6 Prove the MacLaurin expansion of Boolean functions (13.24).
13.7 Given a Boolean function

$$f(x_1, x_2, x_3) = (x_1 \leftrightarrow x_2) \vee (x_3 \rightarrow x_1).$$

Find its MacLaurin expansion.
13.8 A logical function is given as

$$f(x_1, x_2, x_3, x_4) = x_1 \wedge (\neg x_2 \bar{\vee} (x_3 \leftrightarrow x_4)).$$

Calculate the following:
(i) $m_{\partial_1} f$; $m_{\partial_2} f$; $m_{\partial_3} f$; $m_{\partial_4} f$.
(ii) $m_{\partial_{[1]}} f$; $m_{\partial_{[2]}} f$; $m_{\partial_{[3]}} f$; $m_{\partial_{[4]}} f$.
(iii) $m_{\partial_{2,3}} f$; $m_{\partial_{[2,3]}}$.
13.9 Calculate Ψ_n^i for (i) $n = 3$, $i = 2$; (ii) $n = 4$, $i = 1$; (iii) $n = 4$, $i = 3$.
13.10 A Boolean function f has it structure matrix M_f as follows. Find its MacLaurin expansion.
(i) $M_f = \delta_2[1\ 0\ 0\ 1\ 1\ 1\ 1\ 0]$.
(ii) $M_f = \delta_2[0\ 1\ 1\ 0\ 1\ 0\ 0\ 0]$.
(iii) $M_f = \delta_2[0\ 1\ 1\ 0\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 0\ 0]$.
13.11 Denote by $x_{\langle k \rangle}$ the column staking form of

$$S_k = \{x_{i_1} \ltimes \cdots \ltimes x_{i_k} \mid i_1 < i_2 < \cdots < i_k\},$$

arranged in the natural alphabetic order. Precisely,

$$x_{\langle 1 \rangle} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad x_{\langle 2 \rangle} = \begin{bmatrix} x_1 x_2 \\ x_1 x_3 \\ \vdots \\ x_{n-1} x_n \end{bmatrix}, \cdots$$

$$x_{\langle k\rangle} = \begin{bmatrix} x_1x_2\cdots x_k \\ x_1x_2\cdots x_{k+1} \\ \vdots \\ x_{n-k+1}x_{n-k+2}\cdots x_k \end{bmatrix}, \cdots .$$

Using this set of notations, the MacLaurin expansion of a Boolean function can be expressed into its vector form as

$$f(x) = c_0 + c_1x_{\langle 1\rangle} + c_2x_{\langle 2\rangle} + \cdots + c_nx_{\langle n\rangle}.$$

Find the coefficient matrices c_i, $i = 1, 2, \cdots, n$.

13.12 A Boolean mapping $F : \mathcal{D}^3 \to \mathcal{D}^2$ has its structure matrix as

$$M_F = \delta_4[1\ 2\ 4\ 3\ 3\ 4\ 2\ 1].$$

In vector form, its MacLaurin expansion can be expressed as

$$F(x) = C_0 + C_1x_{\langle 1\rangle} + C_2x_{\langle 2\rangle} + C_3x_{\langle 3\rangle} + C_4x_{\langle 4\rangle}.$$

Calculate the coefficients.

13.13 (i) Let $f(x_1, x_2, x_3) = x_1 \leftrightarrow (x_2 \wedge x_3)$. Solve the following equation:

$$\begin{cases} x_1 \leftrightarrow \frac{\partial f}{\partial x_2} = 1 \\ \frac{\partial^2 f}{\partial x_{2,s_3}} \vee x_2 = 0. \end{cases}$$

(ii) Let $f(x_1, x_2, x_3, x_4) = (x_1 \vee x_2) \to (x_3 \wedge x_4)$. Solve the following equation:

$$\begin{cases} \frac{\partial f}{\partial x_4} \vee x_4 = 0 \\ x_4 \leftrightarrow \frac{\partial^2 f}{\partial x_{2,s_3}} = 1 \\ \frac{\partial^3 f}{\partial x_{1,x_3,x_4}} \wedge (x_3 \vee x_4) = 1. \end{cases}$$

13.14 Recall Example 13.2 and let f be as in (13.31). Find the test vector set for the double stuck at $x_1(s-a-0)$, and $x_3(s-a-1)$.

13.15 Solving the following Boolean differential equations:

(i) $F = F(x_1, x_2, x_3)$, $F(0) = 0$, and

$$\begin{cases} \frac{\partial F}{\partial x_1} = x_2 \wedge x_3 \\ \frac{\partial^2 F}{\partial x_2 \partial x_3} = \neg x_1 \\ \frac{\partial F}{\partial x_3} = x_1 \bar{\vee} x_2. \end{cases}$$

(ii) $F = F(x_1, x_2, x_3, x_4)$, $F(0) = 1$, and

$$\begin{cases} \frac{\partial^2 F}{\partial x_1 \partial x_2} = x_3 \vee x_4 \\ \frac{\partial F}{\partial x_1} = \neg(x_2 \to x_3) \\ \frac{\partial^2 F}{\partial x_3 \partial x_4} = x_1 \bar{\vee} x_2 \\ \frac{\partial F}{\partial x_3} = x_1 \leftrightarrow (x_2 \wedge x_4). \end{cases}$$

13.16 Given $\alpha, \beta \in \mathcal{B}_{2^4}$ as

$$\begin{aligned} \alpha &= [1\ 0\ 1\ 0\ 0\ 1\ 0\ 1\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0] \\ \beta &= [1\ 1\ 0\ 0\ 0\ 1\ 0\ 1\ 0\ 0\ 1\ 1\ 1\ 0\ 1\ 0]. \end{aligned}$$

Calculate the following:

(i) (a) $\alpha \dashv_{[2]} \beta$; (b) $\alpha \dashv_{[3]} \beta$.

(ii) (a) $\beta \dashv_{[2]} \alpha$; (b) $\beta \dashv_{[3]} \alpha$.

(iii) Check that in general

$$\alpha \dashv_{[i]} \beta \ \neq \ \beta \dashv_{[i]} \alpha.$$

13.17 Find the primitive functions for the following integrands.

(i)

$$\int (x_1 \to x_2) \vee x_3 d[2];$$

(ii)

$$\int (x_1 \wedge x_2) \leftrightarrow (x_3 \vee x_4) d[3];$$

(iii)

$$\int \int (x_1 \wedge x_2) d[1] d[2].$$

13.18 Check whether the indefinite integrals of the following differential forms exist. If "yes", find them.

(i)

$$\begin{aligned} dh = {} & [\neg x_2 \wedge (x_3 \to x_4)]\, dx_1 + [\neg x_1 \wedge (x_3 \to x_4)]\, dx_2 \\ & + [(x_1 \vee x_2) \wedge \neg x_4]\, dx_3 + [(x_1 \vee x_2) \wedge x_3]\, dx_4; \end{aligned}$$

(ii)

$$\begin{aligned} dh = {} & [\neg x_4 \vee (x_1 \leftrightarrow x_2)]\, dx_1 + [\neg x_3 \wedge x_2]\, dx_2 \\ & + [x_3 \vee x_1]\, dx_3 + [(x_2 \wedge x_1) \vee x_4]\, dx_4; \end{aligned}$$

(iii)

$$\begin{aligned} dh = {} & [x_2 \vee x_3 \vee x_4]\, dx_1 + [x_1 \wedge \neg(x_3 \vee x_4)]\, dx_2 \\ & + [(x_1 \wedge \neg(x_2 \vee x_4)]\, dx_3 + [x_1 \wedge \neg(x_2 \vee x_3)]\, dx_4. \end{aligned}$$

13.19 Calculate the definite integrals

$$\int_S dh, \quad \text{and} \quad \int_S d\bar{h},$$

where dh comes from previous exercise and where the dh is integrable, and

$$S = \left\{ x \in \mathcal{D}^4 | x_1 \wedge x_2 = x_3 \vee x_4 \right\}.$$

Chapter 14

Lattice, Graph, and Universal Algebra

Lattice is a widely used concept. It is applicable to fuzzy logic (Kerre *et al.*, 2004), mathematical logic (Barnes and Mack, 1975), graph theory (Chartrand and Zhang, 2005), and universal algebra (Burris and Sankappanavar, 1981), etc. In this chapter we first investigate lattice and its matrix expressions. Then we give a framework for matrix expression of graph. Finally, a brief introduction to universal algebra is presented.

14.1 Lattice

To begin with, we introduce two equivalent definitions of lattice. They are convenient for certain different situations.

Definition 14.1. A nonempty set L together with two binary operations: joint ($\sqcup$) and meet ($\sqcap$) on L is called a lattice, if it satisfies the following identities:

(1) (Commutative Laws)

$$x \sqcup y = y \sqcup x; \tag{14.1}$$

$$x \sqcap y = y \sqcap x. \tag{14.2}$$

(2) (Associative Laws)

$$x \sqcup (y \sqcup z) = (x \sqcup y) \sqcup z; \tag{14.3}$$

$$x \sqcap (y \sqcap z) = (x \sqcap y) \sqcap z. \tag{14.4}$$

(3) (Idempotent Laws)

$$x \sqcup x = x; \tag{14.5}$$

$$x \sqcap x = x. \tag{14.6}$$

(4) (Absorption Laws)

$$x = x \sqcup (x \sqcap y); \tag{14.7}$$

$$x = x \sqcap (x \sqcup y). \tag{14.8}$$

Example 14.1. The following objects are lattices.

(1) (Boolean algebra) Consider $\mathcal{D}$ with the disjunction as the joint, and the conjunction as the meet:

$$\sqcap := \vee; \quad \sqcap := \wedge.$$

(2) (Natural number) Consider the set of natural numbers $\mathbb{N}$. Let the joint be the least common multiple, and the meet be the greatest common divisor:

$$\sqcup(a, b) = \text{lcm}(a, b);$$
$$\sqcap(a, b) = \gcd(a, b).$$

We leave the verification of the above lattices to the reader.

To introduce the second definition of a lattice, we need the concept of partial order.

Definition 14.2. A binary relation $\leq$ defined on a set A is a partial order on the set A if the following conditions hold identically in A:

(i) (Reflexivity)
$a \leq a$;

(ii) (Antisymmetry)
$a \leq b$ and $b \leq a$ imply $a = b$;

(iii) (Transitivity)
$a \leq b$ and $b \leq c$ imply $a \leq c$.

If, in addition, for every a, b in A

(iv) $a \leq b$ or $b \leq a$,

then we say $\leq$ is a total order on A.

Definition 14.3.

- A nonempty set with a partial order on it is called a partially ordered set, briefly, poset.
- A nonempty set with a total order on it is called a totally ordered set, or linearly ordered set, or a chain.

- In a partially ordered set, if $a \leq b$ but $a \neq b$, then it is said that $a < b$.

Example 14.2.

(1) Let $\mathcal{P}(A)$ denote the power set of A, and the order $\leq$ be the inclusion $\subseteq$. Then $(\mathcal{P}(A), \leq)$ is a partially ordered set.
(2) Let $\mathbb{N}$ be the set of natural numbers, and the order $\leq$ be the relation "divides". For instance, $2 \leq 4$, but $2 \not\leq 3$. Then $(\mathbb{N}, \leq)$ is a partially ordered set.
(3) If $\leq$ has the conventional meaning as $2 \leq 3$, then $(\mathbb{N}, \leq)$ is a totally ordered set (or liner order set, or chain).

Definition 14.4. Let A be a subset of a poset P.

(1) $p \in P$ is an upper bound (a lower bound) of A, if $a \leq p$ ($p \leq a$) for all $a \in A$.
(2) $p \in P$ is the least upper bound of A, or supremum of A (denoted by $p = \sup A$), if p is an upper bound of A, and for any other upper bound of A, say, b, we have $p \leq b$.
(3) $p \in P$ is the greatest lower bound of A, or infimum of A (denoted by $p = \inf A$), if p is a lower bound of A, and and for any other lower bound of A, say, b, we have $p \geq b$.
(4) For $a, b \in P$, we say b covers a, or a is covered by b (denoted by $a \prec b$), if $a < b$, and whenever $a \leq c \leq b$ it follows that $a = c$ or $c = b$.
(5) An interval is defined as: $[a, b] = \{c \in P \mid a \leq c \leq b\}$.
(6) An open interval is defined as: $(a, b) = \{c \in P \mid a < c < b\}$.

Definition 14.5. A finite poset P can be described by a directed graph ($\mathcal{N}$, $\mathcal{E}$), where $\mathcal{N}$ is the set of nodes and $\mathcal{E}$ is the set of edges. The graph is constructed as following:

(i) $\mathcal{N} = P$;
(ii) $\mathcal{E} \subset P \times P$, and $(a, b) \in \mathcal{E}$ (i.e., there is an edge from a to b), if and only if $b < a$.

Such a graph is called the Hasse diagram of poset P.

Fig 14.1 describes Hasse diagrams of four posets A, B, C (M_5) and D (N_5). The graph C is called M_5 and graph D is called N_5, they will be used in the sequel.

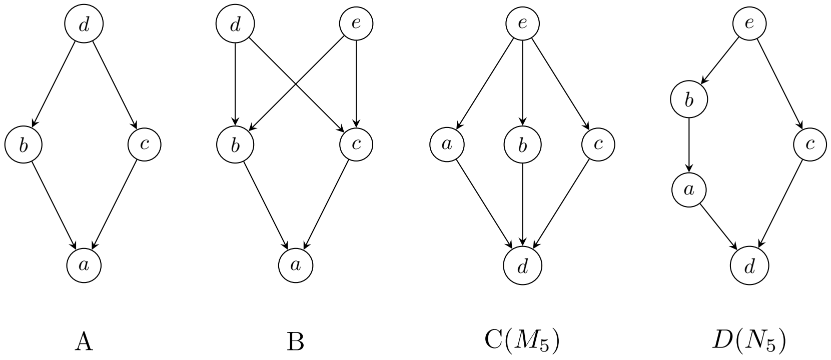


Fig. 14.1 Examples of Hasse diagrams

Now we are ready to introduce the second definition of lattice.

Definition 14.6. A poset L is a lattice, if and only if for any two elements $a, b \in L$ both $\sup\{a, b\}$ and $\inf\{a, b\}$ exist.

Consider Fig 14.1, we can see A, C, D are lattices, but B is not, because in B the $\sup\{b, c\}$ does not exist.

Definition 14.1 and Definition 14.6 are equivalent in the following sense: If L is a lattice by one of the two definitions then we can construct in a simple and uniform fashion on the same set L a lattice by the other definition. We state it as a theorem.

Theorem 14.1.

(i) If L is a lattice by Definition 14.1, define $\leq$ on L as follows: $a \leq b$, if and only if $a = a \sqcap b$, then L satisfies the requirements in Definition 14.6.

(ii) If L is a lattice by Definition 14.6, define the operatios $\sqcup$ and $\sqcap$ by $a \sqcup b = \sup\{a, b\}$, and $a \sqcap b = \inf\{a, b\}$, then L satisfies the requirements in Definition 14.1.

Proof. (i) We need to show that $\leq$ is a partial order and $\sup\{a, b\}$, $\inf\{a, b\}$ exist.

- (Reflexivity) $a \sqcap a = a$ implies $a \leq a$.
- (Antisymmetry) Assume $a \leq b$ and $b \leq a$. Then we have $a = a \sqcap b$, and $b = a \sqcap b$, thus $a = b$.

- (Transitivity) Assume $a \leq b$ and $b \leq c$. Then we have $a = a \sqcap b$ and $b = b \sqcap c$. By associativity,

$$a = a \sqcap (b \sqcap c) = (a \sqcap b) \sqcap c = a \sqcap c.$$

  Hence, $a \leq c$.

We conclude that $\leq$ is partial order.

Next, we prove $\sup\{a, b\}$ and $\inf\{a, b\}$ exist.
Using absorption laws, we have $a = a \sqcap (a \sqcup b)$. So $a \leq a \sqcup b$. Similarly, $b \leq a \sqcup b$. Hence, $a \sqcup b$ is an upper bound of $\{a, b\}$.
For arbitrary upper bound u of $\{a, b\}$, since $a \leq u$, $b \leq u$, we have $a \sqcup u = (a \sqcap u) \sqcup u = u$ (by (14.7)), similarly $b \sqcup u = u$. Then $a \sqcup b \sqcup u = a \sqcup u = u$. Using absorption laws again, we have $(a \sqcup b) \sqcap u = (a \sqcup b) \sqcap [(a \sqcup b) \sqcup u] = a \sqcup b$, then $a \sqcup b \leq u$. Thus $\sup\{a, b\} = a \sqcup b$.
A similar argument shows that $\inf\{a, b\} = a \sqcap b$.

(ii) A straightforward computation shows that the defined joint $\sqcup$ and meet $\sqcap$ satisfy equations (14.1)–(14.8).

□

14.2 Isomorphic Lattices and Sublattices

Definition 14.7. Two lattices L_1 and L_2 are isomorphic if there is a bijective π from L_1 to L_2 such that for every $a, b \in L_1$ the following two equations hold:

(i) $\pi(a \sqcup b) = \pi(a) \sqcup \pi(b)$;
(ii) $\pi(a \sqcap b) = \pi(a) \sqcap \pi(b)$.

Such an π is called an isomorphism.

One would naturally like to reformulate the definition of isomorphism in terms of the corresponding order relations.

Definition 14.8. If P_1 and P_2 are two posets and π is a map from P_1 to P_2, then we say α is order-preserving if $\pi(a) \leq \pi(b)$ holds in P_2 whenever $a \leq b$ holds in P_1.

But a bijection α which is order-preserving may not be an isomorphism, see Fig 14.2 for a counter-example.

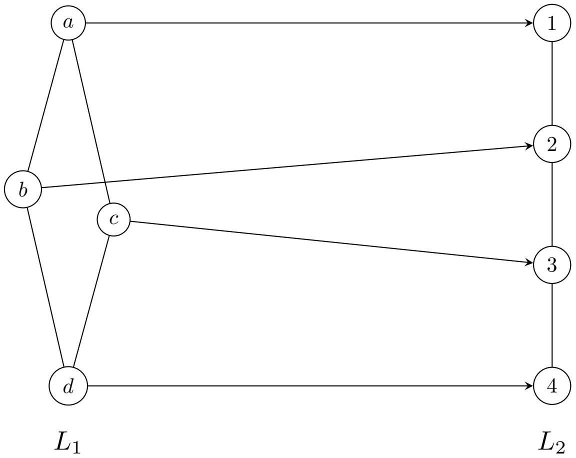


Fig. 14.2 An order-preserving bijection

Theorem 14.2. *Two lattices L_1 and L_2 are isomorphic, if and only if there is a bijection π from L_1 to L_2 such that both π and π^{-1} are order-preserving.*

Proof. (Necessity) For $a \leq b$ in L_1, since π is an isomorphism, $\pi(a) = \pi(a \sqcap b) = \pi(a) \sqcap \pi(b)$. Thus $\pi(a) \leq \pi(b)$, π is order-preserving. As π^{-1} is also an isomorphism, it is also order-preserving.

(Sufficiency) Let π be a bijection from L_1 to L_2 such that both π and π^{-1} are order-preserving. We want to prove $\pi(a \sqcup b) = \pi(a) \sqcup \pi(b)$, that is to say $\pi(a \sqcup b)$ is the supremum of $\{\pi(a), \pi(b)\}$.

Since $a \leq a \sqcup b$ in L_1, we have $\pi(a) \leq \pi(a \sqcup b)$. Similarly, $\pi(b) \leq \pi(a \sqcup b)$. Thus $\pi(a \sqcup b)$ is an upper bound of $\{\pi(a), \pi(b)\}$.

Next, for arbitrary $u \in L_2$ such that $\pi(a) \leq u$, $\pi(b) \leq u$. Since π^{-1} is order-preserving, $a \leq \pi^{-1}(u)$. Similarly, $b \leq \pi^{-1}(u)$. Thus $a \sqcup b \leq \pi^{-1}(u)$, then $\pi(a \sqcup b) \leq u$. This implies that $\pi(a \sqcup b) = \pi(a) \sqcup \pi(b)$. Similarly, it can be argued that $\pi(a \sqcap b) = \pi(a) \sqcap \pi(b)$. □

Definition 14.9. If L is a lattice and $H \neq \varnothing$ is a subset of L such that for every pair of elements $a, b \in H$ both $a \sqcup b$ and $a \sqcap b$ are in H, then we say that H with the same operations (restricted to H) is a sublattice of L.

Definition 14.10. A lattice L_1 can be embedded into a lattice L_2 if there is a sublattice of L_2 isomorphic to L_1; in this case we also say L_2 contains a copy of L_1 as a sublattice.

14.3 Matrix Expression of Finite Lattice

Assume $L = \{v_1, \cdots, v_n\}$ is a finite set and there exists an r-ary operators $\pi : \underbrace{L \times \cdots \times L}_{r} \to L$. To use matrix approach we simply identify

$$v_i \sim \delta_n^i, \quad i = 1, \cdots, n. \tag{14.9}$$

$\delta_n^i \in \Delta_n$ is called the vector form of v_i. Denote

$$\pi(v_{i_1}, \cdots, v_{i_k}) = v_{\pi(i_1,\cdots,i_k)}, \quad 1 \le i_1, \cdots, i_k \le k.$$

Then we can construct a matrix, called the structure matrix of π as

$$M_\pi = \delta_n \left[\pi(1, 1, \cdots, 1)\ \pi(1, 1, \cdots, 2)\ \cdots\ \pi(1, 1, \cdots, n)\ \cdots\ \pi(n, n, \cdots, n)\right]. \tag{14.10}$$

It is easy to check that in vector form we have

$$\pi(x_1, \cdots, x_k) = M_\pi \ltimes_{i=1}^k x_i, \quad x_i \in \Delta_n. \tag{14.11}$$

Example 14.3. Consider Galois field $\mathbb{Z}_5$. We identify

$$i \sim \delta_5^{i+1}, \quad i = 0, 1, 2, 3, 4.$$

Then for addition $\langle + \rangle$, the structure matrix is

$$M_a = \delta_5[1\ 2\ 3\ 4\ 5\ 2\ 3\ 4\ 5\ 1\ 3\ 4\ 5\ 1\ 2\ 4\ 5\ 1\ 2\ 3\ 5\ 1\ 2\ 3\ 4].$$

For product $\langle \times \rangle$, the structure matrix is

$$M_p = \delta_5[1\ 1\ 1\ 1\ 1\ 1\ 2\ 3\ 4\ 5\ 1\ 3\ 5\ 2\ 4\ 1\ 4\ 2\ 5\ 3\ 1\ 5\ 4\ 3\ 2].$$

Now assume $L = \{v_1, \cdots, v_n\}$ is given and there are two binary operators $\sqcup$ and $\sqcap$. Assume the structure matrices of these two operators are $M_\sqcup$ and $M_\sqcap$ respectively. Then we have the following result.

Theorem 14.3. *Let L be described as above. $(L, \sqcup, \sqcap)$ is a lattice, if and only if*

(1) (Commutative Laws)

$$M_\sqcup(I - W_{[n]}) = 0; \tag{14.12}$$

$$M_\sqcap(I - W_{[n]}) = 0. \tag{14.13}$$

(2) (Associative Laws)

$$M_\sqcup(I_n \otimes M_d) = M_\sqcup^2; \tag{14.14}$$

$$M_\sqcap(I_n \otimes M_c) = M_\sqcap^2. \tag{14.15}$$

(3) (Idempotent Laws)

$$M_{\sqcup} M_{r,n} = I; \tag{14.16}$$

$$M_{\sqcap} M_{r,n} = I. \tag{14.17}$$

Note that where $M_{r,n}$ is the power-reducing matrix.

(4) (Absorption Laws)

$$\mathrm{Blk}_i \left(M_{\sqcup}(I_n \otimes M_{\sqcap}) M_{r,n} W_{[n]} \right) = I_n; \quad i = 1, \cdots, n; \tag{14.18}$$

$$\mathrm{Blk}_i \left(M_{\sqcap}(I_n \otimes M_{\sqcup}) M_{r,n} W_{[n]} \right) = I_n. \quad i = 1, \cdots, n. \tag{14.19}$$

Proof. (14.12)–(14.19) are one-to-one corresponding to (14.1)–(14.8). We prove one of them, say, (14.19). Note that in vector form equation (14.19) can be expressed as

$$\begin{aligned} x &= M_{\sqcap} x M_{\sqcup} x y = M_{\sqcap}(I_n \otimes M_{\sqcup}) x^2 y \\ &= M_{\sqcap}(I_n \otimes M_{\sqcup}) M_{r,n} x y = M_{\sqcap}(I_n \otimes M_{\sqcup}) M_{r,n} W_{[n]} y x. \end{aligned}$$

Then we have

$$M_{\sqcap}(I_n \otimes M_{\sqcup}) M_{r,n} W_{[n]} y = I_n, \quad \forall\, y \in \Delta_n.$$

Let $y = \delta_n^i$. Then we have

$$\mathrm{Blk}_i \left(M_{\sqcap}(I_n \otimes M_{\sqcup}) M_{r,n} W_{[n]} \right) = I_n.$$

□

Next, we consider when a Hasse diagram represents a lattice. A Hasse diagram, denoted by $\mathcal{H} = (\mathcal{N}, \mathcal{E})$, can be described by a matrix, denoted by $M_{\mathcal{H}}$ and called its incidence matrix, or Hasse matrix.

Let $|\mathcal{N}| = n$, then $M_{\mathcal{H}} \in \mathcal{B}_{n \times n}$, which is defined by assigning its entries $m_{i,j}$ as follows:

$$m_{i,j} = \begin{cases} 1, & (i,j) \in \mathcal{E} \\ 0, & \text{otherwise.} \end{cases}$$

We consider the incidence matrices of the diagrams in Fig. 14.1.

Example 14.4. Consider the figures A, B, C, and D in Fig. 14.1. We construct the incidence matrix of A first. We identify $a \sim 1$, $b \sim 2$, $c \sim 3$, and $d \sim 4$. Since $b > a$, in the graph we have a side from b to a. Then in the adjacent matrix we set $m_{b,a} = m_{2,1} = 1$. Similarly, since $c > a$ we have

$m_{c,a} = m_{3,1} = 1$, and so on. Finally, we can get the adjacent matrix of A as

$$\mathcal{J}_A = \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{bmatrix} 0&0&0&0 \\ 1&0&0&0 \\ 1&0&0&0 \\ 0&1&1&0 \end{bmatrix} \end{array} \begin{array}{c} \\ a \\ b \\ c \\ d \end{array}.$$

Similarly, the incidence matrices of B, C, and D can be constructed as

$$\mathcal{J}_B = \begin{bmatrix} 0&0&0&0&0 \\ 1&0&0&0&0 \\ 1&0&0&0&0 \\ 0&1&1&0&0 \\ 0&1&1&0&0 \end{bmatrix},$$

$$\mathcal{J}_C = \begin{bmatrix} 0&0&0&1&0 \\ 0&0&0&1&0 \\ 0&0&0&1&0 \\ 0&0&0&0&0 \\ 1&1&1&0&0 \end{bmatrix},$$

$$\mathcal{J}_D = \begin{bmatrix} 0&0&0&1&0 \\ 1&0&0&0&0 \\ 0&0&0&1&0 \\ 0&0&0&0&0 \\ 0&1&1&0&0 \end{bmatrix}.$$

Next, we consider when a Boolean matrix is a Hasse matrix. We have the following result.

Proposition 14.1. *A Boolean matrix $H \in \mathcal{B}_{n\times n}$ is a Hasse matrix, if and only if*

$$h_{i,j}h_{j,i} = 0, \quad i,j = 1,\cdots,n. \tag{14.20}$$

***Proof*.** Denote the corresponding nodes as $N_1, \cdots, N_n$.

(Necessity) If H is a Hasse matrix, then it is clear that (i) $h_{i,i} = 0$, $i = 1,\cdots,n$; (ii) if $h_{i,j} = 1$, then $N_j < N_i$, and hence $h_{j,i} = 0$. Hence, (14.20) is true.

(Sufficiency) If $h_{i,j} = 1$ draw a directed edge from i to j. Since there is no a pair of points, which have more than one edges, the graph is a Hasse one. □

Finally, we consider when a Hasse matrix is a lattice. Precisely, the Hasse graph corresponding to this Hasse matrix is a lattice.

Let $\mathcal{J}$ be a Hasse matrix. Then we define a matrix

$$U_{\mathcal{J}} := \sum_{k=0}^{n-1} \mathcal{J}^{(k)}. \tag{14.21}$$

Note that the power here is defined in (8.24).

Lemma 14.1. *$N_i \geq N_j$, if and only if $u_{i,j} = 1$.*

Proof. Denote by $U_s = \mathcal{J}^{(s)}$. then it is easy to see that $u^s_{i,j} = 1$ means on the graph there is a path, starting from N_i, reaching N_j at sth step. That is, there is a path from N_i to N_j with length s. It follows that $N_i \geq N_j$. The lemma follows immediately. □

Definition 14.11.

(1) Let $X = (x_1, \cdots, x_n) \in \mathcal{B}_n$ (X could be a row or a column). The support of X, denoted by supp(X) is an index set $\{i_1, \cdots, i_k\} \subset \{1, 2, \cdots, n\}$, such that $i_j \in \text{supp}(X)$, if and only if $x_{i_j} = 1$.

(2) Let $W \in \mathcal{M}_{n\times n}$ and $I = \{i_1, \cdots, i_k\} \subset \{1, 2, \cdots, n\}$ be a subindex. then the sub-matrix W_I of W is defined as

$$W_I = \begin{bmatrix} w_{i_1,i_1} & w_{i_1,i_2} & \cdots & w_{i_1,i_k} \\ w_{i_2,i_1} & w_{i_2,i_2} & \cdots & w_{i_2,i_k} \\ \vdots & & & \\ w_{i_k,i_1} & w_{i_k,i_2} & \cdots & w_{i_k,i_k} \end{bmatrix}.$$

Now we are ready to present the condition for a Hasse matrix to be a lattice.

Let $\mathcal{J} \in \mathcal{B}_{n\times n}$ be a Hasse matrix and $U_{\mathcal{J}}$ be defined by (14.21). For any $1 \leq i < j \leq n$ we define two index sets:

$$\begin{aligned} C^{i,j} &:= \text{supp}\left(\text{Col}_i(U_{\mathcal{J}}) \wedge \text{Col}_j(U_{\mathcal{J}})\right); \\ R^{i,j} &:= \text{supp}\left(\text{Row}_i(U_{\mathcal{J}}) \wedge \text{Row}_j(U_{\mathcal{J}})\right). \end{aligned}$$

Using them, we construct two sub-matrices correspondingly as

$$M_{C^{i,j}}, \quad M_{R^{i,j}}.$$

Then we have the following:

Theorem 14.4. *$\mathcal{J}$ is a lattice, if and only if for each pair (i, j) $(i \neq j)$, we have*

(i) $|C^{i,j}| := \alpha \geq 1$, $|R^{i,j}| := \beta \geq 1$;
(ii) $M_{C^{i,j}}$ *has a row, which equals* $\mathbf{1}_{\alpha}^{T}$;
(iii) $M_{R^{i,j}}$ *has a column, which equals* $\mathbf{1}_{\beta}$.

***Proof*.** Consider its corresponding graph, where i corresponds to node N_i, $i = 1, \cdots, n$. We have to show that any two nodes N_i and N_j have the $\sup\{N_i, N_j\}$ and $\inf\{N_i, N_j\}$. We first consider the existence of $\sup\{N_i, N_j\}$.

From the construction it is clear that $s \in C^{i,j}$ implies that N_s is a common upper bound of N_i and N_j. If $|C^{i,j}| = \alpha^{i,j} = 0$, it is obvious that N_i and N_j have no common upper bound. Now assume $\alpha^{i,j} > 0$ and $C^{i,j} = \{u_1, \cdots, u_{\alpha^{i,j}}\}$. Then we can construct the matrix $M_{C^{i,j}}$, which corresponds to the set of common upper bounds of N_i and N_j. Now if a column equals $\mathbf{1}_{\alpha^{i,j}}$ then it corresponds to the least common upper bound, we denote its index by $u_{i,j}$. Note that if such column exists, it is unique. If there is no such a column, which equals $\mathbf{1}_{\alpha^{i,j}}$, then it is clear that the least common upper bound does not exist.

A similar argument shows that $\inf\{N_i, N_j\}$ exists, if and only if there exists a unique row of index $\ell_{i,j}$ in $M_{R^{i,j}}$, which equals $\mathbf{1}^T_{\beta^{i,j}}$. □

From the constructive proof of Theorem 14.4 one sees easily that if matrix H is a lattice, then its structure matrices corresponding to $\sqcup$ and $\sqcap$ are as follows.

$$\begin{aligned} M_{\sqcup} &= \delta_n[u_{11}\ \cdots\ u_{1n}\ \cdots\ u_{nn}]; \\ M_{\sqcap} &= \delta_n[\ell_{11}\ \cdots\ \ell_{1n}\ \cdots\ \ell_{nn}]. \end{aligned} \tag{14.22}$$

We give some examples to illustrate it.

Example 14.5.

(1) Consider graph A in Fig. 14.3. The incidence matrix is

$$\mathcal{J}_A = \begin{bmatrix} 0&0&0&0&0 \\ 0&0&0&0&0 \\ 1&1&0&0&0 \\ 0&1&0&0&0 \\ 1&0&0&1&0 \end{bmatrix}.$$

A direct computation shows that

$$U_{\mathcal{J}_A} = \begin{bmatrix} 1&0&0&0&0 \\ 0&1&0&0&0 \\ 1&1&1&0&0 \\ 0&1&0&1&0 \\ 1&1&0&1&1 \end{bmatrix}.$$

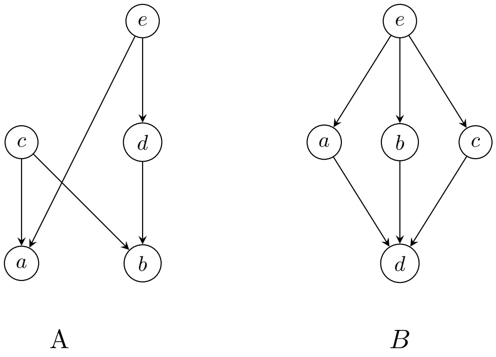


Fig. 14.3 Sample graphs

Then c and e are upper bound of $\{a,b\}$. Since

$$U_{3,5} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

there is no smallest elements of $\{c,e\}$, which means $\{a,b\}$ has no supremum. A is not a lattice.

(2) Consider the graph B in Fig. 14.3. We have

$$\mathcal{J}_B = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

It follows that

$$U_{\mathcal{J}_B} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Thus e is the largest element and d is the smallest. It is easy to see that

$$a \sqcup b = e; \quad a \sqcup c = e; \quad a \sqcup d = a; \quad a \sqcup e = e; \quad b \sqcup c = e$$
$$b \sqcup d = b; \quad b \sqcup e = e; \quad c \sqcup d = c; \quad c \sqcup e = e; \quad d \sqcup e = e.$$

Thus

$$M_{\sqcup} = \delta_5[1\ 5\ 5\ 1\ 5\ 5\ 2\ 5\ 2\ 5\ 5\ 5\ 3\ 3\ 5\ 1\ 2\ 3\ 4\ 5\ 5\ 5\ 5\ 5\ 5].$$

Similarly, we can get

$$M_{\sqcap} = \delta_5[1\ 4\ 4\ 4\ 1\ 4\ 2\ 4\ 4\ 2\ 4\ 4\ 3\ 4\ 3\ 4\ 4\ 4\ 4\ 4\ 1\ 2\ 3\ 4\ 5].$$

14.4 Distributive and Modular Lattices

Definition 14.12. A distributive lattice is a lattice which satisfies either of the distributive laws:

(i)

$$x \sqcap (y \sqcup z) = (x \sqcap y) \sqcup (x \sqcap z); \tag{14.23}$$

(ii)

$$x \sqcup (y \sqcap z) = (x \sqcup y) \sqcap (x \sqcup z). \tag{14.24}$$

This definition is well posed, because we have the following equivalence.

Theorem 14.5. *A lattice satisfies (14.23), if and only if it satisfies (14.24).*

Proof. We prove (14.24) ⇒ (14.23), and leave the proof of (14.23) ⇒ (14.24) to the reader.

Assume (14.24) holds. Then

$$\begin{aligned} x \sqcap (y \sqcup z) &= (x \sqcap (x \sqcup z)) \sqcap (y \sqcup z) && \text{(by (14.8))} \\ &= x \sqcap ((x \sqcup z) \sqcap (y \sqcup z)) && \text{(by (14.4))} \\ &= x \sqcap (z \sqcup (x \sqcap y)) && \text{(by (14.24))} \\ &= (x \sqcup (x \sqcap y)) \sqcap (z \sqcup (x \sqcap y)) && \text{(by (14.7))} \\ &= (x \sqcap y) \sqcup (x \sqcap z). && \text{(by (14.24)).} \end{aligned}$$

□

Remark 14.1.

(1) Every lattice satisfies the following two inequalities:

$$(x \sqcap y) \sqcup (x \sqcap z) \le x \sqcap (y \sqcup z); \tag{14.25}$$

$$x \sqcup (y \sqcap z) \le (x \sqcup y) \sqcap (x \sqcup z). \tag{14.26}$$

We leave the proof to the reader.

(2) For finite lattices, (14.23) and (14.24) have the following equivalent forms (14.27) and (14.28) respectively.

$$M_{\sqcap}(I_n \otimes M_{\sqcup}) = M_{\sqcup} M_{\sqcap}(I_{n^2} \otimes M_{\sqcap})(I_n \otimes W_{[n]})M_{r,n}; \tag{14.27}$$

$$M_{\sqcup}(I_n \otimes M_{\sqcap}) = M_{\sqcap} M_{\sqcup}(I_{n^2} \otimes M_{\sqcap})(I_n \otimes W_{[n]})M_{r,n}. \tag{14.28}$$

Definition 14.13. A modular lattice is any lattice which satisfies the following modular law:

$$x \le y \ \text{ implies } \ x \sqcup (y \sqcap z) = y \sqcap (x \sqcup z). \tag{14.29}$$

Remark 14.2. It is easy to see that every lattice satisfies

$$x \le y \ \text{ implies } \ x \sqcup (y \sqcap z) \le y \sqcap (x \sqcup z).$$

Proposition 14.2. *Every distributive lattice is a modular lattice.*

Proof. Using (14.24) and noting that $x \sqcup y = y$ whenever $x \le y$, the conclusion follows. □

Example 14.6. Recall Fig 14.1, we can check that

(1) In C, (which will be called M_5 in the sequel), $a \sqcup (b \sqcap c) = a \sqcup d = a$, but $(a \sqcup b) \sqcap (a \sqcup c) = e \sqcap e = e$. Hence M_5 is not distributive.
(2) It is easy to verify that M_5 does satisfy the modular law, and hence is a modular.
(3) In D, (which will be called N_5 in the sequel), $a \le b$, $a \sqcup (b \sqcap c) = a \sqcup d = a$, but $b \sqcap (a \sqcup c) = b \sqcap e = b$. Hence N_5 is not modular and therefore, is not distributive.

The following two theorems are important in verifying modular and/or distributive lattice.

Theorem 14.6 (Dedekind (Burris and Sankappanavar, 1981)). *L is a non-modular lattice, if and only if N_5 can be embedded into L.*

Theorem 14.7 (Birkhoff (Burris and Sankappanavar, 1981)). *L is a non-distributive lattice, if and only if M_5 or N_5 can be embedded into L.*

14.5 Graph and Its Adjacency Matrix

Definition 14.14. Let $V = \{v_1, \cdots, v_n\}$ be a finite set, $E \subset V \times V$.

(1) $G(V, E)$ is called a graph, with $V = V(G)$ as its set of vertices and $E = E(G)$ its edges.
(2) $G(V, E)$ is called a simple graph, if (i) there is no duplicated elements (edges) in E; and (ii) $(a, a) \notin E$, $a \in V$.
(3) $G(V, E)$ is called an undirected graph, if $(a, b) \in E$ implies $(b, a) \in E$. Otherwise, it is directed. Directed graph is also called digraph.

Definition 14.15. Let $G(V, E)$ be a graph with $V = \{v_1, \cdots, v_n\}$. A Boolean matrix $M_G \in \mathcal{B}_{n \times n}$ is called the adjacency matrix of G, if its entries are defined as follows:

$$m_{i,j} = \begin{cases} 1, & (i,j) \in E \\ 0, & \text{otherwise.} \end{cases} \tag{14.30}$$

Note that where the adjacency matrix is defined as in (14.30), it means the duplicated edges are not allowed. The following result is an immediate consequence of the definition.

Proposition 14.3.

(1) A graph G is simple, if and only if its adjacency matrix M_G has zero diagonal entries. That is,

$$m_{i,i} = 0, \quad i = 1, \cdots, n. \tag{14.31}$$

(2) A graph G is undirected, if and only if M_G is symmetric.

For statement ease, hereafter we use graph for undirected graph only.

Definition 14.16.

(1) Let G be a graph and $v \in V(G)$. The degree of v is the number of edges incident with v, denoted by $\deg(v)$;

(2) Let G be a digraph and $v \in V(G)$. The in-degree of v is the number of edges toward to v, denoted by $\mathrm{indeg}(v)$, The out-degree of v is the number of edges from v, denoted by $\mathrm{outdeg}(v)$.

We give some examples of useful graphs.

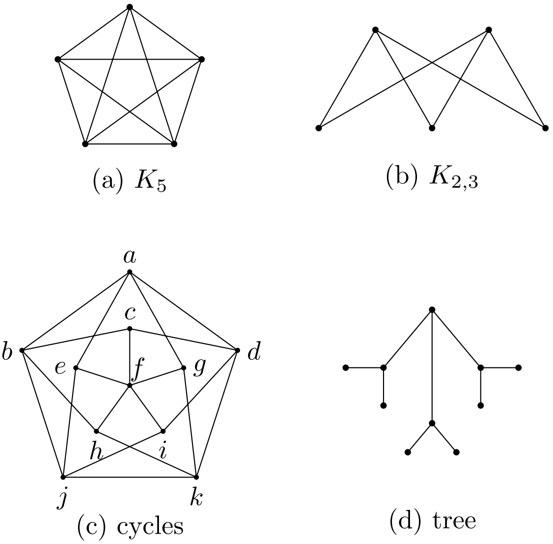
Example 14.7.

(1) Complete graph: A simple graph G in which each pair of distinct vertices are adjacent is called a complete graph, denoted by K_n, where $n = |V(G)|$. (We refer to Fig. 14.4 (a) for K_5.)

(2) Bipartite graph: If a graph G has two disjoint sets of vertices $V(G) = A \cup B$, and each edge joints an edge $a \in A$ and an edge $b \in B$, then G is called a bipartite graph. A bipartite graph is called a complete bipartite graph, if each $a \in A$ is jointed to all $b \in B$ and vice versa. A complete bipartite graph is denoted by $K_{m,n}$, where $m = |A|$ and $n = |B|$. (We refer to Fig. 14.4 (b) for $K_{2,3}$.)

(3) Cycle: Let $\{a_1, a_2, \cdots, a_k\} \subset V(G)$ and $a_i a_{i+1} \in E(G)$, $i = 1, \cdots, k-1$, and $a_k a_1 \in E(G)$. Then $a_1 - a_2 - \cdots - a_k - a_1$ is called a cycle, if there is no repeated edges.

In Fig. 14.4 (c) $a - b - j - k - d - a$ and $c - b - j - i - d - c$ etc. are cycles.



(a) K_5 (b) $K_{2,3}$ (c) cycles (d) tree

Fig. 14.4

(4) Tree: A connected graph with no cycles is called a tree. (We refer to Fig. 14.4 (d) for a tree.)

Definition 14.17.

(1) A graph (digraph) G is Eulerian if there is a cycle, called Eulerian cycle, which consists of all the edges of G.
(2) A graph (digraph) G is Hamiltonian if there is a cycle, called Hamiltonian cycle, which crosses all the vertices of G exactly once.

Eulerian graph and Hamiltonian graph are of particular importance in graph theory. It is easy to verify whether a graph is Eulerian or not. The following result is due to Euler.

Theorem 14.8 (Wilson, 1996). *(1) A connected graph G is Eulerian, if and only if for any $v \in V(G)$ $\deg(v)$ is even.*
(2) A connected digraph G is Eulerian, if and only if, for any $v \in V(G)$, $\operatorname{indeg}(v) = \operatorname{outdeg}(v)$.

Next, we consider when a graph is Hamiltonian. Finding a characterization of Hamiltonian graphs is one of the major unsolved problems of graph theory (Wilson, 1996). In the following we give a numerical method for the verification.

Denote by P_n the set of permutations of $\{1, 2, \cdots, n\}$. For instance,

$$P_3 = \{(1,2,3),\ (1,3,2),\ (2,1,3),\ (2,3,1),\ (3,1,2),\ (3,2,1)\}.$$

We can also use any other set of n objects to replace $\{1, 2, \cdots, n\}$. Say, $\{2, 3, \cdots, n, n+1\}$. Then we denote it as $P_n\{2, 3, \cdots, n, n+1\}$. For instance,

$$P_3\{2,3,4\} = \{(2,3,4),\ (2,4,3),\ (3,2,4),\ (3,4,2),\ (4,2,4),\ (4,3,2)\}.$$

Recall that $\mathbf{S}_n$ is the symmetric group, which consists of all the bijective mappings from $\{1, 2, \cdots, n\}$ to $\{1, 2, \cdots, n\}$. (We refer to Chapter 1 for more details.)

We define a subset $\Theta_n \subset \mathbf{S}_n$ as

$$\Theta_n = \{(1,p) \,|\, p \in P_{n-1}\{2,3,\cdots,n\}\}.$$

For instance,

$$\Theta_4 = \{(1,2,3,4),\ (1,2,4,3),\ (1,3,2,4),\ (1,3,4,2),\ (1,4,2,4),\ (1,4,3,2)\}.$$

Definition 14.18. Let $A = (a_{i,j}) \in \mathcal{B}_{n\times n}$ be a Boolean matrix. The cycle-determinant of A is defined as

$$\operatorname{cdet}(A) = \sum_{\sigma\in\Theta_n} a_{1,\sigma(1)}a_{\sigma(1),\sigma(2)}\cdots a_{\sigma(n-1),\sigma(n)}a_{\sigma(n),1}. \tag{14.32}$$

Theorem 14.9.

(1) A graph G is Hamiltonian, if and only if its adjacency matrix satisfies

$$\operatorname{cdet}(M_G) > 0. \tag{14.33}$$

Moreover, its number of Hamiltonian cycles is $\operatorname{cdet}(M_G)/2$.

(2) A digraph G is Hamiltonian, if and only if the inequality (14.33) holds. Moreover, its number of Hamiltonian cycles is $\operatorname{cdet}(M_G)$.

Proof. It is clear that each element of Θ_n corresponds to a possible Hamiltonian cycle, and only if the corresponding term in $\operatorname{cdet}(M_G)$ equals 1 that cycle becomes a real pass in the graph. So, calculating the cycle-determinant is equal to searching by exhaustion. The conclusion is obvious. It is worth noting that the coefficient 1/2 comes from the consideration that a clockwise cycle and its anti-clockwise counterpart are considered as one cycle in (undirected) graph. □

We give an example for this.

Example 14.8. Consider the Grötzsch graph in Fig. 14.4 (c). Is it Hamiltonian? Label the vertices by numbers as: $a = 1$, $b = 2$, $c = 3$, $d = 4$, $e = 5$, $f = 6$, $g = 7$, $h = 8$, $i = 9$, $j = 10$, and $k = 11$. The adjacency matrix of Grötzsch graph is

$$M_G = \begin{bmatrix} 0&1&0&1&1&0&1&0&0&0&0\\ 1&0&1&0&0&0&0&1&0&1&0\\ 0&1&0&1&0&1&0&0&0&0&0\\ 1&0&1&0&0&0&0&0&1&0&1\\ 1&0&0&0&0&1&0&0&0&1&0\\ 0&0&1&0&1&0&1&1&1&0&0\\ 1&0&0&0&0&1&0&0&0&0&11\\ 0&1&0&0&0&1&0&0&0&0&11\\ 0&0&0&1&0&1&0&0&0&1&0\\ 0&1&0&0&1&0&0&0&1&0&1\\ 0&0&0&1&0&0&1&1&0&1&0 \end{bmatrix}.$$

A brief routine shows that

$$\mathrm{cdet}(M_G) = 20.$$

So there are 10 distinct Hamiltonian cycles, which are

$$\begin{aligned}
C_1 &= 1 \to 2 \to 3 \to 4 \to 9 \to 10 \to 5 \to 6 \to 8 \to 11 \to 7 \to 1;\\
C_2 &= 1 \to 2 \to 8 \to 11 \to 7 \to 6 \to 3 \to 4 \to 9 \to 10 \to 5 \to 1;\\
C_3 &= 1 \to 4 \to 3 \to 2 \to 8 \to 11 \to 7 \to 6 \to 9 \to 10 \to 5 \to 1;\\
C_4 &= 1 \to 4 \to 9 \to 10 \to 5 \to 6 \to 3 \to 2 \to 8 \to 11 \to 7 \to 1;\\
C_5 &= 1 \to 5 \to 6 \to 3 \to 4 \to 9 \to 10 \to 2 \to 8 \to 11 \to 7 \to 1;\\
C_6 &= 1 \to 5 \to 6 \to 8 \to 2 \to 3 \to 4 \to 9 \to 10 \to 11 \to 7 \to 1;\\
C_7 &= 1 \to 5 \to 10 \to 2 \to 3 \to 4 \to 9 \to 6 \to 8 \to 11 \to 7 \to 1;\\
C_8 &= 1 \to 5 \to 10 \to 9 \to 4 \to 11 \to 8 \to 2 \to 3 \to 6 \to 7 \to 1;\\
C_9 &= 1 \to 5 \to 10 \to 9 \to 6 \to 8 \to 2 \to 3 \to 4 \to 11 \to 7 \to 1;\\
C_{10} &= 1 \to 5 \to 10 \to 11 \to 8 \to 2 \to 3 \to 4 \to 9 \to 6 \to 7 \to 1.
\end{aligned}$$

14.6 Vector Space Structure of Graph

Let G be a connected graph and $|E(G)| = m$. We define a vector space structure on 2^E over $\mathbb{Z}_2$ in following way.

(i) If $A, B \in 2^E$ then the plus is defined as

$$A \langle + \rangle B = (A \cup B) \backslash (A \cap B).$$

(ii)

$$1 \langle\times\rangle A := A; \quad 0 \langle\times\rangle A := \varnothing.$$

We denote the vector space of G by V_G. It is clear that $\dim(V_G) = m$, because $E = \{e_1, \cdots, e_m\}$ is a basis.

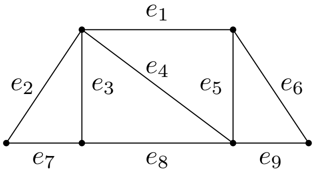


Fig. 14.5 Graph for V_G

Example 14.9. Consider the graph in Fig. 14.5. Let $A = \{e_1, e_5, e_8, e_3\}$, $B = \{e_1, e_5, e_4\}$, $C = \{e_5, e_6, e_9\}$, $D = \{e_5, e_8, e_9\}$. Then

$$\begin{aligned} A \langle+\rangle B &= \{e_8, e_3, e_4\}; \\ B \langle+\rangle C &= \{e_1, e_4, e_6, e_9\}; \\ C \langle+\rangle D &= \{e_8, e_5\}. \end{aligned}$$

Let G be a connected simple graph, $|V(G)| = n$, and $|E(G)| = m$. A spanning tree of G is a tree $T(G)$, which joints all the vertices.

Let G be a connected simple graph. We can choose a cycle and remove any one edge of the cycle and the remaining graph remains connected. Continuing this procedure a spanning tree results. The number of edges removed in this process is called the cycle rank of G, denoted by $\gamma(G)$. It is easy to check that

$$\gamma(G) = m - n + 1.$$

If G has k components, then we can do this on each component. Finally, the remaining graph is called a spanning forest, and we have the cycle rank as

$$\gamma(G) = m - n + k. \tag{14.34}$$

Conversely, if a spanning tree $T(G)$ is given, and $e \in G \backslash T$, then we can add e to T to form a cycle. The cycles obtained in this way are called the fundamental set of cycles associated with T.

It is easy to verify the following factors (Wilson, 1996):

(1) Let C_1 and C_2 be two cycles. Then $C_1 \langle + \rangle C_2$ is a sum of disjointed cycles;

(2) Let G be connected, $T = T(G)$ be a spanning tree of G, and

$$G \backslash T := \{\ell_1,\ \ell_2,\ \cdots,\ \ell_{m-n-1}\}.$$

Then each ℓ_i corresponds a fundamental cycle, denoted by $C(\ell_i)$.

(3) If C is a cycle, then $C \cap (G \backslash T) \neq \varnothing$. Moreover, Assume $C \cap (G \backslash T) = \{\ell_{i_1}, \cdots, \ell_{i_k}\}$, then

$$C = C(\ell_{i_1}) \langle + \rangle C(\ell_{i_1}) \langle + \rangle \cdots \langle + \rangle C(\ell_{i_k}). \tag{14.35}$$

Summarizing the above arguments, we have the following result.

Theorem 14.10. *Let G be a connected graph with T as one of its spanning trees. Denote by*

$$G \backslash T := \{\ell_1,\ \ell_2,\ \cdots,\ \ell_{m-n+1}\}. \tag{14.36}$$

Then the cycle subspace of G, denoted by $C_G \subset V_G$, is a vector subspace of V_G with

$$B := \{C(\ell_1),\ C(\ell_2),\ \cdots,\ C(\ell_{m-n+1})\}$$

as its basis.

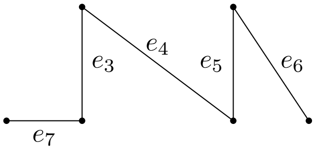


Fig. 14.6 Spanning tree T

Observe Fig. 14.5 again.

Example 14.10. Consider the graph in Fig. 14.5. A spanning tree is: $T = \{e_7,\ e_3,\ e_4,\ e_5,\ e_6\}$ (refer to Fig. 14.6). Then

$$G \backslash T = \{e_1,\ e_2,\ e_8,\ e_9\}.$$

A basis of C_G with respect to T is

$$B = \{C(e_1),\ C(e_2),\ C(e_8),\ C(e_9)\},$$

where

$$C(e_1) = e_1 - e_4 - e_5; \quad C(e_2) = e_2 - e_3 - e_7;$$
$$C(e_8) = e_3 - e_4 - e_8; \quad C(e_9) = e_5 - e_6 - e_9.$$

Now assume a cycle is given as $C = e_3 - e_1 - e_6 - e_9 - e_8$. Hence, $C \cap (G\backslash T) = \{e_1,\ e_8,\ e_9\}$. Then it is ready to check that

$$C = C(e_1)\,\langle + \rangle\,C(e_8)\,\langle + \rangle\,C(e_9).$$

Next, we consider another important subspace.

Definition 14.19. Let G be a connected simple graph.

(1) A subset of edges $D \subset E(G)$ is called a disconnecting set, if removing them disconnects the graph.
(2) A disconnected set is called a cutset, if it has no proper subset which is disconnecting set.

Let G be a connected simple graph with T be its spanning tree. Removing each edge $\ell \in T$ splits the vertices into two parts. all the edges of G which connect the two parts of vertices form a cutset, which is denoted by $C^*(\ell)$. Similar to cycle subspace, we have the following cutset subspace.

Theorem 14.11. *Let G be a connected graph with T as one of its spanning trees. Denote by*

$$T := \{\ell_1,\ \ell_2,\ \cdots,\ \ell_{n-1}\}. \tag{14.37}$$

Then the cutset subspace of G, denoted by $C^*_G \subset V_G$, is a vector subspace of V_G with

$$B := \{C^*(\ell_1),\ C^*(\ell_2),\ \cdots,\ C^*(\ell_{n-1})\}$$

as its basis.

Observe Fig. 14.5 again.

Example 14.11. Consider the graph in Fig. 14.5. A spanning tree is: $T = \{e_7,\ e_3,\ e_4,\ e_5,\ e_6\}$. Then a basis of C^*_G with respect to T is

$$B = \{C^*(e_3),\ C^*(e_4),\ C^*(e_5),\ C^*(e_6),\ C^*(e_7)\},$$

where

$$C^*(e_3) = \{e_2,\ e_3,\ e_8\};\ C^*(e_4) = \{e_1,\ e_4,\ e_8\};\ C^*(e_5) = \{e_1,\ e_5,\ e_9\};$$
$$C^*(e_6) = \{e_6,\ e_9\};\qquad C^*(e_7) = \{e_2,\ e_7\}.$$

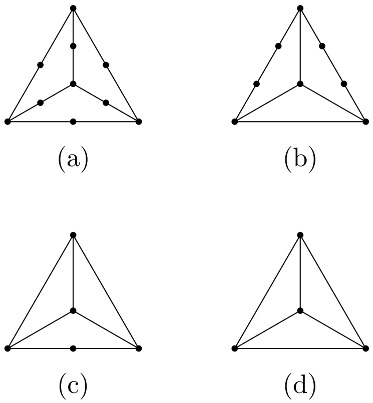
Now assume a cutset is given as $C^* = \{e_2,\ e_3,\ e_4,\ e_5,\ e_6\}$. Hence, $C^* \cap T = \{e_3,\ e_4,\ e_5,\ e_6\}$. Then it is ready to check that

$$\begin{aligned}
&C^*(e_3)\,\langle + \rangle\,C^*(e_4)\,\langle + \rangle\,C^*(e_5)\,\langle + \rangle\,C^*(e_6)\\
&= \{e_2,\ e_3,\ e_8\}\,\langle + \rangle\,\{e_1,\ e_4,\ e_8\}\,\langle + \rangle\,\{e_1,\ e_5,\ e_9\}\,\langle + \rangle\,\{e_6,\ e_9\}\\
&= \{e_2,\ e_3,\ e_4,\ e_5,\ e_6\}\\
&= C^*.
\end{aligned}$$

14.7 Planar Graph and Coloring Problem

Definition 14.20. Two graphs are homeomorphic if they can be obtained from the same graph by adding new vertices of degree 2 into its edges.

Example 14.12. In Fig. 14.7 the four graphs are homeomorphic because (a), (b), and (c) can be obtained from (d) by adding some vertices to its edges.


(a) (b)

(c) (d)

Fig. 14.7 Homeomorphic graphs

Definition 14.21. A planar graph is a graph that can be drawn in the plane without crossings, which is called a plane drawing.

Example 14.13. Consider K_4. Fig. 14.8 gives its two different drawings, where (a) is not a plane drawing and (b) is. So K_4 is a planar graph.

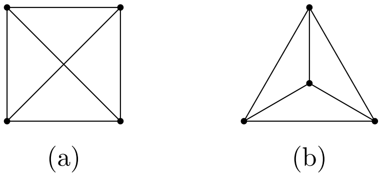

(a) (b)

Fig. 14.8 Different drawings of K_4

Planar graphs are particularly important. In fact, only planar graphs can have a map realization (in dual sense). The following result is essential.

Theorem 14.12 (Wilson, 1996). *A graph is planar, if and only if it contains no subgraph homeomorphic to K_5 or $K_{3,3}$.*

Let G be a planar graph with a plane drawing. Then the plane is divided by the drawing into several disjoined parts, which are called the faces of the graph. The following is the famous Euler's formula.

Theorem 14.13 (Wilson, 1996). *For a plane drawing of a connected planar graph, we have*

$$n - m + f = 2, \tag{14.38}$$

where n is the number of vertices, m is the number of edges, and f is the number of faces.

Next, we consider the dual graph. Let G be a plane drawing of plane graph. We construct its dual graph G^* in two steps:

- Step 1: choose one point inside each face to form the vertices of G^*;
- Step 2: crossing each edge draw a curve connecting the points on both sides of this edge to form the edges of G^*.

Fig. 14.9 shows two graphs with their dual graphs. Note that in the original graph we use $*$ for vertices of G^* within each faces of G, and dotted lines for the edges of G^*, crossing each edges of G.

For dual graphs we have the following result.

Theorem 14.14 (Wilson, 1996). *Let G be a planar connected graph, and G^* its dual graph. Then*

(1)

$$n^* = f, \quad m^* = m, \quad f^* = n;$$

(2)

$$G^* * = G;$$

(3) A set of edges forms a cycle of G, if and only if its corresponding set of edges forms a cutset of G^*.

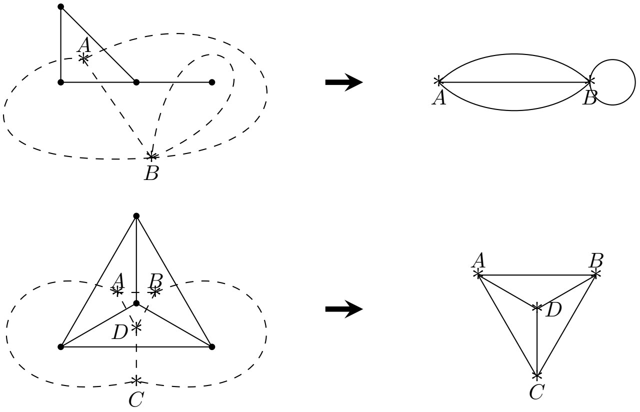


Fig. 14.9 Dual graphs

As an application, we consider famous Four Color Problem (Chartrand and Zhang, 2005): The minimum number of colors for coloring any map in such a way that different colors are used for the adjacent areas, which have a common boundary.

It has been proved via computer that (Chartrand and Zhang, 2005)

Theorem 14.15 (Four Color Theorem). *Each map can be colored by no more that* 4 *colors.*

First, we convert a map into its dual graph, in the following way. Assume in a map we have n different areas: $A_1, \cdots, A_n$. Choose from each area A_i a point v_i, which represents the corresponding area A_i. Let $V = \{v_1, \cdots, v_n\}$ be the vertex set of graph G. Then the edge set E is defined as follows: $(v_i, v_j) \in E$, if and only if A_i and A_j are adjacent. Note that we ignore the infinity face (i.e., the unbounded face).

For a graph the coloring problem colors vertices in such a way, that the adjacent vertices should be colored in different ones.

The coloring problem for a map is equivalent to the coloring problem of the dual graph of the map.

Example 14.14. Fig. 14.10 is a map of Western United States (where 1: Oregon, 2: Idaho, 3: Wyoming, 4: California, 5: Nevada, 6: Utah, 7: Colorado, 8: Arizona, 9: New Mexico). Fig. 14.11 is its dual graph.

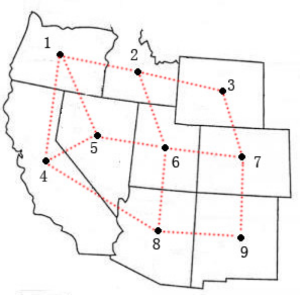


Fig. 14.10 Western United States

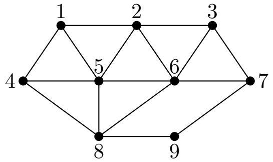


Fig. 14.11 Dual graph of Western United States

It is easy to verify that the adjacent matrix of this graph, denoted by M_G, is

$$M_G = \begin{bmatrix} 0\,1\,0\,1\,1\,0\,0\,0\,0 \\ 1\,0\,1\,0\,1\,1\,0\,0\,0 \\ 0\,1\,0\,0\,0\,1\,1\,0\,0 \\ 1\,0\,0\,0\,1\,0\,0\,1\,0 \\ 1\,1\,0\,1\,0\,1\,0\,1\,0 \\ 0\,1\,1\,0\,1\,0\,1\,1\,0 \\ 0\,0\,1\,0\,0\,1\,0\,0\,1 \\ 0\,0\,0\,1\,1\,1\,0\,0\,1 \\ 0\,0\,0\,0\,0\,0\,1\,1\,0 \end{bmatrix}. \tag{14.39}$$

Definition 14.22. Given a graph $G = (V, E)$, a subset $W \subset U$ is called an independent set, if for any $v_i, v_j \in W$, $(i, j) \not\in E$. W is called a maximum independent set, if it is an independent set and there is no other independent set W', which contains it properly, i.e., $W' \supsetneq W$.

Note that each G can be participated into independent subsets. A trivial partition is

$$V = \cup_{i=1}^{n} \{v_i\}.$$

Hence V can be partitioned into some maximum independent sets. Say,

$$V = \cup_{i=1}^{k} V^i, \tag{14.40}$$

where V^i, $i = 1, \cdots, k$ are maximum independent sets. Assume k is the smallest one such that a partition of (14.33) exists, then G is said to be k-chromatic, denoted by $\chi(G) = k$. The following result comes from definition immediately.

Corollary 14.1. *If a graph G has $\chi(G) = k$, then it is k-colorable, i.e., it can be colored by k colors.*

Observe that the adjacent matrix is vertex-order depending. If we arrange the vertices in the order of $(i_1, i_2, \cdots, i_n)$, which is a permutation of $(1, 2, \cdots, n)$. Then the corresponding structure matrix becomes

$$\tilde{M}_G = T^T M_G T, \tag{14.41}$$

where $T = \delta_n[i_1\ i_2\ \cdots\ i_n] \in \mathcal{L}_{n\times n}$.

Now assume

$$\left\{ \{i_1^s, \cdots, i_{n_s}^s\} \,\middle|\, s = 1, \cdots, k \right\}$$

is the set of maximum independent sets. Precisely,

$$\left\{ v_{i_1^s}, v_{i_2^s}, \cdots, v_{i_{n_s}^s} \right\}, \quad s = 1, 2, \cdots, k$$

are maximum independent sets. Setting

$$T = \delta_n \left[i_1^1\ \cdots\ i_{n_1}^1\ \ \cdots\ \ i_1^k\ \cdots\ i_{n_k}^k \right], \tag{14.42}$$

the corresponding adjacent matrix becomes

$$\tilde{M}_G = T^T M_G T = \begin{bmatrix} \mathbf{0}_{n_1} & \times & \cdots & \times \\ \times & \mathbf{0}_{n_2} & \cdots & \times \\ \vdots & & & \\ \times & \times & \cdots & \mathbf{0}_{n_k} \end{bmatrix}. \tag{14.43}$$

We call (14.43) the canonical form of adjacent matrix. Now the Four Color Theorem is equivalent to that the canonical form of adjacent matrix of any map has at most $k \leq 4$ zero diagonal blocks.

We give the following algorithm for searching the canonical form. To this end, we propose some notations: Let $M \in \mathcal{M}_n$ and $I = \{i_1, i_2, \cdots, i_s\} \subset \{1, 2, \cdots, n\}$. Then M_I is the principle minor of M, consisting of its elements in $i_1, i_2, \cdots, i_s$th rows and $i_1, i_2, \cdots, i_s$th columns. That is,

$$M_I = \begin{bmatrix} m_{i_1,i_1} & m_{i_1,i_2} & \cdots & m_{i_1,i_s} \\ m_{i_2,i_1} & m_{i_2,i_2} & \cdots & m_{i_2,i_s} \\ \vdots & & & \\ m_{i_s,i_1} & m_{i_s,i_2} & \cdots & m_{i_s,i_s} \end{bmatrix}.$$

Algorithm 2. This algorithm consists of two cycles: First cycle to construct a zero diagonal block and second cycle to construct next zero diagonal block.

Loop One

(1) Step 1. Consider the first row (column).
Move all the zero elements to the top-left. Assume $I = \{i_1, \cdots, i_s\}$ and $J = \{j_1, \cdots, j_t\}$ be a partition of $\{2, 3, \cdots, n\}$, such that

$$m_{1,i} = 0, \quad i \in I; \quad m_{1,j} = 1, \quad j \in J.$$

Move all ith row (column) to the top-left.

(2) Step 2. Consider the sub-matrix M_I, where M is the updated M_G and $I = \{2, 3, \cdots, s+1\}$. As in Step 1, move all the zero elements to the top-left. and update M_I too.

(3) Step i. Continuing this process until we get a diagonal zero block, which cannot be enlarged via this process. Then go to Loop Two.

Loop Two

Assume some diagonal zero blocks are already obtained as

$$M_{updated} = \begin{bmatrix} \mathbf{0}_{n_1} & \cdots & \times & \\ & \ddots & & \times \\ \times & \cdots & \mathbf{0}_{n_p} & \\ & \times & & R \end{bmatrix}.$$

If: $R = \varnothing$, then stop, which means the canonical form is obtained.
Else: Set $M = R$, and go back to Loop One.

Remark 14.3.

(1) In Algorithm 2 each row-column moving has to be done for overall updated M_G.

(2) The algorithm can get a diagonal zero-block form, which shows a partition of independent sets. But it may not be the canonical form.

(3) The number of colors required to color the map can be found from the canonical form. Precisely, it equals to the number of zero diagonal blocks. Moreover, how to color the map can be seen from the permutation matrix, T, which convert original M_G to the canonical form.

We use the following example to depict the algorithm.

Example 14.15. Consider the graph in Example 14.14. Observing the adjacent matrix (14.32), we use Algorithm 2 to convert it into canonical form.

Using

$$T_1 = \delta_9[1\ 3\ 6\ 7\ 8\ 9\ 2\ 4\ 5],$$

we can move the zeros on first row (column) at the position of $(3, 6, 7, 8, 9)$th to the top-left as

$$T_1^T M_G T_1 := M_G^1 = \begin{bmatrix} 0&0&0&0&0&0&1&1&1\\ 0&0&1&1&0&0&1&0&0\\ 0&1&0&1&1&0&1&0&1\\ 0&1&1&0&0&1&0&0&0\\ 0&0&1&0&0&1&0&1&1\\ 0&0&0&1&1&0&0&0&0\\ 1&1&1&0&0&0&0&0&1\\ 1&0&0&0&1&0&0&0&1\\ 1&0&1&0&1&0&1&1&0 \end{bmatrix}. \tag{14.44}$$

Next, consider the sub-matrix M_I, where $I = \{2, 3, 4, 5, 6\}$. To move the zeros in second row (first row of M_I) at $(5, 6)$th position to the top-left of M_I, we use

$$T_2 = \delta_9[1\ 2\ 5\ 6\ \ 3\ 4\ 7\ 8\ 9],$$

which produces

$$T_2^T M_G^1 T_2 := M_G^2 = \begin{bmatrix} 0&0&0&0&0&0&1&1&1\\ 0&0&0&0&1&1&1&0&0\\ 0&0&0&1&1&0&0&1&1\\ 0&0&1&0&0&1&0&0&0\\ 0&1&1&0&0&1&1&0&1\\ 0&1&0&1&1&0&0&0&0\\ 1&1&0&0&1&0&0&0&1\\ 1&0&1&0&0&0&0&0&1\\ 1&0&1&0&1&0&1&1&0 \end{bmatrix}. \tag{14.45}$$

It is obvious that M_G^2 is a canonical form.

Now the transfer matrix is

$$T = T_1 T_2 = \delta_9[1\ 3\ 8\ 9\ 6\ 7\ 2\ 4\ 5]. \tag{14.46}$$

Comparing T with the zero-diagonal blocks, one sees that $\{1, 3, 8\}$, $\{9, 6\}$, $\{7, 2, 4\}$, $\{5\}$ form a partition of independent sets. Hence a reasonable coloring is: Using 4 colors for these 4 maximum independent sets.

Before ending this section, we briefly introduce hypergraph (Duchet, 1995).

Definition 14.23. Let $V = \{v_1, \cdots, v_n\}$ be a finite set.

(1) Assume $E = \{e_1, \cdots, e_m\} \subset \mathcal{P}(V)$ satisfies (i) $e_i \neq \varnothing$, $\forall i$; (ii) $\cup_{i=1}^{m} e_i = V$, then (V, E) is called a hypergraph, V is the set of vertices and E is the set of edges.

(2) Assume (V, E) is a hypergraph. It is a kth homogeneous hypergraph, if $|e_i| = k$, $\forall i$.

Note that for a k-homogeneous hypergraph we have

$$E \subset \underbrace{V \times \cdots \times V}_{k}.$$

Hence, when $k = 2$, it becomes a conventional graph (without isolated vertices).

A hypergraph is simple, if

$$e_i \backslash e_j = e_i \cup e_j^c \neq \varnothing, \quad i \neq j.$$

This means e_i is not a subset of e_j. Usually, people are only interested in simple hypergraph.

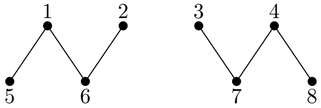


Fig. 14.12 A hypergraph

Example 14.16. Consider Fig. 14.12, where $V = \{1, 2, 3, 4, 5, 6, 7, 8\}$.

(1) If

$$E_1 = \{(5, 1), (1, 6), (6, 2), (3, 7), (7, 4), (4, 8)\},$$

(V, E_1) is a conventional graph.

(2) If

$$E_2 = \{(5, 1, 6), (1, 6, 2), (3, 7, 4), (7, 4, 8)\},$$

(V, E_2) is a 3rd homogeneous hypergraph.

(3) If

$$E_3 = \{(5, 1, 6, 2), (3, 7, 4, 8)\},$$

(V, E_3) is a 4th homogeneous hypergraph.

(4) If

$$E_4 = \{(5,1,6,2),(3,7,4),(4,8)\},$$

(V, E_4) is a simple but not homogeneous hypergraph.

(5) If

$$E_5 = \{(5,1,6,2),(3,7),(3,7,4,8)\},$$

(V, E_2) is not a simple hypergraph.

One sees that unlike the conventional graph, the figure is not enough to describe a hypergraph.

The adjacent matrix of a kth homogeneous hypergraph can be expressed in different ways. Say, let

$$\{i_1,\cdots,i_s\}\cup\{j_1,\cdots,j_t\}=\{1,\cdots,k\}$$

be a partition. Then we can define the adjacent matrix $M_G \in \mathcal{B}_{n^s\times n^t}$ as follows: Label M_G by

$$\mathrm{id}(i_1,\cdots,i_s;\underbrace{n,\cdots,n}_{s})\times \mathrm{id}(j_1,\cdots,j_t;\underbrace{n,\cdots,n}_{t}).$$

Then set the entries of M_G by

$$m_{(\alpha_1,\cdots,\alpha_s)\times(\beta_1,\cdots,\beta_t)} = \begin{cases}1, & (\alpha_1\cdots\alpha_s\beta_1\cdots\beta_t)\in E\\ 0, & \text{otherwise.}\end{cases}$$

Example 14.17. Consider (V, E_2) in Example 14.16.

(i) If we use $\mathrm{id}(i,j;8,8)\times\mathrm{id}(k;8)$ to order the edges. Then the hypergraph matrix becomes

$$M_G = \delta_{64}[\ 0\ \ 6\ \ 0\ 23\ \ 0\ 33\ \ 0\ 52].$$

(ii) If we use $\mathrm{id}(i;8)\times\mathrm{id}(j,k;8,8)$ to order the edges. Then the hypergraph matrix becomes

$$M_G = \delta_8[\underbrace{0\cdots0}_{5}\ 5\ \underbrace{0\cdots0}_{25}\ 7\ \underbrace{0\cdots0}_{9}\ 1\ \underbrace{0\cdots0}_{9}\ 3\ \underbrace{0\cdots0}_{12}].$$

14.8 Universal Algebra

Universal algebra is the most general general description of the algebraic structure on a set. Group, ring, field, algebra, lattice, etc. are all its particular cases. When finite sets (or finite-dimensional vector spaces) are considered, STP can be used to describe all its functions. This section provides only a preliminary.

Definition 14.24. A type of algebras is a set $\mathcal{F}$ of function symbols such that a nonnegative integer n is assigned to each member f of $\mathcal{F}$. This integer n is called the arity of f, and f is said to be an n-ary function symbol. The subset of n-ary function symbols in $\mathcal{F}$ is denoted by $\mathcal{F}_n$.

Definition 14.25. If $\mathcal{F}$ is a type of algebras then an algebra $\mathbf{A}$ of type $\mathcal{F}$ is an ordered pair $\langle A, F\rangle$ where A is a nonempty set and F is a family of finitary operations on A indexed by the type $\mathcal{F}$ such that corresponding to each n-ary function symbol f in $\mathcal{F}$ there is an n-ary operation $f^{\mathbf{A}}$ on A. The set A is call the underlying set of $\mathbf{A} = \langle A, F\rangle$, and the $f^{\mathbf{A}}s$ are called the fundamental operations of $\mathbf{A}$. $\langle A, F\rangle$ with $F \in \mathcal{F}$ is called a universal algebra. In addition,

(i) $\mathbf{A}$ is unary, if all of its operations are unary, and it is mono-unary if it has just one unary operation.
(ii) $\mathbf{A}$ is groupoid, if it has just one binary operation.
(iii) $\mathbf{A}$ is finite if $|A|$ is finite, and trivial if $|A| = 1$.

Example 14.18.

(1) A group $\mathbf{G}$ is an algebra $\langle G, \cdot, ^{-1}, 1\rangle$ with a binary, a unary, and a nullary operations in which the following identities are true:

(i)

$$x \cdot (y \cdot z) = (x \cdot y) \cdot z; \tag{14.47}$$

(ii)

$$x \cdot 1 = 1 \cdot x = x; \tag{14.48}$$

(iii)

$$x \cdot x^{-1} = x^{-1} \cdot x = 1. \tag{14.49}$$

(2) A group $\mathbf{G}$ is Abelian (or commutative) if the following identity is true:

(iv)

$$x \cdot y = y \cdot x. \tag{14.50}$$

(3) A semigroup is a groupoid $\langle G, \cdot \rangle$ in which (14.47) is true.

(4) A ring is an algebra $\langle R, +, \cdot, -, 0 \rangle$, where $+$ and $\cdot$ are binary, $-$ is unary and 0 is nullary, satisfying the following conditions:

(i) $\langle R, +, -, 0 \rangle$ is an Abelian group;

(ii) $\langle R, \cdot \rangle$ is a semigroup;

(iii)

$$x \cdot (y + z) = (x \cdot y) + (x \cdot z), \tag{14.51}$$

$$(x + y) \cdot z = (x \cdot z) + (y \cdot z). \tag{14.52}$$

(5) A semi-lattice is a semigroup $\langle S, \cdot \rangle$ which satisfies the commutative law (14.50) and the idempotent law

$$x \cdot x = x. \tag{14.53}$$

(6) An algebra $\langle L, \sqcup, \sqcap \rangle$ with two binary operations satisfying (14.1)–(14.8) is a lattice.

(7) An algebra $\langle L, \sqcup, \sqcap, 0, 1 \rangle$ with two binary and two nullary operations is a bounded lattice, if it satisfies:

(i) $\langle L, \sqcup, \sqcap \rangle$ is a lattice;

(ii) $x \sqcap 0 = 0$; $x \sqcup 1 = 1$.

(8) A Boolean algebra is an algebra $\langle B, \vee, \wedge, \neg, 0, 1 \rangle$ with two binary, one unary, and two nullary operations which satisfy:

(i) $\langle B, \vee, \wedge \rangle$ is a distributive; lattice

(ii)

$$x \wedge 0 = 0, \quad x \vee 1 = 1; \tag{14.54}$$

(iii)

$$x \wedge (\neg x) = 0, \quad x \vee (\neg x) = 1. \tag{14.55}$$

Definition 14.26. Let $\mathbf{A}$ and $\mathbf{B}$ be two algebras of the same type $\mathcal{F}$. Then a function $\pi : A \to B$ is an isomorphism from $\mathbf{A}$ to $\mathbf{B}$ if π is one-to-one and onto, and for every n-ary $f \in \mathcal{F}$, for $a_1, \cdots, a_n \in A$, we have

$$\pi f^{\mathbf{A}}(a_1, \cdots, a_n) = f^{\mathbf{B}}(\pi(a_1), \cdots, \pi(a_n)). \tag{14.56}$$

We say $\mathbf{A}$ is isomorphic to $\mathbf{B}$, written as $\mathbf{A} \cong \mathbf{B}$, if there is an isomorphism from $\mathbf{A}$ to $\mathbf{B}$. If π is an isomorphism from $\mathbf{A}$ to $\mathbf{B}$ we may simply say "$\pi : \mathbf{A} \to \mathbf{B}$ is an isomorphism".

Definition 14.27. Let $\mathbf{A}$ and $\mathbf{B}$ be two algebras of the same type. If $B \subseteq A$ and every fundamental operation of $\mathbf{B}$ is the restriction of the corresponding operation of $\mathbf{A}$, i.e., for each function symbol f,

$$f^B = f^A\big|_B,$$

then $\mathbf{B}$ is called a subalgebra of $\mathbf{A}$, which is denoted simply by $\mathbf{B} \le \mathbf{A}$.

Consider finite algebras. The following result is obvious.

Proposition 14.4. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two finite algebras of the same type* $\mathcal{F}$*.* $\pi : \mathbf{A} \to \mathbf{B}$ *is an isomorphism, if and only if for each* $f \in \mathcal{F}$ *the structure matrices of* f^A *and* f^B *(corresponding to* $\{a_1, \cdots, a_n\}$ *and* $\{b_1 = \pi(a_1), \cdots, b_n = \pi(a_n)\}$*) are the same.*

We give a simple example for this.

Example 14.19. Consider a group generated by $\{a, b\}$ satisfying

$$a^2 = b^2 = (ab)^3 = 1.$$

Thus, it is easy to verify that this group has only 6 elements (note that $aba = bab$)

$$G = \{1, a, b, ab, ba, aba\},$$

and $1^{-1} = 1$, $a^{-1} = a$, $b^{-1} = b$, $(ab)^{-1} = ba$, $(aba)^{-1} = aba$. As a universal algebra, its has three operators $\cdot$, ${}^{-1}$ and 1, thus we denote $\mathbf{G} = \langle G, \cdot, {}^{-1_G}, 1\rangle$.

Now, we define a bijective $\pi : \mathbf{G} \to \mathbf{S}_3$, where $\mathbf{S}_3$ is the 3rd order symmetric group with operators $\circ, {}^{-1_{S_3}}, \mathrm{id}$, as

$$\begin{array}{lll}
\pi(1) = \mathrm{id} & \pi(a) = (1,2) & \pi(b) = (2,3) \\
\pi(ab) = (1,3,2) & \pi(ba) = (1,2,3) & \pi(aba) = (1,3).
\end{array}$$

Then, by identifying $1 \sim \pi(1) \sim \delta_6^1, a \sim \pi(a) \sim \delta_6^2, \cdots, aba \sim \pi(aba) \sim \delta_6^6$, it is easy to calculate the structure matrices of the operators

$$\begin{aligned}
M_{\cdot} = M_{\circ} = \delta_6[&1\ 2\ 3\ 4\ 5\ 6\ 2\ 1\ 4\ 3\ 6\ 5\ 3\ 5\ 1\ 6\ 2\ 4 \\
&4\ 6\ 2\ 5\ 1\ 3\ 5\ 3\ 6\ 1\ 4\ 2\ 6\ 4\ 5\ 2\ 3\ 1];
\end{aligned}$$

$$M_{^{-1}G} = M_{^{-1}S_3} = \delta_6[1\ 2\ 3\ 5\ 4\ 6];$$

$$M_1 = M_{\mathrm{id}} = \delta_6^1.$$

Hence, $\mathbf{G}$ and $\mathbf{S}_3$ are isomorphic.

In application Proposition 14.4 is not convenient, because in general we are not able to know π in advance. Hence for each f^A and f^B we have only their structure matrices M_{f^A} and M_{f^B} respectively. Now since $\mathbf{A}$ and $\mathbf{B}$ are isomorphic, there is an element rearrangement of A (or B), such that under this new arrangement M_{f^A} and M_{f^B} are the same.

Now assume $(\tilde{a}_1, \cdots, \tilde{a}_n)$ be a rearrangement of $(a_1, \cdots, a_n)$. It means there is a $T \in \mathcal{L}_{n\times n}$ non-singular such that

$$(\tilde{a}_1, \cdots, \tilde{a}_n) = (a_1, \cdots, a_n)T.$$

Now for any element $a \in A$, we have

$$a = (\tilde{a}_1, \cdots, \tilde{a}_n)\tilde{V}_a = (a_1, \cdots, a_n)V_a, \tag{14.57}$$

where $\tilde{V}_a$ and V_a are the vector form of a with respect to different bases $\{\tilde{a}_i\}$ and $\{a_i\}$ respectively. From (14.57) we can easily get that

$$\tilde{V}_a = T^{-1}V_a = T^T V_a. \tag{14.58}$$

Now assume $\tilde{X}$ and X are vector forms of an elements with respect to the two bases respectively, and similarly, we have $\tilde{Y}$ and Y. According to (14.58), we have

$$\tilde{X} *_f \tilde{Y} = T^T X *_f Y.$$

It follows that

$$\begin{aligned}\tilde{M}_{f^A}\tilde{X}\tilde{Y} &= T^T M_{f^A} XY \\ &= T^T M_{f^A} T\tilde{X}T\tilde{Y} \\ &= T^T M_{f^A}(T \otimes T)\tilde{X}\tilde{Y}.\end{aligned}$$

The above argument leads to the following general result.

Theorem 14.16. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two algebras of the same type* $\mathcal{F}$*, and* $|A| = |B| = n$*. They are isomorphic, if and only if there exists a nonsingular transformation* $T \in \mathcal{L}_{n\times n}$*, such that for each* $f \in \mathcal{F}$ *the structure matrices of* f^A *and* f^B *are equivalent with respect to the transformation* T*. Precisely,*

$$\tilde{M}_{f^A} = T^T M_{f^B}(T \otimes T). \tag{14.59}$$

The following example is used to demonstrate this general isomorphic property.

Example 14.20. Consider the power set $\mathcal{P}(A)$ of $A = \{b, c, d\}$. It is natural to let $\cap$ and $\cup$ be two binary operators on $\mathcal{P}(A)$, thus we obtain an universal algebra $\mathcal{P}(\mathbf{A}) = \langle \mathcal{P}(A), \cap, \cup\rangle$.

Identifying

$$\begin{array}{llll} \emptyset \sim \delta_8^1 & \{b\} \sim \delta_8^2 & \{c\} \sim \delta_8^3 & \{d\} \sim \delta_8^4 \\ \{b,c\} \sim \delta_8^5 & \{b,d\} \sim \delta_8^6 & \{c,d\} \sim \delta_8^7 & \{b,c,d\} \sim \delta_8^8, \end{array}$$

we can construct the structure matrices of $\cap$ and $\cup$ as

$$\begin{aligned} M_{\cap} = \delta_8\, [&1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 2\ 2\ 1\ 2 \\ &1\ 1\ 3\ 1\ 3\ 1\ 3\ 3\ 1\ 1\ 1\ 4\ 1\ 4\ 4\ 4 \\ &1\ 2\ 3\ 1\ 5\ 2\ 3\ 5\ 1\ 2\ 1\ 4\ 2\ 6\ 4\ 6 \\ &1\ 1\ 3\ 4\ 3\ 4\ 7\ 7\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8]; \\ M_{\cup} = \delta_8\, [&1\ 2\ 3\ 4\ 5\ 8\ 7\ 8\ 2\ 2\ 5\ 6\ 5\ 6\ 8\ 8 \\ &3\ 5\ 3\ 7\ 5\ 8\ 7\ 8\ 4\ 6\ 7\ 4\ 8\ 6\ 7\ 8 \\ &5\ 5\ 5\ 8\ 5\ 8\ 8\ 8\ 6\ 6\ 8\ 6\ 8\ 6\ 8\ 8 \\ &7\ 8\ 7\ 7\ 8\ 8\ 7\ 8\ 8\ 8\ 8\ 8\ 8\ 8\ 8\ 8]. \end{aligned}$$

Then we consider a lattice **L** whose Hasse diagram is shown as Fig. 14.13.

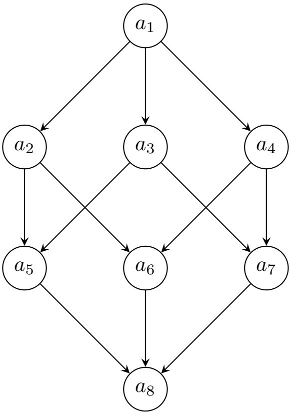


Fig. 14.13 Hasse diagrams of **L**

Thus, identifying $a_i \sim \delta_8^i$, we can also find the structure matrices for $\sqcap$ and $\sqcup$ as

$$\begin{aligned} M_{\sqcap} = \delta_8\, [&1\ 2\ 3\ 4\ 5\ 8\ 7\ 8\ 2\ 2\ 5\ 6\ 5\ 6\ 8\ 8 \\ &3\ 5\ 3\ 7\ 5\ 8\ 7\ 8\ 4\ 6\ 7\ 4\ 8\ 6\ 7\ 8 \\ &5\ 5\ 5\ 8\ 5\ 8\ 8\ 8\ 6\ 6\ 8\ 6\ 8\ 6\ 8\ 8 \\ &7\ 8\ 7\ 7\ 8\ 8\ 7\ 8\ 8\ 8\ 8\ 8\ 8\ 8\ 8\ 8]; \\ M_{\sqcup} = \delta_8\, [&1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 1\ 1\ 2\ 2\ 1\ 2 \\ &1\ 1\ 3\ 1\ 3\ 1\ 3\ 3\ 1\ 1\ 1\ 4\ 1\ 4\ 4\ 4 \\ &1\ 2\ 3\ 1\ 5\ 2\ 3\ 5\ 1\ 2\ 1\ 4\ 2\ 6\ 4\ 6 \\ &1\ 1\ 3\ 4\ 3\ 4\ 7\ 7\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8]. \end{aligned}$$

Then, we can see that $M_\cap = M_\sqcup$ and $M_\cup = M_\sqcap$, thus, $\mathcal{P}(\mathbf{A})$ is isomorphic to $\mathbf{L}$. However, this is inconsistent with our common sense, because when identify a power set with a lattice, we always let $\cap$ be $\sqcap$, and $\cup$ be $\sqcup$.

In fact, if we let

$$T = \delta_8[8\ 7\ 6\ 5\ 4\ 3\ 2\ 1],$$

it is easy to check that

$$M_\cap = T^T M_\sqcap (T \otimes T), \quad M_\cup = T^T M_\sqcup (T \otimes T).$$

By Theorem 14.16, $\mathcal{P}(\mathbf{A})$ and $\mathbf{L}$ are isomorphic, and $\cap$ is identified with $\sqcap$, $\cup$ is identified with $\sqcup$.

14.9 Lattice-Based Logics

We have discussed a lot about various logics. In fact, logic can be defined via lattice. This is an interesting and useful topic. We refer to Cohen (1989) and Barnes and Mack (1975) for details.

Definition 14.28. Let L be a lattice.

(i) An element $1_L \in L$ is called the unit of L, if $p \le 1_L$, $\forall p \in L$;
(ii) An element $0_L \in L$ is called the zero of L, if $p \ge 0_L$, $\forall p \in L$.

Definition 14.29. A lattice L with 1_L, 0_L and a mapping $' :\ L \to L$ is called a logic, if the followings are satisfied:

(i) for all $p \in L$, $p'' = p$ and $p \sqcap p' = 0_L$;
(ii) for $p, q \in L$, if $p \le q$, then $q' \le p'$;
(iii) for $p, q \in L$, if $p \le q$, then $q = p \sqcup (p' \sqcap q)$.

We leave to the reader to verify that (i) classical logic, (ii) multi-valued logic, (iii) fuzzy logic, satisfy the above definition with

$$\sqcap = \wedge, \quad \sqcup = \vee, \quad ' = \neg.$$

As for mix-valued logic, some further concepts such as free algebra etc. are necessary to describe its algebraic structure (Barnes and Mack, 1975).

It is easy to verify that under this definition several basic properties of classical logic remain true. For instance, we have

Theorem 14.17 (De Morgan's Law). *Let L be a logic (as defined in Definition 14.29). Then for any $p, q \in L$*

$$(p \sqcup q)' = p' \sqcap q'; \quad (p \sqcap q)' = p' \sqcup q'. \tag{14.60}$$

We use Definition 14.29 to introduce another useful logic, called the quantum logic (Cohen, 1989; Svozil, 1998).

Definition 14.30. Let L be a logic with $p, q \in L$.

(i) p is said to be orthogonal to q, denoted by $p \perp q$, if $p \leq q'$.
(ii) p and q are called compatible if there exist $u, v, w \in L$, which are pairwise orthogonal, such that

$$p = u \sqcup v, \quad \text{and} \quad q = v \sqcup w.$$

Definition 14.31. A quantum logic is a logic with at least two elements that are not compatible.

Roughly speaking, a quantum logic is built over a Hilbert space H. Precisely, let $S \subset 2^H$ be the set of its closed subspaces, and assume $p, q \in S$. Then we define the operators on S as

- $\sqcap$:

$$p \sqcap q := p \cap q,$$

which is the intersection of p and q;
- $\sqcup$:

$$p \sqcup q := \overline{p \cup q},$$

which is the smallest closed subspace of H, which contains p and q;
- $'$:

$$p' := p^{\perp},$$

which is the orthogonal complement of p in H.

Then we can prove that this logic is a quantum logic (Cohen, 1989).

For a quantum logic L a state is a probability measure $\pi : L \to [0, 1]$, satisfying

(1) for $p, q \in L$ and $p \perp q$,

$$\pi(p \sqcup q) = \pi(p) + \pi(q);$$

(2)

$$\pi(1_L) = 1.$$

Let $\{\pi_1, \cdots, \pi_n\}$ be a finite set of states on L. If there exists $\alpha_i \geq 0$, $i = 1, \cdots, n$, with $\sum\limits_{i=1}^{n} \alpha_i = 1$, such that a state π is defined by

$$\pi(p) = \sum_{i=1}^{n} \alpha_i \pi_i(p), \quad \forall p \in L,$$

then π is called a mixture of the states $\{\pi_1, \cdots, \pi_n\}$.

Now if for any $p \in L$ a state satisfying either $\pi(p) = 0$ or $\pi(p) = 1$, then the state is called a classical state. Otherwise, π is a mixture reflecting an epistemic uncertainty. Quantum logic is a proper model to describe quantum mechanics (Cohen, 1989; Svozil, 1998).

Exercises

14.1 Verify that the two objects in Example 14.1 are lattice.

14.2 For a given lattice prove that (14.23) ⇒ (14.24).

14.3 For any lattice prove (14.25) and (14.26).

14.4 Let S be the set of all subspaces of $\mathbb{R}^n$. Let $s_1, s_2 \in S$. Define

$$s_1 \sqcap s_2 = s_1 \cap s_2,$$

and

$$s_1 \sqcup s_2 = \text{Span}\{s_1, s_2\},$$

that is, it is the subspace spanned by the vectors in $s_1 \cup s_2$.

(i) Show that $\{S, \sqcup, \sqcap\}$ is a lattice.

(ii) Is this a modular lattice?

14.5 Observe Fig. 14.1.

(i) Construct $M_\sqcup$ and $M_\sqcap$ for the lattice described by figure A.

(ii) Construct $M_\sqcup$ and $M_\sqcap$ for the lattice described by figure C.

(iii) Construct $M_\sqcup$ and $M_\sqcap$ for the lattice described by figure D.

14.6 A lattice of n elements is a modular lattice, if one of the following two conditions ((14.61) or (14.62)) is verified. Prove it.

(i)

$$M_\sqcap \left(I_n \otimes M_\sqcup\right) = M_\sqcup M_\sqcap \left(I_{n^2} \otimes M_\sqcap\right) \left(I_n \otimes W_{[n]}\right) M_{r,n}. \tag{14.61}$$

(ii)

$$M_\sqcup \left(I_n \otimes M_\sqcap\right) = M_\sqcap M_\sqcup \left(I_{n^2} \otimes M_\sqcup\right) \left(I_n \otimes W_{[n]}\right) M_{r,n}. \tag{14.62}$$

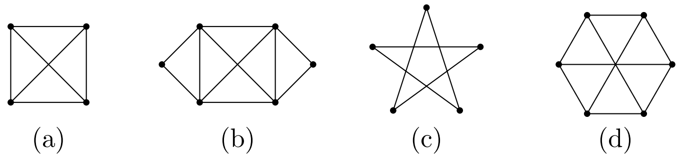


Fig. 14.14 Check "Eulerian"

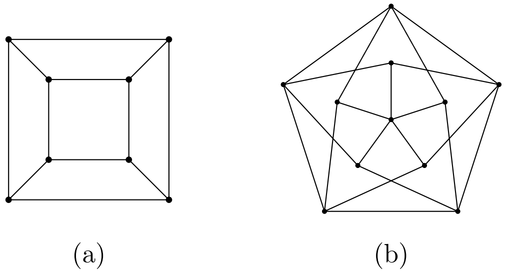


Fig. 14.15 Check "Hamiltonian"

14.7 Calculate the adjacency matrices of the four graphs in Fig. 14.4 (a)–(d).

14.8 In the 4 graphs in Fig. 14.14 verify which one(s) is (are) Eulerian?

14.9 Show that the 2 graphs in Fig. 14.15 are Hamiltonian.

14.10 Draw the dual graph for (i) K_5; (ii) $K_{3,3}$.

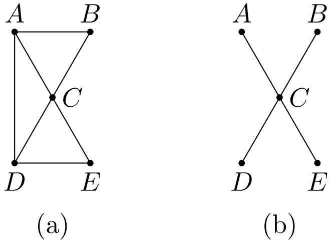


Fig. 14.16 Spanning tree

14.11 Consider the graph Fig. 14.16 (a).

(i) Prove that Fig. 14.16 (b) is a spanning tree of Fig. 14.16 (a).

(ii) Find the set C of the fundamental cycles with respect to this spanning tree.

(iii) Show that each cycle in Fig. 14.16 (a) can be expressed as a linear combination of some cycles in C. For instance, consider (a) $A-B-C-D-A$, (b) $A-B-C-E-D-A$.

14.12 Consider the graph in Fig. 14.16 (a).

(i) Find the set T of the fundamental cutsets with respect to the spanning tree Fig. 14.16 (b).

(ii) Show that each cutset in Fig. 14.16 (a) can be expressed as a linear combination of some cutsets in T. For instance, consider $AC-CB-AD$.

14.13 Consider the bipartite graph $K_{i,j}$. When (for what values i and j) it is a planar graph?

14.14 Consider the hypergraph (V,E_3) in Example 14.16. Express its graph matrix using the order

(i)

$$\mathrm{id}(i,j,k;8,8,8)\times\mathrm{id}(\ell;8);$$

(ii)

$$\mathrm{id}(i,j;8,8)\times\mathrm{id}(k,\ell;8,8);$$

(iii)

$$\mathrm{id}(i;8)\times\mathrm{id}(j,k,\ell;8,8,8).$$

14.15 Assume an algebra $\langle R,+,\times,-,\div,0,1\rangle$ is a field. Give a rigorous definition for field.

14.16 Verify that (i) classical logic, (ii) multi-valued logic, (iii) fuzzy logic satisfy Definition 14.29.

14.17 Let L be a logic and $p,q\in L$. Prove that if $p\perp q$ then $q\perp p$.

Chapter 15

Boolean Network

Boolean network was firstly introduced by Kauffman to formulate the cellular networks. Then, it has been developed by Akutsu *et al.* (2000); Albert and Barabási (2000); Shmulevich *et al.* (2002); Harris *et al.* (2002); Aldana (2003); Samuelsson and Troein (2003); Drossel *et al.* (2005); Kauffman (1993) and many others, and becomes a powerful tool in describing, analyzing, and simulating the cellular networks. Meanwhile, it has also been used for modeling some other complex systems such as neural networks, social and economic networks.

In this chapter, using the matrix expression of logic, discussed in Chapter 6, the dynamics of a Boolean network is converted into an equivalent algebraic form as a standard discrete-time linear system. Then the topological structure (fixed points, cycles, transient period, etc.), subspace, input-state description of Boolean networks and higher-order Boolean networks are investigated.

15.1 An Introduction

When a Boolean network is used to describe a cellular network, gene state is quantized to only two levels: true and false. Then the state of each gene is determined by the states of its neighborhood genes, using logical rules. Precisely speaking, a Boolean network is a directed network graph, $\Sigma = \{\mathcal{N}, \mathcal{E}\}$, consists of a finite set of nodes, $\mathcal{N} = \{x_i \,|\, i = 1, \cdots, n\}$ and a set of edges, denoted by $\mathcal{E} \subset \{x_1, \cdots, x_n\} \times \{x_1, \cdots, x_n\}$. If $(x_i, x_j) \in \mathcal{E}$, there is an edge from x_i to x_j, which means node x_j is affected by node x_i.

The dynamics of a Boolean network can be expressed as a set of logical dynamic equations. We first give a rigorous definition for the dynamics of a Boolean network.

Definition 15.1. (Farrow *et al.*, 2004) A Boolean network consists of a set of nodes $x_1, x_2, \cdots, x_n$, which interact with each other in a synchronous manner. At each given time $t = 0, 1, 2, \cdots$ a node has only one of two possible values: 1 or 0. Thus the network can be described by a set of equations:

$$\begin{cases} x_1(t+1) = f_1(x_1(t), x_2(t), \cdots, x_n(t)) \\ x_2(t+1) = f_2(x_1(t), x_2(t), \cdots, x_n(t)) \\ \vdots \\ x_n(t+1) = f_n(x_1(t), x_2(t), \cdots, x_n(t)), \end{cases} \tag{15.1}$$

where f_i, $i = 1, 2, \cdots, n$ are n-ary logical functions.

We give a simple example to show this.

Example 15.1. Consider a Boolean network, $\Sigma = (\mathcal{N}, \mathcal{E})$ of three nodes as

$$\begin{cases} A(t+1) = B(t) \wedge C(t) \\ B(t+1) = \neg A(t) \\ C(t+1) = B(t) \vee C(t). \end{cases} \tag{15.2}$$

Its set of nodes is $\mathcal{N} = \{x_1 := A,\ x_2 := B,\ x_3 := C\}$, set of edges is $\mathcal{E} = \{(x_1, x_2), (x_2, x_1), (x_2, x_3), (x_3, x_1), (x_3, x_3)\}$. Its network graph is depicted in Fig 15.1.

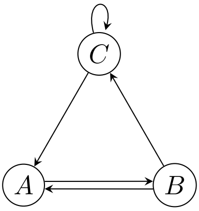


Fig. 15.1 Network graph of (15.1)

Our first purpose is to convert Boolean network dynamics (15.1) into an algebraic form. Precisely, express it as a conventional discrete-time linear system. Recall the technique developed in Chapter 6, we use vector form $x_i(t) \in \Delta$, and define

$$x(t) = x_1(t)x_2(t)\cdots x_n(t) := \ltimes_{i=1}^{n} x_i(t).$$

Using Lemma 6.2, there exist structure matrices, $M_i = M_{f_i}$, $i = 1, \cdots, n$, such that

$$x_i(t+1) = M_i x(t), \quad i = 1, 2, \cdots, n. \tag{15.3}$$

Remark 15.1. Note that usually the right hand side of ith equation of (15.1) may not have all x_j, $j = 1, 2, \cdots, n$. Say, in the previous example, for node A we have

$$A(t+1) = B(t) \wedge C(t).$$

In matrix form it is

$$A(t+1) = M_c B(t) C(t). \tag{15.4}$$

To get the form of (15.3), using the dummy matrix $D_{p,q}$ defined in (7.25), we can rewrite (15.4) as

$$A(t+1) = M_c D_2 A(t) B(t) C(t) = M_c D_2 x(t).$$

Multiplying the equations in (15.3) together yields

$$x(t+1) = M_1 x(t) M_2 x(t) \cdots M_n x(t). \tag{15.5}$$

Using Theorem 7.3, equation (15.5) can be expressed as

$$x(t+1) = L x(t), \tag{15.6}$$

where

$$L = \prod_{i=1}^{n} \left(I_{2^{i-1}} \otimes M_i\right) \left(M_{r,2^n}\right)^{n-1}$$

is called the transition matrix. Thus, we have the following result.

Theorem 15.1. *The dynamics of Boolean network (15.1) is uniquely determined by linear dynamic system (15.6).*

Definition 15.2. Equation (15.3) is called the component-wise algebraic form of network (15.1). Equation (15.6) is called the algebraic form of network (15.1).

Alternatively, a direct computation by using Khatri-Rao product can produce the algebraic form easily. We give a simple example to see how to get algebraic form of the dynamics of Boolean networks.

Example 15.2. Recall the Boolean network in Example 15.1. Its dynamics is described as (15.2). In algebraic form, we have

$$\begin{cases} A(t+1) = M_c B(t) C(t) \\ B(t+1) = M_n A(t) \\ C(t+1) = M_d B(t) C(t). \end{cases} \tag{15.7}$$

Using dummy matrix $D_{p,q}$, it is easy to convert (15.7) to its componentwise algebraic form

$$\begin{cases} A(t+1) = \delta_2[1\ 2\ 2\ 2\ 1\ 2\ 2\ 2]A(t)B(t)C(t) := M_1x(t) \\ B(t+1) = \delta_2[2\ 2\ 2\ 2\ 1\ 1\ 1\ 1]A(t)B(t)C(t) := M_2x(t) \\ C(t+1) = \delta_2[1\ 1\ 1\ 2\ 1\ 1\ 1\ 2]A(t)B(t)C(t) := M_3x(t), \end{cases} \tag{15.8}$$

where $x(t) = A(t)B(t)C(t)$. Using Khatri-Rao product we can get the algebraic form of (15.7) as

$$x(t+1) = Lx(t),$$

where

$$\mathrm{Col}_i(L) = \mathrm{Col}_i(M_1) * \mathrm{Col}_i(M_2) * \mathrm{Col}_i(M_3), \quad i = 1, \cdots, 8.$$

Hence, L can be calculated easily as

$$L = \delta_8[3\ 7\ 7\ 8\ 1\ 5\ 5\ 6].$$

Remark 15.2. Equation (15.6) is a standard linear system with L being a square logical matrix. So all classical methods and conclusions for linear systems can be used for analyzing the dynamics of the Boolean network.

15.2 Fixed Points and Cycles

This section considers the fixed points and cycles of a Boolean network. Since there are only finite nodes in a Boolean network, a trajectory will eventually converge to a cycle, here a fixed point is considered as a cycle of length 1. Hence, the cycles and their regions of attraction are of fundamental importance for describing the topology of a Boolean network. There are some earlier works on calculating cycles. For instance, iteration and scalar form, were developed in Heidel *et al.* (2003) to determine cyclic structure and the transient states that lead to them. In Farrow *et al.* (2004), a linear reduced scalar equation is derived from a more rudimentary nonlinear s-calar equation to get immediate information about both cycle and transient structure of the network. It was pointed out in Zhao (2005) that finding fixed points and cycles of a Boolean network is an NP-complete problem.

In this section, we present some general formulas to calculate cycles. We first give a rigorous definition.

Definition 15.3. (1) Given system (15.6), the pair $(\Delta_n, \mathcal{E})$ where

$$\mathcal{E} = \{(x_i, x_j) | x_j = Lx_i\}$$

forms a directed graph, which is called the state transfer graph (STG).

(2) $(x_0, x_1, x_2, \cdots)$ is called a path of system (15.6), if $x_i = Lx_{i-1}, i \geq 1$.

(3) $(x_0, x_1, \cdots, x_k = x_0)$ is called a cycle of system (15.6) with length k, if $\{x_0, x_1, \cdots, x_{k-1}\}$ are distinct and $x_k = x_0$. A fixed point is a cycle of length 1.

The next two theorems are main results of this section, which show how many fixed points and cycles of different lengths a Boolean network has.

Theorem 15.2. *Consider Boolean network (15.1). $\delta_{2^n}^i$ is its fixed point, if and only if in its algebraic form (15.6) the diagonal element ℓ_{ii} of the network transition matrix L equals 1. It follows that the number of fixed points of network (15.1), denoted by N_e, equals the number of i, for which $\ell_{ii} = 1$. Equivalently,*

$$N_e = \mathrm{tr}(L). \tag{15.9}$$

Proof. Assume $\delta_{2^n}^i$ is its fixed point. Note that $L\delta_{2^n}^i = \mathrm{Col}_i(L)$. It is clear that $\delta_{2^n}^i$ is its fixed point, if and only if $\mathrm{Col}_i(L) = \delta_{2^n}^i$. The conclusion follows immediately. □

For statement ease, if $\ell_{ii} = 1$, the $\mathrm{Col}_i(L)$ is called a diagonal nonzero column of L.

Next, we consider the cycles of Boolean network (15.1). We need a notation: Let $k \in \mathbb{Z}_+$. A positive integer $s \in \mathbb{Z}_+$ is called a proper factor of k if $s < k$ and $k/s \in \mathbb{Z}_+$. The set of proper factors of k is denoted by $\mathcal{P}(k)$. For instance, we have

$$\mathcal{P}(8) = \{1, 2, 4\},$$
$$\mathcal{P}(24) = \{1, 2, 3, 4, 6, 8, 12\},$$

etc.

Using a similar argument as for Theorem 15.2, we can have the following theorem.

Theorem 15.3. *The number of length s cycles, denoted by N_s, is inductively determined by*

$$\begin{cases} N_1 = N_e, \\ N_s = \frac{\mathrm{tr}(L^s) - \sum\limits_{k \in \mathcal{P}(s)} kN_k}{s}, \quad 2 \leq s \leq 2^n. \end{cases} \tag{15.10}$$

The proof is left for exercise.

Next, we consider how to find the cycles. If

$$\mathrm{tr}(L^s) - \sum_{k\in\mathcal{P}(s)} kN_k > 0, \tag{15.11}$$

then we call "s" a nontrivial power.

Assume s is a nontrivial power. Denote by ℓ^s_{ii} the (i,i)th entry of matrix L^s. Then we define

$$C_s = \{i \,|\, \ell^s_{ii} = 1\}, \quad s = 1, 2, \cdots, 2^n,$$

and

$$D_s = C_s \backslash \bigcup_{i\in\mathcal{P}(s)} C_i = \bigcap_{i\in\mathcal{P}(s)} C_i^c,$$

where C_i^c is the compliment of C_i.

From the above argument the following result is obvious.

Proposition 15.1. *Let $x_0 = \delta^i_{2^n}$. Then $(x_0, Lx_0, \cdots, L^s x_0)$ is a cycle with length s, if and only if $i \in D_s$.*

Theorem 15.3 and Proposition 15.1 provide a simple way to construct cycles. We give an example to show the constructing procedure.

Example 15.3. Recall Example 15.1. It is easy to check that

$$\mathrm{tr}(L^t) = 0, \quad t \le 3,$$

and

$$\mathrm{tr}(L^t) = 4, \quad t \ge 4.$$

Using Theorem 15.3, we conclude that there is only one cycle of length 4. Moreover, note that

$$L^4 = \delta_8[1\ 3\ 3\ 1\ 5\ 7\ 7\ 3].$$

Then each diagonal nonzero column can generate the cycle. For instance, choose $Z = \delta^1_8$, then we have

$$LZ = \delta^3_8, \quad L^2 Z = \delta^7_8, \quad L^3 Z = \delta^5_8, \quad L^4 Z = Z.$$

Converting the vector forms back to the scalar forms of $A(t)$, $B(t)$, and $C(t)$, we have the cycle as

$$(1,1,1) \to (1,0,1) \to (0,0,1) \to (0,1,1) \to (1,1,1).$$

We refer to Cheng and Qi (2010a); Cheng *et al.* (2011b) for calculating the transient period and basins of attractors of a Boolean network.

15.3 Invariant Subspace and Input-State Description

15.3.1 *State Space and Subspaces*

This section presents a systematic description of state space and subspaces of a Boolean (control) network. This description makes the state-space approach, similar to that of the modern control theory, applicable to the analysis of Boolean networks and the synthesis of Boolean control systems. This section is based on Cheng and Qi (2010b). Unlike the quantity-based dynamic (control) systems, the logic-based dynamic (control) systems do not have a natural vector space structure. To use the state-space approach, we have to find a proper way to define state space and its various subspaces.

Consider a Boolean network

$$\begin{cases} x_1(t+1) = f_1(x_1(t), \cdots, x_n(t)) \\ \vdots \\ x_n(t+1) = f_n(x_1(t), \cdots, x_n(t)), \quad x_i \in \mathcal{D}, \end{cases} \tag{15.12}$$

or a Boolean control network

$$\begin{cases} x_1(t+1) = f_1(x_1(t), \cdots, x_n(t), u_1(t), \cdots, u_m(t)) \\ \vdots \\ x_n(t+1) = f_n(x_1(t), \cdots, x_n(t), u_1(t), \cdots, u_m(t)), \end{cases} \\ y_j(t) = h_j(x_1(t), \cdots, x_n(t)), \quad j = 1, \cdots, p, \\ \qquad x_i, u_i, y_j \in \mathcal{D}. \tag{15.13}$$

We give the following definitions.

Definition 15.4. Consider Boolean network (15.12) or Boolean control network (15.13).

(1) The state space $\mathcal{X}$ is defined as the set of all logical functions of $x_1, \cdots, x_n$, denoted by

$$\mathcal{X} = \mathcal{F}_\ell\{x_1, \cdots, x_n\}. \tag{15.14}$$

$\{x_1, \cdots, x_n\}$ is called a basis of $\mathcal{X}$.

(2) Let $z_1, \cdots, z_k \in \mathcal{X}$. The subspace $\mathcal{Z}$ generated by $z_1, \cdots, z_k$ is the set of logical functions of $z_1, \cdots, z_k$, denoted by

$$\mathcal{Z} = \mathcal{F}_\ell\{z_1, \cdots, z_k\}. \tag{15.15}$$

$\{z_1, \cdots, z_k\}$ is called a basis of the subspace $\{\mathcal{Z}\}$.

(3) Let $Z = \{z_1, \cdots, z_n\} \subset \mathcal{X}$. For notational ease, we also consider $Z = (z_1, \cdots, z_n)^T$ as a column vector. The mapping $G : \mathcal{D}^n \to \mathcal{D}^n$ defined by $X = (x_1, \cdots, x_n)^T \mapsto Z = (z_1, \cdots, z_n)^T$ is called a coordinate transformation, if G is one-to-one and onto.

(4) A subspace $\mathcal{Z} = \mathcal{F}_\ell\{z_1, \cdots, z_k\} \subset \mathcal{X}$ is called a regular subspace of dimension k if there are $\{z_{k+1}, \cdots, z_n\}$, such that $Z = (z_1, \cdots, z_n)$ is a coordinate frame. Moreover, $\{z_1, \cdots, z_k\}$ is called a regular basis of $\mathcal{Z}$.

(5) Consider system (15.12). If it can be expressed (under a suitable coordinate frame) as

$$\begin{cases} z^1(t+1) = F^1(z^1(t)), & z^1 \in \mathcal{D}^s \\ z^2(t+1) = F^2(z(t)), & z^2 \in \mathcal{D}^{n-s}. \end{cases} \tag{15.16}$$

Then $\mathcal{Z} = \mathcal{F}_\ell\{z^1\} = \mathcal{F}_\ell\{z_1^1, \cdots, z_s^1\}$ is called an invariant subspace of (15.12).

Remark 15.3.

(1) Let $\xi \in \mathcal{X}$. Then ξ is a logical function of $x_1, \cdots, x_n$. Say,

$$\xi = g(x_1, \cdots, x_n).$$

Then it can be uniquely expressed into an algebraic form as

$$\xi = M_g \ltimes_{i=1}^n x_i,$$

where $M_g \in \mathcal{L}_{2\times 2^n}$. Now M_g can be expressed as

$$\delta_2[i_1\ i_2\ \cdots\ i_{2^n}],$$

where i_s can be either 1 or 2. It follows that there are 2^{2^n} different functions. That is,

$$|\mathcal{X}| = 2^{2^n}.$$

(2) Using a set of functions to define a (sub) space is reasonable. For instance, in linear space $\mathbb{R}^n$ with the coordinate frame $\{x_1, \cdots, x_n\}$, we consider all the linear functions over $x_{i_1}, \cdots, x_{i_k}$, it is

$$L_k = \left\{ \sum_{j=1}^k c_j x_{i_j} \,\middle|\, c_1, \cdots, c_k \in \mathbb{R} \right\},$$

which is obviously a k-dimensional subspace. In fact, we can identify L_k with its domain, which is a k-dimensional subspace in state space $\mathbb{R}^n$, called the dual space of L_k.

The logical space (subspace) defined in this section is also in dual sense. Precisely, we consider its domain as a subspace of the state space.

(3) In vector space, the basis of a subspace can always be expanded to the basis of the overall space by adding some vectors to the basis. But this is no longer true in logical space. (Finding a counter-example for this is left for exercise.)

Let $G : \mathcal{D}^n \to \mathcal{D}^n$ be defined by

$$z_i = f_i(x_1, \cdots, x_n), \quad i = 1, \cdots, n. \tag{15.17}$$

The algebraic form of (15.17) is

$$z = T_G x, \quad T_G \in \mathcal{L}_{2^n \times 2^n}, \tag{15.18}$$

where $x = \ltimes_{i=1}^n x_i$, $z = \ltimes_{i=1}^n z_i$. Then the following result about coordinate transformation is obvious.

Theorem 15.4. *G is a coordinate transformation, if and only if its structure matrix T_G is nonsingular.*

Remark 15.4. If a matrix $T \in \mathcal{L}_{s\times s}$ and it is nonsingular, then it is an orthogonal matrix. Hence, if $z = T_G x$ is a coordinate transformation then $x = T_G^T z$.

Next, we consider a logical coordinate transformation of the dynamics of a Boolean network.

Consider a Boolean network in algebraic form as

$$x(t+1) = Lx(t), \quad x \in \Delta_{2^n}. \tag{15.19}$$

Let $z = Tx : \Delta_{2^n} \to \Delta_{2^n}$ be a logical coordinate transformation. Then

$$z(t+1) = Tx(t+1) = TLx(t) = TLT^T z(t).$$

That is, under z coordinate frame Boolean network dynamics (15.19) becomes

$$z(t+1) = \tilde{L}z(t), \tag{15.20}$$

where

$$\tilde{L} = TLT^T. \tag{15.21}$$

Consider a Boolean control system in algebraic form as

$$\begin{cases} x(t+1) = Lu(t)x(t), & x \in \Delta_{2^n},\ u \in \Delta_{2^m} \\ y(t) = Hx(t), & y \in \Delta_{2^p}. \end{cases} \tag{15.22}$$

Let $z = Tx : \Delta_{2^n} \to \Delta_{2^n}$ be a logical coordinate transformation. A straightforward computation shows that (15.22) can be expressed as

$$\begin{cases} z(t+1) = \tilde{L}u(t)z(t), & z \in \Delta_{2^n},\ u \in \Delta_{2^m} \\ y(t) = \tilde{H}z(t), & y \in \Delta_{2^p}, \end{cases} \tag{15.23}$$

where

$$\begin{cases} \tilde{L} = TL(I_{2^m} \otimes T^T) \\ \tilde{H} = HT^T. \end{cases} \tag{15.24}$$

Next, we consider the verification of regular subspaces.

Given a set of functions z_i as

$$z_i = g_i(x_1, \cdots, x_n), \quad i = 1, \cdots, k, \tag{15.25}$$

and let $\mathcal{Z} = \mathcal{F}_\ell\{z_1, \cdots, z_k\}$. We would like to know when $\mathcal{Z}$ is a regular subspace with $\{z_1, \cdots, z_k\}$ as its regular subbasis. Set $z = \ltimes_{i=1}^k z_i$ and $x = \ltimes_{i=1}^n x_i$. We can get its algebraic form as

$$z = Lx := \begin{bmatrix} \ell_{1,1} & \ell_{1,2} & \cdots & \ell_{1,2^n} \\ \vdots & & & \\ \ell_{2^k,1} & \ell_{2^k,2} & \cdots & \ell_{2^k,2^n} \end{bmatrix} x, \tag{15.26}$$

where

$$L = \prod_{i=1}^k (I_{2^{i-1}} \otimes M_i) \left(M_{r,2^k}\right)^{n-1}. \tag{15.27}$$

Theorem 15.5. *Assume there is a set of logical variables $z_1, \cdots, z_k \in \mathcal{X}$ $(k \le n)$, with its algebraic form as (15.26). $\mathcal{Z} = \mathcal{F}_\ell\{z_1, \cdots, z_k\}$ is a regular subspace with regular subbasis $\{z_1, \cdots, z_k\}$, if and only if the corresponding structure matrix L satisfies*

$$\sum_{i=1}^{2^n} \ell_{j,i} = 2^{n-k}, \quad j = 1, 2, \cdots, 2^k. \tag{15.28}$$

Proof. (Sufficiency) Note that condition (15.28) means there are 2^{n-k} different x which make $z = \delta_{2^k}^j$, $j = 1, 2, \cdots, 2^k$. Now we can choose z_{k+1} as follows. Set

$$S_k^j = \{x \,|\, Lx = \delta_{2^k}^j\}, \quad j = 1, 2, \cdots, 2^k.$$

Then the cardinal number $\left|S_k^j\right| = 2^{n-k}$. For half of the elements in S_k^j, define $z_{k+1} = 0$, and for the other half, set $z_{k+1} = 1$. Then it is easy to see that for $\tilde{z} = \ltimes_{i=1}^{k+1} z_i$ the corresponding $\tilde{L}$ satisfies (15.28) with k being replaced by $k+1$.

Continue this process till $k = n$. Then (15.28) becomes

$$\sum_{i=1}^{2^n} \ell_{j,i} = 1, \quad j = 1, 2, \cdots, 2^n. \tag{15.29}$$

(15.29) means the corresponding L contains all the columns of I_{2^n}, *i.e.*, it can be obtained from I_{2^n} via a column permutation. It is, hence, a coordinate change.

(Necessity) Observe that using the swap matrix, it is easy to see that the order of z_i does not affect the property of (15.28). First, we claim that if $\{z_i \,|\, i = 1, \cdots, k\}$ satisfies (15.28), then any of its subset $\{z_{i_t}\} \subset \{z_i \,|\, i = 1, \cdots, k\}$ also satisfies (15.28) with k being replaced by $|\{z_{i_t}\}|$. Since the order does not affect this property, it is enough to show that a $k-1$ subset $\{z_i \,|\, i = 2, \cdots, k\}$ is a proper subbasis, because from $k-1$ we can go to $k-2$ and so on. Assume that $z^2 = \ltimes_{i=2}^{k} z_i = Qx$, and $z_1 = Px$. Using Proposition 6.11, we have

$$\mathrm{Col}_i(L) = \mathrm{Col}_i(P)\,\mathrm{Col}_i(Q), \quad i = 1, \cdots, 2^n. \tag{15.30}$$

Next, we split L into two equal-size blocks as

$$L = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix}.$$

Note that either $\mathrm{Col}_i(P) = \delta_2^1$ or $\mathrm{Col}_i(P) = \delta_2^2$. Using this fact to (15.30), one sees easily that either $\mathrm{Col}_i(L) = \begin{bmatrix} \mathrm{Col}_i(Q) \\ 0 \end{bmatrix}$ (as $\mathrm{Col}_i(P) = \delta_2^1$) or $\mathrm{Col}_i(L) = \begin{bmatrix} 0 \\ \mathrm{Col}_i(Q) \end{bmatrix}$ (as $\mathrm{Col}_i(P) = \delta_2^2$). Hence, $\mathrm{Col}_i(Q) = \mathrm{Col}_i(L_1) + \mathrm{Col}_i(L_2)$. It follows that

$$Q = L_1 + L_2. \tag{15.31}$$

Since L satisfies (15.28), (15.31) ensures that Q satisfies (15.28) too.

Now since $\{z_i \,|\, i = 1, \cdots, k\}$ is a proper subbasis, so there exists $\{z_i \,|\, i = k+1, \cdots, n\}$ such that $\{z_i \,|\, i = 1, \cdots, n\}$ is a coordinate transformation of x, which satisfies (15.28). (Precisely, it satisfies (15.29) with row sum equal to 1.) According to the claim, the subset $\{z_i \,|\, i = 1, \cdots, k\}$ also satisfies (15.29). □

It is easy to see that (15.28) is equivalent to $|\{i \mid \mathrm{Col}_i(L) = \xi\}| = 2^{n-k}$ for any $\xi \in \Delta_{2^k}$.

Example 15.4. Consider the state space of three state variables as $\mathcal{X} = \mathcal{F}_\ell\{x_1, x_2, x_3\}$. Assume there is a subspace $\mathcal{Z} = \mathcal{F}_\ell\{z_1, z_2\} \subset \mathcal{X}$ with

$$\begin{cases} z_1 = x_1 \leftrightarrow x_2 \\ z_2 = x_2 \bar{\vee} x_3. \end{cases} \tag{15.32}$$

Then its algebraic form can be expressed as

$$z_1 z_2 = Mx = \delta_4[2\ 1\ 3\ 4\ 4\ 3\ 1\ 2]x. \tag{15.33}$$

We can see that $\left|\{i \mid \mathrm{Col}_i(M) = \delta_4^j\}\right| = 2$ for $j = 1, 2, 3, 4$, thus $\mathcal{Z}$ is a regular subspace.

Then we want to find z_3 such that $\{z_1, z_2, z_3\}$ form a coordinate frame, that is, to make the columns algebraic form T of the coordinate transformation. Using Khatri-Rao product, we need to construct $M_{z_3} = \delta_2[c_1\ c_2\ c_3\ c_4\ c_5\ c_6\ c_7\ c_8]$ such that

$$\mathrm{Col}_i(M)c_i \neq \mathrm{Col}_j(M)c_j, \quad i \neq j.$$

For this purpose, we only need make M_{z_3} satisfy

$$c_1 \neq c_8;\ c_2 \neq c_7;\ c_3 \neq c_6;\ c_4 \neq c_5.$$

For instance, we can choose

$$M_{z_3} = \delta_2[1\ 0\ 0\ 1\ 0\ 1\ 1\ 0].$$

Then we have

$$z_3 = [x_1 \wedge (x_2 \leftrightarrow x_3)] \vee [\neg x_1 \wedge (x_2 \bar{\vee} x_3)].$$

Next, we consider the verification of invariant subspaces.

Theorem 15.6. *Consider system (15.12) with its algebraic form (15.19). Assume that a regular subspace $\mathcal{Z} = \mathcal{F}_\ell\{z_1, \cdots, z_s\}$ with $z = \ltimes_{i=1}^s z_i$ has the following algebraic form*

$$z = Qx, \tag{15.34}$$

where $Q \in \mathcal{L}_{2^s \times 2^n}$. Then $\mathcal{Z} = \mathcal{F}_\ell\{z_1, \cdots, z_s\}$ is an invariant subspace of system (15.12)*, if and only if*

$$\mathrm{Row}(QL) \subset \mathrm{Span}_{\mathcal{B}}\, \mathrm{Row}(Q), \tag{15.35}$$

where $\mathrm{Span}_{\mathcal{B}}$ *means the coefficients are in $\mathcal{D}$; and L is as in* (15.19)*,* i.e.*, it is the transition matrix of the algebraic form of system* (15.12)*.*

We refer to Cheng *et al.* (2011b) for the proof of this theorem. Note that condition (15.35) is not straightforward verifiable. Since $\{\mathcal{Z}\}$ is a regular subspace, Q is of full row rank. Hence, we give the following equivalent statement.

Corollary 15.1. *$\mathcal{Z}$ is an invariant subspace, if and only if*

$$QL = QLQ^T(QQ^T)^{-1}Q. \tag{15.36}$$

Example 15.5. Consider the following Boolean network

$$\begin{cases} x_1(t+1) = (x_1(t) \wedge x_2(t) \wedge \neg x_4(t)) \vee (\neg x_1(t) \wedge x_2(t)) \\ x_2(t+1) = x_2(t) \vee (x_3(t) \leftrightarrow x_4(t)) \\ x_3(t+1) = (x_1(t) \wedge \neg x_4(t)) \vee (\neg x_1(t) \wedge x_2(t)) \\ \qquad\qquad\quad \vee(\neg x_1(t) \wedge \neg x_2(t) \wedge x_4(t)) \\ x_4(t+1) = x_1(t) \wedge \neg x_2(t) \wedge x_4(t). \end{cases} \tag{15.37}$$

Let $\mathcal{Z} = \mathcal{F}_\ell\{z_1, z_2, z_3\}$, where

$$\begin{cases} z_1 = x_1 \bar{\vee} x_4 \\ z_2 = \neg x_2 \\ z_3 = x_3 \leftrightarrow \neg x_4. \end{cases} \tag{15.38}$$

Set $x = \ltimes_{i=1}^4 x_i$, $z = \ltimes_{i=1}^3 z_i$. Then we have

$$z = Qx,$$

where

$$Q = \delta_8[8\ 3\ 7\ 4\ 6\ 1\ 5\ 2\ 4\ 7\ 3\ 8\ 2\ 5\ 1\ 6],$$

and the algebraic form of (15.37) is

$$x(t+1) = Lx(t),$$

where

$$L = \delta_{16}[11\ 1\ 11\ 1\ 11\ 13\ 15\ 9\ 1\ 2\ 1\ 2\ 9\ 15\ 13\ 11].$$

It is easy to calculate that

$$QL = \delta_8[3\ 8\ 3\ 8\ 3\ 2\ 1\ 4\ 8\ 3\ 8\ 3\ 4\ 1\ 2\ 3] = QLQ^T(QQ^T)^{-1}Q.$$

Hence $\mathcal{Z}$ is an invariant subspace of (15.37).

In fact we can choose $z_4 = x_4$ such that

$$\begin{cases} z_1 = x_1 \bar{\vee} x_4 \\ z_2 = \neg x_2 \\ z_3 = x_3 \leftrightarrow \neg x_4 \\ z_4 = x_4 \end{cases} \tag{15.39}$$

is a coordinate transformation. Moreover, under coordinate frame z, system (15.37) can be expressed into the cascading form (15.16) as

$$\begin{cases} z_1(t+1) = z_1(t) \to z_2(t) \\ z_2(t+1) = z_2(t) \wedge z_3(t) \\ z_3(t+1) = \neg z_1(t) \\ z_4(t+1) = z_1(t) \vee z_2(t) \vee z_4(t). \end{cases} \tag{15.40}$$

15.3.2 *Input-State Description*

Consider Boolean control network (15.13). In this chapter we assume the controls are logical variables satisfying certain logical rule, called the input network, described as

$$\begin{cases} u_1(t+1) = g_1(u_1(t), u_2(t), \cdots, u_m(t)) \\ u_2(t+1) = g_2(u_1(t), u_2(t), \cdots, u_m(t)) \\ \vdots \\ u_m(t+1) = g_m(u_1(t), u_2(t), \cdots, u_m(t)). \end{cases} \tag{15.41}$$

Setting $u(t) = \ltimes_{i=1}^m u_i(t)$, its algebraic form is

$$u(t+1) = Gu(t), \quad G \in \mathcal{L}_{2^m \times 2^m}.$$

Then, we consider the cycles of system (15.13) with (15.41) (the case of free inputs will be considered in next chapter). The state space is $\mathcal{X} = \mathcal{F}_\ell\{x_1, \cdots, x_n\}$, input space is $\mathcal{U} = \mathcal{F}_\ell\{u_1, \cdots, u_m\}$, input-state space is $\mathcal{W} = \mathcal{F}_\ell\{u_1, \cdots, u_m, x_1, \cdots, x_n\}$. It is obvious that $\mathcal{U}$ is the invariant subspace of $\{\mathcal{W}\}$, we want to investigate the relationship between the cycles in $\mathcal{U}$ and the cycles in $\mathcal{W}$.

Assume there is a cycle of length k in the input-state space $\mathcal{W}$. Say, it is

$$\begin{aligned} C_{\mathcal{W}}^k: \quad & w(0) = w_0 = u_0 x_0 \to w(1) = w_1 = u_1 x_1 \to \cdots \\ & \to w(k) = w_k = u_k x_k = w_0. \end{aligned}$$

First of all, one sees easily that since $u_0 = u_k$, in the input space $\mathcal{U}$, the sequence $\{u_0, u_1, \cdots, u_k\}$ contains, say, j folds of a cycle of length ℓ, where $j\ell = k$. Hence $u_\ell = u_0$. Now let us see what condition the $\{x_i\}$ in the cycle $C_{\mathcal{W}}^k$ should satisfy. Define a network transition matrix as

$$\Psi := L(u_{\ell-1})L(u_{\ell-2}) \cdots L(u_1)L(u_0). \tag{15.42}$$

Starting from $w_0 = u_0 x_0$, we have x component of the cycle $C^k_{\mathcal{W}}$ as

$$\begin{array}{l}
x_0 \to x_1 = L(u_0)x_0 \to x_2 = L(u_1)L(u_0)x_0 \to \cdots \to x_\ell = \Psi x_0 \to \\
x_{\ell+1} = L(u_0)\Psi x_0 \to x_{\ell+2} = L(u_1)L(u_0)\Psi x_0 \to \cdots \to x_{2\ell} = \Psi^2 x_0 \to \\
\vdots \\
x_{(j-1)\ell+1} = L(u_0)\Psi^{j-1}x_0 \to x_{(j-1)\ell+2} = L(u_1)L(u_0)\Psi^{j-1}x_0 \to \cdots \\
\to x_{j\ell} = \Psi^j x_0 = x_0.
\end{array} \tag{15.43}$$

We conclude that $x_0 \in \Delta_{2^n}$ is a fixed point of the equation

$$x(t+1) = \Psi^j x(t). \tag{15.44}$$

For convenience, we assume $j > 0$ is the smallest positive integer, which makes x_0 a fixed point of (15.44).

Conversely, assume $x_0 \in \Delta_{2^n}$ is a fixed point of (15.44) and u_0 is a point on a cycle of control space $C^\ell_{\mathcal{U}}$. Then it is obvious that we have the cycle (15.43).

Summarizing above arguments yields

Theorem 15.7. *Consider Boolean control network (15.13) with (15.41). A set $C^k_{\mathcal{W}} \subset \Delta_{2^{k(n+m)}}$ is a cycle in the input-state space $\mathcal{W}$ with length k, if and only if for any point $w_0 = u_0 x_0 \in C^k_{\mathcal{W}}$, there exists an $\ell \le k$ as a factor of k, such that $u_0, u_1 = G u_0, u_2 = G^2 u_0, \cdots, u_\ell = G^\ell u_0 = u_0$ is a cycle in the control space, and x_0 is a fixed point of equation (15.44) with $j = k/\ell$.*

Theorem 15.7 shows how to find all the cycles in the input-state space. First, we can find cycles in the input space. Pick a cycle in the input space, say $C^\ell_{\mathcal{U}}$, then for each point $u_0 \in C^\ell_{\mathcal{U}}$ we can construct an auxiliary system

$$x(t+1) = \Psi x(t). \tag{15.45}$$

Say, $C^\ell_{\mathcal{U}} = (u_0, u_1, \cdots, u_\ell = u_0)$ is a cycle in $\mathcal{U}$, and $C^j_{\mathcal{X}} = (x_0, x_1, \cdots, x_j = x_0)$ is a cycle of (15.45). Then there is a cycle $C^k_{\mathcal{W}}$, $k = \ell j$, in the input-state STP space, which can be constructed by

$$\begin{array}{rl}
w_0 = & u_0 x_0 \to w_1 = u_1 L(u_0)x_0 \to w_2 = u_2 L(u_1)L(u_0)x_0 \to \cdots \to \\
w_\ell = & u_0 x_1 \to w_{\ell+1} = u_1 L(u_0)x_1 \to w_{\ell+2} = u_2 L(u_1)L(u_0)x_1 \to \cdots \to \\
\vdots & \\
w_{(j-1)\ell} = & u_0 x_{(j-1)} \to w_{(j-1)\ell+1} = u_1 L(u_0)x_{(j-1)} \to \\
& w_{(j-1)\ell+2} = u_2 L(u_1)L(u_0)x_{(j-1)} \to \cdots \to \\
w_{j\ell} = & u_0 x_j = u_0 x_0 = w_0.
\end{array} \tag{15.46}$$

We call this $C^k_{\mathcal{W}}$ the composed cycle of $C^\ell_{\mathcal{U}}$ and $C^j_{\mathcal{X}}$, denoted by $C^k_{\mathcal{W}} = C^\ell_{\mathcal{U}} \circ C^j_{\mathcal{X}}$.

Note that from a cycle $C^\ell_{\mathcal{U}}$ we can choose any point as the starting point u_0. Then in equation (15.45) we have different Ψ, which produces different $C^j_{\mathcal{X}}$. It is reasonable to guess that the composed cycle $C^k_{\mathcal{W}} = C^\ell_{\mathcal{U}} \circ C^j_{\mathcal{X}}$ is independent of the choice of u_0. In fact, this is true.

Definition 15.5. Let $C^k_{\mathcal{W}} = \{w(t) | t = 0, 1, \cdots, k\}$ be a cycle in the input-state space, and $C^\ell_{\mathcal{U}}$ be a cycle in the input space. Splitting $w(t) = u(t)x(t)$, we say that $C^k_{\mathcal{W}}$ is attached to $C^\ell_{\mathcal{U}}$ at u_0, if $w(0) = u_0 x_0$, and

(1) $u(t) \in C^\ell_{\mathcal{U}}$, with $u(0) = u_0$;
(2) $x(0) = x_0$ is a fixed point of (15.44) with $j = \frac{k}{\ell} \in \mathbb{Z}_+$.

Proposition 15.2. *The sets of cycles in the input-state space, attached to any point of a given cycle $C^\ell_{\mathcal{U}}$ are the same.*

We refer to Cheng *et al.* (2011b) or Cheng (2009) for the proof of Proposition 15.2. Next, we give an illustrative example.

Example 15.6. Consider a control Boolean network

$$\begin{cases} x_1(t+1) = u(t) \to x_2(t) \\ x_2(t+1) = x_1(t) \vee x_3(t) \\ x_3(t+1) = \neg x_1(t), \end{cases}$$

the control network is

$$u(t+1) = \neg u(t).$$

We have an obvious kernel cycle: $0 \to 1 \to 0$ in $\mathcal{U}$. Then we can easily calculate that

$$\begin{aligned} L(0) &= \delta_8[2\ 2\ 2\ 2\ 1\ 3\ 1\ 3], \\ L(1) &= \delta_8[2\ 2\ 6\ 6\ 1\ 3\ 5\ 7]. \end{aligned}$$

Hence we consider an auxiliary system as

$$x(t+1) = \Psi x(t),$$

where

$$\Psi = L(1)L(0) = \delta_8[2\ 2\ 2\ 2\ 2\ 6\ 2\ 6].$$

A routine calculation shows: (1) nontrivial power of Ψ is 1 and $\mathrm{tr}(\Psi^1) = 2$. So there are two fixed points, which are $\delta_8^1 \sim (1, 1, 0)$ and $\delta_8^6 \sim (0, 1, 0)$. The overall composed cycles are depicted in Fig. 15.2, where the dash lines show the duplicated cycles. Overall, we have a cycle in the input space and two product cycles of length 2 in the input-state space.

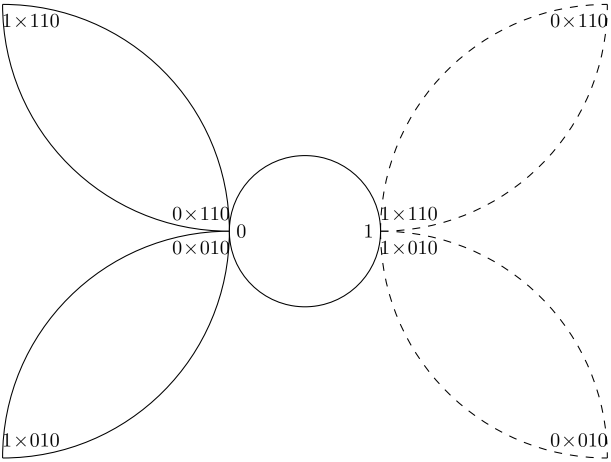


Fig. 15.2 Cycles of a control system

15.4 Higher-Order Boolean Networks

In this section we consider higher-order Boolean networks. This section is based on Li *et al.* (2011).

Definition 15.6. A Boolean network is called a μth-order network if the current states depend on μ-length histories. Precisely, its dynamics can be described as

$$\begin{cases} x_1(t+1) = f_1(x_1(t-\mu+1), \cdots, x_n(t-\mu+1), \cdots, x_1(t), \cdots, x_n(t)), \\ x_2(t+1) = f_2(x_1(t-\mu+1), \cdots, x_n(t-\mu+1), \cdots, x_1(t), \cdots, x_n(t)), \\ \vdots \\ x_n(t+1) = f_n(x_1(t-\mu+1), \cdots, x_n(t-\mu+1), \cdots, x_1(t), \cdots, x_n(t)), \\ \hfill t \geq \mu - 1, \end{cases} \tag{15.47}$$

where $f_i : \mathcal{D}^{\mu n} \to \mathcal{D}$, $i = 1, \cdots, n$ are logical functions.

Note that same as for higher-order discrete-time difference equations, to determine the solution (it is also called a trajectory) we need a set of initial conditions as

$$x_i(j) = a_{ij}, \quad i = 1, \cdots, n; \; j = 0, \cdots, \mu - 1. \tag{15.48}$$

We give an example to illustrate this kind of systems. It is a biochemical network of coupled oscillations in the cell cycle (Goodwin, 1963).

Example 15.7. Consider the following Boolean network

$$\begin{cases} A(t+3) = \neg(A(t) \wedge B(t+1)) \\ B(t+3) = \neg(A(t+1) \wedge B(t)). \end{cases} \tag{15.49}$$

It can be easily converted into the canonical form (15.47) as

$$\begin{cases} A(t+1) = \neg(A(t-2) \wedge B(t-1)) \\ B(t+1) = \neg(A(t-1) \wedge B(t-2)), \quad t \geq 2. \end{cases} \tag{15.50}$$

This is a 3rd-order Boolean network.

Now the first natural question is: Can we use the technique developed in the previous sections of this chapter to investigate the structure of higher-order Boolean networks? The answer is “Yes”. In the following we will discuss two algebraic forms of (15.47).

15.4.1 *First Algebraic Form of Higher-Order Boolean Networks*

Using a vector form, we define

$$\begin{cases} x(t) = \ltimes_{i=1}^{n} x_i(t) \in \Delta_{2^n} \\ z(t) = \ltimes_{i=t}^{t+\mu-1} x(i) \in \Delta_{2^{\mu n}}, \quad t = 0, 1, \cdots . \end{cases}$$

Assume the structure matrix of f_i is $M_i \in \mathcal{L}_{2 \times 2^{\mu n}}$. Then we can express (15.47) into its component-wise algebraic form as

$$x_i(t+1) = M_i z(t-\mu+1), \quad i = 1, \cdots, n;\ t = \mu-1, \mu, \mu+1, \cdots . \tag{15.51}$$

Multiplying the equations in (15.51) together yields

$$x(t+1) = L_0 z(t-\mu+1), \quad t \geq \mu, \tag{15.52}$$

where

$$L_0 = M_1 \ltimes_{j=2}^{n} \left[\left(I_{2^{\mu n}} \otimes M_j \right) M_{r,2^{\mu n}} \right]. \tag{15.53}$$

Note that the L_0 here can be calculated in a standard procedure as we did before. Using some properties of the semi-tensor product of matrices, we have

$$\begin{aligned} z(t+1) &= \ltimes_{i=t+1}^{t+\mu} x(i) \\ &= D_n \ltimes_{i=t}^{t+\mu-1} x(i) (L_0 \ltimes_{i=t}^{t+\mu-1} x(i)) \\ &= D_n (I_{2^{\mu n}} \otimes L_0) M_{r,2^{\mu n}} \ltimes_{i=t}^{t+\mu-1} x(i) \\ &:= L z(t), \end{aligned} \tag{15.54}$$

where the dummy matrix D_n is defined in (7.25).

(15.54) is called the first algebraic form of network (15.47).

In fact, we can prove that the two Boolean networks have the same topological structure, including fixed points, cycles, and transient time, which is for all points to enter the set of cycles. So the first order Boolean network (15.54) provides all such results for higher order Boolean network (15.52).

Definition 15.7. Consider the network (15.47).

(1) $(x_0, x_1, x_2, \cdots)$ is called a path of network (15.47), if $x_i = L_0 \ltimes_{j=i-\mu}^{i-1} x_j, i \geq \mu$.

(2) $(x_0, x_1, \cdots, x_k = x_0)$ is called a cycle of network (15.47) with length k, if $\{x_0, x_1, \cdots, x_{k-1}\}$ are distinct and $x_k = x_0$. A fixed point is a cycle of length 1.

Theorem 15.8. *There is a one-to-one correspondence between the cycles of (15.52) and the cycles of (15.54).*

We refer to Cheng *et al.* (2011b) or Li *et al.* (2011) for the proof.

Example 15.8. Recall Example 15.7. Set $x(t) = A(t)B(t)$. Using vector form, (15.50) can be expressed as

$$x(t+1) = L_0 x(t-2)x(t-1)x(t), \tag{15.55}$$

where

$$\begin{aligned} L_0 = \delta_4[&4\ 4\ 4\ 4\ 2\ 2\ 2\ 2\ 3\ 3\ 3\ 3\ 1\ 1\ 1\ 1\\ &3\ 3\ 3\ 3\ 1\ 1\ 1\ 1\ 3\ 3\ 3\ 3\ 1\ 1\ 1\ 1\\ &2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\\ &1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1]. \end{aligned}$$

Set $z(t) = x(t)x(t+1)x(t+2)$. Then

$$z(t+1) = Lz(t), \tag{15.56}$$

where

$$\begin{aligned} L = \delta_{2^6}[&4\ 8\ 12\ 16\ 18\ 22\ 26\ 30\ 35\ 39\ 43\ 47\ 49\ 53\ 57\ 61\\ &3\ 7\ 11\ 15\ 17\ 21\ 25\ 29\ 35\ 39\ 43\ 47\ 49\ 53\ 57\ 61\\ &2\ 6\ 10\ 14\ 18\ 22\ 26\ 30\ 33\ 37\ 41\ 45\ 49\ 53\ 57\ 61\\ &1\ 5\ 9\ \ 13\ 17\ 21\ 25\ 29\ 33\ 37\ 41\ 45\ 49\ 53\ 57\ 61]. \end{aligned}$$

To find the cycles of (15.55), it is enough to find all the cycles in system (15.56). We can check $\text{tr}(L^k), k = 1, 2, \cdots, 64$. They can be easily calculated as

$$\text{tr}(L^2) = 2, \quad \text{tr}(L^5) = 5, \quad \text{tr}(L^{10}) = 17.$$

Using Theorem 15.3, we conclude that the system does not have fixed point, but it has one cycle of length 2, one cycle of length 5 and one cycle of length 10, which are

$$\delta_{64}^{26} \to \delta_{64}^{29} \to \delta_{64}^{26},$$
$$\delta_{64}^{1} \to \delta_{64}^{4} \to \delta_{64}^{16} \to \delta_{64}^{61} \to \delta_{64}^{49} \to \delta_{64}^{1},$$
$$\delta_{64}^{2} \to \delta_{64}^{8} \to \delta_{64}^{30} \to \delta_{64}^{53} \to \delta_{64}^{17} \to \delta_{64}^{3} \to \delta_{64}^{12} \to \delta_{64}^{47} \to \delta_{64}^{5} \to \delta_{64}^{33} \to \delta_{64}^{2}.$$

Decompose $z(t)$ into $x(t)\cdots x(t+\mu-1)$, we have the cycles of (15.55) as

$$\delta_4^2 \to \delta_4^3 \to \delta_4^2,$$
$$\delta_4^1 \to \delta_4^1 \to \delta_4^1 \to \delta_4^4 \to \delta_4^4 \to \delta_4^1,$$
$$\delta_4^1 \to \delta_4^1 \to \delta_4^2 \to \delta_4^4 \to \delta_4^2 \to \delta_4^1 \to \delta_4^1 \to \delta_4^3 \to \delta_4^4 \to \delta_4^3 \to \delta_4^1.$$

The result coincides with the one in Heidel *et al.* (2003).

15.4.2 *Second Algebraic Form of Higher-Order Boolean Networks*

Define

$$w(\tau) := x(\mu\tau)x(\mu\tau+1)\cdots x(\mu\tau+(\mu-1)) = z(\mu\tau). \tag{15.57}$$

Then we have

$$w(\tau+1) = z(\mu\tau+\mu) = L^{\mu}z(\mu\tau) = L^{\mu}w(\tau),$$

where L is obtained in (15.54). Therefore, we have

$$w(\tau+1) = \Gamma w(\tau), \tag{15.58}$$

where

$$\Gamma = \left[D_n(I_{2^{\mu n}} \otimes L_0)M_{r,2^{\mu n}}\right]^{\mu},$$

with initial value $w(0) = \ltimes_{i=0}^{\mu-1} x(i)$. We call (15.58) the second algebraic form of the μth-order Boolean network (15.47).

In fact, by re-scheduling the sampling time, the second algebraic form provides state variable, $w(\tau)$, $\tau = 0, 1, \cdots$, as a set of non-overlapped segments of $x(t)$. Hence, there is an obvious one-to-one correspondence between the trajectories of (15.47) and the trajectories of (15.58).

Proposition 15.3. *There is a one-to-one correspondence between the trajectories of (15.47) and the trajectories of its second algebraic form (15.58), by*

$$w(\tau) := x(\mu\tau)x(\mu\tau+1)\cdots x(\mu\tau+(\mu-1)), \quad \tau = 0, 1, \cdots .$$

But there is no one-to-one correspondence between the cycles of (15.47) and the cycles of (15.58). It is easy to give an counterexample, we left this for exercise.

15.5 Dynamic-Static Boolean Networks

Consider a Boolean network of n nodes. Assume there are $n-k$ nodes, which satisfy Boolean dynamic models as

$$x_i(t+1) = f_i(x_1, \cdots, x_n), \quad i = 1, \cdots, n-k, \tag{15.59}$$

and the other k nodes are determined by certain static equations as

$$g_j(x_1, \cdots, x_n) = 1, \quad j = 1, \cdots, k. \tag{15.60}$$

Note that the right hand side of (15.60) can be either 0 or 1. Without loss of generality we can set them to be 1, because for $g_j = 0$ we can use $\neg g_j$ to replace g_j.

In vector form set $x^1 = \ltimes_{i=1}^{n-k} x_i$ and $x^2 = \ltimes_{i=n-k+1}^{n} x_i$. The system (15.59)-(15.60) is said to have a normal form, if (15.60) can be expressed as

$$x_j = \phi_j(x_1, \cdots, x_{n-k}), \quad j = n-k+1, \cdots, n. \tag{15.61}$$

It is obvious that (15.61) is very convenient in use, because we can plug (15.61) into (15.59) to get a standard Boolean network, and its properties can be analyzed easily. Hence the problem "when (15.60) can be converted to (15.61)" is interesting. This section is devoted to solving this problem.

(15.60) can be expressed in vector form as

$$M_G x^1 x^2 = \delta_{2^k}^1, \tag{15.62}$$

where $M_G \in \mathcal{L}_{2^k \times 2^n}$.

For any positive integer $s > 1$ define a set of matrices, Ξ_i, as

$$\Xi_i = \left\{ E_i \in \mathcal{L}_{s\times s} \middle| \operatorname{Col}_i(E_i) = \delta_s^1;\ \operatorname{Col}_j(E_i) \neq \delta_s^1,\ j \neq i \right\}, \quad i = 1, 2, \cdots, s. \tag{15.63}$$

Using Ξ_i, we construct a set of types as

$$\mathcal{E}_s := \left\{ [E_1\ E_2\ \cdots\ E_s] \mid E_i \in \Xi_i,\ i = 1, 2, \cdots, s \right\}. \tag{15.64}$$

As discussed in Chapter 12, each type $E \in \mathcal{E}_s$ corresponds to a unique logical mapping $F : \mathcal{D}_s \times \mathcal{D}_s \to \mathcal{D}_s$, which has E as its structure matrix, that is, $M_f = E$.

Then we have the following result:

Lemma 15.1. *Let $X, Y \in \Delta_s$. $X = Y$, if and only if there exists a $E \in \mathcal{E}_s$ such that*

$$EXY = \delta_s^1. \tag{15.65}$$

Proof. Denote

$$E = [E_1 \ E_2 \ \cdots \ E_s]\,,$$

and assume $X = \delta_s^p$ and $Y = \delta_s^q$. A straightforward computation shows that

$$EXY = \mathrm{Col}_q(E_p).$$

Hence (15.65) holds, if and only if $p = q$. □

Now we are ready to present the main result for normalization. We first express (15.60) into its algebraic form as

$$M_G x = \delta_{2^k}^1, \tag{15.66}$$

where M_G is the structure matrix of $G = (g_1, \cdots, g_k) : \mathcal{D}^n \to \mathcal{D}^k$.

The following theorem may be called the theorem of logical implicit function.

Theorem 15.9. *(Theorem of Logical Implicit Function) x_j can be solved as (15.61) from (15.60), if and only if there exists a*

$$E = [E_1 \ E_2 \ \cdots \ E_{2^k}] \in \mathcal{E}_{2^k},$$

such that the structure matrix of G can be expressed as

$$M_G = [M_1 \ M_2, \cdots, M_{2^{n-k}}], \tag{15.67}$$

and

$$M_i \in \{E_1 \ E_2 \ \cdots \ E_{2^k}\}\,, \quad i = 1, \cdots, 2^{n-k}.$$

Proof. (15.61) can be expressed into vector form as $x^2 = M_\phi x^1$. According to Lemma 15.1, (15.60) can be expressed into (15.61), if and only if there exists an $M_F \in \mathcal{E}_{2^k}$ such that $M_F M_\phi x^1 x^2 = \delta_{2^k}^1$. Comparing with (15.66), it is clear that the necessary and sufficient condition becomes that M_G can be expressed as

$$M_G x = M_F M_\phi x^1 x^2, \tag{15.68}$$

where $M_F \in \mathcal{E}_{2^k}$. Now formally consider $x^2 = M_\psi x^2$ with $M_\psi = I_{2^k}$, the conclusion follows from Theorem 12.6 immediately. □

Example 15.9. Consider the follow dynamic-static Boolean network

$$\begin{cases} x_1(t+1) = x_2(t) \to x_4(t) \\ x_2(t+1) = x_1(t) \wedge x_3(t) \\ 1 = (x_3(t) \bar{\vee} x_4(t)) \leftrightarrow (x_1(t) \bar{\vee} x_2(t)) \\ 0 = x_4(t) \bar{\vee} (x_1(t) \vee x_2(t). \end{cases} \tag{15.69}$$

We intend to solve x_3 and x_4 out from the last two equations. First, we convert them to

$$\begin{cases} g_1(x_1,x_2,x_3,x_4) := (x_3(t)\bar{\vee}x_4(t)) \leftrightarrow (x_1(t)\bar{\vee}x_2(t)) = 1 \\ g_2(x_1,x_2,x_3,x_4) := x_4(t) \leftrightarrow (x_1(t) \vee x_2(t)) = 1. \end{cases} \tag{15.70}$$

It is easy to calculate that in vector form we have

$$\begin{cases} g_1(x_1,x_2,x_3,x_4) = M_{g_1}x = \delta_2[1\ 2\ 2\ 1\ 2\ 1\ 1\ 2\ 2\ 1\ 1\ 2\ 1\ 2\ 2\ 1]x \\ g_2(x_1,x_2,x_3,x_4) = M_{g_2}x = \delta_2[1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 2\ 1\ 2\ 1]x. \end{cases} \tag{15.71}$$

Then the structure matrix of $G = (g_1, g_2)$ can be easily calculated as

$$M_G = \delta_4[1\ 4\ 3\ 2\quad 3\ 2\ 1\ 4\quad 3\ 2\ 1\ 4\quad 2\ 3\ 4\ 1]. \tag{15.72}$$

Now we can construct the structure matrix $M_F \in \mathcal{E}_4$ as

$$M_F = \delta_4[1\ 4\ 3\ 2\quad *\ 1\ *\ *\quad 3\ 2\ 1\ 4\quad 2\ 3\ 4\ 1], \tag{15.73}$$

where $2 \leq * \leq 4$ can be arbitrary. Comparing (15.72) with (15.73) yields that

$$M_\phi = \delta_4[1\ 3\ 3\ 4], \tag{15.74}$$

which means

$$x_3(t)x_4(t) = \delta_4[1\ 3\ 3\ 4]x_1(t)x_2(t).$$

It follows that $x_3(t)$ and $x_4(t)$ can be solved from (15.71) uniquely as

$$\begin{cases} x_3(t) = x_1(t) \wedge x_2(t) \\ x_4(t) = x_1(t) \vee x_2(t). \end{cases} \tag{15.75}$$

Plugging (15.75) into (15.69) yields

$$\begin{cases} x_1(t+1) = x_2(t) \rightarrow (x_1(t) \vee x_2(t)) \\ x_2(t+1) = x_1(t) \wedge x_2(t). \end{cases} \tag{15.76}$$

Then the dynamics of the dynamic-static Boolean network (15.69) is determined by (15.76) (with algebraic equation (15.75) for the other two state variables).

Remark 15.5. The method provided above can also be used for dynamic-static Boolean control networks. You have only to replace x^1 by $\{x^1, u\}$ and then use exactly the technique developed above.

Exercises

15.1 Given a Boolean network

$$\begin{cases} x_1(t+1) = x_1 \vee x_2 \\ x_2(t+1) = \neg x_3 \\ x_3(t+1) = x_4 \leftrightarrow x_5 \\ x_4(t+1) = x_5 \wedge x_1 \\ x_5(t+1) = x_2 \bar{\vee} x_4. \end{cases}$$

(i) Draw its network graph.

(ii) Give its adjacent matrix from its graph.

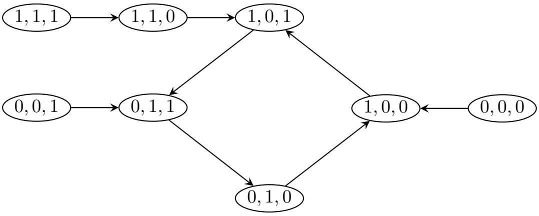


Fig. 15.3 State transition graph

15.2 The state transition graph of a Boolean network with 3 nodes is depicted in Fig. 15.3.

(i) Calculate the algebraic form of the network dynamics.

(ii) Calculate the logical dynamic equation of the network.

15.3 Give the algebraic form of the following Boolean networks:

(i)

$$\begin{cases} x_1(t+1) = x_1 \vee x_2 \\ x_2(t+1) = x_1 \wedge x_2 \\ x_3(t+1) = x_2 \leftrightarrow x_3. \end{cases} \tag{15.77}$$

(ii)

$$\begin{cases} x_1(t+1) = x_2 \rightarrow x_3 \\ x_2(t+1) = x_3 \wedge x_1 \\ x_3(t+1) = x_1 \bar{\vee} x_3. \end{cases} \tag{15.78}$$

15.4 Find all the fixed points and cycles of systems (15.77) and (15.78).

15.5 Find all the cycles on the subspace $\mathcal{F}_\ell\{x_1, x_2\}$ of (15.77). Reveal the "rolling gear" structure of the cycles of (15.77).

15.6 Proof Theorem 15.3.

15.7 Consider a state space $\mathcal{X} = \mathcal{F}_\ell\{x_1, x_2\}$. Let $z = x_1 \wedge x_2 \in \mathcal{X}$. Show that it is impossible to find $w \in \mathcal{X}$ such that $\mathcal{X} = \mathcal{F}_\ell\{z, w\}$. (Hint: This example asserts 3 of Remark 15.3.)

15.8 Consider the following Boolean control network

$$x(t+1) = (u_1(t) \vee u_2(t)) \wedge x(t),$$

where inputs $u_1(t), u_2(t)$ satisfy

$$\begin{cases} u_1(t+1) = u_1(t) \leftrightarrow u_2(t) \\ u_2(t+1) = \neg u_1(t). \end{cases}$$

Find out the cycles of this system.

15.9 Consider the following second order Boolean network

$$\begin{cases} A(t+1) = C(t-1) \vee (A(t) \wedge B(t)) \\ B(t+1) = \neg(C(t-1) \wedge A(t)) \\ C(t+1) = B(t-1) \wedge B(t). \end{cases}$$

Find out its cycles, and the cycles of its second algebraic form. Check whether they have one-to-one correspondence.

15.10 Given a state space $\mathcal{X} = \mathcal{F}_\ell\{x_1, x_2\}$ and $z = x_1 \leftrightarrow x_2$. Can you find a $w \in \mathcal{X}$ such that $(x_1, x_2) \to (z, w)$ is a coordinate transformation?

15.11 Consider a state space $\mathcal{X} = \mathcal{F}_\ell\{x_1, x_2, x_3, x_4\}$.

(i) Let $z = x_1 \wedge (x_2 \bar{\vee} x_3)$. Can we find a two-dimensional subspace $\mathcal{Z} \subset \mathcal{X}$ such that $z \in \mathcal{Z}$. If "yes", find it.

(ii) Let $z = (x_1 \bar{\vee} x_2) \leftrightarrow (x_3 \bar{\vee} x_4)$. Can we find a two-dimensional subspace $\mathcal{Z} \subset \mathcal{X}$ such that $z \in \mathcal{Z}$. If "yes", find it.

15.12 Consider a Boolean network

$$\begin{cases} x_1(t+1) = (x_3(t) \bar{\vee} x_4(t)) \wedge x_2(t) \\ x_2(t+1) = x_1(t) \vee x_4(t) \\ x_3(t+1) = x_1(t) \leftrightarrow x_2(t) \\ x_4(t+1) = x_2(t) \to (\neg x_4(t)), \end{cases} \tag{15.79}$$

and a mapping $\pi : x \to z$ as

$$\begin{cases} z_1 = x_1 \leftrightarrow x_2 \\ z_2 = \neg x_2 \\ z_3 = x_3 \bar{\vee} x_4 \\ z_4 = x_4. \end{cases}$$

(i) Prove that π is a coordinate transformation.

(ii) Express (15.79) under the coordinate frame z.

15.13 Consider a state space $\mathcal{X} = \mathcal{F}_\ell\{x_1, x_2, x_3\}$. $\mathcal{Z} = \mathcal{F}_\ell\{z_1, z_2\} \subset \mathcal{X}$. Is $\mathcal{Z}$ a regular subspace if

(i)

$$\begin{cases} z_1 = x_1 \wedge x_2 \\ z_2 = (x_1 \vee x_2) \to x_3. \end{cases}$$

(ii)

$$\begin{cases} z_1 = \neg x_2 \\ z_2 = x_1 \bar{\vee} x_3. \end{cases}$$

(iii)

$$\begin{cases} z_1 = \neg x_2 \leftrightarrow x_3 \\ z_2 = (x_1 \leftrightarrow x_2) \bar{\vee} (x_2 \leftrightarrow x_3). \end{cases}$$

15.14 A Boolean network has its algebraic form as

$$x(t+1) = Lx(t),$$

where

$$\begin{aligned} L = \delta_{32}[&19\ 21\ \ 5\ \ 5\ \ 6\ \ 6\ 20\ 22\ 23\ 17\ 21\ 21\ 22\ 22\ 24\ 18 \\ &15\ \ 9\ 13\ 13\ 14\ 14\ 16\ 10\ 11\ 13\ 29\ 29\ 30\ 30\ 12\ 14]. \end{aligned}$$

A subspace $\mathcal{G} \subset \mathcal{X}$ is determined by its algebraic form

$$\mathcal{G} = Gx,$$

where

$$\begin{aligned} G = \delta_4[&3\ 3\ 1\ 1\ 1\ 1\ 3\ 3\ 4\ 4\ 3\ 3\ 3\ 3\ 4\ 4 \\ &4\ 4\ 3\ 3\ 3\ 3\ 4\ 4\ 3\ 3\ 1\ 1\ 1\ 1\ 3\ 3]. \end{aligned}$$

Prove that $\mathcal{G}$ is an invariant subspace of L.

15.15 Consider the following input-state Boolean network.

$$\begin{cases} u_1(t+1) = u_1(t) \wedge u_2(t) \\ u_2(t+1) = \neg u_1(t) \\ x_1(t+1) = x_2(t) \vee u_1(t)) \\ x_2(t+1) = x_1(t) \leftrightarrow (x_2(t) \bar{\vee} u_2(t)). \end{cases}$$

(i) Find all the cycles of the system.

(ii) Express each cycle into the form of the product of cycle of u and cycle of x.

15.16 Consider the following Boolean network

$$\begin{cases} x_1(t+1) = [x_1(t) \wedge (x_2(t) \vee x_3(t))] \vee [\neg(x_1(t) \vee x_2(t))] \\ x_2(t+1) = \neg(x_1(t) \bar{\vee} x_3(t)) \\ x_3(t+1) = [(x_1(t) \bar{\vee} x_3(t)) \vee x_2(t)] \wedge x_3(t). \end{cases}$$

Try to find out its invariant subspace and the corresponding coordinate transformation.

15.17 Consider a Boolean network, which has its algebraic form as $x(t+1) = Lx(t)$, where $L \in \mathcal{L}_{2^n \times 2^n}$. A one-dimensional subspace $\xi \in \mathcal{X}$ is called the eigenvector of the system (briefly, of L).

Consider the following system

$$\begin{cases} x_1(t+1) = (x_1(t) \wedge x_2(t)) \vee (\neg x_1(t) \wedge \neg x_2(t) \wedge x_3(t)) \\ x_2(t+1) = [x_1(t) \wedge (x_2(t) \leftrightarrow x_3(t))] \vee \neg x_1(t) \\ x_3(t+1) = x_1(t) \wedge [\neg(x_2(t) \leftrightarrow x_3(t))]. \end{cases} \tag{15.80}$$

Check which ξ spans its eigenvector, if ξ has the following structure matrix: (i) $M_\xi = \delta_2[1\ 0\ 0\ 1\ 1\ 0\ 0\ 1]$; (ii) $M_\xi = \delta_2[1\ 1\ 0\ 0\ 0\ 0\ 1\ 1]$; (iii) $M_\xi = \delta_2[0\ 1\ 1\ 0\ 0\ 1\ 1\ 0]$.

15.18 Consider a Boolean network

$$\begin{cases} x_1(t+2) = x_2(t) \\ x_2(t+1) = x_1(t) \vee x_3(t) \\ x_3(t+2) = x_1(t) \wedge x_3(t+1). \end{cases} \tag{15.81}$$

(i) Convert it into a normal form of (higher order) Boolean network. What is its order?

(ii) Find its First algebraic form.

(iii) Find its Second algebraic form.

15.19 Consider the higher order Boolean network (15.81).

(i) Find its all fixed points.

(ii) Find its all cycles.

(iii) Find the basins of all attractors.

15.20 A dynamic-static Boolean control network satisfies dynamic equation

$$\begin{cases} x_1(t+1) = x_2(t) \wedge x_4(t) \\ x_2(t+1) = (x_1(t) \leftrightarrow x_3(t)) \vee u(t), \end{cases} \tag{15.82}$$

and static equation

$$\begin{cases} (x_1(t) \leftrightarrow u(t)) \wedge x_2(t) \wedge x_4(t) = 0 \\ x_2(t) \vee \neg x_3(t) \vee u(t) = 1. \end{cases} \tag{15.83}$$

(i) Solve $x_3(t)$ and $x_4(t)$ from (15.83).

(ii) Plugging the obtained $x_3(t)$ and $x_4(t)$ into (15.82) to get a reduced Boolean control network with state variables $\{x_1(t), x_2(t)\}$ only.

(iii) Assume the control network satisfies

$$u(t+1) = \neg u(t). \tag{15.84}$$

Find the fixed points and cycles of the network consisting of (15.82), (15.83), and (15.84).

Chapter 16

Boolean Control System

This chapter considers Boolean control networks. Using the algebraic form of logical control systems obtained by the STP and the matrix expression of logic, some fundamental control problems, including controllability and observability, disturbance decoupling, optimal control, etc., are investigated. We refer to Cheng *et al.* (2011b) for a systematic discussion on Boolean network, which was discussed in Chapter 15, and Boolean control network, which is discussed in this chapter.

16.1 Dynamics of Boolean Control Networks

It was pointed out in Ideker *et al.* (2001) that "Gene regulatory networks are defined by trans and cis logic. $\cdots$ Both of these types of regulatory networks have input and output." Hence, the investigation of control problem is essential in the study of cellular network.

A Boolean control network is defined as

$$\begin{cases} x_1(t+1) = f_1(x_1(t), x_2(t), \cdots, x_n(t), u_1(t), \cdots, u_m(t)) \\ x_2(t+1) = f_2(x_1(t), x_2(t), \cdots, x_n(t), u_1(t), \cdots, u_m(t)) \\ \vdots \\ x_n(t+1) = f_n(x_1(t), x_2(t), \cdots, x_n(t), u_1(t), \cdots, u_m(t)), \end{cases} \tag{16.1}$$

and

$$y_j(t) = h_j(x_1(t), x_2(t), \cdots, x_n(t)), \quad j = 1, 2, \cdots, p, \tag{16.2}$$

where $f_i : \mathcal{D}^{n+m} \to \mathcal{D}$, $i = 1, 2, \cdots, n$, and $h_j : \mathcal{D}^n \to \mathcal{D}$, $j = 1, 2, \cdots, p$ are logical functions; $x_i \in \mathcal{D}$, $i = 1, 2, \cdots, n$ are states; $y_j \in \mathcal{D}$, $j = 1, 2, \cdots, p$ are outputs; and $u_\ell \in \mathcal{D}$, $\ell = 1, 2, \cdots, m$ are inputs (or controls).

Denote by $x = \ltimes_{i=1}^{n} x_i$, $u = \ltimes_{i=1}^{m} u_i$, and $y = \ltimes_{i=1}^{p} y_i$. Using vector form, (16.1) and (16.2) can be expressed into their algebraic forms as

$$\begin{cases} x(t+1) = Lu(t)x(t) \\ y(t) = Hx(t). \end{cases} \tag{16.3}$$

Chapter 14 have studied the topological structure of Boolean network and Boolean control network with given input network. In this chapter, we will firstly consider the topological structure of Boolean control network. We first give a rigorous definition for the fixed points and cycles of a Boolean control network.

Definition 16.1. Consider system (16.1). Denote the input-state (product) space by

$$\mathcal{S} = \left\{ (U, X) \,|\, U = (u_1, \cdots, u_m) \in \mathcal{D}^p,\ X = (x_1, \cdots, x_n) \in \mathcal{D}^n \right\}.$$

(It is clear that $|\mathcal{S}| = 2^{m+n}$.)

(1) Let $S_i = (U^i, X^i) \in \mathcal{S}$ and $S_j = (U^j, X^j) \in \mathcal{S}$. Denote by $U^i = (u_1^i, \cdots, u_m^i)$, $X^i = (x_1^i, \cdots, x_n^i)$, etc. (S_i, S_j), is said to be a directed edge, if X^i, U^i, X^j satisfy (16.3). Precisely,

$$x_k^j = f_k(x_1^i, \cdots, x_n^i, u_1^i, \cdots, u_m^i), \quad k = 1, \cdots, n.$$

The set of edges is denoted by $\mathcal{E} \subset \mathcal{S} \times \mathcal{S}$.

(2) The pair $(\mathcal{S}, \mathcal{E})$ forms a directed graph, which is called the input-state transfer graph (ISTG).

(3) $(S_1, S_2, \cdots, S_\ell)$ is called a path, if $(S_i, S_{i+1}) \in \mathcal{E}$, $i = 1, 2, \cdots, \ell - 1$.

(4) A path $(S_1, S_2, \cdots)$ is called a cycle, if $S_{i+\ell} = S_i$ for all i, the smallest $\ell \geq 1$ is called the length of the cycle. Particular, the cycle of length 1 is called a fixed point.

(5) A cycle $(S_1, S_2, \cdots, S_\ell)$ in which $S_i = (U^i, X^i)$ is called a simple cycle, if $X^i \neq X^j$ for $1 \leq i < j \leq \ell$.

Definition 16.2. Denote the vertices of the ISTG of system (16.1) by $\{\delta_{2^{m+n}}^i | i = 1, \cdots, 2^{m+n}\}$. Then the input-state incidence matrix of the Boolean control network (16.1) is defined by

$$\mathcal{J}_{ij} = \begin{cases} 1, & \text{there exists an edge from } \delta_{2^{m+n}}^j \text{ to } \delta_{2^{m+n}}^i, \\ 0, & \text{otherwise.} \end{cases} \tag{16.4}$$

We give a simple example to describe the input-state transfer graph.

Example 16.1. Consider a Boolean control network Σ as

$$\Sigma : \begin{cases} x_1(t+1) = (x_1(t) \vee x_2(t)) \wedge u(t) \\ x_2(t+1) = x_1(t) \leftrightarrow u(t). \end{cases} \tag{16.5}$$

Setting $x(t) = x_1(t) \ltimes x_2(t)$, it is easy to calculate that the algebraic form of Σ is

$$\Sigma : x(t+1) = Lu(t)x(t),$$

where

$$L = \delta_4 \left[1\ 1\ 2\ 4\ 4\ 4\ 3\ 3 \right]. \tag{16.6}$$

According to the dynamic equation (16.5) (equivalently, (16.6)), we can draw the flow of $(u(t), (x_1(t), x_2(t)))$ on the product space $\mathcal{U} \times \mathcal{X}$, called the input-state transfer graph, as in Fig. 16.1.

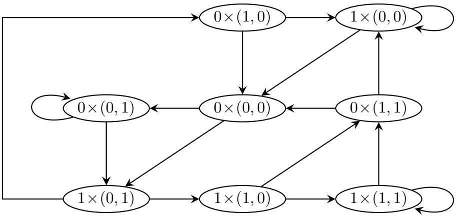


Fig. 16.1 Input-state transfer graph

Then we can find the input-state incidence matrix of (16.5) as

$$\mathcal{J} = \begin{bmatrix} 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1 \\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 0 \\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1 \\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 0 \end{bmatrix}. \tag{16.7}$$

Comparing (16.6) with (16.7), one might be surprised to find that

$$\mathcal{J} = \begin{bmatrix} L \\ L \end{bmatrix}.$$

In fact, this is also true for general case. Consider equation (16.3). Note that, the jth column of L corresponds to the "output" $x(t+1)$ for "input" $u(t)x(t) = \delta_{2^{m+n}}^j$ of the dynamic system. If this column $\mathrm{Col}_j(L) = \delta_{2^n}^i$, then it means that the output $x(t+1)$ is exactly the ith element of Δ_{2^n}. Now since $u(t+1)$ can be arbitrary, it follows that the input-state incidence matrix of system (16.3) is

$$\mathcal{J} = \left.\begin{bmatrix} L \\ L \\ \vdots \\ L \end{bmatrix}\right\} 2^m \in \mathcal{B}_{2^{m+n}\times 2^{m+n}}, \tag{16.8}$$

where the first block corresponds to $u(t+1) = \delta_{2^m}^1$, the second block corresponds to $u(t+1) = \delta_{2^m}^2$, and so on.

We say an $m \times m$ matrix A is row-periodic, if it can be expressed as $A = \mathbf{1}_\tau A_0$, where $A_0 \in \mathcal{M}_{n\times m}$ and $m = \tau n$, then the smallest τ is called the period of A, A_0 is call the basic block of A. By a straightforward computation, it can be verified that if A is row-periodic with period τ, then so is A^s, where s is a positive integer.

The following proposition can help calculating powers of the incidence matrix $\mathcal{J}$.

Proposition 16.1.

$$\mathcal{J}_0^{s+1} = M^s L, \tag{16.9}$$

where

$$M = \sum_{i=1}^{2^m} \mathrm{Blk}_i(L).$$

Proof.

$$\begin{aligned} \mathcal{J}_0^{s+1} &= (\delta_{2^m}^1)^T \mathcal{J}^{s+1} \\ &= \left((\delta_{2^m}^1)^T \mathcal{J}\right) \mathcal{J}^s \\ &= L\left(\mathbf{1}_{2^m} \otimes \mathcal{J}_0^s\right) \\ &= \sum_{i=1}^{2^m} \mathrm{Blk}_i(L)\mathcal{J}_0^s. \end{aligned}$$

□

We consider the physical meaning of $\mathcal{J}^s$. When $s = 1$ we know that $\mathcal{J}_{ij}$ means whether there exists a set of controls such that $\delta^i_{2^{m+n}}$ is reachable from $\delta^j_{2^{m+n}}$ in one step by judging if $\mathcal{J}_{ij} = 1$ or not. Similar argument shows that when $s > 1$ the physical meaning of $\mathcal{J}^s$ is as follows:

Theorem 16.1. *Consider system (16.3). Assume that the (i, j)th element of the sth power of its input-state incidence matrix, $\mathcal{J}^s_{ij} = c$. Then there are c paths from point $P_i = \delta^i_{2^{m+n}}$ reach $P_j = \delta^j_{2^{m+n}}$ at sth step with proper controls.*

Then, similar to Theorems 15.2 and 15.3, we can get the following result about the topological structure of Boolean control networks.

Theorem 16.2. *Consider the state equation of system (16.1) with its input-state incidence matrix $\mathcal{J}$. The number of the fixed points in the input-state dynamic graph is*

$$N_1 = \mathrm{tr}(M). \tag{16.10}$$

The number of length s cycles can be calculated inductively as

$$N_s = \frac{\mathrm{tr}(M^s) - \sum\limits_{k \in \mathcal{P}(s)} kN_k}{s}, \quad s > 1. \tag{16.11}$$

We use an example to depict it.

Example 16.2. Recall Example 16.1. Since

$$M = \mathrm{Blk}_1(L) + \mathrm{Blk}_2(L) = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}.$$

We can calculate that

$$\begin{aligned} &\mathrm{tr}\, M = 3, && \mathrm{tr}\, M^3 = 6, \\ &\mathrm{tr}\, M^4 = 15, && \mathrm{tr}\, M^5 = 33, \\ &\mathrm{tr}\, M^6 = 66, && \mathrm{tr}\, M^7 = 129, \\ &\mathrm{tr}\, M^8 = 255. \end{aligned}$$

Using Theorem 16.2, we conclude that $N_1 = 3$, $N_3 = 1$, $N_4 = 3$, $N_5 = 6$, $N_6 = 10$, $N_7 = 18$, $N_8 = 30$. It is not an easy job to count them from the graph directly.

16.2 Controllability

Controllability is a fundamental topic in modern control theory. Limited by the mathematical tools, there are few known results on control design of Boolean control networks (Akutsu *et al.*, 2007; Datta *et al.*, 2003, 2004). However, the semi-tensor product makes it much easier to investigate the controllability of Boolean control networks. First, we give the definition of controllability.

Definition 16.3. Consider system (16.1). Denote its state space as $\mathcal{X} = \mathcal{D}^n$, and let $X_0 \in \mathcal{X}$.

(1) $X \in \mathcal{X}$ is said to be reachable from X_0 at time $s > 0$, if we can find a sequence of controls $U(0) = \{u_1(0), \cdots, u_m(0)\}$, $U(1) = \{u_1(1), \cdots, u_m(1)\}$, $\cdots$, such that the trajectory of (16.3) with the initial value X_0 and the controls $\{U(t)\}$, $t = 0, 1, \cdots$ will reach X at time $t = s$. The reachable set at time s is denoted by $R_s(X_0)$. The overall reachable set is denoted by

$$R(X_0) = \cup_{s=1}^{\infty} R_s(X_0).$$

(2) System (16.1) is said to be controllable at X_0 if $R(X_0) = \mathcal{X}$. The system is said to be controllable if it is controllable at every $X \in \mathcal{X}$.

Before further studying controllability, we review the Boolean product and power of Boolean matrix, which were discussed in Chapter 8.

Definition 16.4.

(1) Let $A_k = (a_{ij}^k) \in \mathcal{B}_{m\times n}$, $k = 1, \cdots, r$, σ is an r-ary logical operator, then

$$\sigma(A_1, \cdots, A_r) := (\sigma(a_{ij}^1, \cdots, a_{ij}^r)).$$

(2) Let $A \in \mathcal{B}_{m\times n}$, $b \in \mathcal{D}$, the scalar product is define as

$$bA = Ab := b \wedge A.$$

Particularly, if $A = a \in \mathcal{D}$, $ab = ba = a \wedge b$.

(3) Let $A = (a_{ij}), B = (b_{ij}) \in \mathcal{B}_{m\times n}$. Then we define Boolean addition as

$$A(+)B := (a_{ij}(+)b_{ij}) := (a_{ij} \vee b_{ij}).$$

(4) Let $A \in \mathcal{B}_{m\times n}$ and $B \in \mathcal{B}_{n\times p}$. Then the Boolean product is defined as

$$A(\times)B := C \in \mathcal{B}_{m\times p},$$

where

$$c_{ij} = (+)_{k=1}^{n} a_{ik}b_{kj}.$$

If $A \prec_t B$ $(A \succ_t B)$, the Boolean semi-tensor product is defined as

$$A(\ltimes)B := (A \otimes I_t)(\times)B. \quad (A(\ltimes)B := A(\times)(B \otimes I_t).)$$

Particularly, if $A \prec_t A$ $(A \succ_t A)$,

$$A^{(k)} := \underbrace{A(\ltimes)A(\ltimes)\cdots(\ltimes)A}_{k}.$$

Remark 16.1. Boolean addition could be defined in different ways. For instance, in Chapter 11 we define

$$a\,\langle+\rangle\,b := a\bar{\vee}b = a + b \ (\mathrm{mod}\ 2). \tag{16.12}$$

Then we can define matrix operators $A\,\langle+\rangle\,B$, $A\,\langle\times\rangle\,B$, and $A^{\langle k\rangle}$ correspondingly.

We use a simple example to illustrate the Boolean operations.

Example 16.3. Assume

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}; \quad B = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}.$$

(i) We can calculate that

$$\neg A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad A(+)B = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix},$$

$$A(\times)B = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \quad A^{(s)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \ s \geq 1.$$

(ii) If we use definition (16.12), than

$$A\,\langle+\rangle\,B = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad A\,\langle\times\rangle\,B = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$A^{\langle s\rangle} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, s \geq 2.$$

From Theorem 16.1, we know that if $\mathcal{J}_{ij}^s = c$, there are c pathes from $\delta_{2^{m+n}}^j$ to $\delta_{2^{m+n}}^i$ at sth steps. That means, if $\mathcal{J}_{ij}^s > 0$, at sth step, $\delta_{2^{m+n}}^i$ is reachable from $\delta_{2^{m+n}}^j$. When the controllability is considered, we do not need to concern how many paths from one state to the other, but only need to know whether a state can be reached from another one. Hence, we simply use the Boolean power of $\mathcal{J}$, which leads to the following conclusion.

Theorem 16.3. *Consider system (16.3) with its input-state incidence matrix $\mathcal{J}$. Define the controllability matrix as*

$$\mathcal{M}_{\mathcal{C}} := (+)_{s=1}^{2^{m+n}} M^{(s)} \in \mathcal{B}_{2^n \times 2^n}, \tag{16.13}$$

and denote $\mathcal{M}_{\mathcal{C}} = (c_{ij})$. Then

(i) $\delta_{2^n}^i$ is reachable from $\delta_{2^n}^j$, if and only if $c_{ij} = 1$;
(ii) The system is controllable at $\delta_{2^n}^j$, if and only if $\mathrm{Col}_j(\mathcal{M}_{\mathcal{C}}) = \mathbf{1}_{2^n}$;
(iii) The system is controllable, if and only if $\mathcal{M}_{\mathcal{C}} = \mathbf{1}_{2^n \times 2^n}$, where $\mathbf{1}_{2^n \times 2^n}$ is a $2^n \times 2^n$ matrix with all entries equal to 1.

Proof. (i): By Cayley-Hamilton Theorem in linear algebra, it is easy to see that if $\mathcal{J}_{ij}^s = 0$, $s \le 2^{m+n}$, then so are $\mathcal{J}^s$, $\forall\, s$. Thus, we consider only $\{\mathcal{J}^s | s \le 2^{m+n}\}$. Next, we know that $\delta_{2^n}^i$ is reachable from $\delta_{2^n}^j$, if and only if there exist β and s such that $\delta_{2^m}^\alpha \delta_{2^n}^i$ is reachable from $\delta_{2^m}^\beta \delta_{2^n}^j$ at sth step for all $1 \le \alpha \le 2^m$, which means

$$1 = \left[(+)_{s=1}^{2^{m+n}} (+)_{\beta=1}^{2^m} \mathrm{Blk}_\beta(\mathcal{J}_0^{(s)})\right]_{ij} = \left[(+)_{s=1}^{2^{m+n}} \mathcal{M}_{\mathcal{C}}\right]_{ij} = c_{ij}.$$

(ii) and (iii) can be obtained directly from (i). □

Example 16.4. Consider the following Boolean control network

$$\begin{cases} x_1(t+1) = (x_1(t) \leftrightarrow x_2(t)) \vee u_1(t) \\ x_2(t+1) = \neg x_1(t) \wedge u_2(t), \end{cases} \tag{16.14}$$
$$y(t) = x_1(t) \vee x_2(t).$$

Setting $x(t) = \ltimes_{i=1}^2 x_i(t)$, $u = \ltimes_{i=1}^2 u_i(t)$, we have

$$\begin{cases} x(t+1) = Lu(t)x(t) \\ y(t) = Hx(t), \end{cases} \tag{16.15}$$

where

$$L = \delta_4[2\ 2\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 4\ 3\ 1\ 2\ 4\ 4\ 2],$$

$$H = \delta_2[1\ 1\ 1\ 2].$$

For system (16.14), the basic block of its input-state incidence matrix $\mathcal{J}_0 = L$.

(1) Is δ_4^1 reachable from $x(0) = \delta_4^2$?
After a straightforward computation, we have

$$(M^{(1)})_{12} = 0, \quad (M^{(2)})_{12} = 1.$$

That means that $x(2) = \delta_4^1$ is reachable from $x(0) = \delta_4^2$ at 2nd step.

(2) Is the system controllable or controllable at any point?
We check the controllability matrix:

$$\mathcal{C} = (+)_{s=1}^{2^4} M^{(s)} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

According to Corollary 16.3, we conclude that

(i) The system is not controllable. It is controllable at $x_0 = \delta_4^3 \sim (0, 1)$.
(ii) $x_d = \delta_4^3 \sim (0, 1)$ is not reachable from $x_0 = \delta_4^1 \sim (1, 1)$, or $x_0 = \delta_4^2 \sim (1, 0)$, or $x_0 = \delta_4^4 \sim (0, 0)$.

16.3 Observability

In this section we consider the observability of system (16.1) and (16.2). We first give a definition.

Definition 16.5. Consider system (16.1) and (16.2). Denote by $Y(t) = (y_1(t), \cdots, y_p(t)) \in \mathcal{D}^p$.

(1) X_1^0 and X_2^0 are said to be distinguishable, if there exists a control sequence $\{U(0), U(1), \cdots, \}$, and an $s \geq 1$, such that

$$\begin{aligned} &Y^1(s) = y^s(U(s-1), \cdots, U(0), X_1^0) \\ &\neq Y^2(s) = y^s(U(s-1), \cdots, U(0), X_2^0). \end{aligned} \tag{16.16}$$

(2) The system is said to be observable, if any two initial points $X_1^0, X_2^0 \in \mathcal{X}$ are distinguishable.

In the following we introduce a necessary and sufficient condition for observability of controllable Boolean control networks.

Split L into 2^m equal blocks as

$$\begin{aligned} L =& [\mathrm{Blk}_1(L), \mathrm{Blk}_2(L), \cdots, \mathrm{Blk}_{2^m}(L)] \\ :=& [B_1, B_2, \cdots, B_{2^m}], \end{aligned}$$

where $B_i \in \mathcal{L}_{2^n \times 2^n}$, $i = 1, \cdots, 2^m$.

Define a sequence of sets of matrices $\Gamma_i \in \mathcal{L}_{2^p \times 2^n}$, $i = 0, 1, 2, \cdots$, as

$$\begin{cases} \Gamma_0 = \{H\} \\ \Gamma_1 = \{\, HB_i \,|\, i = 1, 2, \cdots, 2^m\} \\ \Gamma_2 = \{\, HB_iB_j \,|\, i, j = 1, 2, \cdots, 2^m\} \\ \vdots \\ \Gamma_s = \{\, HB_{i_1}B_{i_2} \cdots B_{i_s} \,|\, i_1, i_2, \cdots, i_s = 1, 2, \cdots, 2^m\} \\ \vdots \end{cases} \tag{16.17}$$

Note that $\Gamma_s \subset \mathcal{L}_{2^p \times 2^n}$, $\forall\, s$, and $\mathcal{L}_{2^p \times 2^n}$ is a finite set. Then it is easy to see that there exists a smallest s^* such that

$$\Gamma_j \subset \cup_{k=1}^{s^*} \Gamma_k, \quad \forall\, j > s^*.$$

For notational ease, we rewrite $\{\Gamma_1, \cdots, \Gamma_{s^*}\}$ as a set of matrices, i.e.

$$\Gamma_k = \begin{bmatrix} H\,\underbrace{B_1B_1 \cdots B_1}_{k} \\ H\,\underbrace{B_1B_1 \cdots B_2}_{k} \\ \vdots \\ H\,\underbrace{B_{2^m}B_{2^m} \cdots B_{2^m}}_{k} \end{bmatrix}.$$

And then we construct a matrix, called the observability matrix, as

$$\mathcal{O} = \begin{bmatrix} \Gamma_0 \\ \Gamma_1 \\ \vdots \\ \Gamma_{s^*} \end{bmatrix}.$$

Refer to Cheng and Zhao (2011) or Cheng *et al.* (2011b), we have the following result about observability.

Theorem 16.4. *Assume system (16.1)-(16.2) is controllable. Then it is observable, if and only if*

$$\mathrm{rank}\,(\mathcal{O}) = 2^n. \tag{16.18}$$

We give an example to illustrate it.

Example 16.5. Consider the following Boolean control network

$$\begin{cases} x_1(t+1) = \neg x_1(t) \vee x_2(t) \\ x_2(t+1) = u(t) \wedge \neg x_1(t) \vee (\neg u(t) \wedge x_1(t) \wedge \neg x_2(t)), \end{cases} \tag{16.19}$$

$$y(t) = x_1 \bar{\vee} x_2. \tag{16.20}$$

Its algebraic form is

$$\begin{cases} x(t+1) = Lu(t)x(t) \\ y(t) = H(t), \end{cases} \tag{16.21}$$

where

$$\begin{aligned} L &= \delta_4[2\ 4\ 1\ 1\quad 2\ 3\ 2\ 2], \\ H &= \delta_2[2\ 1\ 1\ 2]. \end{aligned}$$

Checking the controllability matrix, we have

$$\mathcal{M}_{\mathcal{C}} = \sum_{s=1}^{2^3}\sum_{i=1}^{2}(\mathrm{Blk}_i(\mathcal{J}_0^{(s)})) = \begin{bmatrix} 1\ 1\ 1\ 1 \\ 1\ 1\ 1\ 1 \\ 1\ 1\ 1\ 1 \\ 1\ 1\ 1\ 1 \end{bmatrix} > 0.$$

Thus the system is controllable.

A straightforward computation shows that the observability matrix is

$$\mathcal{O} = \begin{bmatrix} H \\ HB_1 \\ HB_2 \\ HB_1B_1 \\ \vdots \end{bmatrix} = \begin{bmatrix} 0\ 1\ 1\ 0 \\ 1\ 0\ 0\ 1 \\ 1\ 0\ 0\ 0 \\ 0\ 1\ 1\ 1 \\ 1\ 1\ 1\ 1 \\ 0\ 0\ 0\ 0 \\ 0\ 0\ 1\ 1 \\ 1\ 1\ 0\ 0 \\ \vdots \end{bmatrix}.$$

From part of $\mathcal{O}$, all the columns has already been different, it is enough to see that $\mathrm{rank}(\mathcal{O}) = 4$. Thus, the system is also observable.

There is another necessary and sufficient condition of observability in Cheng and Qi (2009) which is basically the same as Theorem 16.4. But Theorem 16.4 is more convenient to use. However, these conditions need the system to be controllable. Next, we give a sufficient condition of observability which have no assumption about controllability.

For these, we will use the input-state incidence matrix defined in Section 16.1. It has been pointed out that $\mathrm{Blk}_i(\mathcal{J}_0^{(s)})$ corresponds to the input $u(0) = \delta_{2^m}^i$, and $\mathrm{Col}_j[\mathrm{Blk}_i(\mathcal{J}_0^{(s)})]$ corresponds to $x_0 = \delta_{2^n}^j$. We want to exchange the order of the indices i and j, precisely, we define

$$\tilde{\mathcal{J}}_0^{(s)} := \mathcal{J}_0^{(s)} W_{[2^n,2^m]}, \tag{16.22}$$

and then split it into 2^n blocks as

$$\tilde{\mathcal{J}}_0^{(s)} = \left[\mathrm{Blk}_1(\tilde{\mathcal{J}}_0^{(s)}) \quad \mathrm{Blk}_2(\tilde{\mathcal{J}}_0^{(s)}) \quad \cdots \quad \mathrm{Blk}_{2^n}(\tilde{\mathcal{J}}_0^{(s)})\right],$$

where $\mathrm{Blk}_i(\tilde{\mathcal{J}}_0^{(s)}) \in \mathcal{B}_{2^n \times 2^m}$, $i = 1, \cdots, 2^n$. Now we can see that each block $\mathrm{Blk}_i(\tilde{\mathcal{J}}_0^{(s)})$ corresponds to $x_0 = \delta_{2^n}^i$, and each block $\mathrm{Col}_j(\mathrm{Blk}_i(\tilde{\mathcal{J}}_0^{(s)}))$ corresponds to $u(0) = \delta_{2^m}^j$. Using this matrix, we can get the following result.

Theorem 16.5 (Zhao *et al.*, 2010c). *Consider system (16.1)–(16.2) with its algebraic form (16.3). If*

$$\bigvee_{s=1}^{2^{m+n}} \left[\left(H \ltimes \mathrm{Blk}_i(\tilde{\mathcal{J}}_0^{(s)})\right) \bar{\vee} \left(H \ltimes \mathrm{Blk}_j(\tilde{\mathcal{J}}_0^{(s)})\right)\right] \neq 0, \quad 1 \leq i < j \leq 2^n, \tag{16.23}$$

then the system is observable.

Example 16.6. Recall Example 16.4. System (16.14) is not controllable, so we cannot use Theorem 16.4 to verify wether it is observable.

Denote

$$O_{ij} = \vee_{s=1}^{2^4} \left[\left(H \ltimes \mathrm{Blk}_i(\tilde{\mathcal{J}}_0^{(s)})\right) \bar{\vee} \left(H \ltimes \mathrm{Blk}_j(\tilde{\mathcal{J}}_0^{(s)})\right)\right].$$

A straightforward computation yields

$$\begin{aligned} O_{12} &= \begin{bmatrix} 0\,0\,1\,1 \\ 0\,0\,1\,1 \end{bmatrix}, O_{13} = \begin{bmatrix} 0\,0\,0\,1 \\ 1\,0\,0\,1 \end{bmatrix}, O_{14} = \begin{bmatrix} 0\,0\,0\,0 \\ 1\,0\,1\,0 \end{bmatrix}, \\ O_{23} &= \begin{bmatrix} 0\,0\,1\,0 \\ 1\,0\,1\,0 \end{bmatrix}, O_{24} = \begin{bmatrix} 0\,0\,1\,1 \\ 1\,0\,1\,1 \end{bmatrix}, O_{34} = \begin{bmatrix} 0\,0\,0\,1 \\ 0\,0\,1\,1 \end{bmatrix}. \end{aligned}$$

Then by Theorem 16.5, we conclude that the system is observable.

16.4 Disturbance Decoupling

Disturbance decoupling problem (DDP) is a classical problem for control theory. In practice, a control system is inevitably affected by disturbance. To analyze and control the system, we first want to design a control such that the disturbance on the system will not affect the output of the system, which is the purpose of DDP.

Consider the following system,

$$\begin{cases} x_1(t+1) = f_1(x_1(t), \cdots, x_n(t), u_1(t), \cdots, u_m(t), \xi_1(t), \cdots, \xi_q(t)) \\ \vdots \\ x_n(t+1) = f_n(x_1(t), \cdots, x_n(t), u_1(t), \cdots, u_m(t), \xi_1(t), \cdots, \xi_q(t)), \end{cases}$$
$$y_j(t) = h_j(x(t)), \quad j = 1, \cdots, p, \tag{16.24}$$

where $\xi_i(t) \in \mathcal{D}$, $i = 1, \cdots, q$ are disturbances. Let $x(t) = \ltimes_{i=1}^n x_i(t)$, $u(t) = \ltimes_{i=1}^m u_i(t)$, $\xi(t) = \ltimes_{i=1}^q \xi_i(t)$, and $y(t) = \ltimes_{i=1}^p y_i(t)$. Then the algebraic form of (16.24) can be expressed as

$$\begin{aligned} x(t+1) &= Lu(t)\xi(t)x(t), \\ y(t) &= Hx(t), \end{aligned} \tag{16.25}$$

where $L \in \mathcal{L}_{2^n \times 2^{n+m+q}}$, $H \in \mathcal{L}_{2^p \times 2^n}$.

Definition 16.6. Consider system (16.24). The DDP is solvable, if we can find a feedback control $u(t) = \phi(x(t))$, and a coordinate transformation $z = T(x)$, such that under z coordinate frame the closed-loop system becomes

$$\begin{cases} z^1(t+1) = F^1(z(t), \phi(x(t)), \xi(t)) \\ z^2(t+1) = F^2(z^2(t)), \end{cases}$$
$$y(t) = G(z^2(t)) \tag{16.26}$$

where $z(t) = \{z^1(t), z^2(t)\}$.

From the definition we can see that the key issues of solving DDP is: (i) to find a regular subbasis $z^2(t)$ by which the output can be expressed, and (ii) a proper feedback control such that the complement coordinate subbasis z^1 and the disturbances ξ can be deleted from the dynamics of z^2. We begin with finding the coordinate transformation. To this end, we propose a new kind of subspace.

Definition 16.7. Let $\mathcal{X} = \mathcal{F}_\ell\{x_1, \cdots, x_n\}$ be the state space of (16.24), and $Y = \{y_1, \cdots, y_p\} \subset \mathcal{X}$. A regular subspace $\mathcal{Z} \subset \mathcal{X}$ is called a Y-friendly subspace, if $y_i \in \mathcal{Z}$, $i = 1, \cdots, p$. A Y-friendly subspace of minimum dimension is called a minimum Y-friendly subspace.

First, we consider how to find out a minimum Y-friendly subspace. Express the algebraic form of outputs as

$$y = Hx = \delta_{2^p}[i_1 \ i_2 \ \cdots \ i_{2^n}]x. \tag{16.27}$$

Denote by

$$n_j = \left|\{k \mid i_k = j, 1 \le k \le 2^n\}\right|, \quad j = 1, 2, \cdots, 2^p,$$

where $|\cdot|$ is the cardinality of the set. Then we have the following theorem (Cheng, 2011).

Theorem 16.6. *Consider system (16.24). Assume $y = \ltimes_{i=1}^p y_i$ has its algebraic form (16.27).*

(1) There is a Y-friendly subspace of dimension r, if and only if n_j, $j = 1, \cdots, 2^p$ have a common factor 2^{n-r}.
(2) Assume 2^{n-r} is the largest common 2-type factor of n_j, $j = 1, \cdots, 2^p$. Then the minimum Y-friendly subspace is of dimension r.

Note that 2-type factor means a factor of the form 2^r.

The following algorithm for finding y can be considered as a proof of this theorem, we leave the rigorous proof for exercise.

Algorithm 3.

1. For system (16.24), find a common 2-type factor 2^{n-r} of $\{n_j \mid j = 1, \cdots, 2^p\}$, and then find $\{m_j\}$ such that

$$n_j = 2^{n-r} m_j, \quad j = 1, \cdots, 2^p.$$

Denote

$$J_j = \{k | i_k = j\}, \quad j = 1, \cdots, 2^p,$$

and

$$I_j = \left\{ \sum_{\xi=1}^{j-1} m_\xi + 1, \sum_{\xi=1}^{j-1} m_\xi + 2, \cdots, \sum_{\xi=1}^{j} m_\xi \right\}, \quad j = 1, \cdots, 2^p.$$

Split a $2^r \times 2^n$ matrix $T_0 = (t_{r,s})$ to $2^p \times 2^p$ minors as

$$T_0^{i,j} = \{t_{r,s} \mid r \in I_i, s \in J_j\}, \quad i, j = 1, \cdots, 2^p.$$

2. Set

$$T_0^{i,j} = \begin{cases} I_{m_i} \otimes \mathbf{1}_{2^{n-r}}^T, & i = j \\ 0, & \text{otherwise}, \end{cases} \tag{16.28}$$

which is an $m_i \times m_j 2^{n-r}$ matrix.

3. Set

$$z = \ltimes_{i=1}^{r} z_i := T_0 x.$$

And then retrieve z_i, $i = 1, \cdots, r$ from T_0. (We refer to Chapter 6 for retrieving technique.)
Then

$$\{z_i | i = 1, \cdots, r\}$$

is a subbasis of an r-dimensional Y-friendly subspace.

This algorithm provides a way to construct an r-dimensional Y-friendly subspace, but it is obvious that different assignments of 1 in $T_0^{i,i}$ can construct different subbases. One may expect that those subbases form a unique subspace. Unfortunately, the following example shows the minimum Y-friendly subspace is not unique.

Example 16.7. Let $\mathcal{X} = \mathcal{F}_\ell\{x_1, x_2, x_3, x_4\}$.

$$\begin{aligned} y_1 &= h_1(x_1, x_2, x_3, x_4) = (x_1 \leftrightarrow x_3) \wedge (x_2 \bar{\vee} x_4), \\ y_2 &= h_2(x_1, x_2, x_3, x_4) = x_1 \wedge x_3. \end{aligned} \tag{16.29}$$

Setting $x = \ltimes_{i=1}^{4} x_i$, $y = y_1 y_2$, we have its algebraic form as $y = Hx$, where

$$H = \delta_4[3\ 1\ 4\ 4\ 1\ 3\ 4\ 4\ 4\ 4\ 4\ 2\ 4\ 4\ 2\ 4].$$

Thus, $n_1 = n_2 = n_3 = 2$, $n_4 = 10$. Since the largest common 2-type factor is $2 = 2^{4-3}$, we can have the minimum Y-friendly subspace of dimension $r = 3$. To construct T_0 we have

$$\begin{aligned} &J_1 = \{2, 5\}; \quad J_2 = \{12, 15\}; \quad J_3 = \{1, 6\}; \\ &J_4 = \{3, 4, 7, 8, 9, 10, 11, 13, 14, 16\}; \\ &I_1 = \{1\}; \quad I_2 = \{2\}; \quad I_3 = \{3\}; \quad I_4 = \{4, 5, 6, 7, 8\}. \end{aligned}$$

By (16.28), T_0 is obtained as

$$T_0 = \delta_8[3\ 1\ 4\ 4\ 1\ 3\ 5\ 5\ 6\ 6\ 7\ 2\ 7\ 8\ 2\ 8].$$

Correspondingly, we can construct

$$G = \delta_4[1\ 2\ 3\ 4\ 4\ 4\ 4\ 4],$$

such that $GT_0 = H$.

Thus we have obtained a subbasis $\{z_1, z_2, z_3\}$ such that $z = z_1 z_2 z_3 = T_0 x$ and $y = Gz$.

$T_0^{4,4}$ can also be assigned as

$$T_0^{4,4} = \begin{bmatrix} 1\,0\,0\,0\,0\,0\,1\,0 \\ 0\,1\,0\,0\,0\,0\,0\,1 \\ 0\,0\,1\,0\,1\,0\,0\,0 \\ 0\,0\,0\,1\,0\,1\,0\,0 \end{bmatrix},$$

then we can get

$$T_0' = \delta_8[3\ 1\ 5\ 6\ 1\ 3\ 7\ 8\ 7\ 8\ 4\ 2\ 4\ 5\ 2\ 6],$$

which also satisfies $GT_0' = H$. Denoted by $\{z_1', z_2', z_3'\}$ the subbasis satisfying $z' = z_1'z_2'z_3' = T_0'x$, and $\mathcal{Z} = \mathcal{F}_\ell\{z_1, z_2, z_3\}$, $\mathcal{Z}' = \mathcal{F}_\ell\{z_1', z_2', z_3'\}$.

Suppose $\mathcal{Z} = \mathcal{Z}'$, then there exists a nonsingular matrix $P \in \mathcal{L}_{8\times 8}$, such that $T_0 = PT_0'$. Hence

$$P = T_0T_0'^T(T_0'T_0'^T)^{-1} = \begin{bmatrix} 1\,0\,0\,0 & 0 & 0 & 0 & 0 \\ 0\,1\,0\,0 & 0 & 0 & 0 & 0 \\ 0\,0\,1\,0 & 0 & 0 & 0 & 0 \\ 0\,0\,0\,0 & 0.5 & 0.5 & 0 & 0 \\ 0\,0\,0\,0 & 0 & 0 & 0.5 & 0.5 \\ 0\,0\,0\,0 & 0 & 0 & 0.5 & 0.5 \\ 0\,0\,0\,1 & 0 & 0 & 0 & 0 \\ 0\,0\,0\,0 & 0.5 & 0.5 & 0 & 0 \end{bmatrix}.$$

Unfortunately, this P is not a logical matrix. Thus $\mathcal{Z} \neq \mathcal{Z}'$.

Assume a Y-friendly subspace is obtained as z^2. Then we can find z^1, such that $z = \{z^1, z^2\}$ form a new coordinate frame, and the system can be expressed in the new coordinate frame as

$$\begin{cases} z^1(t+1) = F^1(z(t), u(t), \xi(t)) \\ z^2(t+1) = F^2(z(t), u(t), \xi(t)), \end{cases} \\ y(t) = G(z^2(t)). \tag{16.30}$$

We call it Y-friendly form or output-friendly form.

Then it is obvious that if we can find a feedback control $u(t) = u(z(t))$, such that F^2 only depends on z^2, the DDP is solved.

Assume $z^2 = (z_1^2, \cdots, z_k^2)$ is of dimension k, and the feedback control $u(t) = Kz^1(t)$. Then $F^2(z(t), u(t), \xi(t))$ can be expressed as

$$\begin{aligned} &F^2(z(t), u(t), \xi(t)) = M_2u(t)\xi(t)z(t) \\ &\quad = M_2KW_{[2^q,2^{n-k}]}(I_{w^q} \otimes M_r^{2^{n-k}})W_{[2^k,2^{q+n-k}]}z^2(t)\xi(t)z^1(t) \\ &\quad := Qz^2(t)\xi(t)z^1(t), \end{aligned} \tag{16.31}$$

where M_2 is the structure matrix of F^2.

Split Q to 2^k blocks

$$Q = [Q_1, Q_2, \cdots, Q_{2^k}].$$

It is obvious that $F^2(z(t), u(t), \xi(t)) = F^2(z^2(t))$, if and only if for each j, all the columns of Q_j are the same.

To verify whether the DDP is solvable, we should try every subspaces to see if all the columns of Q_j are the same. But there are lots of Y-friendly subspaces of system (16.24). However, the following proposition may help to reduce the searching complexity.

Proposition 16.2. *Let V be a $Y = \{y_1, \cdots, y_p\}$-friendly subspace. Then there exists a minimum Y-friendly subspace W, such that $W \subset V$.*

***Proof*.** Denote by

$$y = \ltimes_{i=1}^{p} y_i = Hx,$$

and $n_i = |\{j|\,\mathrm{Col}_j(H) = \delta_{2^P}^i\}|$, $i = 1, \cdots, 2^p$. We know that 2^s is the largest common 2-type factor of $\{n_i\}$, if and only if the minimum Y-friendly subspace has dimension of $n - s$.

Let $\{v_1, \cdots, v_t\}$ be a basis of V. If $t = n - s$, we are done. So we assume $t > n - s$. Since V is a Y-friendly subspace, denoting $v = \ltimes_{i=1}^{t} v_i$, we can express

$$y = Gv, \quad \text{where } G \in \mathcal{L}_{2^p \times 2^t}.$$

Denote $r_i = |\{j|\,\mathrm{Col}_j(G) = \delta_{2^P}^i\}|$, $i = 1, \cdots, 2^p$. Assume that 2^j is the largest 2-type common factor of r_i, and denote $r_i = m_i 2^j$. Since V is a regular subspace, we denote $v = Ux$, where $U \in \mathcal{L}_{2^t \times 2^n}$ and (15.28) holds for U. Note that

$$y = Gv = GUx,$$

we have to calculate GU. Using the construction of G and the property of U, it is easy to verify that each column of $\delta_{2^p}^i$ yields 2^{n-t} columns of $\delta_{2^p}^i$ in GU. Hence we have

$$r_i \cdot 2^{n-t} = m_i \cdot 2^{n-t+j}.$$

That means, the largest common 2-type factor of $\{n_i\}$ is 2^{n-t+j}. It follows that $n - t + j = s$. Equivalently,

$$j = t - (n - s).$$

Since the dimension of V is t and r_i have largest 2-type common fact 2^j, we can find a minimum Y-friendly subspace of V of the dimension $t - j = t - [t - (n - s)] = n - s$. It follows from the dimension that this Y-friendly minimum subspace of V is also a minimum Y-friendly subspace of $\mathcal{X} = \mathcal{F}_\ell\{x_1, \cdots, x_n\}$. □

We can see that if for a minimum Y-friendly subspace W whose subbasis is $z^2(t)$, there is no such u that $\xi(t)$ can be deleted from $F_2(z^1(t), z^2(t), u(t), \xi(t))$, then neither for other Y-friendly subspaces containing W. Hence, using this proposition, we can start from minimum Y-friendly subspaces to solve DDP. However, enlarging z^2 may make its complement $z^1 = (z^2)^c$ disappear from the dynamics F_2 of z^2. Hence, when a minimum Y-friendly form with certain controls is not affected by disturbances $\xi_i(t)$, enlarging it to avoid the affection of its complement may be necessary.

We give an example to describe how to find a solution of DDP.

Example 16.8. Consider the following system

$$\begin{cases} x_1(t+1) = x_4(t)\bar{\vee}u_1(t) \\ x_2(t+1) = (x_2(t)\bar{\vee}x_3(t)) \wedge \neg\xi(t) \\ x_3(t+1) = [(x_2(t) \leftrightarrow x_3(t)) \vee \xi(t)]\bar{\vee}[(x_1 \leftrightarrow x_5) \vee u_2(t)] \\ x_4(t+1) = [u_1(t) \to (\neg x_2(t) \vee \xi(t))] \wedge (x_2(t) \leftrightarrow x_3(t)) \\ x_5(t+1) = (x_4(t)\bar{\vee}u_1(t)) \leftrightarrow [(u_2(t) \wedge \neg x_2(t)) \vee x_4(t)], \end{cases} \tag{16.32}$$
$$y(t) = x_4(t) \wedge (x_1(t) \leftrightarrow x_5(t)),$$

where $u_1(t), u_2(t)$ are controls, $\xi(t)$ is a disturbance, $y(t)$ is the output.

Setting $x(t) = \ltimes_{i=1}^5 x_i(t)$, $u = u_1(t)u_2(t)$, we express (16.32) into it algebraic form as

$$\begin{cases} x(t+1) = Lu(t)\xi(t)x(t) \\ y(t) = hx(t), \end{cases} \tag{16.33}$$

where

$$
\begin{array}{rl}
L = \delta_{32}[& 30\ 30\ 14\ 14\ 32\ 32\ 16\ 16\ 32\ 32\ 15\ 15\ 30\ 30\ 13\ 13 \\
& 30\ 30\ 14\ 14\ 32\ 32\ 16\ 16\ 32\ 32\ 15\ 15\ 30\ 30\ 13\ 13 \\
& 32\ 32\ 16\ 16\ 20\ 20\ \ 4\ \ 4\ 20\ 20\ \ 3\ \ 3\ 30\ 30\ 13\ 13 \\
& 32\ 32\ 16\ 16\ 20\ 20\ \ 4\ \ 4\ 20\ 20\ \ 3\ \ 3\ 30\ 30\ 13\ 13 \\
& 30\ 26\ 14\ 10\ 32\ 28\ 16\ 12\ 32\ 28\ 16\ 12\ 30\ 26\ 14\ 10 \\
& 26\ 30\ 10\ 14\ 28\ 32\ 12\ 16\ 28\ 32\ 12\ 16\ 26\ 30\ 10\ 14 \\
& 32\ 28\ 16\ 12\ 20\ 24\ \ 4\ \ 8\ 20\ 24\ \ 4\ \ 8\ 30\ 26\ 14\ 10 \\
& 28\ 32\ 12\ 16\ 24\ 20\ \ 8\ \ 4\ 24\ 20\ \ 8\ \ 4\ 26\ 30\ 10\ 14 \\
& 13\ 13\ 29\ 29\ 15\ 15\ 31\ 31\ 15\ 15\ 32\ 32\ 13\ 13\ 30\ 30 \\
& 13\ 13\ 29\ 29\ 15\ 15\ 31\ 31\ 15\ 15\ 32\ 32\ 13\ 13\ 30\ 30 \\
& 13\ 13\ 29\ 29\ \ 3\ \ 3\ 19\ 19\ \ 3\ \ 3\ 20\ 20\ 13\ 13\ 30\ 30 \\
& 13\ 13\ 29\ 29\ \ 3\ \ 3\ 19\ 19\ \ 3\ \ 3\ 20\ 20\ 13\ 13\ 30\ 30 \\
& 13\ \ 9\ 29\ 25\ 15\ 11\ 31\ 27\ 15\ 11\ 31\ 27\ 13\ \ 9\ 29\ 25 \\
& \ 9\ 13\ 25\ 29\ 11\ 15\ 27\ 31\ 11\ 15\ 27\ 31\ \ 9\ 13\ 25\ 29 \\
& 13\ \ 9\ 29\ 25\ \ 3\ \ 7\ 19\ 23\ \ 3\ \ 7\ 19\ 23\ 13\ \ 9\ 29\ 25 \\
& \ 9\ 13\ 25\ 29\ \ 7\ \ 3\ 23\ 19\ \ 7\ \ 3\ 23\ 19\ \ 9\ 13\ 25\ 29],
\end{array}
$$

and

$$h = \delta_2[1\ 2\ 2\ 2\ 1\ 2\ 2\ 2\ 1\ 2\ 2\ 2\ 1\ 2\ 2\ 2\ 2\ 1\ 2\ 2\ 2\ 1\ 2\ 2\ 2\ 1\ 2\ 2\ 2\ 1\ 2\ 2].$$

First, we find a minimum output-friendly subspace. From h we can see $n_1 = 8$ and $n_2 = 24$. Then we have the largest 2-type common factor $2^s = 2^3$, and $m_1 = 1$, $m_2 = 3$. Hence, we know that the minimum output-friendly subspace is of dimension $n - s = 5 - 3 = 2$. Using Algorithm 3, we may choose

$$T_0 = \delta_4[1\ 2\ 3\ 4\ 1\ 2\ 3\ 4\ 1\ 2\ 3\ 4\ 1\ 2\ 3\ 4\ 2\ 1\ 4\ 3\ 2\ 1\ 4\ 3\ 2\ 1\ 4\ 3\ 2\ 1\ 4\ 3],$$

and

$$G = \delta_2[1\ 2\ 2\ 2].$$

From T_0 we can find the output-friendly subbasis, denote it by $\{z_4, z_5\}$, with $z_4 z_5 = T_0 x$. Add $\{z_1, z_2, z_3\}$ to make $z = z^1 z^2 = \ltimes_{i=1}^5 z_i = Tx$ a coordinate transformation. Using Theorem 15.5 and the constructing technique in its proof, we can simply set $z_i = M_i x$, $i = 1, 2, 3$, where M_i are chosen as follows:

$$
\begin{aligned}
M_1 &= \delta_2[1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2]; \\
M_2 &= \delta_2[2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1]; \\
M_3 &= \delta_2[1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 1].
\end{aligned}
$$

It is easy to check that

$$\begin{aligned} T = \delta_{32}[\,&9\ 10\ 11\ 12\ 13\ 14\ 15\ 16\ 5\ 6\ 7\ 8\ 1\ 2\ 3\ 4\ 26\ 25\\ &28\ 27\ 30\ 29\ 32\ 31\ 22\ 21\ 24\ 23\ 18\ 17\ 20\ 19] \end{aligned}$$

is a coordinate transformation. Conversely, we have $x = T^T z$, with

$$\begin{aligned} T^T = [\,&13\ 14\ 15\ 16\ 9\ 10\ 11\ 12\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 30\ 29\\ &32\ 31\ 26\ 25\ 28\ 27\ 18\ 17\ 20\ 19\ 22\ 21\ 24\ 23]. \end{aligned}$$

Now under the coordinate frame z we have

$$\begin{aligned} z(t+1) &= Tx(t+1) = TLu(t)\xi(t)x(t) = TLu(t)\xi(t)T^T z(t)\\ &= TL(I_8 \otimes T^T)u(t)\xi(t)z(t) := \tilde{L}u(t)\xi(t)z(t). \end{aligned}$$

The corresponding Y-friendly form is obtained as

$$\begin{cases} z^1(t+1) = \tilde{L}_1 u(t)\xi(t)z(t)\\ z^2(t+1) = \tilde{L}_2 u(t)\xi(t)z(t), \end{cases}$$
$$y(t) = Gz^2(t),$$

where

$$\begin{aligned} \tilde{L}_1 = \delta_8[&5\ 5\ 1\ 1\ 5\ 5\ 1\ 1\ 5\ 5\ 1\ 1\ 5\ 5\ 1\ 1\ 5\ 5\ 1\ 1\ 5\ 5\ 1\ 1\ 5\ 5\ 1\ 1\ 5\ 5\ 1\ 1\\ &5\ 5\ 1\ 1\ 7\ 7\ 3\ 3\ 5\ 5\ 1\ 1\ 7\ 7\ 3\ 3\ 5\ 5\ 1\ 1\ 7\ 7\ 3\ 3\ 5\ 5\ 1\ 1\ 7\ 7\ 3\ 3\\ &5\ 6\ 1\ 2\ 5\ 6\ 1\ 2\ 5\ 6\ 1\ 2\ 5\ 6\ 1\ 2\ 5\ 6\ 1\ 2\ 5\ 6\ 1\ 2\ 5\ 6\ 1\ 2\ 5\ 6\ 1\ 2\\ &5\ 6\ 1\ 2\ 7\ 8\ 3\ 4\ 5\ 6\ 1\ 2\ 7\ 8\ 3\ 4\ 5\ 6\ 1\ 2\ 7\ 8\ 3\ 4\ 5\ 6\ 1\ 2\ 7\ 8\ 3\ 4\\ &1\ 1\ 5\ 5\ 1\ 1\ 5\ 5\ 1\ 1\ 5\ 5\ 1\ 1\ 5\ 5\ 1\ 1\ 5\ 5\ 1\ 1\ 5\ 5\ 1\ 1\ 5\ 5\ 1\ 1\ 5\ 5\\ &1\ 1\ 5\ 5\ 3\ 3\ 7\ 7\ 1\ 1\ 5\ 5\ 3\ 3\ 7\ 7\ 1\ 1\ 5\ 5\ 3\ 3\ 7\ 7\ 1\ 1\ 5\ 5\ 3\ 3\ 7\ 7\\ &1\ 2\ 5\ 6\ 1\ 2\ 5\ 6\ 1\ 2\ 5\ 6\ 1\ 2\ 5\ 6\ 1\ 2\ 5\ 6\ 1\ 2\ 5\ 6\ 1\ 2\ 5\ 6\ 1\ 2\ 5\ 6\\ &1\ 2\ 5\ 6\ 3\ 4\ 7\ 8\ 1\ 2\ 5\ 6\ 3\ 4\ 7\ 8\ 1\ 2\ 5\ 6\ 3\ 4\ 7\ 8\ 1\ 2\ 5\ 6\ 3\ 4\ 7\ 8]; \end{aligned}$$

$$\begin{aligned} \tilde{L}_2 = \delta_4[&1\ 1\ 1\ 1\ 3\ 3\ 3\ 3\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\ 1\ 1\ 1\ 1\ 3\ 3\ 3\ 3\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\\ &1\ 1\ 1\ 1\ 3\ 3\ 3\ 3\ 3\ 3\ 4\ 4\ 3\ 3\ 4\ 4\ 1\ 1\ 1\ 1\ 3\ 3\ 3\ 3\ 3\ 3\ 4\ 4\ 3\ 3\ 4\ 4\\ &1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\\ &1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\ 3\ 3\ 4\ 4\ 3\ 3\ 4\ 4\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\ 3\ 3\ 4\ 4\ 3\ 3\ 4\ 4\\ &1\ 1\ 1\ 1\ 3\ 3\ 3\ 3\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\ 1\ 1\ 1\ 1\ 3\ 3\ 3\ 3\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\\ &1\ 1\ 1\ 1\ 3\ 3\ 3\ 3\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\ 1\ 1\ 1\ 1\ 3\ 3\ 3\ 3\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\\ &1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\\ &1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4\ 1\ 1\ 2\ 2\ 3\ 3\ 4\ 4]; \end{aligned}$$

For all $K \in \mathcal{L}_{4,8}$, we can not find a K such that each 4×16 block Q_i of $Q = \tilde{L}_2 K W_{[2,8]}(I_2 \otimes M_r^8)W_{[4,16]}$ has the same columns.

However, when we set

$$K = \delta_4[4\ 4\ 4\ 4\ 4\ 4\ 4\ 4],$$

which means $u_1(t) = u_2(t) \equiv 0$, we have

$$\begin{aligned} Q = \delta_4[&1\ 3\ 1\ 3\ 1\ 3\ 1\ 3\ 1\ 3\ 1\ 3\ 1\ 3\ 1\ 3\ 1\ 3\ 1\ 3\ 1\ 3\ 1\ 3\ 1\ 3\ 1\ 3\ 1\ 3\ 1\ 3\\ &2\ 4\ 2\ 4\ 2\ 4\ 2\ 4\ 2\ 4\ 2\ 4\ 2\ 4\ 2\ 4\ 2\ 4\ 2\ 4\ 2\ 4\ 2\ 4\ 2\ 4\ 2\ 4\ 2\ 4\ 2\ 4]. \end{aligned}$$

Then we can see that $z_4(t+1) = z_3(t)$, $z_5(t+1) = z_4(t)$. Thus we want to enlarge the output-friendly subspace to include z_3 to see whether z_3 is affected by ξ, z_1 and z_2. Retrieving the system under the controls $u_1(t) = u_2(t) \equiv 0$, we have

$$\begin{cases} z_1(t+1) = z_4(t) \\ z_2(t+1) = z_3(t) \vee \xi(t) \\ z_3(t+1) = z_5(t) \\ z_4(t+1) = z_3(t) \\ z_5(t+1) = z_4(t), \end{cases} \tag{16.34}$$
$$y(t) = z_4(t) \wedge z_5(t).$$

One sees that the closed-loop system is in such a form that the DDP is solved.

16.5 Some Other Control Problems

This section briefly introduces stabilization, optimal control and identification of Boolean control networks. Most of proofs are ignored, one can refer to Cheng *et al.* (2011b,c); Zhao *et al.* (2010b); Cheng and Zhao (2011) for details.

16.5.1 *Stability and Stabilization*

Consider a Boolean network

$$\begin{cases} x_1(t+1) = f_1(x_1(t), x_2(t), \cdots, x_n(t)) \\ x_2(t+1) = f_2(x_1(t), x_2(t), \cdots, x_n(t)) \\ \vdots \\ x_n(t+1) = f_n(x_1(t), x_2(t), \cdots, x_n(t)), \end{cases} \tag{16.35}$$

or Boolean control network

$$\begin{cases} x_1(t+1) = f_1(x_1(t), x_2(t), \cdots, x_n(t), u_1(t), \cdots, u_m(t)) \\ x_2(t+1) = f_2(x_1(t), x_2(t), \cdots, x_n(t), u_1(t), \cdots, u_m(t)) \\ \vdots \\ x_n(t+1) = f_n(x_1(t), x_2(t), \cdots, x_n(t), u_1(t), \cdots, u_m(t)). \end{cases} \tag{16.36}$$

Their algebraic forms are, respectively,

$$x(t+1) = Lx(t) \tag{16.37}$$

and

$$x(t+1) = Lu(t)x(t). \tag{16.38}$$

Definition 16.8.

(1) System (16.35) is said to be globally stable if it is globally convergent. In other words, it has a fixed point as a global attractor (equivalently, unique attractor).

(2) The global stabilization problem of system (16.36) is to find, if possible, $u(t)$ such that the system becomes globally convergent.

The following results for global stability and stabilization are straightforward.

Theorem 16.7.

(1) System (16.35) with its algebraic form (16.37) is globally stable, if and only if there exist a state $x_0 = \delta_{2^n}^i$ and a positive integer k such that

$$\mathrm{Col}_i(L) = \delta_{2^n}^i, \quad \mathrm{Col}_j(L^k) = \delta_{2^n}^i, j = 1, 2, \cdots, 2^n.$$

(2) System (16.36) with its algebraic form (16.38) can be globally stabilized to $x_0 = \delta_{2^n}^i$ by a closed-loop control $u(t) = Gx(t)$, if and only if there exists a positive integer k such that

$$\mathrm{Col}_i(LGM_{r,2^n}) = \delta_{2^n}^i, \quad \mathrm{Col}_j((LGM_{r,2^n})^k) = \delta_{2^n}^i, j = 1, 2, \cdots, 2^n.$$

(3) System (16.36) with its algebraic form (16.38) can be globally stabilized to $x_0 = \delta_{2^n}^i$ by a free Boolean sequence $\{u(t)\}$, if and only if there exists a positive integer k such that

$$\mathrm{Col}_i(M) = \delta_{2^n}^i, \quad M_{ij}^{(k)} = 1,\ j = 1, 2, \cdots, 2^n,$$

where M is defined in (16.9).

Although these results are necessary and sufficient, it is hard to calculate when n is large. Next, we will propose a sufficient condition of global stability which uses the incidence matrix of the network graph (it is different from the input-state incidence matrix). For this, we firstly introduce vector distance of Boolean matrices.

Definition 16.9.

(1) Recall the network graph of system (16.35), the nodes $\mathcal{N} = \{x_1, x_2, \cdots, x_n\}$, the edges $\mathcal{E} = \{(x_i, x_j) | x_j \text{ is affected by } x_i\}$. Denoted $X = (x_1, x_2, \cdots, x_n)^T \in \mathcal{D}^n$, and the function of the dynamic by $X(t+1) = F(X(t))$. The incidence matrix of the network graph is an $n \times n$ Boolean matrix $\mathcal{I}(F) = (I_{ij})$, where

$$I_{ij} = \begin{cases} 1, & (x_i, x_j) \in \mathcal{E} \\ 0, & \text{otherwise.} \end{cases} \tag{16.39}$$

(2) Let $X = (x_{ij}), Y = (y_{ij}) \in \mathcal{B}_{m\times n}$. We said $X \le Y$ if $x_{ij} \le y_{ij}$, $\forall\, i, j$.

(3) Let $X = (x_{ij}), Y = (y_{ij}) \in \mathcal{B}_{m\times n}$. The vector distance of X and Y, denoted by $D_v(X, Y)$, is defined as

$$D_v(X, Y) = X \bar{\nabla} Y. \tag{16.40}$$

We leave the verifying following properties of vector distance for exercise.

Proposition 16.3.

(1) The $D_v(X, Y)$, defined in (16.40), satisfies

$$\begin{gathered} D_v(X, Y) = D_v(Y, X); \\ D_v(X, Y) = 0 \Rightarrow X = Y; \\ D_v(X, Y) +_{\mathcal{B}} D_v(Y, Z) \ge D_v(X, Z). \end{gathered}$$

(2) Let $A,\ B \in \mathcal{B}_{m\times n}$, and $C \in \mathcal{B}_{n\times p}$, $E \in \mathcal{B}_{q\times m}$. Then

$$D_v(AC, BC) \le D_v(A, B)C, \quad D_v(EA, EB) \le ED_v(A, B). \tag{16.41}$$

In the following, we use the scalar form of state, we can see that $X \in \mathcal{B}_n$, thus we can employ $D_v(X, Y)$ to describe the distance of two states.

Theorem 16.8. *If $\xi \in \mathcal{D}^n$ is a fixed point of Boolean network (16.35), and there exists an integer $k > 0$ such that*

$$[\mathcal{I}(F)]^{(k)} = 0, \tag{16.42}$$

then ξ is a global attractor.

Proof. Firstly, we prove that

$$D_v(F(X), F(Y)) \le \mathcal{I}(F)(\times) D_v(X, Y). \tag{16.43}$$

Since $D_v(F(X),F(Y)) = (D_v(f_1(X),f_1(Y)),\cdots,D_v(f_n(X),f_n(Y)))$, using triangle inequality

$$\begin{aligned} D_v(f_i(X),f_i(Y)) \leq & D_v(f_i(x_1,\cdots,x_n),f_i(y_1,x_2,\cdots,x_n)) \\ & (+)D_v(f_i(y_1,x_2,\cdots,x_n),f_i(y_1,y_2,x_3,\cdots,x_n)) \\ & (+)\cdots \\ & (+)D_v(f_i(y_1,\cdots,y_{n-1},x_n),f_i(y_1,\cdots,y_n)) \\ \leq & (+)_{k=1}^n b_{i,k} D_v(x_k,y_k), \end{aligned}$$

then (16.43) follows.

If $\xi \in \mathcal{D}^n$ is a fixed point and $[\mathcal{I}(F)]^{(k)} = 0$, then using (16.43) we have

$$D_v(F^k(X),\xi) \leq [\mathcal{I}(F)]^{(k)}(\times)D_v(X,\xi) = 0,$$

for any $X \in \mathcal{D}^n$. That means for any initial state X_0, after at most k steps, the state will converge to ξ. □

One can see easily that the condition in Theorem 16.7 is coordinate-independent, since if L^k is a constant mapping (all the columns are the same), then so is $T^{-1}L^kT$ for any coordinate transformation T. But this is not true for the condition in Theorem 16.8. We give an example to illustrate these.

Example 16.9. Consider the following system

$$\begin{cases} x_1(t+1) = [x_1(t) \wedge (x_2(t)\bar{\vee} x_3(x))] \vee (\neg x_1(t) \wedge x_3(t)), \\ x_2(t+1) = [x_1(t) \wedge (\neg x_2(t))] \vee (\neg x_1 \wedge x_2), \\ x_3(t+1) = [x_1(t) \wedge (\neg(x_2(t) \wedge x_3(t)))] \vee [\neg x_1(t) \wedge (x_2(t) \vee x_3(t))]. \end{cases} \tag{16.44}$$

Setting $x(t) = x_1(t)x_2(t)x_3(t)$, we have $x(t+1) = Lx(t)$, where

$$L = \delta_8[8\ 3\ 1\ 5\ 1\ 5\ 3\ 8].$$

Then by Theorem 15.2, we know that the only fixed point of this system is $\delta_8^8 \sim (0,0,0)^T := \xi$. By direct calculation, we have

$$L^3 = \delta_8[8\ 8\ 8\ 8\ 8\ 8\ 8\ 8].$$

Thus, by Theorem 16.7 we conclude that this system is globally stable.

Next, we try to use Theorem 16.8. The incidence matrix of network graph of this system is

$$\mathcal{I}(F) = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}.$$

We can see that $[\mathcal{I}(F)]^{(k)} = \mathbf{1}_{3\times 3}$ for $k \geq 2$, then Theorem 16.8 cannot be employed directly.

However, we consider a coordinate transformation as

$$\begin{cases} z_1 = [x_1 \wedge \neg(x_3)] \vee [(\neg x_1) \wedge (x_2 \bar{\vee} x_3)] \\ z_2 = [x_1 \wedge (x_2 \bar{\vee} x_3)] \vee [(\neg x_1) \wedge x_3] \\ z_3 = x_2. \end{cases}$$

In the vector form, we can easily calculate that

$$z = z_1 z_2 z_3 = Tx,$$

where

$$T = \delta_8[7\ 1\ 6\ 4\ 5\ 3\ 2\ 8].$$

Then in coordinate frame z we have

$$z(t+1) = TLT^T z(t) := \tilde{L} z(t),$$

where

$$\tilde{L} = \delta_8[6\ 6\ 5\ 5\ 7\ 7\ 8\ 8].$$

Retrieving its logical form, we have

$$\begin{cases} z_1(t+1) = 0 \\ z_2(t+1) = z_1(t) \\ z_3(t+1) = z_1(t) \bar{\vee} z_2(t). \end{cases} \tag{16.45}$$

Then its incidence matrix is

$$\mathcal{I}(\tilde{F}) = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix},$$

which is strictly lower triangular, hence $[\mathcal{I}(\tilde{F})]^{(3)} = 0$. And we can see $Z(t) = (z_1(t), z_2(t), z_3(t))^T = (0,0,0)^T$ is the only fixed point of system (16.45), we conclude by Theorem 16.8 that system (16.45) globally converges to $Z_0 = (0,0,0)^T$. Equivalently, system (16.44) globally converges to $X_0 = (0,0,0)^T$.

For a Boolean control network, it is clear that if we can find a control sequence $u(t)$ (or feedback control $u(t) = Gx(t)$) and a proper coordinate transformation $z(t) = Tx(t)$ such that the incidence matrix of network graph of the system under frame $z(t)$ satisfies (16.42), then the system can be stabilized (or stabilized by feedback control). However, so far we have no good method to find out such $u(t)$. Thus in general case, when such control sequence cannot be observed directly, Theorem 16.7 may be more useful.

Next, we give an example for stabilization of Boolean control network.

Example 16.10. Consider the following system

$$\begin{cases} x_1(t+1) = (u_1(t) \leftrightarrow x_2(t)) \wedge x_3(t) \\ x_2(t+1) = \neg x_3(t) \\ x_3(t+1) = (u_2(t)\bar{(}\vee)x_1(t)) \rightarrow x_2(t). \end{cases} \tag{16.46}$$

If we choose $u_1(t) = \neg x_2(t)$ and $u_2(t) = x_1(t)$, then under this feedback control, the system becomes

$$\begin{cases} x_1(t+1) = 0 \\ x_2(t+1) = \neg x_3(t) \\ x_3(t+1) = 1, \end{cases}$$

whose incidence matrix of network graph is

$$\mathcal{I}(F) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

We conclude that the system is globally stabilized to $X = (0,0,1)^T \sim \delta_8^7$ by the given feedback control.

If we can not observe such feedback control, we can also use Theorem 16.7 to verify whether the system is globally stabilizable by open loop control or by feedback control. Here we take the case of open loop control as an example. Setting $x(t) = x_1(t)x_2(t)x_3(t)$ and $u(t) = u_1(t)u_2(t)$, we have

$$x(t+1) = Lu(t)x(t),$$

where

$$L = \delta_8[3\ 5\ 7\ 5\ 3\ 5\ 8\ 6\ 3\ 5\ 8\ 6\ 3\ 5\ 7\ 5 \\ 1\ 5\ 3\ 5\ 7\ 5\ 4\ 6\ 7\ 5\ 4\ 6\ 7\ 5\ 3\ 5].$$

Then,

$$M = \begin{bmatrix} 0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0 \\ 1&0&1&0&1&0&1&0 \\ 0&0&1&0&0&0&1&0 \\ 0&1&0&1&0&1&0&1 \\ 0&0&0&1&0&0&0&1 \\ 1&0&1&0&1&0&1&0 \\ 0&0&1&0&0&0&1&0 \end{bmatrix},$$

and

$$M^{(2)} = \begin{bmatrix} 0\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0 \\ 1\,1\,1\,1\,1\,1\,1\,1 \\ 1\,0\,1\,0\,1\,0\,1\,0 \\ 0\,0\,1\,1\,0\,0\,1\,1 \\ 0\,0\,1\,0\,0\,0\,1\,0 \\ 1\,1\,1\,1\,1\,1\,1\,1 \\ 1\,0\,1\,0\,1\,0\,1\,0 \end{bmatrix}.$$

Hence, we conclude that the system can be stabilized to δ_8^3 or δ_8^7.

16.5.2 *Optimal Control*

This subsection considers the infinite horizon optimal control problem of Boolean networks. For Boolean control network (16.36), our purpose is to find a control sequence $\{u(t)\}$ to maximize the objective function

$$J(u) = \overline{\lim_{T\to\infty}} \frac{1}{T} \sum_{t=1}^{T} P(x(t), u(t)). \tag{16.47}$$

The next result converts the optimal control problem to the problem of finding the cycles of Boolean control network. Then we can use the technique developed in Section 16.1 to solve it.

Theorem 16.9. *For Boolean control network (16.36) with the objective function (16.47), there exists an optimal control $u^*(t)$ such that the objective function is maximized and the trajectory of $s^*(t) = u^*(t)x^*(t)$ will become periodic after a finite time. Moreover, there exists a logical matrix G^*, such that u^* can be expressed as*

$$u^*(t+1) = G^* u^*(t) x^*(t). \tag{16.48}$$

For a length-ℓ cycle, denote

$$P(C) = \frac{1}{\ell} \sum_{t=1}^{\ell} P(x(t), u(t)).$$

A simple cycle is a cycle without duplicated element. For simple cycles we have

Proposition 16.4. *Any cycle C contains a simple cycle C_s such that*

$$P(C_s) \geq P(C). \tag{16.49}$$

According to this proposition, we can find an optimal cycle C^* only from all the simple cycles which can be reached from the initial state. We give an example to describe how to find an optimal control.

Example 16.11. We consider an infinitely repeated game with two players, which is similar to the prisoners' dilemma. Assume player 1 is a machine and player 2 is a person. Their actions can be

0 : the player cooperates with the partner,
1 : the player defects the partner.

The payoff bi-matrix is assumed to be

Table 16.1 Payoff bi-matrix

$P_1\backslash P_2$	0	1
0	3,3	0,5
1	5,0	1,1

Assume player 1 (the machine) fixes its strategy as "One Tit For One Tat", which means player 1 will defect only if player 2 defects in previous step. Denote by $x(t)$ the action of player 1 at tth step, and $u(t)$ the action of player 2. The game can be described by the following Boolean control network.

$$x(t+1) = Lu(t)x(t), \tag{16.50}$$

where

$$L = \delta_2[1\ 1\ 2\ 2].$$

Player 2 (the person) choose his actions $u(t)$, to maximize his own payoff

$$J(u) = \overline{\lim_{T\to\infty}} \frac{1}{T}\sum_{t=1}^{T} P_h(x(t), u(t)),$$

where

$$P_h(x(t), u(t)) = u'(t)\begin{bmatrix}3 & 0\\ 5 & 1\end{bmatrix} x(t).$$

Since the length of a simple cycle can not be bigger than 2, we need only to find all the fixed points and length-2 cycles. By Theorem 16.2 and checking $\mathcal{J}^{(s)}$, we can obtain the simple cycles that can be reached from the initial state $x(0) = \delta_2^1$, which are

$$C_1 = \delta_2 \times \delta_2((1,1)), \quad C_2 = \delta_2 \times \delta_2((2,2)), \quad C_3 = \delta_2 \times \delta_2((1,2),(2,1)).$$

It is easy to see that $P(C_1)$ is the largest. Let $u(0) = \delta_2^1$, the trajectory can enter C_1 directly from initial state.

Finally, we can see that the best strategy for player 2 can be expressed as

$$u^*(t+1) = G^* u^*(t) x^*(t),$$

where

$$G^* = \delta_2[1 \ * \ * \, *], \quad * \text{ is arbitrary in } \mathcal{D}.$$

For example, we can take $G^* = \delta_2[1\ 1\ 1\ 1]$ which means player 2 always cooperates; or $G^* = \delta_2[1\ 2\ 1\ 2]$, which is also "One Tit for One Tat" strategy.

16.5.3 *Identification*

Before the end of this chapter, we consider how to identify a Boolean control network from observed data. First, we give a rigorous definition of identification for Boolean control networks.

Definition 16.10. Assume we have a Boolean control network with its dynamic structure as (16.1)–(16.2). The identification problem is finding the functions f_i, $i = 1, \cdots, n$, h_j, $j = 1, \cdots, p$ (equivalently, finding (L, H) for its algebraic form), via certain input-output data $\{U(0), U(1), \cdots\}$, $\{Y(0), Y(1), \cdots\}$. The identification problem is said to be solvable if f_i and h_j can be uniquely determined by using designed inputs $\{U(0), U(1), \cdots\}$.

Here we use the following notations:

$$\begin{aligned} X(t) &:= (x_1(t), x_2(t), \cdots, x_n(t)); \\ Y(t) &:= (y_1(t), y_2(t), \cdots, y_p(t)); \\ U(t) &:= (u_1(t), u_2(t), \cdots, u_m(t)). \end{aligned}$$

Note that if $z = Tx$ is a coordinate transformation, the system under new coordinate frame becomes

$$\begin{cases} z(t+1) = \tilde{L} u(t) z(t) \\ y(t) = \tilde{H} z(t), \end{cases} \tag{16.51}$$

then (L, H) and $(\tilde{L}, \tilde{H})$ are not distinguishable by any input-output data. So, precisely speaking, we should say: the identification problem is whether the pair (L, H) is identifiable up to a coordinate transformation.

Firstly, assume the state can be observed, or say $H = I_{2^n}$, then we have

Theorem 16.10. *System (16.1) is identifiable by input-state data, if and only if the system is controllable.*

Assume we have enough proper input data $\{U(0), U(1), \cdots, U(T)\}$ and the corresponding state data $\{X(0), X(1), \cdots, X(T)\}$ such that (in set sense)

$$\{U_0 \times X_0, U_1 \times X_1, \cdots, U_{T-1} \times X_{T-1}\} = \mathcal{D}^{n+m}. \tag{16.52}$$

Then L can be identified in the following way: In the vector form, if $u(i)x(i) = \delta^j_{2^{m+n}}$, then $\mathrm{Col}_j(L) = x(i+1)$.

Then the key issue is how to design the input sequence. A reasonable method is to choose inputs randomly, but this method can not ensure (16.52) be satisfied. For this purpose, we have the following method.

Theorem 16.11. *If system (16.1) is identifiable, the logical functions f_i can be determined uniquely by the inputs designed as following:*

$$u(t) = \begin{cases} \delta^i_{2^m}, & \exists s \ s.t. \ x(s) = x(t), u(s) = \delta^{i-1}_{2^m}, \\ & \forall\, t', s < t' < t, x(t') \neq x(t) \\ \delta^1_{2^m}, & otherwise. \end{cases} \tag{16.53}$$

When $x(t)$ enters a cycle, stop the process.

We give an example to illustrate this.

Example 16.12. Consider the following Boolean control network

$$\begin{cases} x_1(t+1) = \neg x_1(t) \vee x_2(t) \\ x_2(t+1) = u(t) \wedge \neg x_1(t) \vee (\neg u(t) \wedge x_1(t) \wedge \neg x_2(t)). \end{cases} \tag{16.54}$$

Setting $x(t) = \ltimes^2_{i=1} x_i(t)$, $u = \ltimes^2_{i=1} u_i(t)$, we have

$$x(t+1) = Lu(t)x(t), \tag{16.55}$$

where

$$L = \delta_4[2\ 4\ 1\ 1\ 2\ 3\ 2\ 2].$$

Checking the controllability matrix, we have

$$\mathcal{C} = (+)^{2^3}_{s=1}(+)^2_{i=1}(\mathrm{Blk}_i(\mathcal{J}^{(s)}_0)) = \begin{bmatrix} 1\ 1\ 1\ 1 \\ 1\ 1\ 1\ 1 \\ 1\ 1\ 1\ 1 \\ 1\ 1\ 1\ 1 \end{bmatrix} > 0.$$

Table 16.2 Input-state data

t	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
$u(t)$	1	1	1	1	1	2	1	1	1	1	2	1	2	1	1	2	1	2	2	1
$x(t)$	1	2	4	1	2	4	2	4	1	2	4	2	4	2	4	1	2	4	2	3

We conclude that the system is identifiable.

We can choose a sequence of controls and the initial state randomly, then the sequence of states can be determined. First, we choose 20 controls, the input-state data are shown in Table 16.2.

Where the number i in $u(t)$ $(x(t))$ means δ_2^i (δ_4^i). We can get L as

$$L = \delta_4[2\ 4 * 1\ 2\ 3 * 2].$$

Some columns of L are not identified, because not all the input-state $\delta_{2^{m+n}}^j$ are reached by the sequence of control chosen randomly.

Then we use (16.53) to obtain the input-state data as in Table 16.3.

Table 16.3 Input-state data

t	0	1	2	3	4	5	6	7	8	9	10	11
$u(t)$	1	1	1	2	2	1	1	1	2	2	2	1
$x(t)$	1	2	4	1	2	3	1	2	4	2	3	2

Hence, L can be identified as

$$L = \delta_4[2\ 4\ 1\ 1\ 2\ 3\ 2\ 2].$$

When we consider input-output data, the case becomes much more complicated. At present, there is no efficient method to identify the system. One can refer to Cheng and Zhao (2011) for an effective algorithm which searches the identification solution one by one in an particular order, but its computation complexity is huge. However, we have a necessary and sufficient condition for identifiability, which is theoretically important.

Theorem 16.12. *The Boolean control network (16.1)–(16.2) is identifiable, if and only if it is controllable and observable.*

Exercises

16.1 Consider a Boolean control system

$$\begin{cases} x_1(t+1) = x_2(t) \vee u(t) \\ x_2(t+1) = x_1(t) \wedge u(t). \end{cases} \tag{16.56}$$

(i) Draw its input-state transfer graph.

(ii) Give its input-state incidence matrix $\mathcal{I}$ from the graph. Check that

$$\mathcal{J} = \begin{bmatrix} L \\ L \end{bmatrix}.$$

16.2 Consider the Boolean control system (16.56). Use formula (16.10) to calculate the number of its fixed points, and use formula (16.11) to calculate its numbers of cycles of length s, $s = 2, 3, \cdots, 8$. Check your result from the input-state transfer graph obtained in previous exercise.

16.3 Given

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Calculate the followings:

(i)

$$A^{(s)}, \quad s = 2, 3, \cdots.$$

(ii)

$$(+)_{i=1}^{s} A^{(i)}, \quad s = 2, 3, \cdots.$$

(iii)

$$A^{\langle s \rangle}, \quad s = 2, 3, \cdots.$$

(iv)

$$\langle + \rangle_{i=1}^{s} A^{\langle i \rangle}, \quad s = 2, 3, \cdots.$$

16.4 Analyze the controllability of the following control systems.

(i)

$$\begin{cases} x_1(t+1) = x_2(t) \\ x_2(t+1) = x_1(t) \wedge u(t). \end{cases} \tag{16.57}$$

(ii)

$$\begin{cases} x_1(t+1) = x_2(t) \\ x_2(t+1) = x_3(t) \\ x_3(t+1) = (x_1(t) \vee u(t)) \bar{\vee} (x_2(t) \vee u(t)). \end{cases} \tag{16.58}$$

16.5 Consider an undirected graph G with n nodes. Assume its adjacency matrix is A. Using the control idea to show that G is connected, if and only if

$$(+)_{i=1}^{n} A^i > 0.$$

Is this result true for undirected graph? (Corresponding connection is assumed to be strongly connected.)

16.6 Consider the system

$$\begin{cases} x_1(t+1) = (x_2(t) \wedge u(t)) \vee \neg(x_1(t) \vee x_2(t) \vee u(t)) \\ x_2(t+1) = \neg(x_1(t) \vee x_2(t)). \end{cases} \tag{16.59}$$

(i) Show that (16.59) is controllable.

(ii) Verify if it is observable with respect to the following outputs: (a) $y = x_1$; (b) $y = x_1 \bar{\vee} x_2$.

16.7 Assume the state space is $\mathcal{X} = \mathcal{F}_\ell(x_1, x_2, x_3)$ and Y is given as follows. Find a minimum Y-friendly subspace.

(i)

$$y = (\neg x_2) \bar{\vee} x_3.$$

(ii)

$$\begin{cases} y_1 = (x_1 \leftrightarrow x_2) \wedge x_3 \\ y_2 = \neg x_3. \end{cases}$$

16.8 Consider the following system

$$\begin{cases} x_1(t+1) = x_1(t) \leftrightarrow (x_3(t) \wedge (u(t) \wedge \xi(t) \\ x_2(t+1) = \neg x_1(t) \\ x_3(t+1) = \xi(t) \wedge u(t) \end{cases} \tag{16.60}$$
$$y(t) = x_1(t) \leftrightarrow x_2(t),$$

where $u(t)$ is control and $\xi(t)$ is disturbance. Solve the disturbance decoupling problem.

16.9 Consider the following system

$$\begin{cases} x_1(t+1) = x_2(t) \leftrightarrow x_3(t) \\ x_2(t+1) = x_3(t) \vee u_1(t) \\ x_3(t+1) = x_1(t) \wedge u_2(t) \end{cases} \tag{16.61}$$
$$y_1(t) = x_1(t)$$
$$y_2(t) = x_2(t) \vee x_3(t).$$

Is the system identifiable?

16.10 Consider system (16.61) again.

(i) How many length-3 cycles does this system have?

(ii) Theorems 16.1 and 16.2 tell us the number of cycles of Boolean control network (16.1). In fact, by checking $\mathcal{J}^{(s)}$, we can also find what the cycles are. Try to find all the length-3 cycles in Exercise 16.1.

16.11 Consider the following system

$$\begin{cases} x_1(t+1) = \neg x_2(t) \vee u_1(t) \\ x_2(t+1) = x_3(t) \leftrightarrow \neg x_1(t) \\ x_3(t+1) = \neg x_1(t) \vee u_2(t). \end{cases} \tag{16.62}$$

Can the system be stabilized? If so, find a control sequence or a feedback control which stabilizes the system.

16.12 Consider a Boolean network, which has its algebraic form as

$$x(t+1) = Lx(t), \text{where } \ L \in \mathcal{L}_{n\times n}. \tag{16.63}$$

The identification of Boolean networks is much easier than that of control Boolean network. Prove the following rule: if we have $x(t) = \delta_n^i$ and $x(t+1) = \delta_n^j$, then

$$\mathrm{Col}_i(L) = \delta_n^j.$$

16.13 Identify the Boolean network with 3 nodes using the following observed data:

(i) There are two sets of observed data as

$$\begin{aligned} D_1 = \{&X(1) = (1,0,0), X(2) = (0,0,0), \ \ X(3) = (1,1,1), \\ &X(4) = (1,0,1), \ \ X(5) = (1,0,0)\}; \\ D_2 = \{&X(1) = (0,1,1), X(2) = (0,1,0), \ \ X(3) = (0,0,1), \\ &X(4) = (1,1,0), \ \ X(5) = (0,0,0)\}. \end{aligned}$$

(ii) There are three sets of observed data as

$$\begin{aligned} D_1 = \{&X(1) = (0,0,0), X(2) = (0,0,1), \ \ X(3) = (0,1,0), \\ &X(4) = (0,1,1), \ \ X(5) = (0,0,0)\}; \\ D_2 = \{&X(1) = (1,0,0), X(2) = (1,0,1), \ \ X(3) = (1,1,0), \\ &X(4) = (0,0,0)\}; \\ D_3 = \{&X(1) = (1,1,1), X(2) = (0,1,1)\}. \end{aligned}$$

(Hint: Use the rule obtained in previous exercise.)

16.14 The Boolean control network (16.1) can also be expressed in algebraic form as

$$x(t+1) = \bar{L}x(t)u(t),$$

where $\bar{L} = LW_{[2^n,2^m]}$. Similar to the construction of input-state incidence matrix, we can construct the state-input incidence matrix of (16.1) as

$$\bar{\mathcal{J}} = \mathbf{1}_{2^m} \rtimes \bar{L},$$

where $\rtimes$ is the right semi-tensor product of matrices. Prove,

(i) There exists $\bar{\mathcal{J}}_0^s$ such that

$$\bar{\mathcal{J}}^s = \mathbf{1}_{2^m} \rtimes \bar{\mathcal{J}}_0^s, s \geq 1.$$

(ii) $\bar{\mathcal{J}}_0^s = \tilde{\mathcal{J}}_0^s$, where $\tilde{\mathcal{J}}_0^s$ is defined in (16.22).

16.15 Prove Theorem 16.6. (Hint: Please refer to Cheng (2011).)

16.16 Prove Proposition 16.3. (Hint: Please refer to Cheng *et al.* (2011c).)

16.17 If $x_i(t), u_j(t), y_\alpha(t) \in \mathcal{D}_k$ in (16.1) and (16.2), we call the system a logical control network. In fact, all the properties of Boolean control networks can be generalized to logical control networks.

Consider the following infinitely repeated game. Both of player 1 and player 2 have three actions, $\{L, M, R\}$. The payoff bi-matrix is assumed to be the Table 16.4.

Table 16.4 Payoff bi-matrix

$P_1\backslash P_2$	L	M	R
L	$3, 3$	$0, 4$	$9, 2$
M	$4, 0$	$4, 4$	$5, 3$
R	$2, 9$	$3, 5$	$6, 6$

Assume player 2's strategy is fixed to play R in the first stage, in the tth stage, if the outcome in the $(t-1)$th stage is (R, R) then plays R, otherwise, plays M.

(i) Represent the game into a logical control network.

(ii) Find a best strategy of player 1 which maximizes his infinite horizon average payoff.

This page intentionally left blank

Chapter 17

Game Theory

Game theory is a young branch of mathematics, formally born in the forties of the last century. Since then it has been growing very fast, and nowadays it is a very important tool in economy, biology, social science, and military competition etc. When the game is infinitely repeated and the strategies of each players are finite, the game can be described as a logical dynamic system via semi-tensor product. This chapter will investigate such dynamic games. Our main goal is to find the Nash equilibriums, sub-Nash equilibriums, and/or local Nash/sub-Nash equilibriums of such dynamic games. We refer to Gibbons (1992); Fudenberg and Tirole (1991) for some basic concepts and fundamental results of game theory.

17.1 An Introduction to Game Theory

Game theory is a tool for studying a wide variety of human, or nature behaviors. It has a very long historical background. Gambling, playing cards or chess etc. may be considered as the origin of game theory. The famous Chinese story of the Tian Ji's strategy on horse racing is a typical example of game theory.

Although some developments occurred before it, the foundation of modern game theory was initiated by the book *Theory of Games and Economic Behavior* by John von Neumann and Oskar Morgenstern in 1944. Neumann et al were mainly interested in cooperative games. But most important games are uncooperative (competitive) ones. This theory was developed extensively in the 1950s by many scholars. The most significant contribution was the Nash equilibrium, named after John Forbes Nash. It becomes a fundamental tool for uncooperative games.

Game theory was later explicitly applied to biology in the 1970s, although similar developments go back at least as far as the 1930s. Game theory has been widely recognized as an important tool in many fields. Eight game theorists, including Nash, have won the Nobel Memorial Prize in economic sciences, and John Maynard Smith was awarded the Crafoord Prize for his application of game theory to biology. Today, “game theory is a sort of umbrella or ‘unified field’ theory for the rational side of social science, where ‘social’ is interpreted broadly, to include human as well as non-human players (computers, animals, plants)” (Aumann 1987).

A game contains at least three basic factors:

(i) Players: Usually, a game contains more than one players. When there is only one player, the game becomes an optimization problem. Throughout this chapter we assume there are only finite players, denoted by $P_1, ..., P_n$, where $1 < n < \infty$.

(ii) Actions : Each player in the game has some playing options, which are called the actions of this player. Denote the set of actions of P_i by A_i. When A_i is a finite set, we denote it by

$$A_i = \left\{a_i^1, \cdots, a_i^{k_i}\right\}, \quad i = 1, \cdots, n.$$

In some books the action is called the strategy. We reserve “strategy” for the way to select actions.

(iii) Payoff functions: The payoff function of player i, denoted by c_i, is the gain of player P_i. It depends on the actions of all players. That is,

$$c_i = c_i(x_1, \cdots, x_n), \quad x_j \in A_j, \quad j = 1, ..., n; \quad i = 1, ..., n.$$

Roughly speaking, Nash equilibrium is the solution for a game. We recall the definition of Nash equilibrium, which was given in Chapter 1:

Definition 17.1. A set of combined actions $(x_1^*, \cdots, x_n^*)$, $x_i^* \in A_i$, $i = 1, \cdots, n$, is called a Nash equilibrium, if

$$c_j(x_1^*, \cdots, x_n^*) \geq c_j(x_1^*, \cdots, x_j, \cdots, x_n^*), \quad \forall\, x_j \in A_j, \quad j = 1, \cdots, n. \tag{17.1}$$

Some examples, including the famous prisoner’s dilemma, have been discussed in Chapter 1. In the following we give some more examples. We use the following example to introduce the best response function.

Example 17.1 (Cournot model of duopoly). Let x and y be the quantities of a product by firms 1 and 2 respectively. Let

$$p(Q) = \begin{cases} a - Q, & Q < a \\ 0, & Q \geq a \end{cases}$$

be the market-clearing price, where $Q = x + y$. Assume the cost for producing unit product is a constant b. Following Cournot, suppose that the firms choose their quantities simultaneously.

Then the payoff functions are

$$\begin{cases} c_1(x,y) = [a-(x+y)]x - bx \\ c_2(x,y) = [a-(x+y)]y - by. \end{cases} \tag{17.2}$$

Now P_1 want to choose x to maximize c_1. To find such x, calculating

$$\frac{\partial c_1}{\partial x} = a - 2x - y - b := 0$$

yields

$$x = \frac{1}{2}(a-b-y). \tag{17.3}$$

(17.3) is called the best response function of P_1, which shows for each action of P_2 what the best reaction of P_1 should be. Similarly, we have the best response function of P_2 as

$$y = \frac{1}{2}(a-b-x). \tag{17.4}$$

Then (assume $x, y < a - b$) the Nash equilibrium is the solution of best response functions

$$\begin{cases} x = \frac{1}{2}(a-b-y) \\ y = \frac{1}{2}(a-b-x). \end{cases}$$

That is,

$$\begin{cases} x = \frac{1}{3}(a-b) \\ y = \frac{1}{3}(a-b). \end{cases}$$

Remark 17.1. In fact, the method for finding Nash equilibrium proposed in Chapter 1 (Example 1.8) is also finding the (discrete) best response functions for each players. Then the common set(s) is(are) the solution(s) of the best response functions.

We give another example.

Example 17.2. A Chinese ancient general Tian Ji was gambling with King Qi Wei Wang via three times horse racing. Both Tian and Qi have 3 horses, denoted by $T = \{t_1, t_2, t_3\}$ and $Q = \{q_1, q_2, q_3\}$ respectively. We know that their corresponding velocities satisfy

$$v_{t_3} < v_{q_3} < v_{t_2} < v_{q_2} < v_{t_1} < v_{q_1}.$$

That is, q_1 is the fastest one. The action sets of P_T and P_Q are the same as

$$A_T = A_Q = \{(123), (132), (213), (231), (312), (321)\},$$

where (123) means that P_T chooses the order of the racing horses as t_1, t_2, t_3, (or P_Q chooses this order), etc. Then we have the payoff bi-matrix in Table 17.1.

Table 17.1 Tian Ji horse racing

$P_T \backslash P_Q$	(123)	(132)	(213)	(231)	(312)	(321)
(123)	-3,$\underline{3}$	-1,1	-1,1	$\underline{1}$,-1	-1,1	-1,1
(132)	-1,1	-3,$\underline{3}$	$\underline{1}$,-1	-1,1	-1,1	-1,1
(213)	-1,1	-1,1	-3,$\underline{3}$	-1,1	-1,1	$\underline{1}$,-1
(231)	-1,1	-1,1	-1,1	-3,$\underline{3}$	$\underline{1}$,-1	-1,1
(312)	$\underline{1}$,-1	-1,1	-1,1	-1,1	-3,$\underline{3}$	-1,1
(123)	-1,1	$\underline{1}$,-1	-1,1	-1,1	-1,1	-3,$\underline{3}$

It is easy to see that there is no Nash equilibrium.

In Example 17.2, each player must guess the other's strategy. If a player know the other's strategy, he can win. This happens in many other games. In the story of Tian Ji horse racing, it is obvious that Tien Ji was on a weak position that his horses are not as good as Qi's. But he knew Qi's strategy, which is (123). Hence he took (312) and won the overall game.

In any game in which each player would like to outguess the other(s), there is no Nash equilibrium because the solution to such a game necessarily involves uncertainty about what the players will do. Then we need to consider strategies with uncertainty. Such strategies are called the mixed strategy. To distinguish this kind of strategies with the previous strategies, we call the strategies without uncertainties pure strategy. We give a rigorous definition for mixed strategy.

Definition 17.2. Assume in a game a player has his action set as $S = \{s_\lambda \mid \lambda \in \Lambda\}$. For a mixed strategy s there is a probability distribution p_λ, $\lambda \in \Lambda$, with $p_\lambda \geq 0$ and $\sum\limits_{\lambda\in\Lambda} p_\lambda = 1$. The player taking this strategy s means he choose action s_λ with probability p_λ. That is,

$$P(s = s_\lambda) = p_\lambda, \quad \lambda \in \Lambda.$$

We give an example to describe it.

Example 17.3 (Matching Pennies). In this game, each player's action set is $S = \{Head, Tail\}$. The payoff bi-matrix is in Table 17.2.

Table 17.2 Matching pennies

$P_1 \backslash P_2$	Head(H)	Tail(T)
Head(H)	-1,1	1,-1
Tail(T)	1,-1	-1,1

To accompany the payoffs in the bi-matrix, imagine that each player has a penny and must choose whether to display in with heads or tails facing up. If the tow pennies match then player 2 (P_2) wins player 1's (P_1) penney, otherwise, P_1 wins $P_2's$ penny.

It is obvious that there is no Nash equilibrium for pure strategies. Then we consider the mixed strategies. Assume P_1 plays H with probability p and T with probability $1-p$. Correspondingly, P_2 plays H with probability q and T with probability $1-q$. Then the expected payoffs for P_1 and P_2, denoted by E_1 and E_2 are, respectively,

$$\begin{cases} E_1 = -pq + p(1-q) + (1-p)q - (1-p)(1-q); \\ E_2 = pq - p(1-q) - (1-p)q + (1-p)(1-q). \end{cases} \tag{17.5}$$

A simple argument shows that the best strategy of P_1, responding to different q, is

$$p = \begin{cases} 0, & q > 0.5 \\ [0,1], & q = 0.5 \\ 1, & q < 0.5. \end{cases} \tag{17.6}$$

We call (17.6) the best response correspondence. The reason we do not call it the best response function is that (17.6) is not a function. Hence, the best response correspondence is a generalization of best response function. Similarly, we have the best response correspondence of P_2 is

$$q = \begin{cases} 1, & p < 0.5 \\ [0,1], & p = 0.5 \\ 0, & p > 0.5. \end{cases} \tag{17.7}$$

Now the only common solution is

$$\begin{cases} p = 0.5 \\ q = 0.5, \end{cases}$$

which is the Nash equilibrium.

The following result is fundamental (Gibbons, 1992).

Theorem 17.1. *In a finite game G, which has finite players, and each player P_i has finite set of actions, there exists at least one Nash equilibrium, possibly involving mixed strategies.*

Theorem 17.1 ensures that a finite game must have at least one Nash equilibrium, but many games have several Nash equilibria. In this case, it is hard to see what the right prediction is. For these, many refinement of Nash equilibria were proposed, such as Pareto-dominant equilibrium, coalition-proof equilibrium and so on (Fudenberg and Tirole, 1991). However, in this chapter, we only consider pure strategies of finite games, so the Nash equilibrium may not exist. Then we may look for a "weaker" solution, which is called a sub-Nash equilibrium.

Definition 17.3.

(1) Given a combined actions $(x_1, x_2, \cdots, x_n)$. Then we can find a non-negative real number $\varepsilon^s \geq 0$, such that

$$c_j(x_1, \cdots, x_j, \cdots, x_n) + \varepsilon^s \geq c_j(x_1, \cdots, x_{j-1}, x'_j, x_{j+1}, \cdots, x_n), \quad \forall\, x'_j \in A_j, j = 1, \cdots, n. \tag{17.8}$$

The smallest $\varepsilon^s \geq 0$, satisfying (17.8), is called a tolerance of $(x_1, x_2, \cdots, x_n)$.

(2) $(x_1, x_2, \cdots, x_n)$ is called a sub-Nash equilibrium if it has the smallest tolerance.

It is easy to see that a Nash equilibrium is a sub-Nash equilibrium with tolerance 0. Thus, sub-Nash equilibrium is a generalization of Nash equilibrium.

We give some examples to illustrate it.

Example 17.4. Consider a game with two players A and B. The payoff bi-matrix is

Table 17.3 Payoff bi-matrix

$A\backslash B$	1	2
1	$\underline{2}$,0	0,$\underline{2}$
2	1,$\underline{2}$	$\underline{2}$,1

It is obvious that there is no Nash equilibrium. It is easy to calculate that the tolerances are as in Table 17.4.

Table 17.4 Tolerances

$A\backslash B$	1	2
1	2	2
2	1	1

Hence $(2,1)$ and $(2,2)$ are sub-Nash equilibriums with tolerance 1. However, It is very likely that A may not be satisfied with $(2,1)$ and B may not be satisfied with $(2,2)$.

Example 17.5. Recall the Tian Ji horse racing in Example 17.2. It is easy to obtain the tolerances are

Table 17.5 Tolerances of Tian Ji horse racing

$P_T \backslash P_Q$	(123)	(132)	(213)	(231)	(312)	(321)
(123)	4	2	2	4	2	2
(132)	2	4	4	2	2	2
(213)	2	2	4	2	2	4
(231)	2	2	2	4	4	2
(312)	4	2	2	2	4	2
(123)	2	4	2	2	2	4

We can see that the minimum tolerance is 2, and there are many sub-Nash equilibria.

17.2 Infinitely Repeated Games

In the games discussed in last section, the players choose their actions simultaneously and the games are played only once. We call this kind of games the static games. Other games are called the dynamic games which contains the following information:

- the set of players.
- the order of moves.
- the players' payoffs as the function of the actions that were made.
- what the actions the players choose when they move.
- what each player knows when he move.
- the probability distributions over any exogenous events.

Here, the point 6 may be hard to understand. For example, consider a game between two players. Assume that they have two payoff bi-matrices, and both of them know the probability of each bi-matrix to be used, but only player 1 knows which bi-matrix is used. Then game can be considered as a game that there is a player called "Nature" who will move firstly to choose a payoff bi-matrix randomly, and then player 1 and player 2 will move simultaneously according to what they know (player 1 knows the chosen bi-matrix but player 2 does not).

When it is player i's turn to move, denote H_i the set of possible historical information the player knows and A_i the set of possible actions the player can choose. A strategy s_i of player i is a mapping $s_i : H_i \to A_i$, that is, the way to choose his action according to his knowledge of historical information. Replacing "actions" by "strategies" in Definition 17.1 and 17.3, we get the concepts of Nash equilibrium and sub-Nash equilibrium for dynamic games.

Example 17.6. Recall the Cournot model in Example 17.1. We now suppose player 1 as the "leader", who chooses his action first, and then player 2 chooses his own action after observing player 1's action. Thus, player 2's strategies are functions $s_2 : X \to Y$ where X and Y are feasible actions set of player 1 and player 2, while player 1's strategies are simply choosing x from $[0, a]$.

After observing player 1's action x, the best strategies of player 2 is to choose y to maximize $f_2(x, y)$. Thus the best strategy for player 2 is the function

$$s_2(x) = \frac{1}{2}(a - b - x).$$

Player 1 also knows player 2's best strategy, thus he only need to maximize $f_1(x, s_2(x))$. Hence we have the Nash equilibrium

$$\begin{cases} x = \frac{a-b}{2} \\ s_2(x) = \frac{1}{2}(a - b - x). \end{cases}$$

And the outcome of this equilibrium is

$$\begin{cases} x = \frac{a-c}{2} \\ y = \frac{a-c}{4}, \end{cases}$$

which is different from the Nash equilibrium of static Cournot model.

Although the prediction in the above example is obtained in a very natural way, it is not the unique Nash equilibrium (we leave it as an exercise to find another Nash equilibrium). To deal with this problem, a refinement of Nash equilibrium for dynamic games named subgame-perfect equilibrium was proposed. And the equilibrium found in the above example is the unique subgame-perfect equilibrium. Roughly speaking, a subgame-perfect equilibrium is a Nash equilibrium for every subgame. Readers can refer to Fudenberg and Tirole (1991); Gibbons (1992) for details.

In the following, we only concern with infinitely repeated games, since it can be described as mix-valued logical (control) networks which can be converted into its algebraic form using semi-tensor product as in Chapter 14.

Definition 17.4. Consider the infinitely repeated game G_∞ of G.

(1) A strategy profile is

$$s = (s_1, \cdots, s_n), \tag{17.9}$$

where s_j is a sequence of logical functions of time, called the strategy of player j, precisely,

$$s_j = \left\{ x_j(0), s_j^t \middle| t = 1, 2, \cdots \right\}, \quad j = 1, \cdots, n,$$

where s_j^t is a function of historical actions, precisely

$$x_j(t) = s_j^t(x_1(0), \cdots, x_n(0), \cdots, x_1(t-1), \cdots, x_n(t-1)).$$

Denote by S_∞ the set of strategy profiles.

(2) The players' payoffs as the function of all the actions that were made. In this chapter, we suppose the payoff functions are the averaged payoffs of all stages

$$J_j(s) = \overline{\lim_{T\to\infty}} \frac{1}{T} \sum_{t=1}^{T} c_j(x_1(t), \cdots, x_n(t)), \quad j = 1, \cdots, n. \tag{17.10}$$

Thus we can use the results of Section 13.5.2.

In this definition, we assume each player knows all the historical information when he moves. If they have finite memories, that is, each player only remember the information of previous finite steps when he moves, we have the following μ-memory strategies.

Definition 17.5. Consider the dynamic game G_∞ of G. A μ-memory strategy is a strategy, where the action $x_j(t+1)$ depends on the past μ historical actions. Precisely, the strategy s_j is generated by

$$\begin{aligned} x_j(t+1) = f_j(&x_1(t), \cdots, x_n(t), \cdots, \\ &x_1(t-\mu+1), \cdots, x_n(t-\mu+1)), \quad t \geq \mu - 1, \end{aligned} \tag{17.11}$$

with initial values

$$x_j(t) = x_j^t, \quad t \leq \mu - 1,\ j = 1, \cdots, n. \tag{17.12}$$

Equivalently, we can also denote the set of initial values as

$$X_0 = \left\{ x_1^0, \cdots, x_n^0,\ \cdots,\ x_1^{\mu-1}, \cdots, x_n^{\mu-1} \right\}.$$

We can see that if the action sets $A_i, i = 1, \cdots, n$ are finite, (17.11) is essentially the dynamics of mix-valued logical control networks. Identifying

$$a_i^j \sim \delta_{k_i}^j, \quad i = 1, \cdots, n \quad j = 1, \cdots, k_i,$$

we have $x_i \in \Delta_{k_i}$. Setting $k = \prod_{i=1}^n k_i$ and $x = \ltimes_{i=1}^n x_i \in \Delta_k$, (17.11) can be converted into their algebraic form as

$$x_j(t+1) = L_j \ltimes_{i=0}^{\mu-1} x(t-i), \quad j = 1, \cdots, n, \tag{17.13}$$

where $L_j \in \mathcal{L}_{k_j \times k^\mu}$ is the structure matrix of f_j. Multiplying the equations in (17.13) together, we have the algebraic form of μ-memory strategies profile

$$x(t+1) = L \ltimes_{i=0}^{\mu-1} x(t-i), \tag{17.14}$$

with initial states X_0. Since (17.11), (17.13), and (17.14) are all equivalent, and it is easy to convert them from one form to another, we simply use $(L_1, \cdots, L_n; X_0)$ or (L, X_0) instead of the corresponding strategy profile $s = (s_1, \cdots, s_n)$.

17.3 Local Optimization of Strategies and Local Nash/Sub-Nash Equilibrium

In Section 13.5.2, we have already investigated the optimization of Boolean control networks. When the scale of network is larger, it is difficult to find out the optimal control. In this section, a distance of strategies is investigated, which is proposed firstly by (Cheng *et al.*, 2010). Using this distance, local optimization of strategies and local Nash/sub-Nash equilibrium are investigated.

Recall Definition 16.9. The vector distance $D_v(A, B)$ of two Boolean matrices A and B are defined. If $A = (a_{ij}) \in \mathcal{B}_{m \times n}$, we denote

$$\|A\| = \sum_{i=1}^m \sum_{j=1}^n a_{ij}. \tag{17.15}$$

It is easy to check that $\|\cdot\|$ is a norm.

Next, we define a distance for two Boolean matrices of the same dimension.

Definition 17.6. Let $A = (a_{ij}), B = (b_{ij}) \in \mathcal{B}_{m \times n}$. Then the distance between A and B, denoted by $d(A, B)$, is defined as

$$d(A, B) := \frac{1}{2} \left\| D_v(A, B) \right\|. \tag{17.16}$$

The following result is an immediate consequence of the definition, we leave it for exercise.

Theorem 17.2. $(\mathcal{B}_{m\times n}, d)$ *is a metric space. That is,*

(i)

$$d(A,B)=0 \Leftrightarrow A=B, \quad \forall\, A,B\in\mathcal{B}_{m\times n};$$

(ii)

$$d(A,B)=d(B,A), \quad \forall\, A,B\in\mathcal{B}_{m\times n};$$

(iii)

$$d(A,C)\le d(A,B)+d(B,C), \quad \forall\, A,B,C\in\mathcal{B}_{m\times n}.$$

Using this distance to the set of μ-memory strategy profiles, it follows that $(\mathcal{L}_{k\times k^\mu}, d)$ is a metric subspace of $(\mathcal{B}_{k\times k^\mu}, d)$. Next, we consider the physical meaning of this d on the set of μ-memory strategy profiles. The following proposition is obvious.

Proposition 17.1. *Let* $A,B\in\mathcal{L}_{k\times k^\mu}$. *Then*

$$|\{i\,|\mathrm{Col}_i(A)\neq \mathrm{Col}_i(B)\,\}| = d(A,B). \tag{17.17}$$

When $\mu=1$, we have more clear description for the geometric meaning of this distance. The proof of the following proposition is left for exercise.

Proposition 17.2. *Assume* $\mu=1$. *Then the distance* $d(L_1,L_2)$ *is the number of different edges between the state transfer graphs of the systems* $x(t+1)=L_1x(t)$ *and* $x(t+1)=L_2x(t)$.

As for $\mu>1$ case, since $x(t+1)$ depends on μ historical strategy profiles, we define a path

$$x(t-\mu+1)\to x(t-\mu+2)\to\cdots\to x(t)\to x(t+1)$$

as a compounded edge. The strategy dynamic graph of a μ-memory strategy profile consists of all such compounded edges. In fact, they form a μ-homogeneous hypergraph. Then the following corollary is clear.

Corollary 17.1. *Assume* $\mu>1$. *Then the distance* $d(L_1,L_2)$ *between two* μ*-memory strategy profiles is the number of different compounded edges between the strategy dynamic hypergraphs of* L_1 *and* L_2.

The results of Theorem 16.9 can be extended to μ-memory mix-valued logical control networks (see Zhao *et al.* (2010b)). The infinite horizon optimization of μ-memory mix-valued logical control network

$$\begin{cases} x_1(t+1) = f_1(x_1(t),\cdots,x_n(t),\cdots,x_1(t-\mu+1),\cdots,x_n(t-\mu+1),\\ \qquad u_1(t),\cdots,u_m(t),\cdots,u_1(t-\mu+1),\cdots,u_m(t-\mu+1))\\ \vdots\\ x_n(t+1) = f_n(x_1(t),\cdots,x_n(t),\cdots,x_1(t-\mu+1),\cdots,x_n(t-\mu+1),\\ \qquad u_1(t),\cdots,u_m(t),\cdots,u_1(t-\mu+1),\cdots,u_m(t-\mu+1)), \end{cases} \tag{17.18}$$

where $x_i \in \mathcal{D}_{k_i}$, $u_i \in \mathcal{D}_{s_j}$ can be considered as an infinitely repeated game in which some players $(x_i, i=1,2,\cdots,n)$ have fixed their strategies as f_i with initial actions $x_i^j, j=0,\cdots,\mu-1$, and other players $(u_j, j=1,\cdots,m)$ can choose their strategies freely and they have a common payoff function

$$J = \overline{\lim_{T\to\infty}}\frac{1}{T}\sum_{t=1}^{T} c(x_i(t),u_j(t)), \quad ,i=1,\cdots,n, \quad ,j=1,\cdots,m. \tag{17.19}$$

Setting $x(t) = \ltimes_{i=1}^n x_i(t), u(t) = \ltimes_{j=1}^m u_j(t), k = \prod_{i=1}^n k_i, s = \prod_{j=1} m s_j$, the algebraic form of (17.18) is

$$x(t+1) = Lu(t-\mu+1)\cdots u(t)x(t-\mu+1)\cdots x(t). \tag{17.20}$$

The optimal control has the form of

$$u(t+1) = Gu(t-\mu+1)\cdots u(t)x(t-\mu+1)\cdots x(t). \tag{17.21}$$

Multiplying both sides of (17.20) and (17.21) together and setting $w(t) = u(t)x(t)$, yields

$$w(t+1) = \Psi(G)w(t). \tag{17.22}$$

We leave the expression of $\Psi(G)$ as an exercise.

For every G, we can calculate $\Psi(G)$, and then find the cycles of (17.18) for every initial $u(0),\cdots,u(\mu-1)$. Thus, by comparing the payoffs we can find the optimal control. But in general, searching all $G \in \mathcal{L}_{s\times sk}$ to find an optimal solution is unrealistic because of the computation complexity. Using the distance of logical matrices, at each step, we can look for only a local optimal solution. That is, look for optimal solution (G, U_0) ($U_0 := \ltimes_{t=0}^{\mu-1} u(t)$) over a neighborhood

$$B_\varepsilon(G^0, U_0^0) = \left\{ (G, U_0) \in \mathcal{L}_{s\times sk} \times \Delta_{s^\mu} \,|\, d(G, G^0) \le \varepsilon, d(U_0, U_0^0) \le \varepsilon \right\}.$$

Set the default $\varepsilon = 1$.

We give an example to illustrate this.

Example 17.7. Consider a Boolean network

$$x(t+1) = Lu(t)x(t),$$

where

$$L = \delta_2[1\ 2\ 2\ 1].$$

Assume

$$c(x(t), u(t)) = u'(t) \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} x(t)$$

and $x^0 = \delta_2^2$. Choosing $G^0 = \delta_2[1\ 2\ 2\ 1]$, $u_0^0 = \delta_2^2$, we can get that in step 1

$$\mathcal{G}^1 = \{(\delta_2[1\ 2\ 2\ 2], \delta_2^1), (\delta_2[1\ 2\ 2\ 2], \delta_2^2)\}$$

is the set of optimal strategies in $B_1(G_0; u_0^0)$. We choose $G^1 = \delta_2[1\ 2\ 2\ 2]$, $u_1^0 = \delta_2^1$, then

$$\begin{aligned} \mathcal{G}^2 = \{&(\delta_2[1\ 2\ 2\ 2], \delta_2^1), (\delta_2[1\ 2\ 2\ 2], \delta_2^2), \\ &(\delta_2[2\ 2\ 2\ 2], \delta_2^1), (\delta_2[2\ 2\ 2\ 2], \delta_2^2)\} \end{aligned}$$

is the set of optimal strategies in $B_1(G^1; u_0^1)$ Thus, $(\delta_2[1\ 2\ 2\ 2], \delta_2^1)$ and $(\delta_2[1\ 2\ 2\ 2], \delta_2^2)$ are local optimal controls. We can check that they are also optimal controls.

Next, we consider the local Nash/sub-Nash equilibrium.

Definition 17.7. For infinitely repeated game G_∞, a μ-memory strategy profile $s = (L; X_0)$ is called a local Nash/sub-Nash equilibrium on $B_\varepsilon(L, X_0)$, if it is a Nash/sub-Nash equilibrium with respect to its ε-neighborhood.

Since there may exits too many Nash/sub-Nash equilibria, in this section we are interested in the initial value independent strategies. We define the common Nash equilibrium.

Definition 17.8. $L^* = ({L_1}^*, {L_2}^*, \cdots, {L_n}^*)$ is called a common Nash equilibrium, if it, combining with any set of initial values, is a Nash equilibrium. Precisely, $\forall\, j = 1, \cdots, n$,

$$\begin{aligned} J_j({L_1}^*, \cdots, {L_j}^*, \cdots, {L_n}^*; X_0) \geq J_j({L_1}^*, \cdots, L_j', \cdots, {L_n}^*; X_0), \\ \forall\, L_j' \in \mathcal{L}_{k_j \times k^\mu}, \forall\, X_0 \in \textstyle\prod_{i=1}^n \mathcal{D}_{k_i}^\mu. \end{aligned} \tag{17.23}$$

It is easy to check that a common Nash equilibrium is a subgame-perfect equilibrium. If the common Nash equilibrium does not exist, we may look for an initial-independent sub-Nash solution. To make it precise, we give the following definition.

Definition 17.9.

(1) For $L = (L_1, L_2, \cdots, L_n)$, we can find a nonnegative real number $\varepsilon \geq 0$, such that, $\forall\, j = 1, \cdots, n$,

$$J_j(L_1, \cdots, L_j, \cdots, L_n; X_0) + \varepsilon \geq J_j(L_1, \cdots, L_j', \cdots, L_n; X_0), \quad \forall\, L_j' \in \mathcal{L}_{k_j \times k^\mu},\ \forall\, X_0 \in \prod_{i=1}^n \mathcal{D}_{k_i}^\mu. \tag{17.24}$$

The smallest $\varepsilon \geq 0$, satisfying (17.24), is called a tolerance of L.

(2) L is called an initial-independent sub-Nash equilibrium if it has the smallest tolerance.

Similar to the local optimization, in each step, we can find a strategy profile with minimum tolerance in the neighborhood. We describe the algorithm as follows.

Algorithm 4.

- Step 0. Choose an initial strategy profile L^0, set $\mathcal{H} = \{L^0\}$.
- $\cdots$
- Step p. On the neighborhood

  $$B_\varepsilon(L^{p-1}) = \{L | d(L, L^{p-1}) \leq \varepsilon\}$$

  search sub-Nash equilibrium(s), denoted as

  $$\mathcal{L}^p = \{L^{p_1}, L^{p_2}, \cdots, L^{p_{k_p}}\}.$$

  – If $L^{p-1} \in \mathcal{L}^p$, choose L^{p-1} as a local sub-Nash equilibrium (the solution) and stop.

  – Else,

    (a) If $\mathcal{L}^p \cap \mathcal{H}^c = \varnothing$:

      ∗ If $p = 1$, no local sub-Nash equilibrium is found (the algorithm fails) and stop.

      ∗ Else, go back to Step $p-1$ to choose another L^{p-1} if possible.

    (b) Else: Choose $L^p \in \mathcal{L}_p \cap \mathcal{H}^c$, and add L^p to $\mathcal{H}$.
- $\cdots$

We give an example to describe this.

Example 17.8. Consider the infinitely repeated game of prisoners' dilemma. The payoff bi-matrix is shown in Table 17.6.

Table 17.6 Payoff bi-matrix

$P_1\backslash P_2$	1	2
1	3,3	0,5
2	5,0	1,1

Choose $L^0 = \delta_4[1\ 1\ 1\ 3]$, using Algorithm 4, we have

$$\begin{aligned} &L^1 = \delta_4[1\ 4\ 1\ 3],\ \ L^2 = \delta_4[1\ 3\ 1\ 3],\ \ L^3 = \delta_4[1\ 3\ 4\ 3],\\ &L^4 = \delta_4[3\ 3\ 4\ 3],\ \ L^5 = \delta_4[3\ 3\ 4\ 2],\ \ L^6 = \delta_4[3\ 4\ 4\ 2],\\ &L^7 = \delta_4[4\ 4\ 4\ 2],\ \ L^8 = \delta_4[1\ 4\ 4\ 2],\ \ L^9 = \delta_4[2\ 4\ 4\ 2],\\ &L^{10} = \delta_4[2\ 4\ 4\ 4],\ L^{11} = \delta_4[1\ 4\ 4\ 4]. \end{aligned}$$

The algorithm terminates at $k = 11$, L^{11} is a locally Nash equilibrium, which is also a 1-memory Nash equilibrium.

The algorithm can not always terminated at a Nash equilibrium. For example, let $L^0 = \delta_4[2\ 1\ 2\ 3]$. Then we can find a locally Nash equilibrium $L = \delta_4[1\ 4\ 2\ 1]$ which is not a global Nash equilibrium.

Remark 17.2. Since the tolerance of a strategy profile depends on the neighborhood, thus the value of tolerance may not be decreasing . Thus, it is possible that the chosen optimal strategy profiles $s_0, s_1, \cdots$, form a cycle, then the algorithm fails. Otherwise, a local sub-Nash equilibrium can be obtained. Theoretically, we have no reason to claim that the algorithm can always terminate properly. But in numerical computations, we did not have experience of failing.

Exercises

17.1 In hand game Rock-Paper-Scissors Rock (R) beats Scissor (S), Scissor beats Paper (P), and Paper beats Rock. Assume "beat" mean winner gets payment 1 from loser. Assume there are two players P_1 and P_2.

(i) Construct the payoff bi-matrix.

(ii) Check that there is no Nash equilibrium for pure strategies.

(iii) Find the Nash equilibrium for mixed strategies.

17.2 Consider the game of Rock-Paper-Scissors again. Assume R beats S and wins 3 dollars from S, S beats P and wins 2 dollars from P, and P beats

R and wins 1 dollars from R. Reconsider the problems (i)–(iii) in previous exercise.

17.3 Consider the infinitely repeated game of Rock-Paper-Scissors. If the strategy of player P_1 is: If he wins, next time he will not change the action, and if he loses next time he will take the opponent's policy.

(i) Describe the dynamic model of the actions of P_1.

(ii) What is the best strategy of P_2 assume he knows P_1's strategy?

17.4 Find another Nash equilibrium for Example 17.6.

17.5 Consider Example 17.4. Assume player P_A uses a mixed strategy as

$$\begin{cases} P(A=1)=p \\ P(A=0)=1-p, \end{cases}$$

and player P_B uses a similar strategy as

$$\begin{cases} P(B=1)=q \\ P(B=0)=1-q. \end{cases}$$

Find the Nash equilibrium for the above strategy.

17.6 Prove Theorem 17.2.

17.7 Prove Proposition 17.2.

17.8 Give the expression of $\Psi(G)$ in (17.22)

17.9 A game of grasping money is as follows: The money available at time t is t dollars. Players A and B are making their decisions by turn. Say, A starts. If A decides to take the money, he gets 1 dollar, and the game terminates. Otherwise, A did not take it and it is B to make his decision. If B decides to take the money, he gets 2 dollars, and the game terminates. Otherwise, A turns to decide whether he takes the 3 dollars, and so on. Give the dynamic model and the payoff functions for this game.

17.10 In Tian Ji horse racing, find the Nash equilibrium for mixed strategies.

17.11 In an infinitely repeated game there are two players P_1 and P_2, their actions are $X=\{x_1=\delta_2^1, x_2=\delta_2^2\}$ and $Y=\{y_1=\delta_3^1, y_2=\delta_3^2, y_3=\delta_3^3\}$ respectively. Assume the payoff bi-matrix is as in Table 17.7. Moreover, the player P_2's strategy is

Table 17.7 Infinite game

$P_1\backslash P_2$	y_1	y_2	y_3
x_1	1,-1	-3, 2	4,-3
x_2	-2, 1	2,-2	-3, 4

$$y(t+1)=\delta_3[2\ 3\ 2\ 3\ 1\ 2]x(t)y(t).$$

To maximize the

$$J_1(s) = \overline{\lim}_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} c_1(x(t), y(t)), \quad j = 1, \cdots, n, \tag{17.25}$$

what is the best strategy of Player P_1?

17.12 Consider a finite time repeated game with players A and B. The payoff bi-matrix is as in Table 17.8.

Table 17.8 Sub-game perfect Nash equilibrium

$A\backslash B$	1	2
1	2,1	0,0
2	-1,-1	1,2

Assume the game is playing as follows: A acts first and then B, after knowing A's action, acts. Assume the payoffs are the sum of each time payoffs. Then we can draw a tree to describe the result as in Fig. 17.1.

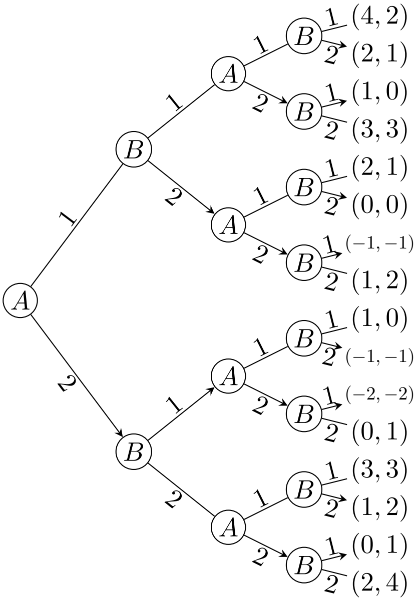


Fig. 17.1 Subgame perfect Nash equilibrium

Searching the best action backward, B can find the best action at each situation. And then A can find its best action. Keep going backward like this, we can finally find a Nash equilibrium.

(i) Find the two times Nash equilibrium from Fig. 17.1;

(ii) As long as the plying time is finite, backward searching can always find a Nash equilibrium. Prove it.

(iii) Prove that the Nash equilibrium obtained in this way is a subgame perfect Nash equilibrium.

17.13 Consider the set of actions of a player, which is $\mathcal{L}_{3\times 9}$. Find $B_\epsilon(L)$ for

(i) $L = \delta_3[2,2,2,2,2,2,2,2,2]$ and $\epsilon = 2$.

(ii) $L = \delta_3[1,2,2,2,3,1,3,1,2]$ and $\epsilon = 1$.

17.14 Consider the set of actions as $\mathcal{L}_{2\times 4}$. Let $L_c \in \mathcal{L}_{2\times 4}$. The radius of L_c is defined as

$$R(L_c) = \max_{L\in\mathcal{L}_{2\times 4}} d(L_c, L).$$

Find L_c^*, called the center of action set, such that

$$R(L_c^*) = \min_{L_c\in\mathcal{L}_{2\times 4}} R(L_c).$$

17.15 Prove that that a common Nash equilibrium is a subgame-perfect Nash equilibrium.

17.16 A game with two players has the payoff bi-matrix shown in Table 17.9.

Table 17.9 Infinite game

$P_1\backslash P_2$	y_1	y_2	y_3
x_1	1,-1	-3, 2	4,-3
x_2	-2, 1	2,-2	-3, 4
x_2	-1, 2	1,-1	-1, 2

(i) For each pair (x_i, y_j) find its tolerance.

(ii) Find the sub-Nash equilibrium(s).

17.17 Recall Example 17.8.

(i) Show that $L = \delta_4[1\ 4\ 4\ 4]$ is a Nash equilibrium.

(ii) Choose $L^0 = \delta_4[2\ 2\ 2\ 2]$ and use Algorithm 4 to find a local Nash equilibrium. Is it a Nash equilibrium?

Chapter 18

Multi-Variable Polynomials

This chapter considers the matrix expression of multi-variable polynomials via semi-tensor product. Under this expression the differential of functional matrices is introduced and its calculation formulas are obtained. Then we consider the expression of polynomials under two different generators, and the conversion formulas are presented. As an application the Taylor expansion of multi-variable functions is investigated. Finally, the formulas are presented for calculating Lie derivatives.

This expression provides a convenient tool to solve nonlinear problems via matrices. Moreover, under this expression the nonlinear computations can easily be realized by computer.

18.1 Matrix Expression of Multi-Variable Polynomials

Let $x = (x_1, \cdots, x_n)^T$ be a set of coordinate variables on $\mathbb{R}^n$. Denote the set of kth degree homogeneous polynomials by B_n^k. Denote by $B_n^0 = \mathbb{R}$, which represents the zero-degree polynomials, that is, constants.

It is easy to see that the components of x^k form a generator of B_n^k, that is, this set contains a basis of B_n^k. Precisely speaking, let $f(x) \in B_n^k$. Then there exists a matrix $F \in \mathcal{M}_{1\times n^k}$ such that

$$f(x) = F \ltimes x^k. \tag{18.1}$$

But this generator is not a basis, because it contains some redundant elements. In other words, the elements in this generator are not linearly independent. It follows that the F in (18.1) is not unique.

Example 18.1. Let $f(x) \in B_2^3$ be

$$f(x) = x_1^3 + x_1^2 x_2 - 2x_1 x_2^2 - x_2^3. \tag{18.2}$$

Note that

$$x = (x_1,\ x_2)^T, \quad x^2 = (x_1^2,\ x_1x_2,\ x_2x_1,\ x_2^2)^T,$$

$$x^3 = (x_1^3,\ x_1^2x_2,\ x_1x_2x_1,\ x_1x_2^2,\ x_2x_1^2,\ x_2x_1x_2,\ x_2^2x_1,\ x_2^3)^T, \quad \cdots$$

Then $f(x)$ can be expressed as

$$f(x) = (1\ 1\ 0\ \ -2\ 0\ 0\ 0\ \ -1) \ltimes x^3. \tag{18.3}$$

Since the generator has redundant elements, the coefficient matrix is not unique. For instance, an alternative expression of $f(x)$ can be

$$f(x) = (1\ 1\ 0\ \ -1\ 0\ \ -1\ 0\ \ -1) \ltimes x^3. \tag{18.4}$$

An obvious advantage of the above STP-based matrix expression of multi-variable polynomial is that the semi-tensor product has many nice properties such as associativity etc., which provide us a convenient way to manipulate the polynomials. For instance, we may factorize the polynomials in a similar way as that of single-variable polynomials. Let us see the following example.

Example 18.2. Assume $x = (x_1, \cdots, x_n)^T \in \mathbb{R}^n$, and the corresponding coefficient matrices have proper dimensions. Then we have

(1)

$$F_3 \ltimes x^3 + F_5 \ltimes x^5 = (F_3 + F_5 \ltimes x^2) \ltimes x^3.$$

(2)

$$(A \ltimes x)^2 - 1 = (A \ltimes x + 1)(A \ltimes x - 1).$$

Since x^k has redundant elements, which makes the coefficient matrix F non-unique, we may try to pose some restrictions on F. Recall that in linear algebra a quadratic function $f(x) = \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} a_{i,j}x^ix^j$ can be expressed as

$$f(x) = x^T Ax,$$

where $A = (a_{i,j})$ is not unique. But if we require A being symmetric, then it is unique. Using a symmetric matrix to express a quadratic form is a convention in linear algebra. Similarly, we can define a symmetric expression of (18.1).

Definition 18.1. Let $f(x) \in B_n^k$ be expressed as in (18.1) with coefficient matrix F. Assume the elements of F are labeled by multi-index $\mathrm{id}(i_1, \cdots, i_k; n, \cdots, n)$. F is said to be a symmetric coefficient matrix, if

$$F_{i_{\sigma(1)},\cdots,i_{\sigma(k)}} = F_{i_1,\cdots,i_k}, \quad \forall\, \sigma \in \mathbf{S}_k,$$

where $\mathbf{S}_k$ is the kth order symmetric group.

Remark 18.1.

(1) If the F in (18.1) is symmetric, then it is easy to verify that it is unique.

(2) If $f(x)$ is a quadratic form and F is symmetric, i.e., $x \in \mathbb{R}^n$ and

$$f(x) = Fx^2 = (F_{11}, F_{12}, \cdots, F_{1n}, \cdots, F_{n1}, F_{n2}, \cdots, F_{nn})x^2,$$

where $F_{ij} = F_{ji}$, then $V_c^{-1}(F) = V_r^{-1}(F)$ is a symmetric matrix. That is, the symmetric coefficient matrix is a generalization of symmetric matrix.

(3) In fact, we may consider $f(x) \in B_n^k$ as a tensor $f \in \mathcal{T}^k(\mathbb{R}^n)$. That is,

$$f(z_1, \cdots, z_k) := F \ltimes z_1 \ltimes \cdots \ltimes z_k,$$

where $z_1 = \cdots = z_k = x$. Hence, a homogeneous multi-variable polynomial has tensor structure. The advantage of using STP to tensor calculation has already been discussed in Chapter 2.

In the following discussion symmetric group is a fundamental tool. Symmetric group has been discussed in Chapter 1. Briefly speaking, $\mathbf{S}_n$ consists of all the permutations of n elements. Hence $|S_n| = n!$. For instance, $|\mathbf{S}_5| = 5! = 120$. In the following we give an example for the permutation and the index permutation.

Example 18.3.

(1) Consider two permutations $\sigma_1, \sigma_2 \in \mathbf{S}_6$ as

$$\sigma_1 = (123)(456), \quad \sigma_2 = (15642).$$

Then the product

$$\sigma_2\sigma_1 = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 3 & 1 & 5 & 6 & 4 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 1 & 3 & 5 & 6 & 4 & 2 \end{bmatrix}.$$

It can be briefly denoted as $\sigma_2\sigma_1 = (23546)$.

(2) Let

$$F = \{ f_{i_1,i_2,i_3,i_4,i_5,i_6} | \, i_j = 1, \cdots, n; \; j = 1, \cdots, 6\}$$

be a set of data with $|F| = n^6$. The elements are labeled by multi-index $\mathrm{id}(i_1, i_2, i_3, i_4, i_5, i_6; n^6)$. For instance, an element $f_{i_1,i_2,i_3,i_4,i_5,i_6} = f_{1,3,2,3,1,1}$, that is, $(i_1, i_2, i_3, i_4, i_5, i_6) = (1, 3, 2, 3, 1, 1)$.

Then we use f_σ to represent the elements after subscript index permutation $f_{i_{\sigma(1)},\cdots,i_{\sigma(k)}}$. For instance, consider $a = f_{1,3,2,3,1,1}$ and take $\sigma = \sigma_1$. Since $\sigma_1(1) = 2$, and $i_2 = 3$, then correspondingly, a has the first index "3"; similarly, since $\sigma_1(2) = 3$, and $i_3 = 2$, the second index is "2", and so on. Finally, we have $a = f_{\sigma_1} = f_{3,2,1,1,1,3}$. Similarly, we have $a = f_{\sigma_2} = f_{1,1,2,3,1,3}$, $a = f_{\sigma_2\sigma_1} = F_{1,2,1,1,3,3}$.

Next, we consider how to convert the $f(x)$ in (18.1) to symmetric form. A simple way is to collect terms by hand. But to deal with theoretical analysis or to use computer to carry out the conversion we need a general formula. To this end, we need some new concepts and notations.

Definition 18.2. Let $I = (I_1, I_2, \cdots, I_k) \in \mathrm{id}(i_1, i_2, \cdots, i_k; n^k)$. Then

(1) the permutating set of I, denoted by P_I, is defined as

$$\begin{aligned} P_I = \{J \in \mathrm{id}(i_1, i_2, \cdots, i_k; n^k) \mid \\ (J_1, \cdots, J_k) = (I_{\sigma(1)}, \cdots, I_{\sigma(k)}),\ \sigma \in \mathbf{S}_k\}; \end{aligned}$$

(2) the index frequency of I, denoted by $C_I = (c_1, \cdots, c_n) \in \mathbb{Z}_+^n$, is defined as follows: c_j is the times of j appearing into $\{I_1, I_2, \cdots, I_k\}$. Note that $1 \le j \le n$.

Denote the cardinality (size) of P_I by $|P_I|$, and set

$$C_I! = \prod_{j=1}^{n} c_j!.$$

Then we have

$$|P_I| = \frac{k!}{C_I!}. \tag{18.5}$$

Example 18.4. Let $k = 4$, $n = 5$, and

$$I = (2, 5, 2, 3) \in id(i_1, i_2, i_3, i_4; n^4).$$

Then

$$P_I = \left\{ \begin{matrix} (2,2,3,5)\ (2,2,5,3)\ (2,3,2,5)\ (2,3,5,2)\ (2,5,2,3)\ (2,5,3,2) \\ (3,2,2,5)\ (3,2,5,2)\ (3,5,2,2)\ (5,2,2,3)\ (5,2,3,2)\ (5,3,2,2) \end{matrix} \right\}.$$

Since there is no 1 in I, we have $c_1 = 0$. Similarly, we have $c_2 = 2$, $c_3 = 1$, $c_4 = 0$, $c_5 = 1$. Finally, we have $C_I = (0, 2, 1, 0, 1)$, and hence

$$|P_I| = \frac{k!}{C_I!} = \frac{4!}{0!2!1!0!1!} = 12.$$

The following result comes from the above definitions and notations.

Proposition 18.1. *Let $P(x) = Fx^k$, $x \in \mathbb{R}^n$ be a kth homogeneous polynomial with its symmetric form as $P(x) = \tilde{F}x^k$. Assume both F and $\tilde{F}$ are labeled by*
$id(i_1, i_2, \cdots, i_k; n^k)$, then for

$$I = (i_1, \cdots, i_k) \in \mathrm{id}(i_1, i_2, \cdots, i_k; n^k)$$

we have

$$\tilde{F}_I = \frac{1}{|P_I|} \sum_{J \in P_I} F_J. \tag{18.6}$$

Definition 18.3. Let $F = \{F_{i_1, \cdots, i_k}\} \in \mathbb{R}^{n^k}$ be a set of n^k numbers, labeled by the multi-index $id(i_1, \cdots, i_k;\ n^k)$. Equation (18.6) defines a mapping $\tilde{F} = \psi_n^k(F) : \mathbb{R}^{n^k} \to \mathbb{R}^{n^k}$. Then the mapping ψ is called a symmetrization on $\mathbb{R}^{n^k}$.

Note that it is easy to verify by definition that both ψ_n^1 and ψ_n^0 are identity mapping.

The following proposition is an immediate consequence of the definition.

Proposition 18.2. *Let*

$$P(x) = P_0 + P_1 x + P_2 x^2 + \cdots + P_k x^k, \quad x \in \mathbb{R}^n$$

be a kth degree polynomial. $P(x) \equiv 0$, if and only if,

$$\psi_n^i(P_i) = 0, \quad i = 0, 1, \cdots, k.$$

In fact, ψ is a linear mapping, hence it can be expressed by a matrix. When $s \geq 2$, we define an $n^s \times n^s$ matrix Ψ_n^s as following: using indices $\{i_1, \cdots, i_s\}$ to label its rows and columns in the order of $\mathrm{id}(i_1, \cdots, i_s; n^k)$. Let $J = (J_1, \cdots, J_s)$ be its row index and $I = (I_1, \cdots, I_s)$ its column index. Then we assign the elements of $\Psi_n^s = (\psi_{J,I})$ by

$$\psi_{J,I} = \begin{cases} 0, & J \notin P_I, \\ \frac{1}{|P_I|}, & J \in P_I. \end{cases}$$

As for $s \leq 1$ we set $\Psi_n^0 = 1$ and $\Psi_n^1 = I_n$. Then these Ψ_n^k are the matrices of the mappings ψ_n^k.

Proposition 18.3. *A kth degree polynomial*

$$P(x) = P_0 + P_1 x + P_2 x^2 + \cdots + P_k x^k, \quad x \in \mathbb{R}^n$$

has its symmetric expression as

$$P(x) = \tilde{P}_0 + \tilde{P}_1 x + \tilde{P}_2 x^2 + \cdots + \tilde{P}_k x^k,$$

where

$$\tilde{P}_i = P_i \Psi_n^i, \quad i = 0, 1, \cdots, k. \tag{18.7}$$

Moreover, $P(x) \equiv 0$*, if and only if*

$$P_i \Psi_n^i = 0, \quad i = 0, 1, \cdots, k.$$

Example 18.5.

(1) Let $n = 3$ and $k = 2$. Then Ψ_3^2 is

$$\Psi_3^2 = \begin{array}{c} \begin{array}{ccccccccc} (11) & (12) & (13) & (21) & (22) & (23) & (31) & (32) & (33) \end{array} \\ \left[\begin{array}{ccccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \end{array} \begin{array}{c} \\ (11) \\ (12) \\ (13) \\ (21) \\ (22) \\ (23) \\ (31) \\ (32) \\ (33) \end{array}.$$

(2) Let $n = 2$ and $k = 3$. Then Ψ_2^3 is

$$\Psi_3^2 = \begin{array}{c} \begin{array}{cccccccc} 111 & 112 & 121 & 122 & 211 & 212 & 221 & 222 \end{array} \\ \left[\begin{array}{cccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \end{array} \begin{array}{c} \\ (111) \\ (112) \\ (121) \\ (122) \\ (211) \\ (212) \\ (221) \\ (222) \end{array}.$$

(3) Consider the polynomial (18.2). It is easy to check that both (18.3) and (18.4) are not symmetric. Choosing any one and using its coefficient matrix F, and using formula (18.7), we can get the symmetric coefficient matrix of (18.2) as

$$f(x) = \begin{bmatrix} 1 & \frac{1}{3} & \frac{1}{3} & -\frac{2}{3} & \frac{1}{3} & -\frac{2}{3} & -\frac{2}{3} & -1 \end{bmatrix} \ltimes x^3.$$

Let $P[x]$, $x \in \mathbb{R}^n$ be the vector space of real polynomials. Then its is a direct sum of the spaces of kth homogeneous polynomials with $k = 0, 1, \cdots$. That is,

$$P[x] = B_n^0 \oplus B_n^1 \oplus B_n^2 \oplus \cdots . \tag{18.8}$$

Proposition 18.4. *Let $P(x) \in B_n^p$, $Q(x) \in B_n^q$ with $P(x) = M_P x^p$ and $Q(x) = M_Q x^q$. Then their product $P(x)Q(x) \in B_n^{p+q}$ has its coefficient matrix as*

$$M_{PQ} = M_P \ltimes M_Q. \tag{18.9}$$

Proof.

$$\begin{aligned} M_P \ltimes M_Q \ltimes x^{p+q} &= M_P \ltimes M_Q \ltimes x^q \ltimes x^p = M_P \ltimes (M_Q \ltimes x^q) \ltimes x^p \\ &= (M_Q \ltimes x^q) \ltimes (M_P \ltimes x^p) = M_{PQ} \ltimes x^{p+q}. \end{aligned}$$

□

The following is an immediate consequence of Proposition 18.4.

Proposition 18.5. *Let $M_P \in \mathbb{R}^{n^p}$ and $M_Q \in \mathbb{R}^{n^q}$ be two row vectors. Then*

$$M_P \ltimes M_Q \ltimes x^{p+q} = M_Q \ltimes M_P \ltimes x^{p+q}. \tag{18.10}$$

Since x^k is not a basis of B_n^k, (18.10) does not imply that $M_P \ltimes M_Q = M_Q \ltimes M_P$.

According to Proposition 18.4, all the algebraic formulas in elementary algebra can be extended to the corresponding multi-variable case. For instance, since $(a + b)^2 = a^2 + 2ab + b^2$, it is also true that

$$(F_1 \ltimes x + F_2 \ltimes x^2)^2 = F_1^2 \ltimes x^2 + 2F_1 \ltimes F_2 x^3 + F_2^2 \ltimes x^4.$$

Equality (18.10) has the following generalization:

Proposition 18.6. *Let $F \in R^{n^\alpha}$ be a row vector and $t > \alpha$. Then for any row vector $G \in R^{n^{t-\alpha}}$*

$$(G \ltimes F) \ltimes x^t = F \ltimes x^\alpha \ltimes G \ltimes x^{t-\alpha}. \tag{18.11}$$

We give an example to demonstrate the advantage of this expression.

Example 18.6. Given a polynomial

$$P(x) = A_0 + A_1 x + A_2 x^2 + \cdots + A_t x^t, \quad x \in \mathbb{R}^n.$$

We want to check whether it has a linear factor $Q(x) = C - Dx$, where $C \neq 0$.

For simplicity, we set $C = 1$. Assume we have

$$\begin{aligned}&(1 - Dx)(B_0 + B_1 x + \cdots + B_{t-1}x^{t-1})\\ &= A_0 + A_1 x + A_2 x^2 + \cdots + A_t x^t.\end{aligned} \tag{18.12}$$

Comparing the coefficients on both sides of (18.12) yields

$$\begin{cases} B_0 = A_0, \\ B_k = A_k + DB_{k-1}, \quad k = 1, \cdots, t-1. \end{cases} \tag{18.13}$$

To satisfy (18.12), if and only if the inductively defined B_{t-1} satisfies

$$-DB_{t-1}x^t = A_t x^t. \tag{18.14}$$

Calculating B_{t-1} from (18.13) and plugging it into (18.14), we have

$$(A_t + DA_{t-1} + D^2 A_{t-2} + \cdots + D^t A_0)x^t = 0. \tag{18.15}$$

Using Example 18.6 and Proposition 18.3, we have the following proposition.

Proposition 18.7. *The polynomial* $P(x) = A_0 + A_1 x + A_2 x^2 + \cdots + A_t x^t$ *has a left linear factor* $1 - Dx$, *if and only if*

$$(A_t + DA_{t-1} + D^2 A_{t-2} + \cdots + D^t A_0)\Psi_n^t = 0. \tag{18.16}$$

Since the expression of a polynomial is not unique, in the above factorization we have to make some additional calculations. Note that for a zero term, it could be expressed into the sum of some nonzero terms. But it does not affect the final result, because in the overall calculation process there are only additions and productions. Then the zero terms remain zero at the final result. Hence the above factorization is independent of the choice of equivalent coefficients. Here we say the coefficients F_1^t F_2^t of x^t are equivalent, if $\psi_n^t(F_1^t - F_2^t) = 0$, that is, $(F_1^t - F_2^t)\Psi_n^t = 0$, because in this case we have $F_1^t x^t = F_2^t x^t$.

Note that this result is similar to the case of single variable case. The only difference is the "scalar product" in single-variable case is replaced by semi-tensor product for multi-variable case.

The following is a numerical example.

Example 18.7. Given a polynomial

$$\begin{aligned}P(\xi) = 1 &- (3x + 2y - z) + 2x^2 - 3xy + z^2 + 6x^2 y - 2x^2 z\\ &+7xy^2 - 9xyz + xz^2 - 2y^2 z + y^3 + z^3,\end{aligned}$$

$$Q(\xi) := 1 - D(\xi) = 1 - (x + y - z),$$

where $\xi = (x, y, z)^T \in \mathbb{R}^3$. Then

$$D = (1 \;\; 1 \;\; -1),$$

$$A_1 = (-3 \;\; -2 \;\; 1), \quad A_2 = (2 \;\; -3 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 1),$$

$$A_3 = (0\; 6\; -2\; 0\; 7\; -9\; 0\; 0\; 1\; 0\; 0\; 0\; 0\; 1\; -2\; 0\; 0\; 0\; 0\; 0\; 0\; 0\; 0\; 0\; 0\; 0\; 1).$$

Assume there is an $R(\xi)$ such that $Q(\xi)R(\xi) = P(\xi)$. Then $R(\xi) = 1 + B_1\xi + B_2\xi^2$. According to (18.13), we can calculate B_1 and B_2 as follows:

$$B_1 = A_1 + D = (-3 \;\; -2 \;\; 1) + (1 \;\; 1 \;\; -1) = (-2 \;\; -1 \;\; 0),$$

$$B_2 = A_2 + DB_1 = (0 \;\; -5 \;\; 2 \;\; -1 \;\; -1 \;\; 1 \;\; 0 \;\; 0 \;\; 1).$$

Hence

$$\begin{aligned} DB_2 = (&0\; 0\; 0\; -5\; -5\; 5\; 2\; 2\; -2\; -1\; -1\; 1\; -1\; -1\; 1 \\ &1\; 1\; -1\; 0\; 0\; 0\; 0\; 0\; 0\; 1\; 1\; -1). \end{aligned}$$

It is clear that $(A_3 + DB_2)\Psi_3^3 = 0$. Then we have

$$\begin{aligned} R(\xi) = 1 + B_1\xi + B_2\xi^2 &= 1 - (2x + y) \\ &\quad + (-6xy - y^2 + 2xz + yz + z^2). \end{aligned}$$

Next, we consider a polynomial mapping. An m-dimensional and kth order polynomial mapping can be expressed as

$$P(x) = A_0 + A_1x + A_2x^2 + \cdots + A_kx^k; \quad x \in \mathbb{R}^n,\; P(x) \in \mathbb{R}^m,$$

where $A_j \in \mathcal{M}_{m\times n^j}$, $j = 0, \cdots, k$.

In the follow we consider the semi-tensor product of two polynomial mappings.

Proposition 18.8. *Let* $P(x) = A_0 + A_1x + A_2x^2 + \cdots + A_px^p$, $Q(x) = B_0 + B_1x + B_2x^2 + \cdots + B_qx^q$. *Then*

$$P(x) \ltimes Q(x) = \sum_{i=0}^{p+q} \sum_{j=0}^{i} (A_j \otimes B_{i-j})\, x^i. \tag{18.17}$$

Proof. Recall (2.19) in Example 2.4, for two vectors x and y we have

$$(Ax) \ltimes (By) = (A \otimes B)(x \ltimes y).$$

Using this formula and the distributive law of the semi-tensor product, (18.17) follows. □

The following example shows how to use STP to deal with the polynomial matrices.

Example 18.8. Let $A(x)$ be a square matrix with its entries as tth degree polynomials, where $x \in \mathbb{R}^n$. Let $B(x) = I - Dx$ be a linear form, where $D \in \mathcal{M}_{n\times n^2}$. The question is: when there exists a polynomial matrix $C(x)$, such that $A(x) = B(x)C(x)$? We are looking a similar result as in Proposition 18.7.

First, we show that $A(x)$ can be expressed as

$$A(x) = A_0 + A_1 x + \cdots + A_t x^t. \tag{18.18}$$

To this end, we first express each entries of A into the sum of homogeneous matrix forms as before. That is, $A(x)$ can be expressed as

$$A(x) = A_0 + A^1 + \cdots + A^t,$$

where

$$A^k = \begin{bmatrix} a_{11}x^k & \cdots & a_{1n}x^k \\ \vdots & \ddots & \vdots \\ a_{n1}x^k & \cdots & a_{nn}x^k \end{bmatrix}, \quad a_{ij} \in \mathcal{M}_{1\times n^k}, \quad k = 1, 2, \cdots, t.$$

It is easy to verify that

$$A^k = E_k(I_n \otimes x^k) = E_k \rtimes x^k, \tag{18.19}$$

where

$$E_k = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \in \mathcal{M}_{n\times n^{k+1}}.$$

Using (4.21) of Proposition 4.3 to (18.19), we have

$$E_k \rtimes x^k = E_k \ltimes W_{[n^k,n]} \ltimes x^k \ltimes W_{[n,1]}.$$

Note that

$$W_{[n,1]} = W_{[1,n]} = I_n. \tag{18.20}$$

Using (18.20) and setting

$$A_k = E_k \ltimes W_{[n^k,n]}, \quad k = 1, \cdots, t$$

yield (18.18).

Now similar to polynomial case, we assume there exists $C(x) = C_0 + C_1 x + \cdots + C_{t-1}x^{t-1}$, such that $A(x) = B(x)C(x) = (I - Dx)C(x)$. Comparing the coefficients of both sides of

$$(I - Dx)(C_0 + C_1 x + \cdots + C_{t-1}x^{t-1}) = A_0 + A_1 x + \cdots + A_t x^t,$$

we have

$$\begin{cases} C_0 = A_0, \\ C_k x^k = A_k x^k + DxC_{k-1}x^{k-1}, \quad k = 1, \cdots, t-1. \end{cases} \tag{18.21}$$

Note that

$$xC_{k-1} = (I_n \otimes C_{k-1}) \ltimes x.$$

Plugging it into (18.21) yields that

$$\begin{cases} C_0 = A_0, \\ C_k = A_k + D \rtimes C_{k-1}, \quad k = 1, \cdots, t-1. \end{cases} \tag{18.22}$$

An iterative calculation shows that

$$C_{t-1} = \sum_{i=1}^{t} D^{\rtimes(i-1)} \rtimes A_{t-i},$$

where $D^{\rtimes i} = \underbrace{D \rtimes D \rtimes \cdots \rtimes D}_{i}$. Finally, matching the coefficients of x^t yields that

$$Dx \ltimes C_{t-1}x^{t-1} = (D \rtimes C_{t-1})x^t + A_t x^t = 0.$$

That is,

$$\left[\sum_{i=0}^{t} D^{\rtimes i} \rtimes A_{t-i}\right] x^t = 0, \tag{18.23}$$

which assures that $B(x)$ is a left factor of $A(x)$.

Next, we consider when there is a matrix $C(x)$, such that $A(x) = C(x)B(x)$? A similar argument shows that the condition is

$$\left[\sum_{i=0}^{t} A_{t-i} \rtimes D^{\rtimes i}\right] x^t = 0. \tag{18.24}$$

Summarizing the results obtained in the above example, we have the following conclusion.

Theorem 18.1. *Assume $A(x)$ is a square matrix with its entries as tth degree polynomials with $x \in \mathbb{R}^n$. Let $B(x) = I - Dx$ be a linear form. Then $B(x)$ is a left factor of $A(x)$, if and only if*

$$\left[\sum_{i=0}^{t} D^{\rtimes i} \rtimes A_{t-i}\right] \Psi_n^t = 0. \tag{18.25}$$

$B(x)$ is a right factor of $A(x)$, if and only if

$$\left[\sum_{i=0}^{t} A_{t-i} \rtimes D^{\rtimes i}\right] \Psi_n^t = 0. \tag{18.26}$$

***Proof*.** In fact (18.25) and (18.26) follow from (18.23) and (18.24) respectively. The only thing which is different from the polynomial case is how to multiply the symmetrization matrix Ψ_n^t. What we need now is to multiply it to each entries of the matrix. That is, it must be multiplied to $(x_{i,1}, x_{i,n+1}, x_{i,2n+1}, \cdots)$. We, hence, need

$$\left[\sum_{i=0}^{t} D^{\rtimes i} \rtimes A_{t-i}\right] (\Psi_n^t \otimes I_n) = 0,$$

etc. According to Proposition 2.4, the above equation is equivalent to (18.25). □

18.2 Differential Form of Functional Matrices

Definition 18.4.

(1) Let $f(x) : \mathbb{R}^n \to \mathbb{R}$ be a smooth function. Its differential is defined as

$$Df(x) = \left(\frac{\partial f(x)}{\partial x_1}, \cdots, \frac{\partial f(x)}{\partial x_n}\right); \tag{18.27}$$

and its gradient is defined as

$$\nabla f(x) = (Df(x))^T. \tag{18.28}$$

(2) Let $M(x) \in M_{p\times q}$ be a functional matrix with its entries as functions of $x \in \mathbb{R}^n$. The differential of $M(x)$, denoted by $DM(x) \in \mathcal{M}_{p\times nq}$, is obtained by replacing each entries $m_{i,j}$ of $M(x)$ by their differentials $\left(\frac{\partial dm_{i,j}}{\partial x_1}, \cdots, \frac{\partial m_{i,j}}{\partial x_n}\right)$. That is,

$$DM(x) = \begin{bmatrix} \frac{\partial m_{11}(x)}{\partial x_1} \cdots \frac{\partial m_{11}(x)}{\partial x_n} & \cdots & \frac{\partial m_{1q}(x)}{\partial x_1} \cdots \frac{\partial m_{1q}(x)}{\partial x_n} \\ \vdots \qquad\qquad \vdots & & \vdots \qquad\qquad \vdots \\ \frac{\partial m_{p1}(x)}{\partial x_1} \cdots \frac{\partial m_{p1}(x)}{\partial x_n} & \cdots & \frac{\partial m_{pq}(x)}{\partial x_1} \cdots \frac{\partial m_{pq}(x)}{\partial x_n} \end{bmatrix}; \tag{18.29}$$

and its gradient, denoted by $\nabla M(x) \in \mathcal{M}_{np\times q}$, is obtained by replacing each entries $m_{i,j}$ of $M(x)$ by their gradients $\left(\frac{\partial dm_{i,j}}{\partial x_1}, \cdots, \frac{\partial m_{i,j}}{\partial x_n}\right)^T$.

The higher order differentials or gradients can be defined inductively as

$$D^{k+1}M = D(D^kM) \in M_{p\times n_{k+1}q}, \quad k \geq 1,$$
$$\nabla^{k+1}M = \nabla(\nabla^kM) \in M_{pn_{k+1}\times q}, \quad k \geq 1.$$

To show the usefulness of these definitions we consider its application to the pseudo Hamiltonian control system (van der Schaft, 2000).

Example 18.9.

(1) In $\mathbb{R}^n$ an $n\times n$ functional matrix $M(x)$ is given. It can be used to define a pseudo Poisson bracket: $C^\infty(\mathbb{R}^n) \times C^\infty(\mathbb{R}^n) \to C^\infty(\mathbb{R}^n)$, which has $M(x)$ as its structure matrix, as

$$\{f(x), g(x)\} = Df(x)M(x)\nabla g(x), \quad f(x), g(x) \in C^\infty(\mathbb{R}^n). \quad (18.30)$$

$\mathbb{R}^n$ with the structure matrix $M(x)$ is called a pseudo Poisson manifold.

(2) In a pseudo Poisson manifold let $f(x)$ be a smooth function with $x \in \mathbb{R}^n$. A pseudo Hamilton vector field, with $f(x)$ as its pseudo Hamiltonian function is defined as

$$X_F(x) = M(x)\nabla F(x). \quad (18.31)$$

(3) The following dynamic control system is called a pseudo Hamiltonian system:

$$\dot{x} = X_f(x), \quad x \in \mathbb{R}^n. \quad (18.32)$$

$$(18.33)$$

Consider a control system as

$$\begin{cases} \dot{x} = X_f(x) + \sum\limits_{i=1}^m X_{g_i}(x)u_i(x), \quad x \in \mathbb{R}^n \\ y = h(x) = (\{g_1, f\}, \{g_2, f\}, \cdots, \{g_m, f\})^T. \end{cases} \quad (18.34)$$

If we can find controls $u_i(x)$, $i = 1, \cdots, m$, such that the closed-loop system is a pseudo Hamilton system with the original structure matrix $M(x)$ unchanged, then it is called a pseudo Hamilton realization.

It is clear that what do we need is

$$\sum_{i=1}^m X_{g_i}(x)u_i(x) = M(x)\nabla\phi(x). \quad (18.35)$$

(4) Assume dg_i, $i = 1, \cdots, m$ are linearly independent. Then we can find $\{z_j(x),\ j = 1,\ \cdots,\ n-m\}$, such that $\{dg_i,\ i = 1, \cdots, m,\ dz_j(x),\ j = 1, \cdots, n-m\}$ are linearly independent. Then, we can locally express any smooth function as $\phi(x) = \phi(g_i, z_j)$.

In addition, assume $M(x)$ is nonsingular, then (18.35) can be expressed as

$$\nabla\phi(x) = \sum_{i=1}^{m} \nabla g_i \frac{\partial\phi(g_i, z_j)}{\partial g_i} + \sum_{j=1}^{n-m} \nabla z_j \frac{\partial\phi(g_i, z_j)}{\partial z_j} = \sum_{i=1}^{m} \nabla g_i u_i.$$

Hence

$$\begin{cases} \dfrac{\partial\phi(g_i, z_j)}{\partial z_j} = 0, \quad j = 1, \cdots, n-m, \\ u_i = \dfrac{\partial\phi(g_i, z_j)}{\partial g_i}, \quad i = 1, \cdots, m. \end{cases}$$

We conclude that assume $M(x)$ is nonsingular, then the system (18.34) has a local pseudo Hamilton realization, if and only if there exists a smooth function $\phi = \phi(g_1, \cdots, g_m)$, such that

$$u_i = \frac{\partial\phi(g_1, \cdots, g_m)}{\partial g_i}, \quad i = 1, \cdots, m. \tag{18.36}$$

When $M(x)$ is allowed to be singular, (18.36) becomes sufficient only.

(5) Assume the system (18.34) is output detectable (Isidori, 1995), that is the output satisfies $\lim_{t\to\infty} h(x(t)) = 0, \ \lim_{t\to\infty} x(t) = 0$. Then we can assume there is a Lyapunov function $L = h^T(x)P(x)h(x)$, where $P(x) > 0$ is positive definite. Then for the closed-loop system we have

$$\dot{L} = D(h^T(x)P(x)h(x))M(x)(\nabla F(x) + \nabla\phi(G(x))).$$

Hence if we can find $P(x) > 0$ and $\phi(g_1, \cdots, g_m)$, such that $\dot{L} < 0$, then the system is stabilized at the origin by the controls (18.36).

Above example shows that the differential of functional matrix is very useful. In the following we consider the differential of product of functional matrices. First, we consider $D(A(x) \ltimes B(x))$, where $A(x) \succ B(x)$.

Proposition 18.9. *Assume $A(x) \succ_t B(x)$, where $x \in \mathbb{R}^n$. Then*

$$D(A(x) \ltimes B(x)) = DA(x) \ltimes B(x) + A(x) \ltimes DB(x) \ltimes (I_s \otimes W_{[t,n]}), \tag{18.37}$$

where s is the number of columns of $B(x)$.

Proof. Without loss of generality, we assume

$$A(x) = (a_{11}, \cdots, a_{1t}, \cdots, a_{q1}, \cdots, a_{qt})$$

is a row vector. First, we assume $s = 1$, then $B(x) = (b_1, \cdots, b_q)^T$ is a column vector. It follows that

$$A \ltimes B = \left(\sum_{k=1}^{q} a_{k1}b_k, \cdots, \sum_{k=1}^{q} a_{kt}b_k\right).$$

$$\begin{aligned} &D(A \ltimes B) = \\ &\quad\Big(\sum_{k=1}^{q} \frac{\partial a_{k1}}{\partial x_1}b_k + \sum_{k=1}^{q} a_{k1}\frac{\partial b_k}{\partial x_1}, \cdots, \sum_{k=1}^{q} \frac{\partial a_{k1}}{\partial x_n}b_k + \sum_{k=1}^{q} a_{k1}\frac{\partial b_k}{\partial x_n}, \cdots, \\ &\qquad \sum_{k=1}^{q} \frac{\partial a_{kt}}{\partial x_1}b_k + \sum_{k=1}^{q} a_{kt}\frac{\partial b_k}{\partial x_1}, \cdots, \sum_{k=1}^{q} \frac{\partial a_{kt}}{\partial x_n}b_k + \sum_{k=1}^{q} a_{kt}\frac{\partial b_k}{\partial x_n}\Big) \\ &\quad= \Big(\sum_{k=1}^{q} \frac{\partial a_{k1}}{\partial x_1}b_k, \cdots, \sum_{k=1}^{q} \frac{\partial a_{k1}}{\partial x_n}b_k, \cdots, \sum_{k=1}^{q} \frac{\partial a_{kt}}{\partial x_1}b_k, \cdots, \sum_{k=1}^{q} \frac{\partial a_{kt}}{\partial x_n}b_k\Big) \\ &\qquad + \Big(\sum_{k=1}^{q} a_{k1}\frac{\partial b_k}{\partial x_1}, \cdots, \sum_{k=1}^{q} a_{k1}\frac{\partial b_k}{\partial x_n}, \cdots, \sum_{k=1}^{q} a_{kt}\frac{\partial b_k}{\partial x_1}, \cdots, \sum_{k=1}^{q} a_{kt}\frac{\partial b_k}{\partial x_n}\Big) \\ &\quad := I + II. \end{aligned} \tag{18.38}$$

A straightforward computation shows that

$$(DA)\ltimes B = \left(\sum_{k=1}^{q} \frac{\partial a_{k1}}{\partial x_1}b_k, \cdots, \sum_{k=1}^{q} \frac{\partial a_{k1}}{\partial x_n}b_k, \cdots, \sum_{k=1}^{q} \frac{\partial a_{kt}}{\partial x_1}b_k, \cdots, \sum_{k=1}^{q} \frac{\partial a_{kt}}{\partial x_n}b_k\right),$$

which is the first part (I) of (18.38).

Similarly, we can calculate that

$$A\ltimes DB = \left(\sum_{k=1}^{q} a_{k1}\frac{\partial b_k}{\partial x_1}, \cdots, \sum_{k=1}^{q} a_{kt}\frac{\partial b_k}{\partial x_1}, \cdots, \sum_{k=1}^{q} a_{k1}\frac{\partial b_k}{\partial x_n}, \cdots, \sum_{k=1}^{q} a_{kt}\frac{\partial b_k}{\partial x_n}\right).$$

Now both the $A \ltimes DB$ and the second part (II) of (18.38) consist of the elements of $\sum_{k=1}^{q} a_{ki}\frac{\partial b_k}{\partial x_j}$. But in the prior the elements are arranged in the order of $\mathrm{id}(j, i; n, t)$, and in the later they are arranged in the order of $\mathrm{id}(i, j; t, n)$. We, therefore, have

$$(II)^T = W_{[n,t]}(A \ltimes DB)^T.$$

It follows that

$$II = (A \ltimes DB)W_{[t,n]}.$$

As for the general case, i.e., $s > 1$, denote the ith row of B by B_i, then

$$\begin{aligned}
&D(A \ltimes B)\\
&= (D(A \ltimes B_1), \cdots, D(A \ltimes B_s))\\
&= (DA \ltimes B_1 + (A \ltimes DB_1)W_{[n,t]}, \cdots, DA \ltimes B_s + (A \ltimes DB_s)W_{[n,t]})\\
&= (DA \ltimes B_1, \cdots, DA \ltimes B_s) + ((A \ltimes DB_1)W_{[n,t]}, \cdots, (A \ltimes DB_s)W_{[n,t]})\\
&= (DA \ltimes B) + ((A \ltimes DB_1), \cdots, (A \ltimes DB_s))(I_s \ltimes W_{[n,t]})\\
&= (DA \ltimes B) + (A \ltimes DB)(I_s \ltimes W_{[n,t]}).
\end{aligned}$$

Hence (18.37) holds. □

We use the following example to verify the above formula.

Example 18.10. Assume

$$A = \begin{bmatrix} x_1^2 & x_1x_2 & x_1x_2 & x_2^2 \\ x_2^2 & x_1x_2 & x_1x_2 & x_1^2 \end{bmatrix}, \quad B = \begin{bmatrix} \sin(x_1+x_2) & \cos(x_1+x_2) \\ -\cos(x_1+x_2) & \sin(x_1+x_2) \end{bmatrix}.$$

Then

$$DA = \begin{bmatrix} 2x_1 & 0 & x_2 & x_1 & x_2 & x_1 & 0 & 2x_2 \\ 0 & 2x_2 & x_2 & x_1 & x_2 & x_1 & 2x_1 & 0 \end{bmatrix},$$

$$DB = \begin{bmatrix} \cos(x_1+x_2) & \cos(x_1+x_2) & -\sin(x_1+x_2) & -\sin(x_1+x_2) \\ \sin(x_1+x_2) & \sin(x_1+x_2) & \cos(x_1+x_2) & \cos(x_1+x_2) \end{bmatrix}.$$

For notational brevity, we denote $S := \sin(x_1+x_2)$, $C := \cos(x_1+x_2)$. Then

$$\begin{aligned}
&D(A(x) \ltimes B(x))\\
&\quad = \begin{bmatrix} 2x_1 & 0 & x_2 & x_1 & x_2 & x_1 & 0 & 2x_2 \\ 0 & 2x_2 & x_2 & x_1 & x_2 & x_1 & 2x_1 & 0 \end{bmatrix} \ltimes \begin{bmatrix} S & C \\ -C & S \end{bmatrix}\\
&\qquad + \begin{bmatrix} x_1^2 & x_1x_2 & x_1x_2 & x_2^2 \\ x_2^2 & x_1x_2 & x_1x_2 & x_1^2 \end{bmatrix} \ltimes \begin{bmatrix} C & C & -S & -S \\ S & S & C & C \end{bmatrix} \ltimes \left(I_2 \otimes W_{[2]}\right) := (a_{ij}),
\end{aligned}$$

where, we have

$$\begin{array}{ll}
a_{11} = 2x_1S - x_2C + x_1^2C + x_1x_2S & a_{12} = -x_1C + x_1^2C + x_1x_2S \\
a_{13} = x_2S + x_1x_2C + x_1^2S & a_{14} = x_1C + 2x_2S + x_1x_2C + x_2^2S \\
a_{15} = 2x_1C + x_2S - x_1^2S + x_1x_2C & a_{16} = x_1S - x_1^2S + x_1x_2C \\
a_{17} = x_2C - x_1x_2S + x_2^2C & a_{18} = x_1C + 2x_2S - x_1x_2S + x_2^2C \\
a_{21} = -x_2C + x_2^2C + x_1x_2S & a_{22} = 2x_2S - x_1C + x_2^2C + x_1x_2S \\
a_{23} = x_2S - 2x_1C + x_1x_2C + x_1^2S & a_{24} = x_1S + x_1x_2C + x_1^2S \\
a_{25} = x_1S - x_2^2S + x_1x_2C & a_{26} = 2x_2C + x_1S - x_2^2S + x_1x_2C \\
a_{27} = x_2C + 2x_1S - x_1x_2S + x_1^2C & a_{28} = x_1C - x_1x_2S + x_1^2C.
\end{array}$$

To verify the above result a direct calculation yields

$$\begin{aligned} & A(x) \ltimes B(x) \\ & \quad = \begin{bmatrix} x_1^2S - x_1x_2C & x_1x_2S - x_2^2C & x_1^2C + x_1x_2S & x_1x_2C + x_2^2S \\ x_1^2S - x_1x_2C & x_1x_2S - x_1^2C & x_2^2C + x_1x_2S & x_1x_2C + x_2^2S \end{bmatrix}. \end{aligned}$$

Differentiating it, we have the same result as in the above.

Applying Proposition 18.9 to conventional matrix product yields the following result.

Corollary 18.1. *For conventional product of functional matrices we have*

$$D(A(x)B(x)) = DA(x) \ltimes B(x) + A(x)DB(x). \tag{18.39}$$

Proof. Note that $W_{[1,k]} = W_{[k,1]} = I_k,\ k > 0$. Then it is clear that (18.39) is a special case of (18.37). □

Next, we consider the case of $A(x) \prec_t B(x)$, $t > 1$. We intend to use the formula $A(x) \ltimes B(x) = (A(x) \otimes I_t)B(x)$.

First, we give a lemma, which itself is useful.

Lemma 18.1. *Assume $A(x) \in M_{p\times q}$, $x \in \mathbb{R}^n$. Then*

$$D(A \otimes I_k) = (DA \otimes I_k)(I_q \otimes W_{[k,n]}). \tag{18.40}$$

Proof. First we split the columns of both $D(A \otimes I_k)$ and $(DA \otimes I_k)$ into $p \times q$ equal blocks, and denote them as $D(A \otimes I_k) = E = \{E_{ij}\}$ and $(DA \otimes I_k) = F = \{F_{ij}\}$ respectively. Then

$$E_{ij} = D(a_{ij}I_k), \quad F_{ij} = da_{ij} \otimes I_k.$$

Assume the columns of F_{ij} are labeled by (p, q) and arranged in the order of $\mathrm{id}(p, q; n, k)$. Then it is easy to verify that E_{ij} consists of the same set of columns, indexed by (p, q) but in the order of $\mathrm{id}(q, p; k, n)$. According to Proposition 2.9 we have $E_{ij} = F_{ij}W_{[k,n]}$. Now set

$$W = \mathrm{diag}(\underbrace{W_{[k,n]}, \cdots, W_{[k,n]}}_{q}).$$

Then it is clear that $E = FW$, which leads to (18.40). □

Proposition 18.10. *Let $A(x) \in M_{p\times q}$, and $A(x) \prec_t B(x)$, $x \in \mathbb{R}^n$. Then*

$$D(A(x) \ltimes B(x)) = DA(x) \ltimes (I_q \otimes W_{[t,n]}) \ltimes B(x) + A \ltimes DB(x). \tag{18.41}$$

Proof. Since $A(x) \ltimes B(x) = (A(x) \otimes I_t)B(x)$, using (18.39) and (18.40), we have

$$\begin{aligned} D(A(x) \ltimes B(x)) &= D[(A(x) \otimes I_t)B(x)] \\ &= (DA \otimes I_t)(I_q \otimes W_{[t,n]}) \ltimes B(x) + A(x) \ltimes DB(x). \end{aligned}$$

Equation (18.41) follows. □

Following example shows how to use STP and the differential of functional matrix to calculate the Lie derivatives of vector fields.

Example 18.11. Let $f(x), g(x) \in V(M)$ be two vector fields on M.

(1) The Lie derivative of $g(x)$ with respect to $f(x)$ can locally expressed (in a coordinate chart) as (Isidori, 1995)

$$\mathrm{ad}_f(g) = [f, g] = Dgf - Dfg. \tag{18.42}$$

(2) Consider the second order Lie derivative. Using (18.37), we have

$$\begin{aligned} \mathrm{ad}_f^2(g) &= D(Dgf - Dfg)f - Df(Dgf - Dfg) \\ &= D^2 g \ltimes f^2 + DgDff - D^2 f \ltimes g \ltimes f - 2DfDgf + (Df)^2 g. \end{aligned}$$

In the following we consider the higher order differential of products. We need some preparations. First, we define an $(i+1)$th index as

$$S(i,k) = \left\{ d = (d_1, \cdots, d_{i+1}) \in \mathbb{Z}_+^{i+1} \;\middle|\; \sum_{j=1}^{i+1} d_j = k \right\}.$$

If $d = (d_1, \cdots, d_{i+1}) \in S(i,k)$, We define the differential $D_d(B)$ as

$$D_d B = D[(\cdots D(D(B \otimes I_{n^{d_1}}) \otimes I_{n^{d_2}}) \cdots \otimes I_{n^{d_i}}) \otimes I_{n^{d_{i+1}}}]. \tag{18.43}$$

Using it, we define

$$D^{S(i,k)} B = \sum_{d \in S(i,k)} D_d B. \tag{18.44}$$

With these notations we have

Corollary 18.2. *For conventional matrix product we have*

$$D^k(A(x)B(x)) = \sum_{j=0}^{k} D^j A(x) D^{S(k-j,j)}(B(x)). \tag{18.45}$$

***Proof*.** We prove it by mathematical induction. When $k = 1$, it reduces to (18.39). Assume (18.45) holds for k. For $k + 1$ we consider the terms of $D^{j+1}AD^{S(k-j,j+1)}B$, which are produced from $D^{k+1}(A(x)B(x))$. They come from two groups of terms of $D^k(A(x)B(x))$, namely, $D^jAD^{S(k-j,j)}B$ and $D^{j+1}AD^{S(k-j-1,j+1)}B$. Differentiating the $A(x)$ of the terms in first group and the $B(x)$ of the terms in the second group, yields the required terms. Then it suffices to prove that

$$D^{S(k-j,j)}B \otimes I_n + D[D^{S(k-j-1,j+1)}B] = D^{S(k-j,j+1)}B. \tag{18.46}$$

Note that for

$$d = (d_1, \cdots, d_{k-j}, d_{k-j+1} + 1), \quad d_1 + \cdots + d_{k-j+1} = j,$$

the first group of the left hand side contains all D_d, meanwhile, for

$$d = (d_1, \cdots, d_{k-j}, 0), \quad d_1 + \cdots + d_{k-j} = j + 1,$$

the second group of the left hand side also contains all D_d. The non-overlapped part of these two groups of d forms the set of $S(k - j, j + 1)$. Hence, one sees easily that (18.45) follows. □

We give an example for the calculation.

Example 18.12. Assume $k = 3$. We calculate the differential

$$\begin{cases} D^{S(3,0)}B = D^3B, \\ D^{S(2,1)}B = D^2(B \otimes I_n) + D(DB \otimes I_n) + D^2(B) \otimes I_n, \\ D^{S(1,2)}B = D(B \otimes I_{n^2}) + D(B \otimes I_n) \otimes I_n + DB \otimes I_{n^2}, \\ D^{S(0,3)}B = D^3B \otimes I_{n^3}. \end{cases}$$

Hence,

$$\begin{aligned} D^3(AB) = AD^3B & \\ & +DA \ltimes (D^2(B \otimes I_n) + D(DB \otimes I_n) + D^2(B) \otimes I_n) \\ & +D^2A \ltimes (D(B \otimes I_{n^2}) + D(B \otimes I_n) \otimes I_n + DB \otimes I_{n^2}) \\ & +D^3A(B \otimes I_{n^3}). \end{aligned}$$

Next, we consider the gradient.

Lemma 18.2.

(1) Let $x = (x_1, \cdots, x_n)^T \in \mathbb{R}^n$ be a column vector. Then

$$x^T = V_r^T(I_n) \ltimes x. \tag{18.47}$$

(2) Let $x(x_1, \cdots, x_n) \in \mathbb{R}^n$ be a row vector. Then

$$x^T = x \ltimes V_r(I_n). \tag{18.48}$$

A straightforward computation can verify this lemma, and we leave it to the reader. Using this lemma, we have the following formulas, which convert the differential to gradient and vice versa.

Proposition 18.11. *Assume $A(x) \in M_{p\times q}$, where $x \in \mathbb{R}^n$. Then*

$$DA(x) = (I_p \otimes V_r^T(I_n)) \ltimes \nabla A(x). \tag{18.49}$$

Conversely, we have

$$\nabla A(x) = DA(x) \ltimes (I_q \otimes V_r(I_n)). \tag{18.50}$$

***Proof*.** Splitting both $DA(x)$ and $\nabla A(x)$ into $p \times q$ equal blocks in a natural way. Then what we have to do is to transpose each block, which is an n-dimensional row vector, into a column vector, or vice versa. Using (18.47) and (18.48) and the block multiplication, both formulas can be easily verified. □

Before the end of this section we give the formula for the gradient of product of matrices.

Proposition 18.12. *Given two functional matrices $A(x)$ and $B(x)$, where $x \in \mathbb{R}^n$.*

(1) If $A(x) \prec B(x)$, then

$$\nabla(A(x)B(x)) = (\nabla A(x))B(x) + A(x)(\nabla B(x)). \tag{18.51}$$

(2) if $A(x) \succ_t B(x)$, precisely, let $A(x) \in M_{m\times tp}$, $B(x) \in M_{p\times q}$. Then

$$\nabla(A(x)B(x)) = (\nabla A(x))B(x) + A(x)DB(x)[I_{qt} \otimes V_r(I_n)]. \tag{18.52}$$

***Proof*.** Using the block multiplication rule of semi-tensor product, we can verify (18.51) by straightforward computation. Consider (18.52): the first term can be obtained by block multiplication; as for the second term, without loss of generality, we can assume $A(x) = A = \text{const}$, Since $A \succ B(x)$, then

$$\nabla(AB(x)) = D(AB(x)) \ltimes (I_{qt} \otimes V_r(I_n)) = ADB(x)[I_{qt} \otimes V_r(I_n)].$$

□

18.3 Conversion of Generators

As we discussed before that x^k is a generator of B_n^k, but not a basis. It follows that the coefficient matrix F in (18.1) is not unique. Hence, we need a basis. A basis of B_n^k, called the natural basis and denoted by N_n^k, is defined as

$$N_n^k := \left\{ x_1^{d_1} \cdots x_n^{d_n} \;\middle|\; \sum_{j=1}^n d_j = k \right\}.$$

For convenience N_n^k is also used for the matrix consists of its elements arranging in the alphabetic order. That is,

$$x_1^{d_1} \;\cdots\; x_n^{d_n} \prec x_1^{t_1} \;\cdots\; x_n^{t_n},$$

if and only if there is an index $i \le n$ such that $d_s = t_s,\ s < i$ and $d_i > t_i$. That is,

$$N_n^k = [x_1^k \; x_1^{k-1} x_2 \;\cdots\; x_1^{k-1} x_n \;\cdots\; x_n^{k-1} x_1 \; x_n^{k-1} x_2 \;\cdots\; x_n^k].$$

Briefly denote $\mathrm{id}(I; n^k) := \mathrm{id}(i_1, \cdots, i_k; \underbrace{n, \cdots, n}_{k})$. It is easy to see that the

$$x_{i_1} \;\cdots\; x_{i_k} \in x^k$$

is arranged in the order of $\mathrm{id}(I; n^k)$.

Define an index subset of $\mathrm{id}(I; n^k)$ as

$$\mathrm{is}(I; n^k) = \{(i_1, \cdots, i_k) \in \mathrm{id}(I; n^k) \,|\, i_1 \le i_2 \le \cdots \le i_k\},$$

which is called the symmetric index set. Note that for the set of elements in a generator, symmetric index can pick out different elements and ignore repeated elements. Precisely, if we pick out only the elements which have non-decreasing multi-index, we pick out all the different elements without repeats. For instance, $x_1x_1x_2$, $x_1x_2x_1$, and $x_2x_1x_1$ are components (or elements) of x^3, where $x \in \mathbb{R}^2$. In other words, (112) (121) (211) $\in \mathrm{id}(I; 2^3)$. But only (112) $\in \mathrm{is}(I; 2^3)$.

Recall Definition 18.2, and let $I = (i_1, \cdots, i_k) \in \mathrm{id}(I; n^k)$. Then the index frequency of I, denoted by $C_I \in \mathbb{Z}_+^n$, is defined as $C_I(k)$ is the times for k appearing into I. For instance, assume $I = (11424) \in \mathrm{id}(I; 6^5)$, then we have $C_I = (210200)$.

Now we define a mapping $\xi : \mathrm{id}(I; n^k) \to \mathbb{Z}_+^n$ by $I \mapsto C_I$. The following proposition follows from definition immediately.

Proposition 18.13.

(1) Denote the region of ξ as

$$\mathcal{I}m(\xi) := R(n^k) = \left\{ (c_1, \cdots, c_n) \in \mathbb{Z}_+^n \;\middle|\; \sum_{j=1}^{n} c_j = k \right\}. \tag{18.53}$$

(2) If ξ is restricted on $\mathrm{is}(I; n^k)$, then $\xi|_{\mathrm{is}(I;n^k)} : \mathrm{is}(I; n^k) \to R(n^k)$ is an bijective.

(3) Define the order at $R(n^k)$, denoted by $\prec$ as

$$(c_1, \cdots, c_n) \prec (d_1, \cdots, d_n),$$

if there is an $s \le n$ such that $c_1 = d_1$, $\cdots$, $c_{s-1} = d_{s-1}$ and $c_s > d_s$. That is,

$$R(n^k) = \{(k\ 0\ \cdots\ 0), (k-1\ 1\ \cdots\ 0), \cdots, (0\ \cdots\ 0\ 1\ k-1), (0\ \cdots\ 0\ k)\}.$$

It is reasonable to assume that the order in $\mathrm{is}(I; n^k)$ is inherited from $\mathrm{id}(I; n^k)$. Under this order it is easy to verify that $\xi : \mathrm{is}(I; n^k) \to R(n^k)$ is an order-reserve mapping.

Note that $R(n^k)$ is also a set of indices, which are used to label the natural basis N_n^k. For instance, assume $I = (11335) \in \mathrm{id}(I; 6^5)$, then it is used to label an element in B_6^5, which is $x_1x_1x_3x_3x_5$. If we use the symmetric index set $Is(I; 6^5)$ to label the elements in B_6^5, the indices and the elements have one-one correspondence. But it is still not convenient because the monomial forms are not very natural. According to the Proposition 18.13, we can use C_I to label the elements in N_n^k. For instance, for $I = (11335)$ we have $C_I(I) = (2, 0, 2, 0, 1, 0)$, which labels the element $x_1^2x_2^0x_3^2x_4^0x_5^2x_6^0 = x_1^2x_3^2x^5$. Using this set of indices, we denote the basis of N_n^K as

$$\left\{ x^{C_I} = \prod_{j=1}^{n} x_j^{c_j} \;\middle|\; I \in \mathrm{is}(I; n^k) \right\}.$$

Proposition 18.14. *The cardinality (size) of $\mathrm{is}(I; n^k)$ is*

$$|\,\mathrm{is}(I; n^k)| = \frac{(n+k-1)!}{k!(n-1)!}, \quad k \ge 0,\ n \ge 1. \tag{18.54}$$

Proof. Since $\xi : \mathrm{is}(I; n^k) \to R(n^k)$ is a bijective, we need only to consider the size of

$$R(n^k) = \left\{ (c_1, \cdots, c_n) \in \mathbb{Z}_+^n \;\middle|\; \sum_{j=1}^{n} c_j = k \right\}.$$

Let c_1 runs from 0 to k, it obvious that

$$|R(n^k)| = \sum_{j=0}^{k} \left|R((n-1)^j)\right|.$$

Using equality

$$\binom{n-1}{0} + \binom{n}{1} + \cdots + \binom{n+k-1}{k} = \binom{n+k}{k}, \tag{18.55}$$

then the conclusion can be proved by mathematical induction. □

It is interesting that the cardinality of $\text{is}(I; n^k)$ can be checked from the following Table 18.1. In this table all the elements in the first row and first column are 1. All other elements are the sum of their upper element and left element.

Table 18.1 Cardinality of $\text{is}(I; n^k)$

$k\backslash \dim n$	1	2	3	4	5	6	$\cdots$
0	1	1	1	1	1	1	
1	1	2	3	4	5		
2	1	3	6	10			
3	1	4	10				
4	1	5					
5	1						
$\vdots$							

Next, we consider how to convert the coefficient of a polynomial with respect to generator B_n^k to the coefficient with respect to natural basis $T_N(n, k)$ and vice versa.

Denote by $s = |\,\text{is}(I; n^k)|$, $t = n^k$, then $T_B(n, k) \in M_{s\times t}$, $T_N(n, k) \in M_{t\times s}$ are constructed as follows.

(1) Constructing $T_B(n, k)$:
Step 1. Label the columns of an $s \times t$ matrix by $\text{id}(J; n^k)$, and the rows by $\text{is}(I; n^k)$.
Step 2 . Assign its entries as follows:
Assume the row index is $I = (i_1, i_2, \cdots, i_k) \in \text{is}(I; n^k)$, where $i_1 \le i_2 \le \cdots \le i_k$. For any element in this column, whose row multi-index is $J \in P_I$, that is, $J = (i_{\sigma(1)}, i_{\sigma(2)}, \cdots, i_{\sigma(k)})$, for certain $\sigma \in S_k$, precisely, the column multi-index is a permutation of its row multi-index, then set the entry to be

$$\alpha_I = \frac{C_I!}{k!} = \frac{\prod_{j=1}^{n} C_I(j)!}{k!},$$

otherwise, set it to be 0.

That is, the matrix $T_B(n,k)$ is constructed by setting its elements as

$$\beta_{I,J}=\begin{cases}\alpha_I, & J\in P_I,\\ 0, & \text{otherwise},\end{cases}$$

where P_I is the set of permutations of I.

(2) Constructing $T_N(n,k)$:

Step 1. Let T be a $t\times s$ matrix. Label its rows by $\mathrm{id}(J;n^k)$, and its columns by $Is(I;n^k)$.

Step 2. Assign its entries as follows:

Assume the column multi-index is $J=(j_1,j_2,\cdots,j_k)\in \mathrm{is}(J;n^k)$. For each element in this column, if its row index $I\in P_J$, set it to be 1, otherwise, set it to be 0.

That is, the matrix $T_N(n,k)$ is constructed by setting its elements as

$$n_{I,J}=\begin{cases}1, & I\in P_J,\\ 0, & \text{otherwise}.\end{cases}$$

We simply denote the basis N_n^k by $x_{(k)}$. For instance, when $n=3$, $x_{(2)}$ is expressed as $x_{(2)}=(x_1^2,\ x_1x_2,\ x_1x_3,\ x_2^2,\ x_2x_3,\ x_3^2)^T$.

The following proposition is an immediate consequence of the construction.

Proposition 18.15.

(1) $x_{(k)}$ and x^k satisfy the following relation:

$$\begin{cases}x_{(k)}=T_B(n,k)x^k,\\ x^k=T_N(n,k)x_{(k)}.\end{cases}\tag{18.56}$$

(2) Assume $p(x)\in B_n^k$ is a kth degree homogeneous polynomial, and $p(x)=Fx^k=Sx_{(k)}$, then

$$S=FT_N(n,k).\tag{18.57}$$

Moreover, $ST_B(n,k)$ is a symmetric expression of F. Particularly, if F is symmetric, then $F=ST_B(n,k)$.

Example 18.13. Let $n=2$ and $k=3$.

(1) The matrix $T_B(n,k)$ is

$$T_B(2,3) = \begin{array}{c} \begin{array}{cccccccc} (111) & (112) & (121) & (122) & (211) & (212) & (221) & (222) \end{array} \\ \left[\begin{array}{cccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \end{array} \begin{array}{c} \\ (111) \\ (112) \\ (122) \\ (222) \end{array} . \tag{18.58}$$

Similarly, we have

$$T_N(2,3) = \begin{array}{c} \begin{array}{cccc} (111) & (112) & (122) & (222) \end{array} \\ \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right] \end{array} \begin{array}{c} \\ (111) \\ (112) \\ (121) \\ (122) \\ (211) \\ (212) \\ (221) \\ (222) \end{array} . \tag{18.59}$$

(2) Assume $f(x) = x_1^3 + 3x_1^2x_2 - x_1x_2^2 + x_2^3$, it can be rewritten as $f(x) = (1\ 3\ {-1}\ 1)x_{(3)}$. Using (18.58), we have

$$f(x) = (1\ 3\ {-1}\ 1)T_B(2,3)x^3 = \left(1\ 1\ 1\ {-\frac{1}{3}}\ 1\ {-\frac{1}{3}}\ {-\frac{1}{3}}\ 1\right)x^3.$$

(3) Assume $f(x) = (1\ 2\ 1\ 1\ {-1}\ {-1}\ {-2}\ {-1})x^3$, using (18.59), we have

$$f(x) = (1\ 2\ 1\ 1\ {-1}\ {-1}\ {-2}\ {-1})T_N(2,3)x_{(3)} = (1\ 2\ {-2}\ {-1})x_{(3)}.$$

Then we can have the symmetric expression of $f(x)$ as

$$\begin{aligned} (1\ 2\ \ -2\ \ -1)x_{(3)} &= (1\ 2\ {-2}\ \ -1)T_B(2,3)x^3 \\ &= \left(1\ \frac{2}{3}\ \frac{2}{3}\ \ -\frac{2}{3}\ \frac{2}{3}\ {-\frac{2}{3}}\ {-\frac{2}{3}}\ {-1}\right)x^3. \end{aligned}$$

18.4 Taylor Expansion of Multi-Variable Functions

The differential of functional matrices provides a condensed form of Taylor expansion of multi-variable functions. The following expansion is easily verifiable.

Theorem 18.2 (Taylor Series Expansion). *Assume $F:\mathbb{R}^n\to\mathbb{R}^m$ is an analytic mapping, then its Taylor series expansion is*

$$F(x)=F(x_0)+\sum_{k=1}^{\infty}\frac{1}{k!}D^kF(x_0)\ltimes(x-x_0)^k,\quad x\in\mathbb{R}^n. \tag{18.60}$$

There is a similar expression in Abraham and Marsden (1978), where terms of $(x-x_0)^k$ are understood as tensor product of vectors. But without semi-tensor product, it can hardly be used in both theoretical analysis and numerical computation. An obvious advantage of (18.60) is: it has the same expression as single variable mappings. Another advantage is that within each term there is only one product, i.e., the STP. Hence the factors are associative, etc.

As an application of the Taylor series expansion and the calculation of semi-tensor form of multi-variable polynomials, we consider the Taylor series expansion of the inverse mapping of a local diffeomorphism.

Let $F:\mathbb{R}^n\to\mathbb{R}^n$ be an analytic mapping with $F(0)=0$. Otherwise, replace F by $F-F(0)$. Using Taylor series expansion, we can express it as

$$y=F_1x+F_2x^2+F_3x^3+\cdots, \tag{18.61}$$

where

$$F_k=\frac{1}{k!}D^kF|_0,\quad k\geq 1.$$

Using (18.17), we have

$$\begin{bmatrix}y\\y^2\\\vdots\\y^k\end{bmatrix}=\begin{bmatrix}A_{11}&A_{12}&A_{13}&\cdots&A_{1k}\\0&A_{22}&A_{23}&\cdots&A_{2k}\\\vdots&&&&\\0&0&0&\cdots&A_{kk}\end{bmatrix}\begin{bmatrix}x\\x^2\\\vdots\\x^k\end{bmatrix}+O(\|x\|^{k+1}), \tag{18.62}$$

where $A_{1i}=F_i$, $i=1,2,\cdots$, and the A_{ki}, $k>1$ can be calculated as

$$A_{ki}=\sum_{j_1+j_2+\cdots+j_k=i}(F_{j_1}\otimes F_{j_2}\otimes\cdots\otimes F_{j_k}),\quad i\geq k. \tag{18.63}$$

Assume $y=F(x)$ is a local diffeomorphism, then $F_1=A_{11}=J_F$, as the Jacobi matrix of F at the origin, is invertible. Moreover, because $A_{kk}=\underbrace{F_1\otimes\cdots\otimes F_1}_{k}$, it is also invertible. Then it follows from (18.62) that

$$\begin{bmatrix}x\\x^2\\\vdots\\x^k\end{bmatrix}=\begin{bmatrix}B_{11}&B_{12}&B_{13}&\cdots&B_{1k}\\0&B_{22}&B_{23}&\cdots&B_{2k}\\\vdots&&&&\\0&0&0&\cdots&B_{kk}\end{bmatrix}\begin{bmatrix}y\\y^2\\\vdots\\y^k\end{bmatrix}+R_{k+1}. \tag{18.64}$$

Denote the coefficient matrix on the right hand side of (18.64) by B^k, then B^k can be inductively expressed as

$$\begin{cases} B^1 = B_{11} = F_1^{-1}, \\ B^{t+1} = \begin{bmatrix} B^t & -B^t A^{t,t+1} A^{-1}_{t+1,t+1} \\ 0 & A^{-1}_{t+1,t+1} \end{bmatrix}, \quad t \geq 1, \end{cases} \tag{18.65}$$

where

$$A^{-1}_{t+1,t+1} = \underbrace{F_1^{-1} \otimes \cdots \otimes F_1^{-1}}_{t+1}, \quad A^{t,t+1} = \begin{bmatrix} A_{1,t+1} \\ \vdots \\ A_{t,t+1} \end{bmatrix}.$$

Theorem 18.3. *Let $y = F(x) : \mathbb{R}^n \to \mathbb{R}^n$, $F(0) = 0$ be a local diffeomorphism around the origin. Then its inverse mapping $x = F^{-1}y$ has the following Taylor series expansion as*

$$x = B_{11}y + B_{12}y^2 + \cdots + B_{1k}y^k + O(\|y\|^{k+1}),$$

where $B_{11}, \cdots, B_{1k}$ are shown in the first row of B^k, which is calculated by (18.65).

Proof. In fact, this form is a summary of the above discussion. The only thing we need to notice is: the remaining R_{k+1} of (18.64). Note that for a local diffeomorphism $y = F(x)$, $y(0) = 0$, we have $O(\|x\|^k) = O(\|y\|^k)$, which yields that $R_{k+1} = O(\|y\|^{k+1})$. □

In solving practical engineering problems it is convenient to express a Taylor series expression of a multi-variable mapping over the natural form, which has no redundant terms. To this end, we define two matrices as follows.

$$\begin{cases} T^N(n,k) = \text{diag}(I_n, T_N(n,2), T_N(n,3), \cdots, T_N(n,k)), \\ T^B(n,k) = \text{diag}(I_n, T_B(n,2), T_B(n,3), \cdots, T_B(n,k)). \end{cases}$$

Using some properties of the matrices $T_B(n,k)$ and $T_N(n,k)$, it is easy to see that if

$$\begin{bmatrix} x \\ x^2 \\ \vdots \\ x^k \end{bmatrix} = B^k \begin{bmatrix} y \\ y^2 \\ \vdots \\ y^k \end{bmatrix},$$

then under the natural basis it becomes

$$\begin{bmatrix} x \\ x_{(2)} \\ \vdots \\ x_{(k)} \end{bmatrix} = T^B(n,k)B^kT^N(n,k)\begin{bmatrix} y \\ y_{(2)} \\ \vdots \\ y_{(k)} \end{bmatrix}. \tag{18.66}$$

Example 18.14. Consider a mapping $y = F(x)$ as

$$\begin{cases} y_1 = \sin(x_1) + x_2 - x_2^2, \\ y_2 = \log(1 + x_1 - x_2). \end{cases} \tag{18.67}$$

Using Taylor series expansion, (18.67) can be expressed as

$$\begin{cases} y_1 = x_1 + x_2 - x_2^2 - \frac{1}{6}x_1^3 + O(\|x\|^4), \\ y_2 = x_1 - x_2 - \frac{1}{2}(x_1 - x_2)^2 + \frac{1}{3}(x_1 - x_2)^3 + O(\|x\|^4). \end{cases} \tag{18.68}$$

A straightforward computation yields the following result.

Table 18.2 Coefficients of the Taylor series expansion

y\\x	x_1	x_2	x_1^2	x_1x_2	x_2^2	x_1^3	$x_1^2x_2$	$x_1x_2^2$	x_2^3	$\cdots$
y_1	1	1	-1	0	0	-1/6	0	0	0	$\cdots$
y_2	1	-1	-1/2	1	-1/2	1/3	-1	1	-1/3	$\cdots$
y_1^2	0	0	1	2	1	0	-2	0	-2	$\cdots$
y_1y_2	0	0	1	0	-1	-1/2	1/2	1/2	-1/2	$\cdots$
y_2^2	0	0	1	-2	1	-1	3	-3	1	$\cdots$
y_1^3	0	0	0	0	0	1	3	3	1	$\cdots$
$y_1^2y_2$	0	0	0	0	0	1	1	-1	-1	$\cdots$
$y_1y_2^2$	0	0	0	0	0	1	-1	-1	1	$\cdots$
y_2^3	0	0	0	0	0	1	-3	3	1	$\cdots$
$\vdots$										

Then we have the coefficient matrix of the inverse mapping as in 18.3.

From the first block of the Table 18.3 we can find the inverse mapping of $y = F(x)$, defined in (18.67), as follows.

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = F^{-1}(y)$$

$$= \begin{bmatrix} 0.5y_1 + 0.5y_2 + 0.125y_1^2 + 0.25y_1y_2 + 0.375y_2^2 \\ +0.0208y_1^3 + 0.125y_1^2y_2 + 0.0625y_1y_2^2 + 0.2083y_2^3 \\ \\ 0.5y_1 - 0.5y_2 + 0.125y_1^2 + 0.25y_1y_2 - 0.125y_2^2 \\ +0.764y_1^3 - 0.0417y_1^2y_2 + 0.2292y_1y_2^2 - 0.0139y_2^3 \end{bmatrix} + O(\|y\|^4).$$

Table 18.3 The coefficients of the inverse mapping

x$\backslash y$	y_1	y_2	y_1^2	y_1y_2	y_2^2	y_1^3	$y_1^2y_2$	$y_1y_2^2$	y_2^3
x_1	0.5	0.5	0.125	0.25	0.375	0.0208	0.125	0.0625	0.2083
x_2	0.5	-0.5	0.125	0.25	-0.125	0.0764	-0.0417	0.2294	-0.0139
x_1^2	0	0	0.25	0.5	0.25	0.0833	0	0.25	0.1667
x_1x_2	0	0	0.25	0	-0.25	0.25	-0.5	0.5	-0.5
x_2^2	0	0	0.25	-0.5	0.25	0.0833	0	-0.25	0.1667
x_1^3	0	0	0	0	0	0.0833	0.5	0.25	0.1667
$x_1^2x_2$	0	0	0	0	0	0.1667	0	0	-0.1667
$x_1x_2^2$	0	0	0	0	0	0.0833	0	-0.25	0.1667
x_2^3	0	0	0	0	0	0.1667	-0.5	0.5	-0.1667
$\vdots$									

18.5 Fundamental Formula of Differential

Consider an analytic function $h(x) \in C^\omega(\mathbb{R}^n)$. Using Taylor series expansion, we have

$$h(x) = h_0 + h_1 x + h_2 x^2 + \cdots ,$$

where h_k are $1 \times n^k$ constant matrices (precisely, row vectors). Hence, the differential of $h(x)$ can be expressed as

$$Dh(x) = h_1 + h_2 D(x^2) + h_3 D(x^3) + \cdots . \tag{18.69}$$

Similarly, consider an analytic vector field $X(x) \in V^\omega(M)$. Using Taylor series expansion, we have

$$X(x) = X_0 + X_1 x + X_2 x^2 + \cdots ,$$

where X_k are $n \times n^k$ constant matrices. Then the Jacobian matrix of $X(x)$ is expressed as

$$J_X(x) = X_1 + X_2 D(x^2) + X_3 D(x^3) + \cdots . \tag{18.70}$$

Observing the above expressions, one sees easily that in geometric calculations the differential $D(x^k)$ plays a key role. This section aims on its formula.

Lemma 18.3.

$$\begin{aligned} D(x^k) = {} & W_{[n^{k-1},n]} x^{k-1} + x W_{[n^{k-2},n]} x^{k-2} + \cdots \\ & + x^{k-2} W_{[n,n]} x + x^{k-1} \otimes I_n, \qquad k \geq 2. \end{aligned} \tag{18.71}$$

***Proof*.** We prove it by mathematical induction. It is trivial that

$$Dx = I_n.$$

Using (18.37), we have

$$\begin{aligned} D(x^2) &= Dx \ltimes (1 \otimes W_{[n,n]}) \ltimes x + x \ltimes I_n \\ &= I_n \ltimes W_{[n,n]} \ltimes x + (x \otimes I_n) I_n = W_{[n,n]} \ltimes x + x \otimes I_n. \end{aligned}$$

Assume (18.71) holds for k. Invoking (18.40), we have

$$D(x \otimes I_{n^k}) = (I_n \otimes I_{n^k})(1 \otimes W_{[n^k,n]}) = I_{n^{k+1}} W_{[n^k,n]} = W_{[n^k,n]}.$$

Hence

$$\begin{aligned} D(x^{k+1}) &= D[(x \otimes I_{n^k})x^k] = D(x \otimes I_{n^k})x^k + (x \otimes I_{n^k})D(x^k) \\ &= W_{[n^k,n]}x^k + (x \otimes I_{n^k})(W_{[n^{k-1},n]}x^{k-1} \\ &\quad + xW_{[n^{k-2},n]}x^{k-2} + \cdots + x^{k-1} \otimes I_n) \\ &= W_{[n^k,n]}x^k + xW_{[n^{k-1},n]}x^{k-1} + \cdots + x^k \otimes I_n. \end{aligned}$$

□

The following theorem provides a fundamental differential formula.

Theorem 18.4. *Let $x = (x_1, \cdots, x_n)^T \in \mathbb{R}^n$. The differential of x^m satisfies the following formula.*

$$D(x^{k+1}) = \Phi_k^n x^k, \quad k \geq 0, \tag{18.72}$$

where

$$\Phi_k^n = \sum_{s=0}^{k} I_{n^s} \otimes W_{[n^{k-s},n]}. \tag{18.73}$$

***Proof*.** Invoking Lemma 18.3 and using the following swap formula

$$x^p W_{[n^s,n]} = (I_{n^p} \otimes W_{[n^s,n]})x^p,$$

(18.72) follows immediately. □

Remark 18.2. Since $I_1 = 1$ is a scalar and $W[1,n] = I_n$, It is easy to see that

$$\Phi_0^n = I_n.$$

If there is no possible confusion, Φ_k^n can be denoted briefly as Φ_k.

18.6 Lie Derivative

Lie derivatives of functions, vector fields, and forms with respect to vector fields are fundamental in nonlinear control theory. To avoid concerning the differentiable orders, all the objects are assumed to be analytic. They are denoted respectively by $C^\omega(\mathbb{R}^n)$, $V^\omega(\mathbb{R}^n)$, and $V^{*\omega}(\mathbb{R}^n)$ respectively. Moreover, we consider only the calculating formulas on $\mathbb{R}^n$, which are also available for general manifolds but within a coordinate chart.

Definition 18.5. Let $F : M = \mathbb{R}^n \to N = \mathbb{R}^n$ be a diffeomorphism.

(1) For a smooth function $h(x) \in C^\omega(N)$, F deduces a mapping $F^* : C^\omega(N) \to C^\omega(M)$, defined as

$$F^*(h) = h \circ F \in C^\omega(M).$$

(2) For a vector field $X \in V^\omega(M)$, F deduces a mapping $F_* : V^\omega(M) \to V^\omega(N)$, defined as

$$F_*(X)(h) = X(h \circ F), \quad \forall\, h \in C^\omega(N).$$

(3) For a co-vector field $\alpha \in V^{*\omega}(N)$, F deduces a mapping $F^* : V^{*\omega}(N) \to V^{*\omega}(M)$, defined as

$$\langle F^*(\alpha), X\rangle = \langle \alpha, F_*(X)\rangle, \quad \forall\, X \in V^r(M).$$

If F is a local diffeomorphism, all the above mappings are also locally defined.

Consider a vector field $X \in V^\omega(\mathbb{R}^n)$, Its integral curve with initial value $x(0) = x_0$ is denoted as $\phi_t^X(x_0)$. For convenience, we assume X is complete, that is, it is defined for all $t \in \mathbb{R}$. Then for each fixed t, the $\phi_t^X : \mathbb{R}^n \to \mathbb{R}^n$ is a diffeomorphism (Boothby, 1986).

Definition 18.6. Let $X \in V^\omega(M)$ and $h \in C^\omega(M)$. Then the Lie derivative of h with respect to X, denoted by $L_X(h)$, is defined by

$$L_X(h) = \lim_{t\to 0} \frac{1}{t}\left[(\phi_t^X)^* f(x) - f(x)\right]. \tag{18.74}$$

Proposition 18.16. *Under local coordinates (18.74) can be expressed as*

$$L_X(h) = \langle dh, X\rangle = \sum_{i=1}^n X_i \frac{\partial h}{\partial x_i}. \tag{18.75}$$

Proof. According to the definition, we have $(\phi_t^X)^* h(x) = h(\phi_t^X(x))$. Hence its Taylor expansion with respect to t is

$$h\left(\phi_t^X(x)\right) = h(x) + t df \cdot X(x) + O(t^2).$$

Plugging it into (18.74) yields (18.75). □

Definition 18.7. Let $X,\ Y \in V(M)$. The Lie derivative of Y with respect to X, denoted by $\mathrm{ad}_X(Y)$, is defined as

$$\mathrm{ad}_X(Y) = \lim_{t\to 0} \frac{1}{t}\left[(\phi_{-t}^X)_* Y(\phi_t^X(x)) - Y(x)\right]. \tag{18.76}$$

Proposition 18.17. *Under local coordinates (18.76) can be expressed by*

$$\mathrm{ad}_X(Y) = J_Y X - J_X Y = [X, Y]. \tag{18.77}$$

Where J_Y is the Jacobian matrix of Y. Precisely,

$$J_Y = \begin{bmatrix} \dfrac{\partial Y_1}{\partial x_1} & \cdots & \dfrac{\partial Y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial Y_n}{\partial x_1} & \cdots & \dfrac{\partial Y_n}{\partial x_n} \end{bmatrix}.$$

Proof. Using Taylor series expansion, we have

$$\phi_t^X(x) = x + (tX) + O(t^2). \tag{18.78}$$

$$Y(\phi_t^X(x)) = Y(x) + J_Y(tX) + O(t^2). \tag{18.79}$$

Using (18.78), the Jacobian matrix of ϕ_{-t}^X is

$$J_{\phi_{-t}^X} = I - tJ_X + O(t^2). \tag{18.80}$$

Invoking (18.78)∼(18.80), we have

$$\begin{aligned}(\phi_{-t}^X)_* Y(\phi_t^X(x)) &= (I - tJ_X + O(t^2))(Y(x) + J_Y(tX) + O(t^2)) \\ &= Y(x) + t(J_Y X - J_X Y) + O(t^2).\end{aligned}$$

Plugging it into (18.76) yields (18.77). □

Definition 18.8. Let $X \in V(M)$ and $\alpha \in V^{*\omega}(M)$. The Lie derivative of the co-vector field α with respect to X, denoted by $L_X(\alpha)$, is defined as

$$L_X(\alpha) = \lim_{t\to 0} \frac{1}{t}\left[(\phi_t^X)^* \alpha(e_t^X(x)) - \omega(x)\right]. \tag{18.81}$$

Proposition 18.18. *Under local coordinates (18.78) can be expressed as*

$$L_X(\alpha) = (J_{\alpha^T} X)^T + \alpha J_X. \tag{18.82}$$

Proof. Similar to the proof of Proposition 18.17, we first use Taylor series expansion to have

$$\begin{aligned}(\phi_t^X)^*\alpha(\phi_t^X(x)) &= (\alpha(x)+t(J_{\alpha^T}X)^T+O(t^2))(I+tJ_x+O(t^2))\\ &= \alpha(x)+t(J_{\alpha^T}X)^T+t\alpha(x)J_X+O(t^2).\end{aligned}$$

Here the transpose comes from the following convention: in local coordinate frame a co-vector field is always expressed as a row vector. Plugging the above equation into (18.81) yields (18.82). □

The higher order Lie derivatives can be defined inductively as follows:

$$L_X^{k+1}h = L_X^k(L_Xh), \quad k\ge 1; \tag{18.83}$$

$$\mathrm{ad}_X^{k+1}Y = \mathrm{ad}_X^k(\mathrm{ad}_X Y), \quad k\ge 1; \tag{18.84}$$

$$L_X^{k+1}\alpha = L_X^k(L_X\alpha), \quad k\ge 1. \tag{18.85}$$

Next, we consider the numerical calculation of Lie derivatives. First, we express function $h\in C^\omega(M)$, vector fields $X,Y\in V^\omega(M)$, and co-vector field $\alpha\in V^{*\omega}(M)$ into their Taylor series expansions as

$$\begin{aligned}h &= h_0+h_1x+h_2x^2+\cdots;\\ X &= X_0+X_1x+X_2x^2+\cdots;\\ Y &= Y_0+Y_1x+Y_2x^2+\cdots;\\ \alpha^T &= \alpha_0+\alpha_1x+\alpha_2x^2+\cdots.\end{aligned}$$

To give the formulas we need the following lemma, which itself is interesting. We leave its proof to the reader.

Lemma 18.4. *Let $X\in\mathbb{R}^n$ be a column vector. Then*

$$X^T = V_c^T(I_n)X; \tag{18.86}$$

$$X = X^TV_c^T(I_n). \tag{18.87}$$

Now we are ready to present the Taylor series expansions of Lie derivatives.

Proposition 18.19.

(1)

$$L_Xh = \sum_{i=0}^\infty c_ix^i, \tag{18.88}$$

where

$$c_i = \sum_{k=0}^i h_{k+1}\Phi_k^n(I_{n^k}\otimes X_{i-k}).$$

(2)

$$L_X Y = \sum_{i=0}^{\infty} d_i x^i, \tag{18.89}$$

where

$$d_i = \sum_{k=0}^{i} \left[Y_{k+1}\Phi_k^n (I_{n^k} \otimes X_{i-k}) - X_{k+1}\Phi_k^n (I_{n^k} \otimes Y_{i-k}) \right].$$

(3)

$$(L_X \alpha)^T = \sum_{i=0}^{\infty} e_i x^i, \tag{18.90}$$

where

$$e_i = \sum_{k=0}^{i} \left[\alpha_{k+1}\Phi_k^n (I_{n^k} \otimes X_{i-k}) - V_c^T \left(I_{n^k} \otimes X_{k+1}\Phi_k^n\right)^T \left(I_{n^k} \otimes \alpha_{i-k}\right) \right].$$

Proof. We prove equation (18.90) only. The proves of the other two formulas are similar. Invoking equation (18.82), we can obtain

$$(L_X \alpha)^T = \frac{\partial \alpha^T}{\partial x} X + \left(\frac{\partial X}{\partial x} \right)^T \alpha^T. \tag{18.91}$$

Consider its first term, which is

$$\begin{aligned} \tfrac{\partial \alpha^T}{\partial x} X &= \left(\sum_{i=1}^{\infty} \alpha_i \Phi_{i-1}^n x^{i-1} \right) \left(\sum_{i=0}^{\infty} X_i x^i \right) \\ &= \sum_{k=0}^{\infty} \left[\sum_{j=0}^{k} \alpha_{j+1} \Phi_j^n (I_{n^j} \otimes X_{k-j}) \right] x^k. \end{aligned} \tag{18.92}$$

$$\begin{aligned} (\tfrac{\partial X}{\partial x})^T &= \sum_{i=1}^{\infty} (x^{i-1})^T (\Phi_{i-1}^n)^T X_i^T \\ &= V_c^T (I_{n^{i-1}}) x^{i-1} (\Phi_{i-1}^n)^T X_i^T \\ &= V_c^T (I_{n^{i-1}}) \left[I_{n^{i-1}} \otimes X_i \Phi_{i-1}^n \right]^T x^{i-1}. \end{aligned}$$

Hence

$$\begin{aligned} (\tfrac{\partial X}{\partial x})^T \alpha^T &= \sum_{i=1}^{\infty} \sum_{k=0}^{i} \left[V_c^T I_{n^k} \left(I_{n^k} \otimes X_{k+1} \Phi_k^n\right)^T x^k \alpha_{i-k} x^{i-k} \right] \\ &= \sum_{i=1}^{\infty} \left[\sum_{k=0}^{i} V_c^T I_{n^k} \left(I_{n^k} \otimes X_{k+1} \Phi_k^n\right)^T \left(I_{n^k} \otimes \alpha_{i-k}\right) \right] x^i. \end{aligned} \tag{18.93}$$

Plugging (18.92) and (18.93) into (18.91) yields (18.90). □

Remark 18.3. This chapter involves many concepts and notations in differential geometry and control theory. We refer to van der Schaft (2000) and Cheng *et al.* (2000) for geometric concepts, and to Isidori (1995) for related control concepts.

Exercises

18.1 Assume $x = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ and

$$f(x_1, x_2, x_3) = x_2^3 - 2x_3^3 + 2x_1^2 x_2 - x_1^2 x_2 + x_1 x_2 x_3 \in B_3^3.$$

(i) Find a (non-symmetric) F such that $f(x) = Fx^3$.
(ii) Find a symmetric $\tilde{F}$ such that $f(x) = \tilde{F}x^3$.
(iii) Find a matrix G such that $f(x) = Gx_{(3)}$, where

$$x_{(3)} = (x_1^3, x_1^2 x_2, x_1^2 x_3, x_1 x_2^2, x_1 x_2 x_3, x_2^3, x_2^2 x_3, x_2 x_3^2, x_3^3)^T.$$

18.2 Assume $k = 5$, $n = 4$, and $I = (2, 2, 4, 4, 3) \in \mathrm{id}(i_1, i_2, i_3, i_4, i_5; 4^5)$.
(i) Calculate C_I, $|P_I|$.
(ii) List all the elements of P_I.

18.3 Calculate the matrix Ψ_3^3. Then use it to check that in Example 18.7 $(A_3 + DB_2)\Psi_3^3 = 0$ holds.

18.4 Given $f(x_1, x_2, x_3) = x_1 - x_2 + x_3 + 2x_1 x_2 - x_3^2$.
(i) Find F_1 and F_2 such that

$$f(x) = F_1 x + F_2 x^2,$$

where $x = (x_1, x_2, x_3)^T \in \mathbb{R}^3$.
(ii) Using F_1 and F_2 obtained above, calculate

$$F_1^2, \quad F_1 F_2, \quad F_2^2.$$

(iii) Check that

$$f^2(x) = F_1^2 x^2 + 2F_1 F_2 x^3 + F_2 x^4.$$

18.5 Give a similar necessary and sufficient condition as Proposition 18.7 for the polynomial $P(x) = A_0 + A_1 x + A_2 x^2 + \cdots + A_t x^t$ to have a right linear factor $1 - Dx$.

18.6 Mimic to Theorem 18.1 and using same notations, give a necessary and sufficient condition for square polynomial matrix $A(x)$ to have $B(x) = I - Dx$ as its right factor.

18.7 Assume $x = (x_1, \cdots, x_n)^T \in \mathbb{R}^n$, calculate (a) Dx^2; (b) $D(Fx + Gx^2 + Hx^3)$; (c) $\nabla(Fx + Gx^2 + Hx^3)$.

18.8 Let $x = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ and

$$A(x) = \begin{bmatrix} x_1^2 + x_2 & x_2 x_3 - x_3^2 \\ 1 - x_2^2 & x_1 x_3 + x_2^2 \end{bmatrix}.$$

Calculate (i) $D^k(A(x))$, $k = 1, 2, \cdots$. (ii) $\nabla^k(A(x))$, $k = 1, 2, \cdots$.

18.9 Let $x \in \mathbb{R}^n$ and $A(x) \in \mathcal{M}_{p\times q}$. Find formulas to convert $DA(x)$ to $\nabla A(x)$ and vice versa.

18.10 Calculate $T_N(3, 3)$ and $T_B(3, 3)$.

18.11 Given

$$A(x) = \begin{bmatrix} x_1^2 & x_1 \\ x_2 & x_2^2 \end{bmatrix}, \quad B(x) = \begin{bmatrix} e^{x_1} & e^{x_2} \\ e^{x_2} & e^{x_1} \end{bmatrix}.$$

(i) Calculate $(DA)B + A(DB)$.

(ii) Calculate $D(AB)$, compare it with previous result.

18.12 Let $A(x) \in \mathcal{M}_{m\times n}$, $B(x) \in \mathcal{M}_{n\times p}$, $C(x) \in \mathcal{M}_{p\times q}$. Deduce a formula for $D(ABC)$.

18.13 Let $A_i \in \mathcal{M}_{n\times n}$, $i = 1, \cdots, k$. Deduce a formula for

$$D\left(\prod_{i=1}^{k} A_i\right).$$

18.14 A double index $I = (i_1, i_2)$ and an integer $n = 3$ are given.

(i) Calculate the index set $\mathrm{id}(i_1, i_2; n^2)$.

(ii) Construct its subset $\mathrm{is}(i_1, i_2; n^2)$.

(iii) Construct the set $R(n^2) = \{(c_1, c_2, c_3) \mid c_1 + c_2 + c_3 = 2\}$.

(iv) For each $e \in \mathrm{is}(i_1, i_2; n^2)$ figure out $\xi(e) \in R(n^2)$. Check that this is a bijective (i.e., one to one and onto) mapping.

18.15 A mapping $F : \mathbb{R}^3 \to \mathbb{R}^3$ is defined by

$$\begin{cases} f_1(x_1, x_2, x_3) = e^{x_1+x_2+x_3} \\ f_2(x_1, x_2, x_3) = x_1 + \ln(1 + x_2^2) - x_3 \\ f_3(x_1, x_2, x_3) = \sin(x_1 - x_2 + x_3). \end{cases}$$

Find the Taylor expansion of F till quadratic term as

$$F(x_1, x_2, x_3) = F_0 + F_1 x + F_2 x^2 + R(\|x\|^3).$$

18.16 Consider the formula (18.72)–(18.73). Show that when $n = 1$ it degenerated to (one-variable derivative formula)

$$D(x^{n+1}) = (x^{n+1})' = (n + 1)x^n.$$

18.17 Assume a mapping $\pi : x \to y$ is defined as

$$\begin{cases} y_1 = x_1 + x_2^2, \\ y_2 = \sin(x_1 + x_2). \end{cases}$$

Give the Taylor series expansion of the inverse mapping π^{-1} up to cubic terms. Precisely, find A_1, A_2, A_3, such that

$$x = A_1 y + A_2 y^2 + A_3 y^3 + O(\|y\|^4).$$

18.18 $f(x) = (f_1(x), \cdots, f_n(x))^T \in V(\mathbb{R}^n)$ is a smooth vector field on $\mathbb{R}^n$. f is called a k-homogeneous vector field if its components $f_i(x) \in B_n^k$, $i = 1, \cdots, n$. The set of k-homogeneous vector fields is denoted by $V_n^k(\mathbb{R}^n)$ (briefly, V_n^k).

(i) Show that $f \in V_n^k$, if and only if there exists a matrix $F \in \mathcal{M}_{n \times n^k}$ such that $f(x) = Fx^k$.

(ii) Let $f = Ax \in V_n^1$, where $A \in \mathcal{M}_n$. Define $\mathrm{ad}_f : g \mapsto \mathrm{ad}_f\, g$. Show that

$$\mathrm{ad}_f|_{V_n^k}$$

is a linear mapping from V_n^k to V_n^k.

(iii) Let $g = Gx^k \in V_n^k$ and $f = Ax$. Then

$$\mathrm{ad}_f(g) := Hx^k \in V_n^k.$$

Deduce the formula of $H = H(G, A)$.

18.19 Let $f = Fx^p \in V_n^p$ and $g = Gx^q \in V_n^q$.

(i) Show that $[f, g] \in V_n^{p+q-1}$.

(ii) Find $H \in \mathcal{M}_{n \times n^{p+q-1}}$ such that

$$[f, g] = Hx^{p+q-1}.$$

18.20 Assume $h(x) \in C^\omega(\mathbb{R}^n)$, $X, Y \in V^\omega(\mathbb{R}^n)$, $\alpha \in V^{*\omega}(\mathbb{R}^n)$. Using their Taylor series expansions to calculate the second order Lie derivatives of $L_X^2 h(x)$, $\mathrm{ad}_X^2\, Y$ and $L_X^2 \alpha$.

18.21 Assume $h(x) \in C^\omega(\mathbb{R}^n)$, $X, Y \in V^\omega(\mathbb{R}^n)$, $\alpha \in V^{*\omega}(\mathbb{R}^n)$.

(i) Deduce the formula for $D(\mathrm{ad}_X\, Y)$.

(ii) Deduce the formula for $D(L_X h)$.

(iii) Deduce the formula for $D(L_X \alpha)$.

This page intentionally left blank

Chapter 19

Some Applications to Differential Geometry and Algebra

This chapter presents some applications of semi-tensor product to differential geometry and algebra. In many mathematical problems we need to deal with multiple-dimensional data, or data labeled by multi-index. In this case STP could be a proper tool for formula deduction or numerical calculation. The first part of this chapter considers some geometric problems. First, we consider the calculation of connection, curvature, and Riemann curvature tensor etc. The converting formulas are provided to deal with coordinate transformations. Some further properties are investigated. Then, we prove a formula for tensor contraction, which comes from books of relativity. Second part considers the applications to algebra. The general description of algebras over $\mathbb{R}$ is presented by their structure matrices. Then the classification and some properties of algebras with dimension 2 and dimension 3 are investigated. Finally, the product algebras are discussed.

19.1 Calculation of Connection

Connection is a fundamental concept in geometry as well as in physics. In this section we will consider some connection-related calculations using STP. First, we give the definition (Boothby, 1986).

Definition 19.1. Let $f, g \in V(M)$ be two smooth vector fields on a smooth manifold M. An $\mathbb{R}$-bilinear mapping $\nabla\colon\ V(M) \times V(M) \to V(M)$ is called a connection, if it satisfies

(1)

$$\nabla_{rf} sg = rs\nabla_f g, \quad r, s \in \mathbb{R}; \tag{19.1}$$

(2)

$$\nabla_{hf}g = h\nabla_f g, \quad \nabla_f(hg) = L_f(h)g + h\nabla_f g, \quad h \in C^\infty(M). \tag{19.2}$$

Note that in this chapter "smooth" means infinitely differentiable, denoted by C^∞.

According to the definition, it is easy to verify that if a connection is defined over a basis of $V(M)$, it is overall well defined. Hence, we need to define the action of connection over basis vectors. Under a local coordinate chart x, the action of a connection is determined by the following forms.

$$\nabla_{\partial/\partial x_i}\left(\frac{\partial}{\partial x_j}\right) = \sum_{k=1}^n \gamma_{ij}^k \frac{\partial}{\partial x_k},$$

where γ_{ij}^k is called the Christoffel symbol. Using Christoffel symbol, we can construct a matrix

$$\Gamma = \begin{bmatrix} \gamma_{11}^1 & \cdots & \gamma_{1n}^1 & \cdots & \gamma_{n1}^1 & \cdots & \gamma_{nn}^1 \\ \vdots & & \vdots & & \vdots & & \vdots \\ \gamma_{11}^n & \cdots & \gamma_{1n}^n & \cdots & \gamma_{n1}^n & \cdots & \gamma_{nn}^n \end{bmatrix},$$

which is called the Christoffel matrix.

Next, we give a matrix form of the connection.

Proposition 19.1. *Let* $f = \sum\limits_{i=1}^n f_i \frac{\partial}{\partial x_i}$, $g = \sum\limits_{j=1}^n g_j \frac{\partial}{\partial x_j}$, *and denote their vector form as* $f = (f_1, f_2, \cdots, f_n)^T$, $g = (g_1, g_2, \cdots, g_n)^T$. *The connection under vector form is expressed as*

$$\nabla_f g = Dgf + \Gamma fg. \tag{19.3}$$

Proof. According to the definition (19.1)–(19.2), we can calculate that

$$\begin{aligned} \nabla_f g &= \sum_{i=1}^n f_i \left[\sum_{j=1}^n L_{\partial/\partial x_i} g_j \frac{\partial}{\partial x_j} + \sum_{j=1}^n \sum_{k=1}^n g_j \gamma_{ij}^k \frac{\partial}{\partial x_k}\right] \\ &= Dg \ltimes f + \Gamma \ltimes f \ltimes g. \end{aligned} \tag{19.4}$$

Then (19.3) follows immediately. □

Now assume $y = y(x)$ is another local coordinate frame, we intend to present the Chrostoffel matrix Γ under this new coordinate frame. Denote by $\tilde{\Gamma}$ and $\tilde{\gamma}_{ij}^k$ for new Γ and its entries γ_{ij}^k respectively, then we have

Lemma 19.1. *Under new coordinate frame y the Chrostoffel matrix can be calculated as*

$$\begin{bmatrix} \tilde{\gamma}_{ij}^1 \\ \vdots \\ \tilde{\gamma}_{ij}^n \end{bmatrix} = \begin{bmatrix} \dfrac{\partial^2 x_1}{\partial y_j \partial y_1} & \cdots & \dfrac{\partial^2 x_1}{\partial y_j \partial y_n} \\ \vdots & & \\ \dfrac{\partial^2 x_n}{\partial y_j \partial y_1} & \cdots & \dfrac{\partial^2 x_n}{\partial y_j \partial y_n} \end{bmatrix} \begin{bmatrix} \dfrac{\partial x_1}{\partial y_i} \\ \vdots \\ \dfrac{\partial x_n}{\partial y_i} \end{bmatrix} + \Gamma \ltimes \begin{bmatrix} \dfrac{\partial x_1}{\partial y_i} \\ \vdots \\ \dfrac{\partial x_n}{\partial y_i} \end{bmatrix} \ltimes \begin{bmatrix} \dfrac{\partial x_1}{\partial y_j} \\ \vdots \\ \dfrac{\partial x_n}{\partial y_j} \end{bmatrix}. \quad (19.5)$$

Proof. Set

$$f = \frac{\partial}{\partial y_i} = \sum_{s=1}^n \frac{\partial}{\partial x_s} \frac{\partial x_s}{\partial y_i},$$

and

$$g = \frac{\partial}{\partial y_j} = \sum_{t=1}^n \frac{\partial}{\partial x_t} \frac{\partial x_t}{\partial y_j}.$$

Recall the definition of γ_{ij}^k, we have

$$\sum_{k=1}^n \tilde{\gamma}_{ij}^k \frac{\partial}{\partial y_k} = \nabla_f g.$$

Applying (19.3) to the above equation, equation (19.5) follows immediately. □

Theorem 19.1. *Under new coordinate frame y we have*

$$\tilde{\Gamma} = D^2 x D x + \Gamma \ltimes Dx (I \otimes Dx). \quad (19.6)$$

Proof. A direct calculation shows that

$$D^2 x \ltimes Dx = \left[\begin{matrix} \displaystyle\sum_{s=1}^n \frac{\partial^2 x_1}{\partial y_s \partial y_1} \frac{\partial x_s}{\partial y_1} & \cdots & \displaystyle\sum_{s=1}^n \frac{\partial^2 x_1}{\partial y_s \partial y_n} \frac{\partial x_s}{\partial y_1} \\ \vdots & \ddots & \vdots \\ \displaystyle\sum_{s=1}^n \frac{\partial^2 x_n}{\partial y_s \partial y_1} \frac{\partial x_s}{\partial y_1} & \cdots & \displaystyle\sum_{s=1}^n \frac{\partial^2 x_n}{\partial y_s \partial y_n} \frac{\partial x_s}{\partial y_1} \end{matrix} \right.$$

$$\left. \begin{matrix} \cdots & \displaystyle\sum_{s=1}^n \frac{\partial^2 x_1}{\partial y_s \partial y_1} \frac{\partial x_s}{\partial y_n} & \cdots & \displaystyle\sum_{s=1}^n \frac{\partial^2 x_1}{\partial y_s \partial y_n} \frac{\partial x_s}{\partial y_n} \\ \ddots & \vdots & \ddots & \vdots \\ \cdots & \displaystyle\sum_{s=1}^n \frac{\partial^2 x_n}{\partial y_s \partial y_1} \frac{\partial x_s}{\partial y_n} & \cdots & \displaystyle\sum_{s=1}^n \frac{\partial^2 x_n}{\partial y_s \partial y_n} \frac{\partial x_s}{\partial y_n} \end{matrix} \right].$$

If we label $\mathrm{Col}(D^2x \ltimes Dx)$ by (ij) in the order of $\mathrm{id}(i,j;n,n)$, then it is easy to verify that its (i,j)th column is the first term of the right hand side of (19.5).

Denote by $J_i = \mathrm{Col}_i(Dx)$ the ith column of Dx, then

$$\Gamma \ltimes Dx = (\Gamma \ltimes J_1, \Gamma \ltimes J_2, \cdots, \Gamma \ltimes J_n).$$

We also have $I \otimes Dx = \mathrm{diag}(J, \cdots, J)$, where $J = (J_1, \cdots, J_n)$. Hence

$$\begin{aligned}&\Gamma \ltimes Dx \ltimes (I \otimes Dx)\\ &\quad = (\Gamma \ltimes J_1 \ltimes J_1, \cdots, \Gamma \ltimes J_1 \ltimes J_n, \cdots, \Gamma \ltimes J_n \ltimes J_1, \cdots, \Gamma \ltimes J_n \ltimes J_n)\,.\end{aligned}$$

It is clear that the (ij)th column of the above, $\mathrm{Col}_{i,j}(\Gamma \ltimes Dx \ltimes (I \otimes Dx))$ is the second term of the right hand side of (19.5). □

Remark 19.1. Using right semi-tensor product, (19.6) can also be expressed as

$$\tilde{\Gamma} = D^2x \ltimes Dx + (\Gamma \ltimes Dx) \rtimes Dx. \tag{19.7}$$

It is worth noting that since there is no associativity between left and right semi-tensor product, the parentheses in (19.7) cannot be omitted. To avoid possible confusion, we use right semi-tensor product rarely.

Let M be a Riemannian manifold, its Reimannian metric is determined by a positive definite two-form, which has its structure matrix $G = (g_{ij})_{n\times n}$. The fundamental theorem of Riemannian geometry says that, on M there exists an unique connection. Moreover, its Christoffel symbol is determined by the following formula (Abraham and Marsden, 1978).

$$\gamma_{ij}^k = \frac{1}{2}\sum_{s=1}^n g^{ks}\left(\frac{\partial g_{si}}{\partial x_j} - \frac{\partial g_{ij}}{\partial x_s} + \frac{\partial g_{js}}{\partial x_i}\right), \tag{19.8}$$

where g^{ij} is the (i,j)th element of G^{-1}.

Using this uniquely determined connection, on a Riemannian manifold we have (Boothby, 1986)

$$[f,g] = \nabla_f g - \nabla_g f. \tag{19.9}$$

A Christoffel matrix is said to be symmetric, if

$$\gamma_{ij}^k = \gamma_{ji}^k, \quad \forall\, i,j,k. \tag{19.10}$$

Then we have the following result.

Theorem 19.2. *If a manifold N has a connection with symmetric Christoffel matrix, then (19.9) holds.*

Proof. It is easy to see that if Christoffel matrix Γ is symmetric, then $\Gamma W_{[n]} = \Gamma$. That is

$$\Gamma fg = \Gamma gf, \quad \forall\, f, g \in V(N).$$

Using (19.3), we have

$$\nabla_f g - \nabla_g f = Dgf - Dfg = [f, g].$$

□

According to (19.8), it is clear that for a Riemannian manifold the Christoffel matrix is symmetric. Hence (19.9) holds. Therefore, Theorem 19.2 is a more general result.

In both formula deduction and the numerical calculation via computer matrix form is often more convenient. A straightforward computation shows that (19.8) has its matrix form as

$$\Gamma = \frac{1}{2} G^{-1} \left(DG + DGW_{[n]} - (DV_r(G))^T \right). \tag{19.11}$$

A related topic is the geodesic. Let $r(t)$ be a curve on a Riemannian manifold M. $r(t)$ is a geodesic, if and only if (Abraham and Marsden, 1978)

$$\ddot{r}^i = \sum_{j=1}^{n} \sum_{k=1}^{n} \Gamma^i_{j,k} \dot{r}^j \dot{r}^k. \tag{19.12}$$

Now under a local coordinate frame $r(t)$ is expressed in its component form as $r(t) = (x_1(t), \cdots, x_n(t))^T$. Then (19.12) has its matrix form as

$$\begin{bmatrix} \ddot{x}_1 \\ \vdots \\ \ddot{x}_n \end{bmatrix} = \Gamma \begin{bmatrix} \dot{x}_1 \\ \vdots \\ \dot{x}_n \end{bmatrix}^2 . \tag{19.13}$$

Finally, we consider the curvature operator and the Riemannian curvature tensor.

Definition 19.2 (Boothby, 1986). *(1) The curvature operator $R(X, Y)$ is an operator determined by two C^∞ vector fields X and Y, for each C^∞ vector field Z it assigned a C^∞ vector field $R(X, Y) \cdot Z$ as*

$$R(X, Y) \cdot Z = \nabla_X(\nabla_Y Z) - \nabla_Y(\nabla_X Z) - \nabla_{[X,Y]} Z. \tag{19.14}$$

(2) The Riemannian curvature tensor is a 4th order C^∞ covariant tensor field defined by

$$\mathcal{R}(X, Y, Z, W) = (R(X, Y) \cdot Z, W). \tag{19.15}$$

We calculate the structure matrix of the curvature operator.

Assume under a local coordinate frame $x = (x_1, \cdots, x_n)$ we have $E_i := \frac{\partial}{\partial x_i}$. Denote the three vector fields in vector form as $X = (\alpha_1, \cdots, \alpha_n)$, $Y = (\beta_1, \cdots, \beta_n)$, and $Z = (\gamma_1, \cdots, \gamma_n)$. Then (Boothby, 1986)

$$R(X,Y)\cdot Z = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \alpha_i \beta_j \gamma_k R(E_i, E_j)\cdot E_k. \tag{19.16}$$

Since

$$\begin{aligned}\nabla_{E_i}(\nabla_{E_j} E_k) &= \nabla_{E_i}\left(\sum_{t=1}^{n} \gamma_{jk}^t E_t\right)\\ &= \sum_{t=1}^{n} L_{E_i}(\gamma_{jk}^t)E_t + \sum_{t=1}^{n}\gamma_{jk}^t \sum_{\ell=1}^{n}\gamma_{it}^{\ell} E_{\ell} \qquad (19.17)\\ &= \sum_{t=1}^{n}\left(L_{E_i}(\gamma_{jk}^t) + \sum_{\ell=1}^{n}\gamma_{i\ell}^t\gamma_{jk}^{\ell}\right)E_t.\end{aligned}$$

Similarly,

$$\nabla_{E_j}(\nabla_{E_i} E_k) = \sum_{t=1}^{n}\left(L_{E_j}(\gamma_{ik}^t) + \sum_{\ell=1}^{n}\gamma_{j\ell}^t\gamma_{ik}^{\ell}\right)E_t. \tag{19.18}$$

Using (19.17) and (19.18) to (19.14) and (19.16), we can construct the structure matrix as follows: Define

$$m_{ijk}^t := \left(L_{E_i}(\gamma_{jk}^t) + \sum_{\ell=1}^{n}\gamma_{i\ell}^t\gamma_{jk}^{\ell}\right) - \left(L_{E_j}(\gamma_{ik}^t) + \sum_{\ell=1}^{n}\gamma_{j\ell}^t\gamma_{ik}^{\ell}\right),$$

then the structure matrix of R is

$$M_R = \left(m_{ijk}^t\right), \tag{19.19}$$

where the elements are arranged by $\mathrm{id}(t; n) \times \mathrm{id}(i, j, k; n, n, n)$. It follows that

$$R(X,Y)\cdot Z = M_R XYZ. \tag{19.20}$$

Note that the above deduction is formal. To prove (19.20) we need to show that R is multilinear with respect to X, Y, and Z. That is, for any smooth functions f, g, and h, we have

$$R(fX, gY, hZ) = fghR(X,Y,Z).$$

In fact, this is correct (Boothby, 1986).

Next, we consider the Riemannian curvature tensor. Note that a bilinear form $p^T Gq$ can be equivalently expressed as

$$p^T Gq = \sum_{i,j} g_{ij}p_i q_j = V_r^T(G)pq = V_c^T(G)qp.$$

For the Riemannian curvature tensor observe that

$$\mathcal{R}(X,Y,Z,W) = W^T GR(X,Y)\cdot Z.$$

Using the above bilinear form, we have

$$\mathcal{R}(X,Y,Z,W) = V_c^T(G)M_{\mathcal{R}}XYZW. \tag{19.21}$$

19.2 Contraction of Tensor Field

This section considers the contraction of tensor field. The contraction of tensor field plays a key role in physics, particularly in relativity (Foster and Nightngale, 1995; Sachs and Wu, 1977).

In physics, multi-index has often been used. It is said in Foster and Nightngale (1995) that "It is also sometimes convenient to use matrix methods to handle the summations over repeated suffixes. These methods are restricted to quantities carrying either one or two suffixes. enabling them to be arranged as either one-dimensional arrays (row vectors or column vectors) or two-dimensional arrays (matrices)." From this statement one sees that conventional matrix product can be used only for one- or two-dimensional data. Using STP, the tensor field, which is used to deal with multiple-dimensional data, can be treated in matrix form via matrix calculations. Contraction of tensor fields is a perfect example for this.

First, we review the structure matrix of a tensor. Let $\sigma \in \mathcal{T}^r_s(V)$ be an (r,s)-type tensor on an n-dimensional vector space V, and $1 \le p \le r$, $1 \le q \le s$. Fix the basis of V as $\{d_1, \cdots, d_n\}$ and its dual basis $\{e^1, \cdots, e^n\}$ of V^*. For $\sigma \in \mathcal{T}^r_s(V)$ define

$$\omega^{i_1\cdots i_r}_{j_1\cdots j_s} = \sigma(d_{i_1}, \cdots, d_{i_r}; e^{j_1}, \cdots, e^{j_s}), \quad i_1, \cdots, i_r, j_1, \cdots, j_s = 1, \cdots, n.$$

Then this set of n^{r+s} data can be arranged into a matrix as

$$M_\sigma = \begin{bmatrix} \omega^{1\cdots 11}_{1\cdots 11} & \omega^{1\cdots 12}_{1\cdots 11} & \cdots & \omega^{n\cdots nn}_{1\cdots 11} \\ \omega^{1\cdots 11}_{1\cdots 12} & \omega^{1\cdots 12}_{1\cdots 12} & \cdots & \omega^{n\cdots nn}_{1\cdots 12} \\ \vdots & \vdots & & \vdots \\ \omega^{1\cdots 11}_{n\cdots nn} & \omega^{1\cdots 12}_{n\cdots nn} & \cdots & \omega^{n\cdots nn}_{n\cdots nn} \end{bmatrix}, \tag{19.22}$$

which is called the structure matrix of σ.

Using structure matrix, we have

$$\begin{aligned} &\sigma(\sigma_1, \cdots, \sigma_s; X_1, \cdots, X_r) \\ &\quad = \mu_s \ltimes \cdots \ltimes \mu_1 \ltimes M_\sigma \ltimes X_1 \ltimes \cdots \ltimes X_r, \end{aligned} \tag{19.23}$$

where $\mu_i \in V^*$, $i = 1, \cdots, s$, and $X_i \in V$, $i = 1, \cdots, r$.

In the following we define the contraction $\pi^p_q : \mathcal{T}^r_s(V) \to \mathcal{T}^{r-1}_{s-1}(V)$ by constructing its structure matrix. Assume the bases of V and V^* are fixed and σ has its structure matrix as in (19.22). Then the elements of the structure matrix of $\pi^p_q(\sigma)$ are defined as

$$\omega^{i_1\cdots\hat{i_p}\cdots i_r}_{j_1\cdots\hat{j_q}\cdots j_s} = \sum_{i_p=j_q} \omega^{i_1\cdots i_p\cdots i_r}_{j_1\cdots j_q\cdots j_s}\ , \tag{19.24}$$

where “$\hat{\cdot}$” means the corresponding index is omitted.

Since the definition depends on the choice of basis, we need to prove the definition is well posed, that is, it is independent on the choice of basis. We will first deduce a formula for the structure matrix of the contracted tensor, and then prove that this formula defines a tensor, which is independent of the choice of the coordinates.

To deduce the structure matrix $\pi_q^p(\sigma)$ set $\xi = n^{s-1}$, $\eta = n^{r-1}$, and split the structure matrix M_σ into $\xi \times \eta$ block form as

$$M_\sigma = \begin{bmatrix} M_{11} & \cdots & M_{1\eta} \\ \vdots & & \vdots \\ M_{\xi 1} & \cdots & M_{\xi\eta} \end{bmatrix}, \tag{19.25}$$

where each block M_{ij} is an $n \times n$ matrix.

Assume p and q correspond to last vector and last co-vector arguments, then a straightforward computation shows the following result.

Lemma 19.2. *Assume $p = r$ and $q = s$, then*

$$M_{\pi_s^r(\sigma)} = \begin{bmatrix} \mathrm{tr}(M_{11}) & \cdots & \mathrm{tr}(M_{1\eta}) \\ \vdots & & \vdots \\ \mathrm{tr}(M_{\xi 1}) & \cdots & \mathrm{tr}(M_{\xi\eta}) \end{bmatrix} := \mathcal{T}_r(M_\sigma), \tag{19.26}$$

where the operator $\mathcal{T}_r$ is used to calculate the traces of each blocks.

For general case we need to interchange the index p with r, and q with s. Note that the swap of two elements can be realized by a sequence of swaps of two adjacent elements. Based on this consideration, we can use Proposition 2.7 to realize the required reordering. We leave the detailed proof to the reader and present the corresponding structure matrix of σ as

$$\begin{aligned} \tilde{M}_\sigma &= \textstyle\prod_{t=0}^{s-q-1}(I_{n^{s-2-t}} \otimes W_{[n]} \otimes I_{n^t}) M_\sigma \prod_{t=0}^{r-p-1}(I_{n^{r-2-t}} \otimes W_{[n]} \otimes I_{n^t}) \\ &:= \Pi_1 M_\sigma \Pi_2. \end{aligned} \tag{19.27}$$

Similar to the case of $p = r$ and $q = s$, we replace M_σ by $\tilde{M}_\sigma$, split the later into $\xi \times \eta$ block form, and denote each $n \times n$ blocks by $\tilde{M}_{ij}$. Then we have

Proposition 19.2. *The structure matrix of $\pi_q^p(\sigma)$ is*

$$M_{\pi_q^p(\sigma)} = \mathcal{T}_r(\tilde{M}_\sigma) = \mathcal{T}_r(\Pi_1 M_\sigma \Pi_2). \tag{19.28}$$

We give a numerical example to depict the contraction.

Example 19.1. Let $n = 2$, $r = 2$, and $s = 3$. We consider $\pi_1^1(\sigma)$. Denote

$$M_\sigma = \begin{bmatrix} a_{111}^{11} & a_{111}^{12} & a_{111}^{21} & a_{111}^{22} \\ a_{112}^{11} & a_{112}^{12} & a_{112}^{21} & a_{112}^{22} \\ a_{121}^{11} & a_{121}^{12} & a_{121}^{21} & a_{121}^{22} \\ a_{122}^{11} & a_{122}^{12} & a_{122}^{21} & a_{122}^{22} \\ a_{211}^{11} & a_{211}^{12} & a_{211}^{21} & a_{211}^{22} \\ a_{212}^{11} & a_{212}^{12} & a_{212}^{21} & a_{212}^{22} \\ a_{221}^{11} & a_{221}^{12} & a_{221}^{21} & a_{221}^{22} \\ a_{222}^{11} & a_{222}^{12} & a_{222}^{21} & a_{222}^{22} \end{bmatrix}.$$

$$\begin{aligned} \Pi_1 &= \textstyle\prod_{t=0}^{1} I_{2^{1-t}} \otimes W_{[2]} \otimes I_{2^t} = (I_2 \otimes W_{[2]})(W_{[2]} \otimes I_2) \\ &= \begin{bmatrix} 1&0&0&0&0&0&0&0 \\ 0&0&0&0&1&0&0&0 \\ 0&1&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0 \\ 0&0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&1&0 \\ 0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&1 \end{bmatrix}. \end{aligned}$$

$$\Pi_2 = W_{[2]} = \begin{bmatrix} 1&0&0&0 \\ 0&0&1&0 \\ 0&1&0&0 \\ 0&0&0&1 \end{bmatrix}.$$

Then

$$\Pi_1 M_\sigma \Pi_2 = \begin{bmatrix} a_{111}^{11} & a_{111}^{21} & a_{111}^{12} & a_{111}^{22} \\ a_{211}^{11} & a_{211}^{21} & a_{211}^{12} & a_{211}^{22} \\ a_{112}^{11} & a_{112}^{21} & a_{112}^{12} & a_{112}^{22} \\ a_{212}^{11} & a_{212}^{21} & a_{212}^{12} & a_{212}^{22} \\ a_{121}^{11} & a_{121}^{21} & a_{121}^{12} & a_{121}^{22} \\ a_{221}^{11} & a_{221}^{21} & a_{221}^{12} & a_{221}^{22} \\ a_{122}^{11} & a_{122}^{21} & a_{122}^{12} & a_{122}^{22} \\ a_{222}^{11} & a_{222}^{21} & a_{222}^{12} & a_{222}^{22} \end{bmatrix}.$$

Using the above form and the Proposition 19.2, we can obtain that

$$M_{\pi_1^1(\sigma)} = \begin{bmatrix} a_{111}^{11} + a_{211}^{21} & a_{111}^{12} + a_{211}^{22} \\ a_{112}^{11} + a_{212}^{21} & a_{112}^{12} + a_{212}^{22} \\ a_{121}^{11} + a_{221}^{21} & a_{121}^{12} + a_{221}^{22} \\ a_{122}^{11} + a_{222}^{21} & a_{122}^{12} + a_{222}^{22} \end{bmatrix}.$$

Finally we prove that the contracted tensor defined by (19.27) is independent of the choice of the coordinates. Assume we have a coordinate transformation $z = z(x)$, and its Jacobi matrix is $J = \frac{\partial z}{\partial x}$.

The following lemma can be verified by straightforward computations.

Lemma 19.3.

(1) Assume $P \in \mathcal{M}_{s\times m}$, $Q \in \mathcal{M}_{n\times n}$, and $A_i \in \mathcal{M}_{n\times rn}$, $i = 1, \cdots, m$. Set

$$\tilde{A} = \begin{bmatrix} \tilde{A}_1 \\ \vdots \\ \tilde{A}_m \end{bmatrix} = (P \otimes Q) \begin{bmatrix} A_1 \\ \vdots \\ A_m \end{bmatrix} := (P \otimes Q)A,$$

then

$$\begin{bmatrix} \tilde{A}_1 \\ \vdots \\ \tilde{A}_m \end{bmatrix} = P \ltimes \begin{bmatrix} QA_1 \\ \vdots \\ QA_m \end{bmatrix}.$$

Moreover, we also have

$$\mathcal{T}_r(\tilde{A}) = P \cdot \mathcal{T}_r \left(\begin{bmatrix} QA_1 \\ \vdots \\ QA_m \end{bmatrix} \right).$$

(2) Let $P \in M_{m\times s}$, $Q \in M_{n\times n}$, and $A_i \in M_{nr\times n}$, $i = 1, \cdots, m$. Assume

$$\tilde{A} = \left[\tilde{A}_1, \cdots, \tilde{A}_m\right] = [A_1, \cdots, A_m]\,(P \otimes Q) = A(P \otimes Q),$$

then we have

$$\left[\tilde{A}_1, \cdots, \tilde{A}_m\right] = [A_1Q, \cdots, A_mQ]\,P.$$

Moreover, we also have

$$\mathcal{T}_r(\tilde{A}) = \mathcal{T}_r\left(A_1Q, \cdots, A_mQ\right)P.$$

Now we are ready to prove the key issue: the contraction defined above is properly defined. That is, the definition is independent of the choice of the coordinates.

Theorem 19.3. *The contracted tensor defined by (19.24) is independent of the choice of coordinates.*

Proof. Let

$$M_\sigma = \begin{bmatrix} M_{11} & \cdots & M_{1\eta} \\ \vdots & & \vdots \\ M_{\xi 1} & \cdots & M_{\xi\eta} \end{bmatrix}.$$

Using Proposition 19.2, we have

$$M_{\pi_q^p(\sigma)} = \mathcal{T}_r \left(\Pi_1 M_\sigma \Pi_2 \right).$$

Now consider a coordinate change $z = z(x)$. Under this new coordinate frame M_σ becomes

$$\bar{M}_\sigma = \underbrace{J^{-1} \otimes \cdots \otimes J^{-1}}_{s} M_\sigma \underbrace{J \otimes \cdots \otimes J}_{t}.$$

Note that Π_1 is commutative with $\underbrace{J^{-1} \otimes \cdots \otimes J^{-1}}_{s}$ and Π_2 is commutative with $\underbrace{J \otimes \cdots \otimes J}_{t}$. Applying Proposition 19.2 to $\bar{M}_\sigma$ yields

$$\begin{aligned}
\bar{M}_{\pi_q^p(\sigma)} &= \mathcal{T}_r \left(\Pi_1 \underbrace{(J^{-1} \otimes \cdots \otimes J^{-1})}_{s} M_\sigma \underbrace{(J \otimes \cdots \otimes J)}_{t} \Pi_2 \right) \\
&= \mathcal{T}_r \left(\underbrace{(J^{-1} \otimes \cdots \otimes J^{-1})}_{s} (\Pi_1 M_\sigma \Pi_2) \underbrace{(J \otimes \cdots \otimes J)}_{t} \right) \\
&= \mathcal{T}_r \left((\underbrace{(J^{-1} \otimes \cdots \otimes J^{-1})}_{s-1} \otimes J^{-1})(\tilde{M}_\sigma)(\underbrace{(J \otimes \cdots \otimes J)}_{t-1} \otimes J) \right) \\
&= (\underbrace{J^{-1} \otimes \cdots \otimes J^{-1}}_{s-1}) \mathcal{T}_r (J^{-1}(\tilde{M}_\sigma) J)(\underbrace{J \otimes \cdots \otimes J}_{t-1}) \\
&= (\underbrace{J^{-1} \otimes \cdots \otimes J^{-1}}_{s-1}) \mathcal{T}_r (\tilde{M}_\sigma)(\underbrace{J \otimes \cdots \otimes J}_{t-1}) \\
&= (\underbrace{J^{-1} \otimes \cdots \otimes J^{-1}}_{s-1}) M_{\pi_q^p(\sigma)} (\underbrace{J \otimes \cdots \otimes J}_{t-1}).
\end{aligned}$$

Note that the last three equalities are obtained by using Lemma 19.3. □

19.3 Structure Matrix of Finite-Dimensional Algebra

Definition 19.3 (Hungerford, 1974). (1) An n-dimensional algebra over $\mathbb{R}$ is an n-dimensional vector space $\mathcal{L}$ over $\mathbb{R}$ with a multiplication

$* : \mathcal{L} \times \mathcal{L} \to \mathcal{L}$, satisfying distributive rule. That is,

$$\begin{array}{l}(aX+bY)*Z = a(X*Z)+b(Y*Z),\\ Z*(aX+bY) = a(Z*X)+b(Z*Y), \quad X,Y,Z \in \mathcal{L},\ a,b \in \mathbb{R}.\end{array} \tag{19.29}$$

(2) An algebra is called a Lie algebra, if it satisfies

(i) skew-symmetry,

$$X*Y = -Y*X; \tag{19.30}$$

(ii) Jacobi Identity,

$$(X*Y)*Z + (Y*Z)*X + (Z*X)*Y = 0, \quad \forall X,Y,Z \in \mathcal{L}. \tag{19.31}$$

Definition 19.4. Let $\{e_1, \cdots, e_n\}$ be a basis of an algebra $\mathcal{L}$.

(1) Assume

$$e_i * e_j = \sum_{k=1}^n \alpha_{ij}^k e_k, \quad i,j = 1, \cdots, n. \tag{19.32}$$

Then α_{ij}^k are called the structure constants of the algebra.

(2) The structure matrix of $\mathcal{L}$ (with product $*$) is defined as

$$M_{\mathcal{L}} = \begin{bmatrix} \alpha_{11}^1 & \cdots & \alpha_{1n}^1 & \cdots & \alpha_{n1}^1 & \cdots & \alpha_{nn}^1 \\ \alpha_{11}^2 & \cdots & \alpha_{1n}^2 & \cdots & \alpha_{n1}^2 & \cdots & \alpha_{nn}^2 \\ \vdots & & \vdots & & \vdots & & \vdots \\ \alpha_{11}^n & \cdots & \alpha_{1n}^n & \cdots & \alpha_{n1}^n & \cdots & \alpha_{nn}^n \end{bmatrix}. \tag{19.33}$$

Fix this basis and let vectors $X = \sum\limits_{i=1}^n x_i e_i$, $Y = \sum\limits_{i=1}^n y_i e_i$ etc. be expressed in vector form as $X = (x_1, \cdots, x_n)^T$, $Y = (y_1, \cdots, y_n)^T$ etc. Then the following proposition is easily verifiable.

Proposition 19.3. *Let $Z = X * Y$. Then the vector form of Z can be calculated as*

$$Z = M_{\mathcal{L}} \ltimes X \ltimes Y = M_{\mathcal{L}} XY. \tag{19.34}$$

It is clear that all the properties of an algebra are determined by it structure matrix. In the following we investigate some basic properties of an algebra via its structure matrix. For notational ease, the symbol $\ltimes$ is omitted.

Definition 19.5. Given an algebra $\mathcal{L}$.

(1) $\mathcal{L}$ is said to be symmetric, if

$$X * Y = Y * X, \quad \forall X, Y \in \mathcal{L}; \tag{19.35}$$

(2) $\mathcal{L}$ is said to be skew-symmetric, if

$$X * Y = -Y * X, \quad \forall X, Y \in \mathcal{L}; \tag{19.36}$$

(3) $\mathcal{L}$ is said to be associative, if

$$(X * Y) * Z = X * (Y * Z), \quad \forall X, Y, Z \in \mathcal{L}. \tag{19.37}$$

Proposition 19.4. *Given an n-dimensional algebra $\mathcal{L}$.*

(1) $\mathcal{L}$ is symmetric, if and only if

$$M_{\mathcal{L}}(W_{[n]} - I_{n^2}) = 0; \tag{19.38}$$

(2) $\mathcal{L}$ is skew symmetric, if and only if

$$M_{\mathcal{L}}(W_{[n]} + I_{n^2}) = 0; \tag{19.39}$$

(3) $\mathcal{L}$ is associative, if and only if

$$M_{\mathcal{L}}(M_{\mathcal{L}} \otimes I_n - I_n \otimes M_{\mathcal{L}}) = 0. \tag{19.40}$$

To prove this proposition we need a lemma, which itself is useful.

Recalling Definition 2.6 and Proposition 2.13, we know that let V_1 and V_2 be two vector spaces with dimensions n_1 and n_2 respectively. The tensor product space $V = V_1 \otimes V_2$ is generated by $\{x \otimes y \,|\, x \in V_1, y \in V_2\}$. Furthermore, assume $\Phi : V_1 \times V_2 \to W$ is a bilinear mapping, then there exists a unique mapping $\Psi : V \to W$, which makes the graph in Fig. 19.1 commutative.

$$\begin{tikzcd}
V_1 \times V_2 \arrow[rr, "\Phi"'] \arrow[d, "\otimes"] & & W \\
V \arrow[urr, "\Psi"'] & &
\end{tikzcd}$$

Fig. 19.1 Tensor product space

This concept can be extended to the tensor product of k vector spaces deduced by kth fold linear mapping.

Let V_i, $i = 1, \cdots, k$ be n_i-dimensional vector spaces with their bases $\{d_1^i, \cdots, d_{n_i}^i\}$ and W an m-dimensional vector space with a fixed basis. Corresponding to these basis a multilinear mapping $\Phi : V_1 \times \cdots \times V_k \to W$ has its structure matrix $M_\Phi \in \mathcal{M}_{m\times n}$, where $n = \prod_{i=1}^k n_i$. Tensor product space $V = \otimes_{i=1}^k V_i$ has basis

$$\begin{aligned}\{e_1, \cdots, e_n\} := \big\{ e_{i_1}^1 \otimes e_{i_1}^2 \otimes \cdots \otimes e_{i_k}^k \;\big|\; i_k = 1, \cdots, n_k, \\ i_{k-1} = 1, \cdots, n_{k-1}, \cdots, i_1 = 1, \cdots, n_1 \big\}.\end{aligned} \tag{19.41}$$

Then there exists a unique deduced mapping $\Psi : V \to W$ such that Φ and $\Psi \circ \otimes$ are commutative.

Using these concepts, we can prove the following lemma.

Lemma 19.4.

(1) Under the basis defined in (19.41), Ψ and Φ have the same structure matrices, i.e.,

$$M_\Psi = M_\Phi. \tag{19.42}$$

(2) If

$$\Phi(X_1, \cdots, X_k) = 0, \quad \forall\, X_i \in V_i,$$

then

$$\Psi(Y) = 0, \quad \forall\, Y \in V.$$

Proof. (1) From the construction of the mapping one sees that

$$\Psi(e_{i_1}^1 \otimes \cdots \otimes e_{i_k}^k) = \Phi(e_{i_1}^1, \cdots, e_{i_k}^k), \quad \forall\, 1 \le i_t \le n_t,\ t = 1, \cdots, k.$$

Note that

$$\big\{ e_{i_1}^1 \otimes \cdots \otimes e_{i_k}^k \;\big|\; 1 \le i_t \le n_t, t = 1, \cdots, k \big\}$$

form a basis of V. That is, (19.42) holds on the basis of V. Since $\Psi : V \to W$ is a linear mapping, (19.42) holds for overall V.

(2) It is a direct consequence of (1). In fact, it comes from (1) by setting $M_\Phi = 0$.

□

Proof of Proposition 19.4. We prove (19.40) only. The proof of (19.38) or (19.39) is similar.

Using (19.34), equation (19.37) can be expressed in matrix form as

$$MM(XY)Z = MX(MYZ), \quad \forall\, X, Y, Z \in \mathcal{L}.$$

Using associativity, we have

$$M^2XYZ = M(XM)YZ = M(I_n \otimes M)XYZ, \quad \forall X, Y, Z \in \mathcal{L}.$$

Then we have

$$(M^2 - M(I_n \otimes M))XYZ = 0, \quad \forall X, Y, Z \in \mathcal{L}. \tag{19.43}$$

Note that though

$$S = \{XYZ \mid X, Y, Z \in \mathcal{L}\}$$

is not a vector space, Lemma 19.4 assures the correctness of (19.43) for any X, Y, Z in the vector space. (19.40) follows. □

Next, we consider when an algebra is a Lie algebra.

Proposition 19.5. *An algebra $\mathcal{L}$ is a Lie algebra, if and only if its structure matrix satisfies*

(i) sky-symmetry: That is, (19.39) holds;
(ii) Jacobi identity: That is,

$$M_{\mathcal{L}}^2(I_{n^2} + W_{[n,n^2]} + W_{[n^2,n]}) = 0. \tag{19.44}$$

***Proof*.** It is well know that an algebra is a Lie algebra, if and only if it is skew symmetric and satisfies Jacobi identity (19.31). According to Proposition 19.4, the skew symmetry is equivalent to (19.39), which leads to (i).

As for (ii), using structure matrix, (19.31) can be expressed as

$$M_{\mathcal{L}}^2(XYZ + YZX + ZXY) = 0. \tag{19.45}$$

Using the proposition of swap matrix, we have

$$W_{[n,n^2]}XYZ = YZX, \quad W_{[n^2,n]}XYZ = ZXY.$$

Plugging it into (19.45), an argument similar to the proof of Proposition 19.4 shows (19.31) is correct. □

Example 19.2. In $\mathbb{R}^3$ the cross product is defined as follows: Let $X = x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k}$ and $Y = y_1\mathbf{i} + y_2\mathbf{j} + y_3\mathbf{k}$. Then

$$X \bowtie Y = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix}.$$

Its structure matrix is easily calculated as

$$M_{\bowtie} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{19.46}$$

A straightforward computation shows that

$$M_{\bowtie}\left(I_9 + W_{[3]}\right) = 0,$$

and

$$M_{\bowtie}^2\left(I_{27} + W_{[3.9]} + W_{[9,3]}\right) = 0.$$

Hence $\mathbb{R}^3$ with cross product is a Lie algebra.

Next, we consider the structure matrix of general linear algebra $gl(n, \mathbb{R})$. First, we give a formula for the column stacking form of the product of two matrices, which itself is useful.

Lemma 19.5. *Let $A \in \mathcal{M}_{m\times n}$ and $B \in \mathcal{M}_{n\times p}$. Then*

$$V_c(AB) = \Psi^n_{mp} V_c(A) V_c(B), \tag{19.47}$$

where

$$\Psi^n_{mp} = \begin{bmatrix} I_m \otimes (\delta_p^1 \delta_n^1)^T & I_m \otimes (\delta_p^1 \delta_n^2)^T & \cdots & I_m \otimes (\delta_p^1 \delta_n^n)^T \\ I_m \otimes (\delta_p^2 \delta_n^1)^T & I_m \otimes (\delta_p^2 \delta_n^2)^T & \cdots & I_m \otimes (\delta_p^2 \delta_n^n)^T \\ \vdots & \vdots & & \vdots \\ I_m \otimes (\delta_p^p \delta_n^1)^T & I_m \otimes (\delta_p^p \delta_n^2)^T & \cdots & I_m \otimes (\delta_p^p \otimes \delta_n^n)^T \end{bmatrix}, \tag{19.48}$$

where $\delta_p^i = \mathrm{Col}_i(I_p)$.

Proof. Observing Table 3.2, it is easy to verify that

$$V_c(AB) = (I_p \otimes A) V_c(B).$$

Hence we have to calculate $I_p \otimes A$. A straightforward computation shows

$$I_p \otimes A = \Psi^n_{mp} V_c(A).$$

□

Now we are ready to consider the structure matrix of $gl(n, \mathbb{R})$. As a shorthand we denote $\Psi_n := \Psi^n_{nn}$.

Example 19.3. Consider $gl(n, \mathbb{R})$. Choose a basis as $\{M_{I,J},\ I = 1, \cdots, n;\ J = 1, \cdots, n\}$, where $M_{I,J}$ is defined by

$$(M_{I,J})_{i,j} = \begin{cases} 1, & i = I \text{ and } j = J, \\ 0, & \text{otherwise.} \end{cases}$$

Define the product, called the Lie bracket, as

$$[A,B]=AB-BA.$$

We construct its structure matrix. Using Lemma 19.5, we have

$$V_c(AB)=\Psi_n V_c(A)V_c(B),$$

and

$$V_c(BA)=\Psi_n V_c(B)V_c(A)=\Psi_n W_{[n^2]}V_c(A)V_c(B).$$

Hence,

$$V_c([A,B])=(\Psi_n-\Psi_n W_{[n^2]})V_c(A)V_c(B),$$

It follows that the structure matrix of $gl(n,\mathbb{R})$ is

$$M=\Psi_n(I_{n^4}-W_{[n^2]}). \tag{19.49}$$

We leave to the reader to verify that M satisfies (19.39) and (19.44).

19.4 Two-Dimensional Algebras

This section considers the classification of two-dimensional algebras. Therefore, we need an equivalence over the set. To this end, we consider the topological structure of the n-dimensional real algebras. Note that an algebra is uniquely determined by its structure matrix, we therefore can pose the conventional topological structure of $\mathbb{R}^{n\times n^2}$ to n-dimensional algebra. Since the structure matrix depends on the coordinate frame. The different structure matrices caused by coordinate transformations should be equivalent. The quotient topology under this equivalence then becomes a proper description of the space of algebras.

Let $M_{\mathcal{L}_n}$ be the structure matrix of an n-dimensional algebra $\mathcal{L}_n$ with respect to the basis $E_1=(e_1^1\ e_2^1\ \cdots\ e_n^1)$. Let $E_2=(e_1^2\ e_2^2\ \cdots\ e_n^2)$ be another basis of $\mathbb{R}^n$ and there exists a linear transformation $T\in GL(n,\mathbb{R})$, such that $E_2=E_1T$. Here $GL(n,\mathbb{R})$ is the n-dimensional general linear group. Now, consider two vectors $V_1=E_1X_1$ and $V_2=E_1X_2$. Using new basis, their vector forms become $\tilde{X}_i=T^{-1}X_i$, $i=1,2$. Then we have

$$\begin{aligned}&T^{-1}M_{\mathcal{L}_n}X_1X_2=\tilde{M}_{\mathcal{L}_n}T\tilde{X}_1T\tilde{X}_2\\&=\tilde{M}_{\mathcal{L}_n}T^{-1}(I_n\otimes T^{-1})X_1X_2=\tilde{M}_{\mathcal{L}_n}(T^{-1}\otimes T^{-1})X_1X_2.\end{aligned}$$

Therefore, we have

$$\tilde{M}_{\mathcal{L}_n}=T^{-1}M_{\mathcal{L}_n}(T\otimes T). \tag{19.50}$$

Applying this result to two algebras, we have

Proposition 19.6. *Consider two n-dimensional algebras $\mathcal{L}_n^1$ and $\mathcal{L}_n^2$. They are equivalent, if and only if there exists a $T \in GL(n, \mathbb{R})$ such that*

$$M_{\mathcal{L}_n^1} = T^{-1} M_{\mathcal{L}_n^2} (T \otimes T). \tag{19.51}$$

Hence the topology of n-dimensional algebras is

$$\mathbb{R}^{n \times n^2} / GL(n, \mathbb{R}).$$

Under this equivalence we consider the classification of two-dimensional real algebras with identity, which can be always expressed as

$$\mathcal{A} = \{a + b\xi \mid a, b \in \mathbb{R}\}. \tag{19.52}$$

Then 1 is the identity and the product $*$ is defined in a natural way as

$$(a + b\xi) * (c + d\xi) = ac + (ad + bc)\xi + bd\xi^2, \quad a, b, c, d \in \mathbb{R}. \tag{19.53}$$

It is natural that $\{1, \xi\}$ is chosen as the required basis. ξ^2 can be always expressed as

$$\xi^2 = \alpha + \beta\xi. \tag{19.54}$$

First, we list three classes of two-dimensional real algebras with identity.

(i) Complex numbers ($\mathbb{C}$) : Corresponding to this we have $\xi^2 = -1$.

(ii) Dual numbers ($\mathbb{D}$) : Corresponding to this we have $\xi^2 = 0$.
Dual numbers were invented by Clifford in 1873 (Clifford, 1873). It has many applications in mechanics (Cheng and Thompson, 1997; Azariadis and Aspragathos, 2001).

(iii) Hyperbolic numbers ($\mathbb{H}$) : Corresponding to this we have $\xi^2 = 1$.
Hyperbolic number is also called split-complex number or hypercomplex number. The use of hyperbolic numbers dates back to 1848 when J. Cockle revealed his tessarines. We refer to Borota *et al.* (2000) for its basic properties and some applications to functions of space-time variables.

Theorem 19.4. *There are only three classes of two-dimensional algebras with identity over $\mathbb{R}$, which are $\mathbb{C}$, $\mathbb{D}$, and $\mathbb{H}$.*

Proof. The structure matrix of two-dimensional algebra with identity must be

$$M_{\mathcal{A}} = \begin{bmatrix} 1 & 0 & 0 & \alpha \\ 0 & 1 & 1 & \beta \end{bmatrix}. \tag{19.55}$$

By Proposition 19.6, to classify the algebras we have to consider the equivalent classis under linear transformations. To keep $1 \sim \delta_2^1$ unchanged, the linear transformation T should be of the following form:

$$T = \begin{bmatrix} 1 & u \\ 0 & v \end{bmatrix}, \quad v \neq 0.$$

Then

$$T^{-1} = \begin{bmatrix} 1 & -\frac{u}{v} \\ 0 & \frac{1}{v} \end{bmatrix}.$$

Under the linear transformation (19.51),

$$\begin{aligned} T^{-1} M_{\mathcal{A}} T (I_2 \otimes T) &= \begin{bmatrix} 1 & -\frac{u}{v} \\ 0 & \frac{1}{v} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & \alpha \\ 0 & 1 & 1 & \beta \end{bmatrix} \begin{bmatrix} 1 & u \\ 0 & v \end{bmatrix} \left(I_2 \otimes \begin{bmatrix} 1 & u \\ 0 & v \end{bmatrix} \right) \\ &= \begin{bmatrix} 1 & 0 & 0 & -u^2 - \beta uv + \alpha v^2 \\ 0 & 1 & 1 & 2u + \beta v \end{bmatrix}. \end{aligned} \tag{19.56}$$

It follows that

$$\begin{cases} \tilde{\alpha} = -u^2 + \alpha v^2 - \beta uv \\ \tilde{\beta} = 2u + \beta v. \end{cases} \tag{19.57}$$

From (19.57) we conclude that for the case of two-dimensional with identity, each algebra is equivalent to an algebra with $\beta = 0$. That is, with an algebra, that has its structure matrix as

$$M_{\mathcal{A}} = \begin{bmatrix} 1 & 0 & 0 & \alpha \\ 0 & 1 & 1 & 0 \end{bmatrix}. \tag{19.58}$$

Now we have only to consider how many classes of algebras with its structure matrix as in (19.58). Recall, (19.57) again. Now since $\beta = \tilde{\beta} = 0$ we have $u = 0$. Hence we have

$$\tilde{\alpha} = v^2 \alpha, \quad v \neq 0. \tag{19.59}$$

It follows that there exists a linear transformation, which converts one to the other, if and only if α and $\tilde{\alpha}$ have the same sign. It follows that there are three classes of two-dimensional algebras with identity, which corresponding to $\alpha = -1$, $\alpha = 0$, and $\alpha = 1$. It is easy to verify that they correspond to $\mathbb{C}$, $\mathbb{D}$, and $\mathbb{H}$ respectively. □

19.5 Three-Dimensional Algebras

This section considers three-dimensional real algebras with identity. We also require the algebra to be associative and commutative. Note that in two-dimensional case, identity assures association and commutativity. But it is no longer true for three-dimensional case.

Similar to two-dimensional case, we assume the algebra is expressed as

$$\mathcal{A} = \{r_0 + r_1\alpha + r_2\beta \mid r_0, r_1, r_2 \in \mathbb{R}\}. \tag{19.60}$$

Taking $B = \{1, \alpha, \beta\}$ as the basis, to assure the commutativity, we need only $\alpha\beta = \beta\alpha$. Ignoring the associativity tentatively, we can have the structure matrix of $\mathcal{A}$ with respect to this basis as

$$M_{\mathcal{A}} = \begin{bmatrix} 1 & 0 & 0 & 0 & a & d & 0 & d & h \\ 0 & 1 & 0 & 1 & b & e & 0 & e & i \\ 0 & 0 & 1 & 0 & c & f & 1 & f & j \end{bmatrix}. \tag{19.61}$$

To meet the associative requirement, we need $(x * y) * z = x * (y * z)$, which can be expressed in vector form as

$$M_{\mathcal{A}}(M_{\mathcal{A}}xy)z = M_{\mathcal{A}}x(M_{\mathcal{A}}yz) = M_{\mathcal{A}}(I_3 \otimes M_{\mathcal{A}})xyz.$$

It follows that

$$M_{\mathcal{A}}^2 = M_{\mathcal{A}}(I_3 \otimes M_{\mathcal{A}}). \tag{19.62}$$

Plugging (19.61) into (19.62), a straightforward computation shows that to verify (19.62) the parameters must satisfy the following algebraic equations:

$$\begin{cases} ci = d + ef \\ db + ch = ae + df \\ a + bf + cj = ce + f^2 \\ e^2 + fi = h + bi + ej \\ de + fh = ai + dj. \end{cases} \tag{19.63}$$

We look for some non-zero solutions. It is natural that we assume $\alpha^2, \beta^2 \in \mathbb{R}$. That is, we assume

A-1

$$b = c = i = j = 0. \tag{19.64}$$

To see A-1 is reasonable, recall that all $\mathbb{C}$, $\mathbb{D}$, $\mathbb{H}$ and quaternion satisfy this.

Under assumption A-1 (19.63) becomes

$$\begin{cases} d+ef=0 \\ ae+df=0 \\ a=f^2 \\ e^2=h \\ de+fh=0. \end{cases} \tag{19.65}$$

Case 1 $f=0$:

It follows that $a=d=0$. We have the product rule as

$$\begin{cases} \alpha^2=0 \\ \alpha\beta=\beta\alpha=e\alpha \\ \beta^2=e^2. \end{cases} \tag{19.66}$$

Now let $x=r_1+r_2\alpha+r_3\beta$ and $y=r_4+r_5\alpha+r_6\beta$. Then

$$x*y=(r_1r_4+r_3r_6e^2)+[(r_2r_6+r_3r_5)e+r_1r_5+r_2r_4]\alpha+(r_1r_6+r_3r_4)\beta.$$

There are two sub-cases: **Case 1.1** $e=0$, and **Case 1.2** $e\neq 0$.

Case 2 $f\neq 0$:

It follows that

$$\begin{cases} a=f^2 \\ e=-\frac{d}{f} \\ h=\frac{d^2}{f^2}. \end{cases}$$

Now e can also be chosen freely.

Case 2.1 $e=0$:

Then the product rule is:

$$\begin{cases} \alpha^2=f^2 \\ \alpha\beta=\beta\alpha=f\beta \\ \beta^2=0. \end{cases} \tag{19.67}$$

This is the same as **Case 1.1**.

Case 2.2 $e\neq 0$:

Then the product rule is:

$$\begin{cases} \alpha^2=f^2 \\ \alpha\beta=\beta\alpha=d-\frac{d}{f}\alpha+f\beta \\ \beta^2=\frac{d^2}{f^2}. \end{cases} \tag{19.68}$$

Let $x = r_1 + r_2\alpha + r_3\beta$ and $y = r_4 + r_5\alpha + r_6\beta$. Then

$$x * y = r_1r_4f^2 + r_3r_6\frac{d^2}{f^2} - (r_2r_6 + r_3r_5)\frac{d}{f}\alpha + (r_2r_6 + r_3r_5)f\beta.$$

Since the particular values of f etc. are meaningless. (Re-scheduling the sizes of α and β can get any values of f and e.) Hence we have the following result.

Theorem 19.5. *There are three typical three-dimensional associative and commutative algebras with identity, satisfying assumption* **A-1**. *They are*

$$\mathcal{A} = \{r_1 + r_2\alpha + r_3\beta \mid r_1, r_2, r_3 \in \mathbb{R}\}, \tag{19.69}$$

with the product rules respectively as

- **R-1***:*

$$\begin{cases} \alpha^2 = 0 \\ \alpha\beta = \beta\alpha = 0 \\ \beta^2 = 0. \end{cases} \tag{19.70}$$

- **R-2***:*

$$\begin{cases} \alpha^2 = 1 \\ \alpha\beta = \beta\alpha = \beta \\ \beta^2 = 0. \end{cases} \tag{19.71}$$

- **R-3***:*

$$\begin{cases} \alpha^2 = 1 \\ \alpha\beta = \beta\alpha = 1 - \alpha + \beta \\ \beta^2 = 1. \end{cases} \tag{19.72}$$

Next, we consider when a three-dimensional commutative and associative algebra with identity can be expressed as one of the above two forms. Similar to two-dimensional case, we want to find a coordinate transformation to convert a general form to satisfy **A-1**.

Recall that

$$\begin{cases} \alpha^2 = a + b\alpha + c\beta \\ \beta^2 = h + i\alpha + j\beta \\ \alpha\beta = \beta\alpha = d + e\alpha + f\beta. \end{cases} \tag{19.73}$$

Assume we have a transformation as

$$\begin{cases} \tilde{\alpha} = x_1 + x_2\alpha + x_3\beta \\ \tilde{\beta} = y_1 + y_2\alpha + y_3\beta. \end{cases} \tag{19.74}$$

Note that we do not change real part. Then (19.74) is a coordinate change, if and only if

$$\det \begin{pmatrix} x_2 & y_2 \\ x_3 & y_3 \end{pmatrix} \neq 0. \tag{19.75}$$

Using (19.73), a straightforward computation shows that

$$\begin{aligned} \tilde{\alpha}^2 &= (x_1^2 + ax_2^2 + hx_3^2 + 2dx_2x_3) \\ &\quad +(bx_2^2 + 2x_1x_2 + 2ex_2x_3 + ix_3^2)\alpha \\ &\quad +(cx_2^2 + 2fx_2x_3 + 2x_1x_3 + jx_3^2)\beta, \\ \tilde{\beta}^2 &= (y_1^2 + ay_2^2 + hy_3^2 + 2dy_2y_3) \\ &\quad +(by_2^2 + 2y_1y_2 + 2ey_2y_3 + iy_3^2)\alpha \\ &\quad +(cy_2^2 + 2fy_2y_3 + 2y_1y_3 + jy_3^2)\beta. \end{aligned} \tag{19.76}$$

Note that at least one of $\{b, c, i, j\}$ is non-zero. Otherwise, α and β need no change. Without loss of generality, we assume $c \neq 0$. Then $x_3 \neq 0$, otherwise x_2 and x_3 are both zero simultaneously since we want $cx_2^2 + 2fx_2x_3 + 2x_1x_3 + jx_3^2 = 0$. For the same reason, $y_3 \neq 0$. Denote by $z = \frac{x_2}{x_3}$ and $\mu = \frac{x_1}{x_3}$ (or $z = \frac{y_2}{y_3}$ and $\mu = \frac{y_1}{y_3}$). Then it is clear from (19.76) that the coordinate transformation, which meets assumption **A-1**, exists, if and only if the following equation has two distinguish solutions:

$$\begin{cases} bz^2 + 2(\mu + e)z + i = 0 \\ cz^2 + 2fz + (2\mu + j) = 0. \end{cases} \tag{19.77}$$

Solving μ from the second equation of (19.77),

$$\mu = -\frac{1}{2}(cz^2 + 2fz + j).$$

Plugging it into the first equation yields

$$cz^3 + (2f - b)z^2 + (j - 2e)z - i = 0. \tag{19.78}$$

We simply express it as

$$F(z) := z^3 + \sigma_1 z^2 + \sigma_2 z + \sigma_3 = 0, \tag{19.79}$$

where

$$\sigma_1 = \frac{2f - b}{c} \quad, \sigma_2 = \frac{j - 2e}{c} \quad, \sigma_3 = -\frac{i}{c}.$$

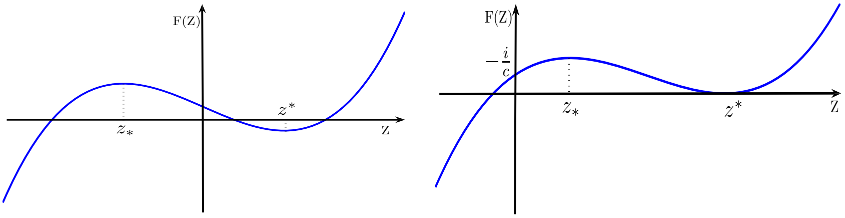
Setting

$$F'(z) = 3z^2 + 2\sigma_1 z + \sigma_2 = 0,$$

we have solutions

$$z = \frac{-\sigma_1 \pm \sqrt{\sigma_1^2 - 3\sigma_2}}{3}.$$

According to the properties of cubic curve, we know that when $\sigma_1^2 - 3\sigma_2 \le 0$, $F(z)$ is strictly increasing, and there is unique solution. So only if $\sigma_1^2 - 3\sigma_2 > 0$ we may have more than one real solutions.



(a) $F(z_*)F(z^*) < 0$

(b) $F(z_*)F(z^*) = 0$

Fig. 19.2 $F(z)$

Using above notation, we have

Theorem 19.6. *A three-dimensional commutative associative algebra with identity is convertible to a canonical form (19.70), (19.71) or (19.72), if and only if*

(i)

$$\sigma_1^2 - 3\sigma_2 > 0,$$

and

(ii)

$$F(z_*)F(z^*) \le 0,$$

where

$$z_* = \frac{-\sigma_1 - \sqrt{\sigma_1^2 - 3\sigma_2}}{3} < z^* = \frac{-\sigma_1 + \sqrt{\sigma_1^2 - 3\sigma_2}}{3}.$$

Proof. Since $\sigma_1^2 - 3\sigma_2 > 0$, we have $F(z_*) > F(z^*)$. If $F(z_*)F(z^*) > 0$, then either $F(z_*) > F(z^*) > 0$ or $0 > F(z_*) > F(z^*)$, there is only one solution of (19.79).

When $F(z_*)F(z^*) < 0$, as Fig. 19.2 (a) shows, or $F(z_*)F(z^*) = 0$ (Fig. 19.2 (b)), it is easy to see there must be at least two distinguish solutions. □

Remark 19.2. In the previous procedure, we have assumed $c \neq 0$. If $c = 0$, we can assume a member in $\{b, i, j\}$ which is nonzero. Then repeat above argument, we can obtain a similar result.

19.6 Lower-Dimensional Lie Algebra and Invertible Algebra

The structure of Lie algebras has been discussed before. This section considers the structure and properties of lower-dimensional Lie algebras. First, we consider the case of $n = 2$. To assure the skew symmetry, a 2-dimensional Lie algebra has its structure matrix as

$$M_{\mathcal{L}_2} = \begin{bmatrix} 0 & a & -a & 0 \\ 0 & b & -b & 0 \end{bmatrix}. \tag{19.80}$$

A simple computation shows that

$$M_{\mathcal{L}_2}^2 = \begin{bmatrix} 0 & 0 & -ab & a^2 & ab & -a^2 & 0 & 0 \\ 0 & 0 & -b^2 & ab & b^2 & -ab & 0 & 0 \end{bmatrix}. \tag{19.81}$$

We also have

$$I_8 + W_{[2,4]} + W_{[4,2]} = \begin{bmatrix} 3&0&0&0&0&0&0&0 \\ 0&1&1&0&1&0&0&0 \\ 0&1&1&0&1&0&0&0 \\ 0&0&0&1&0&1&1&0 \\ 0&1&1&0&1&0&0&0 \\ 0&0&0&1&0&1&1&0 \\ 0&0&0&1&0&1&1&0 \\ 0&0&0&0&0&0&0&3 \end{bmatrix}. \tag{19.82}$$

It is ready to check that

$$M_{\mathcal{L}_2}^2 (I_8 + W_{[2,4]} + W_{[4,2]}) = 0.$$

Hence we have the following result:

Proposition 19.7. *Any 2-dimensional skew symmetric algebra is a Lie algebra.*

Next, we consider the case of $n = 3$. To assure skew symmetry, its structure matrix must be as

$$M_{\mathcal{L}_3} = \begin{bmatrix} 0 & a & d & -a & 0 & g & -d & -g & 0 \\ 0 & b & e & -b & 0 & h & -e & -h & 0 \\ 0 & c & f & -c & 0 & i & -f & -i & 0 \end{bmatrix}. \tag{19.83}$$

A simple computer routine can calculate

$$M_{\mathcal{L}_3}^2(I_{27}+W_{[3,9]}+W_{[9,3]}),$$

which is a 3×27 matrix. Fortunately, it has only a few nonzero elements, which are

$$\begin{aligned}
m_{1,6}&=m_{1,16}=m_{1,22}=-m_{1,8}=-m_{1,12}=-m_{1,20}=bg+gf-ah-di;\\
m_{2,6}&=m_{2,16}=m_{2,22}=-m_{2,8}=-m_{2,12}=-m_{2,20}=ae-bd+hf-ei;\\
m_{3,6}&=m_{3,16}=m_{3,22}=-m_{3,8}=-m_{3,12}=-m_{3,20}=af+bi-cd-ch.
\end{aligned}$$

We conclude that

Theorem 19.7. *A 3-dimensional algebra is a Lie algebra, if and only if its structure matrix has the form of (19.83), where the nonzero elements satisfy the following equations:*

$$\begin{cases} bg+gf-ah-di=0,\\ ae-bd+hf-ei=0,\\ af+bi-cd-ch=0. \end{cases} \tag{19.84}$$

Example 19.4. According to Theorem 19.7, we are able to construct many 3-dimensional Lie algebras.

(1) Set

$$a=b=d=f=h=i=0,\quad c=g=1,\quad e=-1.$$

It is easy to check that this is a solution of (19.84). In fact, this solution corresponds to the standard cross product on $\mathbb{R}^3$, which is a known Lie algebra.

(2) To get another nontrivial solution of (19.84), we convert it into a matrix form.

$$\begin{bmatrix} -h & g & 0\\ e & -d & 0\\ f & i & -d-h \end{bmatrix}\begin{bmatrix} a\\ b\\ c\end{bmatrix}=\begin{bmatrix} di-gf\\ ei-hf\\ 0\end{bmatrix}. \tag{19.85}$$

To find a nonzero solution of (19.84) it suffices to choose d,e,f,g,h,i, such that the coefficient matrix of (19.85) is nonsingular. Then we can uniquely solve the corresponding a,b,c. For instance, we choose $d=-e=1$, $f=-g=2$, $h=-i=3$, Then we have

$$\begin{bmatrix} -3 & -2 & 0\\ -1 & -1 & 0\\ 2 & -3 & -4 \end{bmatrix}\begin{bmatrix} a\\ b\\ c\end{bmatrix}=\begin{bmatrix} 1\\ -3\\ 0\end{bmatrix}.$$

Its solution is $a = -7$, $b = 10$, $c = -11$. Hence, we can have another Lie algebra as

$$\mathcal{L}_3 = \{\alpha I + \beta J + \gamma K \mid \alpha, \beta, \gamma \in \mathbb{R}\},$$

with its product $*$ defined as

$$\begin{aligned}
&I * I = J * J = K * K = 0\\
&I * J = -J * I = -7I + 10J - 11K\\
&I * K = -K * I = I - J + 2K\\
&J * K = -K * J = -2I + 3J - 3K.
\end{aligned}$$

We continue to consider the case of $n = 4$. To assure skew symmetry its 4×16 structure matrix, denoted by $M_{\mathcal{L}_4} = (m_{ij})$, should satisfy

$$\begin{cases}
m_{i,j} = 0, \quad j = 1, 6, 11, 16;\\
m_{i,2} = -m_{i,5} := x_i,\\
m_{i,3} = -m_{i,9} := x_{4+i},\\
m_{i,4} = -m_{i,13} := x_{8+i},\\
m_{i,7} = -m_{i,10} := x_{12+i},\\
m_{i,8} = -m_{i,14} := x_{16+i},\\
m_{i,12} = -m_{i,15} := x_{20+i},\\
\qquad\qquad\qquad i = 1, 2, 3, 4.
\end{cases} \tag{19.86}$$

Using Matlab, a simple routine converts (19.44) to its equivalent form as

$$\begin{cases}
-x_1x_{14} + x_2x_{13} - x_4x_{21} - x_5x_{15} + x_7x_{13} + x_8x_{17} - x_9x_{16} = 0,\\
x_1x_6 - x_2x_5 - x_4x_{22} - x_6x_{15} + x_7x_{14} + x_8x_{18} - x_{10}x_{16} = 0,\\
x_1x_7 + x_2x_{15} - x_3x_5 - x_3x_{14} - x_4x_{23} + x_8x_{19} - x_{11}x_{16} = 0,\\
x_1x_8 + x_2x_{16} - x_4x_5 - x_4x_{14} - x_4x_{24} + x_7x_{16} - x_8x_{15}\\
\quad + x_8x_{20} - x_{12}x_{16} = 0,\\
-x_1x_{18} + x_2x_{17} + x_3x_{21} - x_5x_{19} - x_9x_{20} + x_{11}x_{13} + x_{12}x_{17} = 0,\\
x_1x_{10} - x_2x_9 + x_3x_{22} - x_6x_{19} - x_{10}x_{20} + x_{11}x_{14} + x_{12}x_{18} = 0,\\
x_1x_{11} + x_2x_{19} - x_3x_9 - x_3x_{18} + x_3x_{23} - x_7x_{19} + x_{11}x_{15}\\
\quad - x_{11}x_{20} + x_{12}x_{19} = 0,\\
x_1x_{12} + x_2x_{20} + x_3x_{24} - x_4x_9 - x_4x_{18} - x_8x_{19} + x_{11}x_{16} = 0,\\
-x_1x_{22} - x_5x_{23} + x_6x_{17} + x_7x_{21} - x_9x_{24} - x_{10}x_{13} + x_{12}x_{21} = 0,\\
-x_2x_{22} + x_5x_{10} - x_6x_9 + x_6x_{18} - x_6x_{23} + x_7x_{22} - x_{10}x_{14}\\
\quad - x_{10}x_{24} + x_{12}x_{22} = 0,\\
-x_3x_{22} + x_5x_{11} + x_6x_{19} - x_7x_9 - x_{10}x_{15} - x_{11}x_{24} + x_{12}x_{23} = 0,\\
-x_4x_{22} + x_5x_{12} + x_6x_{20} + x_7x_{24} - x_8x_9 - x_8x_{23} - x_{10}x_{16} = 0,\\
x_1x_{21} - x_5x_{17} + x_9x_{13} - x_{13}x_{18} - x_{13}x_{23} + x_{14}x_{17}\\
\quad + x_{15}x_{21} - x_{17}x_{24} + x_{20}x_{21} = 0,\\
x_2x_{21} - x_6x_{17} + x_{10}x_{13} - x_{14}x_{23} + x_{15}x_{22} - x18x_{24} + x_{20}x_{22} = 0,\\
x_3x_{21} - x_7x_{17} + x_{11}x_{13} + x_{14}x_{19} - x_{15}x_{18} - x_{19}x_{24} + x_{20}x_{23} = 0,\\
x_4x_{21} - x_8x_{17} + x_{12}x_{13} + x_{14}x_{20} + x_{15}x_{24} - x_{16}x_{18}\\
\quad - x_{16}x_{23} = 0.
\end{cases} \tag{19.87}$$

Then we have

Theorem 19.8. *A* 4*-dimensional algebra is a Lie algebra, if and only if its structure matrix satisfies (19.86) with its nonzero parameters* $x_1, \cdots, x_{24}$ *satisfy (19.87).*

Example 19.5. In fact, (19.87) has many solutions. For instance, choosing

$$x_i = 0, \quad i > 8,$$

then (19.87) becomes

$$\begin{cases} x_1x_6 - x_2x_5 = 0, \\ x_1x_7 - x_3x_5 = 0, \\ x_1x_8 - x_4x_5 = 0. \end{cases} \tag{19.88}$$

It is obvious that $\{x_1, x_2, \cdots, x_8\}$ is a solution of (19.87), if and only if

$$x_1 : x_2 : x_3 : x_4 = x_5 : x_6 : x_7 : x_8.$$

For instance, a simple solution is $x_1 = 1$, $x_2 = -1$, $x_3 = 2$, $x_4 = -2$, $x_5 = -1$, $x_6 = 1$, $x_7 = -2$, $x_8 = 2$. Using it, we have a Lie algebra as

$$\mathcal{L} = \{aI + bJ + cK + dH \mid a, b, c, d \in \mathbb{R}\},$$

with its product $*$ satisfying

$$\begin{cases} I * I = J * J = K * K + H * H = 0, \\ I * J = -J * I = -I + J - 2K + 2H, \\ I * K = -K * J = I - J + 2K - 2H, \\ I * H = -H * I = J * K = -K * J = J * H \\ \quad = -H * J = K * H = -H * K = 0. \end{cases}$$

We consider a particular Lie algebra $gl(2, \mathbb{R})$.

Example 19.6. Consider $gl(2, \mathbb{R})$. The product was defined in Example 19.3. We choose its basis as $\{e_1, e_2, e_3, e_4\}$, where

$$e_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad e_3 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad e_4 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

A straightforward computation shows that its structure matrix is

$$M = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Finally, we consider the general case where the dimension is n. Let the structure matrix of an n-dimensional algebra be expressed as

$$M_{\mathcal{L}_n} = \begin{bmatrix} W_{11} \cdots W_{1n} & \cdots & W_{n1} \cdots W_{nn} \end{bmatrix},$$

where each W_{ij} is a column with w_{ij}^s as its elements. Then we can calculate that

$$M_{\mathcal{L}_n} = \left[\sum_{s=1}^{n} w_{ij}^s W_{sk} \,\middle|\, i,j,k = 1,\cdots,n \right]. \tag{19.89}$$

Note that the columns of the matrix in (19.89) are arranged by the order of $\mathrm{id}(i,j,k;n^3)$. When right multiply it by $W_{[n^2,n]}$ (or $W_{[n,n^2]}$), its order becomes $\mathrm{id}(j,k,i;n^3)$ (correspondingly, $\mathrm{id}(k,i,j;n^3)$). Taking this into consideration, we have

$$\sum_{s=1}^{n} w_{ij}^s W_{sk} + \sum_{s=1}^{n} w_{jk}^s W_{si} + \sum_{s=1}^{n} w_{ki}^s W_{sj} = 0.$$

Because of the skew symmetry we need only to consider the case when i, j, k are distinct to each other. Hence, we have

$$\begin{aligned} &\sum_{s<k} w_{ij}^s W_{sk} - \sum_{s>k} w_{ij}^s W_{ks} + \sum_{s<i} w_{jk}^s W_{si} - \sum_{s>i} w_{jk}^s W_{is} + \sum_{s<j} w_{ki}^s W_{sj} \\ &\quad - \sum_{s>j} w_{ki}^s W_{js} = 0, \quad 1 \le i < j < k \le n. \end{aligned} \tag{19.90}$$

Hence we have

$$\frac{k!n}{3!(k-3)!}$$

independent equations. The set of Lie algebra is the algebraic variety generated by this set of polynomial equations (Hartshorne, 1977).

Another interesting topic is the invertibility of algebra.

Definition 19.6. Given an algebra $\mathcal{L}$ with its multiplication $*$. $\mathcal{L}$ is said to be invertible if

(i) there exists a unitary element e, such that

$$e * x = x, \quad \text{and} \quad x * e = x, \quad \forall\, x \in \mathcal{L}; \tag{19.91}$$

(ii) for any element $x \neq 0$, there exists unique element, denoted by $x^{-1} \in \mathcal{L}$, such that

$$x * x^{-1} = e. \tag{19.92}$$

In addition, if

(iii) $\mathcal{L}$ is commutative, that is,

$$x * y = y * x, \quad \forall\, x, y \in \mathcal{L}, \tag{19.93}$$

then $\mathcal{L}$ is called a field.

Note that since $0 \neq e \in \mathcal{L}$, it can be considered as a basis element. In fact, the one-dimensional subspace generated by e is isomorphic to $\mathbb{R}$. For convenience, if there is an unitary e, we always choose e as the first element of a basis of $\mathcal{L}$. That is, a conventional basis is $B = \{e_1 = e, e_2, e_3, \cdots, e_n\}$. In vector form we have $e = \delta_n^1$.

Assume the basis B is fixed, with respect to B we have the structure matrix M of $\mathcal{L}$. Then we consider how to check the condition (i) and (ii) of Definition 19.6. Split M as

$$M = \left[M_1 \; M_2 \; \cdots \; M_n\right],$$

where $M_i = \mathrm{Blk}_i(M)$ are $n \times n$ matrices. First, we consider condition (i). The following result is an immediate consequence of the definition.

Lemma 19.6. *There exists the unitary $e = \delta_n^1$, if and only if $M_1 = I_n$,* $\mathrm{Col}_1(M_j) = \delta_n^j$.

Proof.

$$e * x = Mex = [M_1, \cdots, M_n]\, \delta_1^n x = M_1 x.$$

Then $e * x = x$ leads to

$$M_1 x = x, \quad \forall\, x \in \mathbb{R}^n.$$

It follows that $M_1 = I_n$.

Similarly, since

$$x * e = MW_{[n]}ex = \left[M_1^1, \cdots, M_n^1\right] x = x, \quad \forall\, x \in \mathbb{R}^n,$$

which means $\left[M_1^1, \cdots, M_n^1\right] = I_n$. □

Next, we consider condition (ii). What we need to verify is: for any $x \neq 0$, there exists a unique y, such that

$$Mx * y = \delta_n^1. \tag{19.94}$$

Then, it is not difficult to verify that

Lemma 19.7. *Assume $\mathcal{L}$ has unitary $e = \delta_n^1$, then for any $x \neq 0$ there is a unique inverse element x^{-1}, if and only if*

$$\det(Mx) \neq 0, \quad \forall\, x \neq 0. \tag{19.95}$$

We investigate some examples.

Example 19.7. Consider the set of complex numbers $\mathbb{C}$, which can be considered as a vector space over $\mathbb{R}$, which has basis $\{1, i\}$. Then its structure matrix is

$$M_{\mathbb{C}} = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

We verify the commutativity, condition (i), and condition (ii).

(1) Commutativity: Note that

$$W_{[2]} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

It is easy to verify that

$$M_{\mathbb{C}} W_{[2]} = M_{\mathbb{C}}.$$

(2) Condition (i): Let $x = (\alpha, \beta)^T \sim \alpha + \beta i \in \mathbb{C}$. Then

$$\begin{aligned} x * e = M_{\mathbb{C}} x e &= \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \end{aligned}$$

(3) Condition (ii):

$$\det(M_{\mathbb{C}} x) = \det\left(\begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix} \right) = \alpha^2 + \beta^2.$$

It follows that $\det(M_{\mathbb{C}} x) = 0$, if and only if $x = 0$.

Example 19.8. Consider the quaternion (Greub, 1981), its standard basis is $\{1, I, J, K\}$. Under this basis its structure matrix is

$$M_Q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}. \tag{19.96}$$

It is easy to verify condition (i). As for condition (ii), assume $x = \begin{bmatrix} a\ b\ c\ d \end{bmatrix}^T \neq 0$, we have

$$M_Q x = \begin{bmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{bmatrix},$$

and

$$\begin{aligned} E :&= \det(M_Q x) \\ &= a^4 + b^4 + c^4 + d^4 + 2(a^2b^2 + a^2c^2 + a^2d^2 + b^2c^2 + b^2d^2 + c^2d^2) \\ &= (a^2 + b^2 + c^2 + d^2)^2 > 0. \end{aligned} \tag{19.97}$$

Hence the quaternion is invertible. Moreover, we have

$$x^{-1} = (M_Q X)^{-1} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} := \frac{1}{E} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix},$$

where

$$\alpha = \det\left(\begin{bmatrix} a & -d & c \\ d & a & -b \\ -c & b & a \end{bmatrix} \right) = a^3 + a(b^2 + c^2 + d^2);$$

$$\beta = -\det\left(\begin{bmatrix} b & -d & c \\ c & a & -b \\ d & b & a \end{bmatrix} \right) = -b^3 - b(a^2 + c^2 + d^2);$$

$$\gamma = \det\left(\begin{bmatrix} b & a & c \\ c & d & -b \\ d & -c & a \end{bmatrix} \right) = -c^3 - c(a^2 + b^2 + d^2);$$

$$\delta = -\det\left(\begin{bmatrix} b & a & -d \\ c & d & a \\ d & -c & b \end{bmatrix} \right) = -d^3 - d(a^2 + b^2 + c^2).$$

It is not a field, because it is not commutative.

Remark 19.3. Note that if $\mathcal{L}$ is commutative, then $x * x^{-1} = e$ implies $x^{-1} * x = e$. In general, Definition 19.6 defines only the right inverse. If the right (left) inverse exists, the commutativity assures the exists of left (right) inverse, and the equivalence of the two inverses.

In the following we give a weak condition for the existence and equivalence of the left and right inverses.

Proposition 19.8. *Assume $\mathcal{L}$ is an associative algebra with right (left) unitary, and its nonzero elements have right (left) inverse. Then the right (left) inverse of each nonzero element is also its left (right) inverse.*

Proof. It is obvious that all the nonzero elements with the algebra product form a group. Then the conclusion comes from the corresponding property of group. □

The above property motivates a natural question: when an algebra is associative? We have the following result.

Proposition 19.9. *An n-dimensional algebra $\mathcal{L}$ is associative, if and only if its structure matrix satisfies the following condition.*

$$M_{\mathcal{L}}^2 = M_{\mathcal{L}}(I_n \otimes M_{\mathcal{L}}). \tag{19.98}$$

Proof.

$$(x * y) * z = M_{\mathcal{L}}(M_{\mathcal{L}}xy)z = M_{\mathcal{L}}^2 xyz,$$

$$x * (y * z) = M_{\mathcal{L}}x(M_{\mathcal{L}}yz) = M_{\mathcal{L}}xM_{\mathcal{L}}yz = M_{\mathcal{L}}(I_n \otimes M_{\mathcal{L}})xyz.$$

The conclusion follows. □

Example 19.9. Consider the quaternion in Example 19.8. It is easy to verify that its structure matrix M_Q satisfies (19.98). Hence the left inverse of the quaternion is also the right inverse.

In the following we consider whether we can find a new algebra over $\mathbb{R}$, which is a field.

First, we consider the case when the dimension $n = 2$. Assume $\{1, \xi\}$ is a basis, which makes

$$\mathcal{F}_2 := \{a + b\xi \mid a, b \in \mathbb{R}\}$$

a field. We need to define a product. Since it needs to satisfy the commutativity requirement and the condition (i), its structure matrix must have the form as

$$M = \begin{bmatrix} 1 & 0 & 0 & \alpha \\ 0 & 1 & 1 & \beta \end{bmatrix}. \tag{19.99}$$

Now we consider the condition (ii). Then for any $x = (a, b)^T$ we have

$$\det(Mx) = a^2 + \beta ab - \alpha b^2.$$

To assure $\det(Mx) > 0$, $\forall\, x \neq 0$, we need

$$\Delta = \beta^2 + 4\alpha < 0. \tag{19.100}$$

We then have the following result.

Theorem 19.9. *A 2-dimensional algebra over $\mathbb{R}$, which is a field, if and only if for a standard basis $(1, \xi)$, its structure matrix has the form of (19.99), where $\alpha, \beta \in \mathbb{R}$ satisfy*

$$|\beta| < 2\sqrt{-\alpha}.$$

Remark 19.4. The following two facts come from equation (19.99):

(1)

$$\xi^2 = \alpha + \beta\xi. \tag{19.101}$$

(2) If only $x = (a, b)^T \neq 0$, its inverse is

$$(a + b\xi)^{-1} = \frac{1}{a^2 + \beta ab - \alpha b^2} \left[(a + \beta b) - b\xi\right]. \tag{19.102}$$

Example 19.10. Define a 2-dimensional algebra $\mathcal{J}$ as

$$\mathcal{J} = \{a + b\mathbf{j} \mid \alpha, \beta \in \mathbb{R}\},$$

with its structure matrix as

$$M = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 1 \end{bmatrix}. \tag{19.103}$$

According to Theorem 19.9, it is a field. Using (19.101), we have

$$\mathbf{j}^2 = -1 + \mathbf{j}. \tag{19.104}$$

Now the product of two numbers in $\mathcal{J}$ can be calculated via (19.104). For instance,

$$(3 + 2\mathbf{j})(2 - \mathbf{j}) = 6 + \mathbf{j} - 2\mathbf{j}^2 = 8 - \mathbf{j}.$$

The division can be performed by using (19.102). For instance, consider $(3 + 2\mathbf{j})/(2 - \mathbf{j})$, using (19.102), we have

$$\frac{1}{2 - \mathbf{j}} = \frac{1}{3}(1 + \mathbf{j}).$$

Hence,

$$\frac{3 + 2\mathbf{j}}{2 - \mathbf{j}} = \frac{1}{3} + \frac{7}{3}\mathbf{j}.$$

In fact, we did not find a new field, because of the following proposition.

Proposition 19.10. *Any two 2-dimensional fields are isomorphic.*

Proof. Let $\mathcal{J} = \{a + b\mathbf{j} \mid a, b \in \mathbb{R}\}$ be a two-dimensional field with its structure matrix as

$$M_{\mathcal{J}} = \begin{bmatrix} 1 & 0 & 0 & \alpha \\ 0 & 1 & 1 & \beta \end{bmatrix}.$$

Define a linear mapping $\Phi : \mathcal{J} \to \mathbb{C}$ as

$$1 \mapsto 1, \quad \mathbf{j} \mapsto \frac{\beta}{2} + \frac{\mathbf{i}}{2}\sqrt{-(4\alpha + \beta^2)}.$$

It is easy to verify that Φ is an isomorphism. Hence, any 2-dimensional field over $\mathbb{R}$ is isomorphic to the field of complexity $\mathbb{C}$. □

Example 19.11. Given a 3-dimensional algebra $\mathcal{L}_3$ over $\mathbb{R}$. Assume it is a field, then for a standard basis $\{e, I, J\}$, where e is the unitary. Assume it is symmetric, then its structure matrix should have the following form.

$$M_{\mathcal{L}_3} = \begin{bmatrix} 1 & 0 & 0 & 0 & a & d & 0 & d & g \\ 0 & 1 & 0 & 1 & b & e & 0 & e & h \\ 0 & 0 & 1 & 0 & c & f & 1 & f & i \end{bmatrix}. \tag{19.105}$$

Now we check when the condition (ii) is satisfied. Assume $w = (x, y, z)^T \in \mathcal{L}_3$, then

$$\det\left(M_{\mathcal{L}_3} w\right) = \det\left(\begin{bmatrix} x & ay + dz & dy + gz \\ y & x + by + ez & ey + hz \\ z & cy + fz & z + fy + iz \end{bmatrix}\right) = x^3 + LDT(x),$$

where $LDT(x)$ express the lower order terms of x. Now it is clear that it cannot be positive definite. We conclude that there is no three-dimensional field.

In fact, Weierstrass proved in 1861 that a finite-dimensional algebra over $\mathbb{R}$, which satisfies associativity and commutativity, can only be either the field of real numbers $\mathbb{R}$ or field of complex numbers $\mathbb{C}$.

In the following, we give an almost invertible 4-dimensional commutative algebra with identity.

Example 19.12. Let $\mathcal{L}$ be a 4-dimensional commutative algebra with identity. Moreover, its structure matrix has the following form

$$M_{\mathcal{L}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}. \tag{19.106}$$

Let $\xi = (x, y, z, w)^T \in \mathcal{L}$. A straightforward computation shows that

$$\det(M_{\mathcal{L}}\xi) = (x^2 - z^2)^2 + (y^2 - w^2)^2 + 2(xy + zw)^2 + 2(xw + yz)^2.$$

Hence, $x + iy + jz + kw$ is invertible, if and only if

$$\begin{bmatrix} x \\ y \end{bmatrix} \neq \pm \begin{bmatrix} z \\ w \end{bmatrix}.$$

$\mathcal{L}$ is invertible except a zero-measure set.

$$\left\{ (x, y, z, w)^T \in \mathbb{R}^4 \,\middle|\, (x, y) = \pm(z, -w) \right\}.$$

Finally, we consider in general how to calculate $\det(M\xi)$ (as required in (19.95)) .

Assume an n-dimensional algebra has its structure matrix as

$$M_n = \left[W_{11} \cdots W_{1n} \quad \cdots \quad W_{n1} \cdots W_{nn}\right], \tag{19.107}$$

where $W_{ij} \in \mathbb{R}^n$ are the columns of M_n arranged in the order of $\mathrm{id}(i,j;n^2)$. We deduce the formula for $\det(M_n\xi)$.

Denote by $k_1, \cdots, k_n$ a set of nonnegative integers satisfying $k_1 + k_2 + \cdots + k_n = n$, then we define $P(k_1, \cdots, k_n)$ as the set of permutations of

$$\left\{ \underbrace{1, \cdots, 1}_{k_1}, \underbrace{2, \cdots, 2}_{k_2}, \cdots, \underbrace{n, \cdots, n}_{k_n} \right\}$$

Using (19.107), we have

$$\det(M_n\xi) = \sum_{k_1+\cdots+k_n=n} \mu(k_1, \cdots, k_n) x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n}, \tag{19.108}$$

where

$$\mu(k_1, \cdots, k_n) = \sum_{(\alpha_1, \cdots, \alpha_n) \in P(k_1, \cdots, k_n)} \det\left(\left[W_{11} \cdots W_{1\alpha_1} \quad \cdots \quad W_{n\alpha_n}\right]\right).$$

For instance, assume $n = 3$, $k_1 = 0$, $k_2 = 2$, and $k_3 = 1$, then

$$P(k_1, k_2, k_3) = Perm\{2, 2, 3\} = \{(2, 2, 3), (2, 3, 2), (3, 2, 2)\}.$$

where $Perm\{S\}$ is the set of all permutations of S.

It follows that the coefficient of monomial $x_2^2 x_3$ is

$$\begin{aligned} \mu(0, 2, 3) &= \det\left(\left[W_{21}\; W_{22}\; W_{33}\right]\right) + \det\left(\left[W_{21}\; W_{32}\; W_{23}\right]\right) \\ &\quad + \det\left(\left[W_{31}\; W_{22}\; W_{2,3}\right]\right). \end{aligned}$$

19.7 Tensor Product Algebra

In Section 19.3, we have defined the tensor product space $V \otimes W$ of vector spaces V and W, as the generation of

$$V \ltimes W := \{z = x \otimes y \mid x \in V \text{ and } y \in W\}.$$

If V and W are algebras with $*_v$ and $*_w$ respectively, it is natural to ask how to define the algebra on $V \otimes W$. First, for the subset $V \ltimes W$, we can define the product of $z_1 = x_1 \otimes y_1$, $z_2 = x_2 \otimes y_2 \in V \ltimes W$ as

$$z_1 * z_2 := (x_1 *_v x_2) \ltimes (y_1 *_w y_2). \tag{19.109}$$

Note that $x \otimes y = x \ltimes y$ when x and y are vectors, then we know that the structure matrix of this product on $V \ltimes W$, denoted by $M_{V\ltimes W}$, satisfies

$$\begin{aligned} M_{V\ltimes W} z_1 z_2 &= (M_V x_1 x_2) \ltimes (M_W y_1 y_2) \\ &= M_V \left(I_{m^2} \otimes M_W\right) x_1 x_2 y_1 y_2 \\ &= M_V \left(I_{m^2} \otimes M_W\right) x_1 W_{[n,m]} y_1 x_2 y_2 \\ &= M_V \left(I_{m^2} \otimes M_W\right) \left(I_m \otimes W_{[n,m]}\right) x_1 y_1 x_2 y_2 \\ &= M_V \left(I_{m^2} \otimes M_W\right) \left(I_m \otimes W_{[n,m]}\right) z_1 z_2, \end{aligned} \tag{19.110}$$

where M_V and M_W are the structure matrices of $*_v$ and $*_w$ respectively. It follows that

$$M_{V\ltimes W} = M_V \left(I_{m^2} \otimes M_W\right) \left(I_m \otimes W_{[n,m]}\right). \tag{19.111}$$

Then, similar to the proof of Lemma 19.4, the product $*_{v\ltimes w} : (V \ltimes W) \times (V \ltimes W) \to (V \ltimes W)$ can be extended linearly to $*_{v\otimes w} : (V \otimes W) \times (V \otimes W) \to (V \otimes W)$ such that the structure matrices $M_{V\ltimes W} = M_{V\otimes W}$. We leave this as an exercise.

In the following we consider some examples.

Dual quaternion is a useful tool in analysis and control of the motion of rigid body (Han *et al.*, 2008; Pennestri and Valentini, 2009). We give its structure matrix as follows.

Example 19.13 (Dual Quaternion). *The structure matrix of quaternion is*

$$M_q = \begin{bmatrix} 1&0&0&0&0&-1&0&0&0&0&-1&0&0&0&0&-1 \\ 0&1&0&0&1&0&0&0&0&0&0&1&0&0&-1&0 \\ 0&0&1&0&0&0&0&-1&1&0&0&0&0&1&0&0 \\ 0&0&0&1&0&0&1&0&0&-1&0&0&1&0&0&0 \end{bmatrix}.$$

Using formula (19.111), we have the structure matrix of $D \times Q$ as

$$M_{d\times q} = M_d \left(I_4 \otimes M_q\right) \left(I_2 \otimes W_{[4,2]}\right) \in \mathcal{M}_{8\times 64}. \tag{19.112}$$

Its non-zero elements are listed as follows.

$m_{1,1} = 1;$ $m_{2,2} = 1;$ $m_{3,3} = 1;$ $m_{4,4} = 1;$ $m_{5,5} = 1;$ $m_{6,6} = 1;$
$m_{7,7} = 1;$ $m_{8,8} = 1;$ $m_{2,9} = 1;$ $m_{1,10} = -1;$ $m_{4,11} = 1;$ $m_{3,12} = -1;$
$m_{6,13} = 1;$ $m_{5,14} = -1;$ $m_{8,15} = 1;$ $m_{7,16} = -1;$ $m_{3,17} = 1;$ $m_{4,18} = -1;$
$m_{1,19} = -1;$ $m_{2,20} = 1;$ $m_{7,21} = 1;$ $m_{8,22} = -1;$ $m_{5,23} = -1;$ $m_{6,24} = 1;$
$m_{4,25} = 1;$ $m_{3,26} = 1;$ $m_{2,27} = -1;$ $m_{1,28} = -1;$ $m_{8,29} = 1;$ $m_{7,30} = 1;$
$m_{6,31} = -1;$ $m_{5,32} = -1;$ $m_{5,33} = 1;$ $m_{6,34} = 1;$ $m_{7,35} = 1;$ $m_{8,36} = 1;$
$m_{6,41} = 1;$ $m_{5,42} = -1;$ $m_{8,43} = 1;$ $m_{7,44} = -1;$ $m_{7,49} = 1;$ $m_{8,50} = -1;$
$m_{5,51} = -1;$ $m_{6,52} = 1;$ $m_{8,57} = 1;$ $m_{7,58} = 1;$ $m_{6,59} = -1;$ $m_{5,60} = -1.$

In next example we consider the product of special linear algebra $sl(2,\mathbb{R})$ with orthogonal algebra $so(3,\mathbb{R})$. (We refer to Varadarajan (1984) for their definitions and the corresponding Lie groups.) In this example the algebras do not have unique default bases. Hence, to see the product is well defined the arguments in last section are essential.

Example 19.14. Consider the product algebra of $sl(2,\mathbb{R})$ and $o(3,\mathbb{R})$. Choose

$$e_1=\begin{bmatrix}0&1\\0&0\end{bmatrix},\quad e_2=\begin{bmatrix}\frac{1}{2}&0\\0&-\frac{1}{2}\end{bmatrix},\quad e_3=\begin{bmatrix}0&0\\1&0\end{bmatrix}$$

as a basis of $sl(2,\mathbb{R})$. Then under this basis it is easy to calculate that the structure matrix of $sl(2,\mathbb{R})$ is

$$M_{sl}=\begin{bmatrix}0&1&0&-1&0&0&0&0&0\\0&0&-2&0&0&0&2&0&0\\0&0&0&0&0&1&0&-1&0\end{bmatrix}.$$

Choose

$$\sigma_1=\begin{bmatrix}0&1&0\\-1&0&0\\0&0&0\end{bmatrix},\quad \sigma_2=\begin{bmatrix}0&0&1\\0&0&0\\-1&0&0\end{bmatrix},\quad \sigma_3=\begin{bmatrix}0&0&0\\0&0&1\\0&-1&0\end{bmatrix}$$

as a basis of $o(3,\mathbb{R})$. Then under this basis the structure matrix of $o(3,\mathbb{R})$ is

$$M_o=\begin{bmatrix}0&0&0&0&0&-1&0&1&0\\0&0&1&0&0&0&-1&0&0\\0&-1&0&1&0&0&0&0&0\end{bmatrix}.$$

Using formula (19.111), we have the structure matrix of $sl(2,\mathbb{R})\times o(3,\mathbb{R})$ as

$$M_{sl\times o}=M_{sl}\left(I_9\otimes M_o\right)\left(I_3\otimes W_{[3,3]}\right)\in\mathcal{M}_{9\times 81}. \tag{19.113}$$

Its non-zero elements are listed as follows.

$$\begin{array}{llllll}
m_{3,5}=-1; & m_{2,6}=1; & m_{6,8}=2; & m_{5,9}=-2; & m_{3,13}=1; & m_{1,15}=-1;\\
m_{6,16}=-2; & m_{4,18}=2; & m_{2,22}=-1; & m_{1,23}=1; & m_{5,25}=2; & m_{4,26}=-2;\\
m_{3,29}=1; & m_{2,30}=-1; & m_{9,35}=-1; & m_{8,36}=1; & m_{3,37}=-1; & m_{1,39}=1;\\
m_{9,43}=1; & m_{7,45}=-1; & m_{2,46}=1; & m_{1,47}=-1; & m_{8,52}=-1; & m_{7,53}=1;\\
m_{6,56}=-2; & m_{5,57}=2; & m_{9,59}=1; & m_{8,60}=-1; & m_{6,64}=2; & m_{4,66}=-2;\\
m_{9,67}=-1; & m_{7,69}=1; & m_{5,73}=-2; & m_{4,74}=2; & m_{8,76}=1; & m_{7,77}=-1.
\end{array}$$

Exercises

19.1 Consider a Riemannian manifold (upper half plane)

$$\mathbb{R}^2_u = \{(x,y) \in \mathbb{R}^2 \mid y > 0\}, \tag{19.114}$$

with Riemannian metric determined by its structure matrix

$$G = \begin{bmatrix} \frac{1}{y^2} & 0 \\ 0 & \frac{1}{y^2} \end{bmatrix}. \tag{19.115}$$

(i) Calculate the Christoffel matrix Γ.

(ii) Let

$$f(x,y) = \begin{bmatrix} x+y \\ x-y \end{bmatrix}, \quad g(x,y) = \begin{bmatrix} y \\ \frac{1}{y} \end{bmatrix}. \tag{19.116}$$

Calculate $\nabla_f g$.

19.2 Consider the Riemannian manifold (19.114)–(19.115) again. A coordinate transformation $\psi : \mathbb{R}^2_u \to \mathbb{R}^2_u$ is defined by

$$\begin{cases} u = x + y \\ v = \ln(y). \end{cases}$$

(i) Calculate the new Christoffel matrix $\tilde{\Gamma}$ under new coordinate frame.

(ii) Let $\tilde{f}$ and $\tilde{g}$ be the same vector fields as in (19.116) but expressed under new coordinate frame. Calculate $\nabla_{\tilde{f}}\tilde{g}$ using $\tilde{\Gamma}$. Compare the result with the previous result.

19.3 On $\mathbb{R}^n$ assume the metric is conventional one. That is, $G = I_n$. Calculate Γ and use (19.13) to show that the geodesic under conventional metric is straight lines.

19.4 On $\mathbb{R}^2$ assume the metric is determined by

$$G = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}.$$

Calculate the geodesic between $(0,0)$ and $(1,1)$.

19.5 Assume the structure matrix of Riemannian curvature $\mathcal{R}(X,Y,Z,W)$ under coordinate x is $M_{\mathcal{R}}$, that is

$$\mathcal{R}(X,Y,Z,W) = M_{\mathcal{R}} XYZW.$$

Let $z = F(x)$ be a coordinate transformation. Find the structure matrix of $\mathcal{R}$ under coordinate z.

19.6 Assume $\alpha, \beta \in V^*(\mathbb{R}^3)$, $X, Y, Z \in V(\mathbb{R}^3)$.

$$\sigma(X,Y,Z;\alpha,\beta) := \alpha(X \bowtie Y)\beta(Z). \tag{19.117}$$

(i) Show that $\sigma \in \mathcal{T}_2^3(\mathbb{R}^3)$.

(ii) Using conventional base, (that is, $\{d_1 = \delta_3^1, d_2 = \delta_3^2, d_3 = \delta_3^3\}$ for $V(\mathbb{R}^3)$ and $\{e^1 = (\delta_3^1)^T, e^2 = (\delta_3^2)^T, e^3 = (\delta_3^3)^T\}$ for $V^*(\mathbb{R}^3)$,) give the structure matrix M_σ of σ.

(iii) Assume $X = (1, 0, -1)^T$, $Y = (2, 1, -3)^T$, $Z = (4, 1, -1)^T$, $\alpha = (2, 1, -1)$, $\beta = (3, -3, 3)^T$, calculate $\sigma(X, Y, Z; \alpha, \beta)$.

19.7 Consider the tensor σ defined in (19.117). Assume another basis of V is chosen as

$$\begin{bmatrix} \tilde{d}_1 \\ \tilde{d}_2 \\ \tilde{d}_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix},$$

and its dual basis

$$\begin{bmatrix} \tilde{e}_1 \\ \tilde{e}_2 \\ \tilde{e}_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 1 & 2 & 1 \end{bmatrix}^{-T} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}.$$

Find the structure matrix of σ as $\tilde{M}_\sigma$ with respect to the new bases.

19.8 Consider the tensor σ defined in (19.117).

(i) Find the structure matrix of $\pi_2^3(\sigma)$.

(ii) Find the structure matrix of $\pi_1^2(\sigma)$.

19.9 Let $\omega = (\omega_x, \omega_y, \omega_z)^T \in \mathbb{R}^3$ be vector of angular velocities of a solid body. Then the matrix

$$\omega^\times := \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix}$$

plays an important role in investigating the rotation of a solid body.

(i) Find a matrix M_ω such that

$$\omega^\times = M_\omega \omega.$$

(ii) Find a matrix M_{ω^2} such that

$$\left[\omega^\times\right]^2 = M_{\omega^2} \omega^2.$$

19.10 A three-dimensional algebra has the following structure matrix. Check whether it is symmetric or skew-symmetric?

(i)

$$M_{\mathcal{L}_1} = \begin{bmatrix} 0 & 1 & 2 & 1 & 0 & 3 & 2 & 3 & 1 \\ 1 & -2 & 1 & -2 & 0 & -3 & 1 & -3 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \end{bmatrix}. \tag{19.118}$$

(ii)

$$M_{\mathcal{L}_2} = \begin{bmatrix} 1 & 4 & 1 & -4 & 2 & 1 & -1 & -1 & 1 \\ 2 & -2 & 3 & 2 & 1 & -7 & -3 & 7 & 3 \\ 1 & -8 & -1 & 8 & 0 & 10 & 1 & -10 & 1 \end{bmatrix}. \tag{19.119}$$

(iii)

$$M_{\mathcal{L}_3} = \begin{bmatrix} 0 & -7 & 1 & 7 & 0 & -2 & -1 & 2 & 0 \\ 0 & 10 & -1 & -10 & 0 & 3 & 1 & -3 & 0 \\ 0 & -11 & 1 & 11 & 0 & -3 & -2 & 3 & 0 \end{bmatrix}. \tag{19.120}$$

19.11 Check if the algebra in (19.118) (or (19.119) or (19.120)) is associative.

19.12 Check if the algebra in (19.118) (or (19.119) or (19.120)) is a Lie algebra.

19.13 Construct structure matrix of $gl(2,\mathbb{R})$ with respect to the following bases.

(i)

$$e_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad e_3 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad e_4 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

(ii)

$$e_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad e_3 = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}, \quad e_4 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

19.14 Prove the two algebras $\mathcal{L}_1$ and $\mathcal{L}_2$ with the following structure matrices are equivalent.

$$M_{\mathcal{L}_1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 3 \end{bmatrix}; \quad M_{\mathcal{L}_2} = \begin{bmatrix} 1 & 0 & 0 & -2 & 0 & 0 & 2 & 1 & 1 \\ 0 & 1 & 0 & -1 & -1 & 1 & 1 & 2 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & 1 & 2 \end{bmatrix}.$$

19.15 Assume an algebra is expressed as

$$A_1 = \{r_0 + r_1\alpha + r_2\beta | r_0, r_1, r_2 \in R\}$$

with the product rules as

$$\alpha^2 = \beta, \quad \alpha\beta = \beta\alpha = \alpha, \quad \beta^2 = \beta.$$

Convert it to a canonical form in Theorem 19.5.

19.16 Similar to cross product on $\mathbb{R}^3$, we intend to define a cross product on $\mathbb{R}^4$, satisfying

(1) $$\delta_4^1 \times_c \delta_4^2 = \delta_4^3; \tag{19.121}$$

(2)
$$\delta_4^2 \times_c \delta_4^3 = \delta_4^4; \tag{19.122}$$

(3)
$$\delta_4^3 \times_c \delta_4^4 = \delta_4^1; \tag{19.123}$$

(4)
$$\delta_4^4 \times_c \delta_4^1 = \delta_4^2. \tag{19.124}$$

(i) Is there a skew-symmetric algebra, which satisfies (19.121)–(19.124)? If so give an example. If not explain why?

(ii) Is there a Lie algebra, which satisfies (19.121)–(19.124)? If so give an example. If not explain why?

19.17 Prove Lemma 19.7.

19.18 For two algebras $(V, *_v)$ and $(W, *_w)$, prove that the product $*_{v \ltimes w}$: $(V \ltimes W) \times (V \ltimes W) \to (V \ltimes W)$ can be extended linearly to $*_{v \otimes w} : (V \otimes W) \times (V \otimes W) \to (V \otimes W)$ such that the structure matrices $M_{V \ltimes W} = M_{V \otimes W}$. (Hint: Please refer to Lemma 19.4)

19.19 Find the structure matrix for the following tensor product algebras.

(i) dual numbers and cross product in $\mathbb{R}^3$.

(ii) dual numbers $\mathbb{D}$ with $sl(2, \mathbb{R})$;

(iii) hyperbolic numbers $\mathbb{H}$ with $gl(2, \mathbb{R})$.

Chapter 20

Morgan's Problem

Consider a control system, which has multiple inputs and multiple outputs. Morgan's problem is also called the input-output decoupling problem. It plays an important role in control design. When the number of inputs equals the number of outputs, the problem has been solved perfectly (Falb and Wolovich, 1967; Wonham and Morse, 1970). But as the number of inputs is greater than the number of outputs, it becomes a long standing open problem. The problem has been claimed to be solved several times, but the conclusions have then been proved wrong by counter-examples. In this chapter we consider only the static feedback control. Some later developments can be seen in Glumineau and Moog (1992); Di Benedetto *et al.* (1994). As for the dynamic feedback case, a necessary and sufficient condition has been provided in Descusse and Morg (1985).

Using semi-tensor product, this chapter provides a numerical solution to the problem.

20.1 Input-Output Decomposition

Consider a linear control system

$$\begin{cases} \dot{x} = Ax + \sum\limits_{i=1}^{m} b_i u_i := Ax + Bu, \quad x \in \mathbb{R}^n,\ u \in \mathbb{R}^m, \\ y = Cx, \quad y \in \mathbb{R}^p, \end{cases} \tag{20.1}$$

where u_i, $i = 1, \cdots, m$ are controls and y_j, $j = 1, \cdots, p$ are outputs, and $m \geq p$. For statement ease, we simply assume $\text{rank}(B) = m$, $\text{rank}(C) = p$. The input-output decomposition problem, which is also called the Morgan's problem, means finding, if possible, a partition of inputs, such that each block of inputs control the corresponding output without affecting other outputs. We give a precise definition.

Definition 20.1. Consider the system (20.1). The input-output decomposition problem (Morgan's problem) is solvable, if there exists a state feedback

$$u = kx + Hv, \quad v \in \mathbb{R}^p, \tag{20.2}$$

and a partition of v, say $v^1, \cdots, v^p$, such that v^i controls y_i and v^i does not affect y_j, $j \neq i$. Equivalently, the transfer matrix from v to y has full row rank (Wonham and Morse, 1970).

For statement ease, we introduce some concepts.

Let $y_j = c_j x$, its relative degree, denoted by ρ_j, is defined as

$$\rho_j = \min\{i \,|\, c_j A^{i-1} B \neq 0\}, \quad j = 1, \cdots, p. \tag{20.3}$$

We assume $\rho_j < \infty$, $j = 1, \cdots, p$, exist. Otherwise, y_j is not affected by any input.

Using relative degree vector $\rho = (\rho_1, \cdots, \rho_p)$, we define a $p \times m$ matrix, called the decoupling matrix, as

$$D = \begin{bmatrix} c_1 A^{\rho_1 - 1} B \\ c_2 A^{\rho_2 - 1} B \\ \vdots \\ c_p A^{\rho_p - 1} B \end{bmatrix}. \tag{20.4}$$

When $m = p$, we have the following classical result as:

Theorem 20.1 (Falb and Wolovich, 1967). *When $m = p$, the Morgan's problem is solvable, if and only if the decoupling matrix D is nonsingular.*

Corresponding to linear case, the Morgan's problem for nonlinear control systems has also been discussed widely. Consider a nonlinear control system

$$\begin{cases} \dot{x} = f(x) + \sum\limits_{i=1}^{m} g_i(x) u_i := f(x) + G(x) u, \quad x \in \mathbb{R}^n, \ u \in \mathbb{R}^m, \\ y_j = h_j(x), \quad j = 1, \cdots, p, \end{cases} \tag{20.5}$$

where, $f(x)$, $g_i(x)$ $i = 1, \cdots, m$ are smooth vector fields, $h_j(x)$, $j = 1, \cdots, p$ are smooth functions. For system (20.5), the relative degree vector $\rho =$

$(\rho_1, \cdots, \rho_p)^T$ is defined as

$$L_{g_i} L_f^k h_j(x) = 0, \quad x \in U,$$

$$i = 1, \cdots, m; \; k = 0, 1, \cdots, \rho_j - 2.$$

$$L_{g_i} L_f^{\rho_j - 1} h_j(x_0) \neq 0, \quad \exists\, i \in \{1, 2, \cdots, m\},$$

where U is a neighborhood of x_0, which is the concerned point. For simplicity, set $x_0 = 0$. Assume the relative degree vector is well defined, we define the decoupling matrix as

$$D(x) = \begin{bmatrix} L_{g_1} L_f^{\rho_1 - 1} h_1(x) & \cdots & L_{g_m} L_f^{\rho_1 - 1} h_1(x) \\ L_{g_1} L_f^{\rho_2 - 1} h_2(x) & \cdots & L_{g_m} L_f^{\rho_2 - 1} h_2(x) \\ \vdots & & \vdots \\ L_{g_1} L_f^{\rho_p - 1} h_p(x) & \cdots & L_{g_m} L_f^{\rho_p - 1} h_p(x) \end{bmatrix}. \tag{20.6}$$

As a generalization of linear case, we have the following result:

Theorem 20.2 (Freund, 1975). *When $m = p$, the Morgan's problem is locally solvable at x_0, if and only if the decoupling matrix is nonsingular at x_0.*

Denote

$$\xi(x) = (L_f^{\rho_1} h_1(x), L_f^{\rho_2} h_2(x), \cdots, L_f^{\rho_p} h_p(x))^T,$$

which is called the decoupling vector.

Then it is easy to check that when $n = p$ the following control solves the local input-output decoupling problem. (Refer to, e.g., Qiao *et al.* (2011) for detail.)

$$u = -D^{-1}(x)\xi(x) + D^{-1}(x)v. \tag{20.7}$$

Precisely, for the closed-loop system obtained by control (20.7), each control v_i controls y_i and does not affect y_j $(j \neq i)$.

Note that when a linear control system is considered, the decoupling vector becomes

$$\xi(x) := \xi x = \begin{bmatrix} c_1 A^{\rho_1} \\ \vdots \\ c_p A^{\rho_p} \end{bmatrix} x.$$

Then the corresponding control becomes

$$u = -D^{-1}\xi x + D^{-1}v. \tag{20.8}$$

When the number of inputs equals the number of outputs and D is nonsingular, the control (20.8) solves the input-output decoupling problem globally.

In the rest of this chapter, we consider only the linear case.

20.2 Problem Formulation

From previous section it is clear that the challenging case for Morgan's problem is when $m > p$. According to Theorem 20.1, we have the following lemma.

Lemma 20.1. *Morgan's problem is solvable, if and only if there exist* $K \in \mathcal{M}_{m\times n}$, $H \in \mathcal{M}_{m\times p}$, $1 \le \rho_i \le n$, $i = 1, \cdots, p$ *such that*

$$c_i(A + BK)^{t_i} BH = 0, \quad t_i = 0, \cdots, \rho_i - 2, \quad i = 1, \cdots, p. \tag{20.9}$$

Moreover, the decoupling matrix for the closed-loop system,

$$D = \begin{bmatrix} c_1(A + BK)^{\rho_1 - 1} BH \\ \vdots \\ c_p(A + BK)^{\rho_p - 1} BH \end{bmatrix} \tag{20.10}$$

is nonsingular.

According to the above lemma, there are two designable feedback matrices K and H. The purpose of this section is to give an equivalent condition, which reduce the unknown matrices to one.

First, note that when $\rho_i = 1$ (20.9) disappears. We then denote

$$\Lambda = \{i \,|\, \rho_i = 2\}, \quad C_\Lambda = col\{c_i \,|\, i \in \Lambda\}.$$

By definition of relative degree, when $1 \le \rho_i \le 2$, $i = 1, \cdots, p$, (20.9) becomes

$$H \subset (C_\Lambda B)^\perp.$$

Denote

$$\Lambda^c = \{1, \cdots, p\} \backslash \Lambda.$$

Then $i \in \Lambda^c$ implies that $\rho_i = 1$.

Hence we have

Corollary 20.1. *For* $1 \le \rho_i \le 2$, $i = 1, \cdots, p$, *Morgan's problem is solvable, if and only if there exists* $K \in \mathcal{M}_{m\times n}$, *such that*

$$D = \begin{bmatrix} C_{\Lambda^c} B \\ c_\Lambda (A + BK) B \end{bmatrix} (C_\Lambda B)^\perp \tag{20.11}$$

has full row rank.

In general case, it is not so easy to eliminate H. Further works are necessary.

Define

$$W(K) := \begin{bmatrix} c_1B \\ c_1(A+BK)B \\ \vdots \\ c_1(A+BK)^{\rho_1-2}B \\ \vdots \\ c_p(A+BK)^{\rho_p-2}B \end{bmatrix}, \quad T(K) := \begin{bmatrix} c_1(A+BK)^{\rho_1-1}B \\ c_2(A+BK)^{\rho_2-1}B \\ \vdots \\ c_p(A+BK)^{\rho_p-1}B \end{bmatrix}.$$

Then (20.9) becomes

$$W(K)H = 0, \tag{20.12}$$

and (20.10) becomes

$$D = T(K)H. \tag{20.13}$$

Since $1 \leq \rho_i \leq n$, $i = 1, \cdots, p$, for fixed ρ_i, $i = 1, \cdots, p$ we may consider the solvability of Morgan's problem, Because we need only to check finite (precisely, n^p) cases. In the remaining part of this chapter we consider the solvability of Morgan's problem under a set of fixed ρ_i, unless elsewhere stated.

Lemma 20.2. *Morgan's problem is solvable, if and only if there exists $K \in \mathcal{M}_{m\times n}$ such that*

(1)

$$\mathcal{I}m(T^T(K)) \cap \mathcal{I}m(W^T(K)) = \{0\}; \tag{20.14}$$

(2) $T(K)$ has full row rank.

Proof. We prove the following statements are equivalent:

(i) there exists H such that $T(K)H$ is nonsingular and $W(K)H = 0$;
(ii) $T(K)((W(K)^T)^{\perp}) = \mathbb{R}^p$;
(iii) $[(T^T(K))^{-1}(W(K)^T)]^{\perp} = \mathbb{R}^p$;
(iv) $(T^T(K))^{-1}(W(K)^T) = 0$;
(v) Conditions (1) and (2) in Lemma 20.2.

Where $T(K)$ and $(T^T(K))^{-1}$ are considered as functional mappings (Wonham, 1979).

(i)⇒(ii): If $\dim(T(k)(W(K)^T)^{\perp}) < p$, since $\mathcal{I}m(H) \subset (W(K)^T)^{\perp}$, then $\text{rank}(T(K)H) < p$, which leads to a contradiction;

(ii)⇒(i): Choosing p vectors $h_i \in (W(K)^T)^\perp$, such that

$$T(K)\mathcal{I}m(h_1, \cdots, h_p) = \mathbb{R}^p.$$

Then we can set $H = (h_1, \cdots, h_p)$;

(ii)⇔(iii): Refer to Wonham (1979) page 23;

(iii)⇔(iv): It is obvious;

(iv)⇔(v): It is easy to verify that both (iv) and (v) are equivalent to the following statement: If $Y \in \mathbb{R}^p$ and $T^T(K)Y \in \mathcal{I}m((W(K))^T)$, then $Y = 0$. □

From the above lemma we can prove the following theorem easily.

Theorem 20.3. *For fixed $\rho_i' s$ the Morgan's problem is solvable, if and only if there exists $K_0 \in \mathcal{M}_{m\times n}$, such that*

$$\operatorname{rank}\left(\begin{bmatrix} T(K_0) \\ W(K_0) \end{bmatrix}\right) = p + \operatorname{rank}(W(K_0)). \tag{20.15}$$

Consider the case when $\rho_i \leq 2$, $\forall\, i$, if the following assumption holds:

A1. $C_\Lambda B = 0$, then we need not consider $W(K_0)$. Hence we have

Corollary 20.2. *Assume $1 \leq \rho_i \leq 2$, $i = 1, \cdots, p$, and A1 holds. Moreover, if there exists $K_0 \in \mathcal{M}_{m\times n}$ such that $T(K_0)$ has full row rank, then Morgan's problem is solvable.*

Definition 20.2. Assume $A(K)$ is a matrix with its entries $a_{ij}(K)$ as the polynomial of K, where $K \in \mathcal{M}_{m\times n}$. Define the essential rank of $A(K)$, denoted by $\operatorname{rank}_e(A(K))$, as

$$\operatorname{rank}_e(A(K)) = \max_{K \in \mathcal{M}_{m\times n}} \operatorname{rank}(A(K)).$$

Now under the fixed $\{\rho_i\}$ denote

$$\operatorname{rank}_e(T(K)) = t, \quad \operatorname{rank}_e(W(K)) = s, \quad \operatorname{rank}_e\left(\begin{bmatrix} T(K) \\ W(K) \end{bmatrix}\right) = q.$$

Because the essential rank is easily computable, the following corollary is convenient in certain cases.

Corollary 20.3. *Morgan's problem is solvable, if $q = p + s$.*

Since both $T(K)$ and $W(K)$ are polynomial matrices of K, the essential rank can be reached on all $K \in \mathcal{M}_{m\times n} \sim \mathbb{R}^{mn}$ except a zero-measure set of K, it is easy to calculate the essential rank via computer. Hence, the condition of Corollary 20.3 is easily verifiable.

20.3 Numerical Expression of Solvability

We first calculate $T(K)$ and $W(K)$. Denote by $Z = V_r(K) \in \mathbb{R}^{mn}$, We first express $T(K)$ and $W(K)$ as polynomials of Z via STP.

Lemma 20.3. *Given a matrix $A \in \mathcal{M}_{n\times m}$.*

(1) If $x \in \mathbb{R}^n$ is a row vector, then

$$xA = V_r^T(A) \ltimes x^T. \tag{20.16}$$

(2) If $Y \in \mathcal{M}_{p\times n}$, then

$$YA = (I_p \otimes V_r^T(A)) \ltimes V_r(Y). \tag{20.17}$$

Proof. A straightforward computation shows that

$$\begin{aligned} xA &= V_r^T(A)x^T \\ &= \left(\sum_{i=1}^{n} a_{i1}x_i, \cdots, \sum_{i=1}^{n} a_{im}x_i\right). \end{aligned}$$

Using (20.16), we have

$$YA := \begin{bmatrix} Y^1 \\ \vdots \\ Y^p \end{bmatrix} A = \begin{bmatrix} V_r^T(A)(Y^1)^T \\ \vdots \\ V_r^T(A)(Y^p)^T \end{bmatrix} = (I_p \otimes V_r^T(A))V_r(Y).$$

□

Next, we expand $(A+BK)^t$ as follows:

$$(A+BK)^t = \sum_{i=0}^{2^t-1} P_i(A, BK),$$

where $P_i(x,y)$ is used to present an ith homogeneous monomial of x and y, which are defined as follows: Convert i into a binary number of length t, (add zero at ahead if necessary). Then use "x" to replace "0" and use "y" to replace "1". For instance, since

$$0 = \underbrace{0\,0\,\cdots\,0}_{i}, \quad 1 = \underbrace{0\,0\,\cdots\,0}_{i-1}\,1, \quad 2 = \underbrace{0\,0\,\cdots\,0}_{i-2}\,1\,0, \quad \cdots,$$

we have

$$\begin{aligned} &P_0(x,y) = \underbrace{x\,x\,\cdots\,x}_{i}, \qquad P_1(x,y) = \underbrace{x\,x\,\cdots\,x}_{i-1}\,y, \\ &P_2(x,y) = \underbrace{x\,x\,\cdots\,x}_{i-2}\,y\,x, \cdots. \end{aligned}$$

Then we collect terms with respect to different powers of "K". Finally, we can have the following expression.

$$c_k(A+BK)^tB = \sum_{i=0}^{t}\sum_{j=1}^{T_i} S_0^{ij}KS_1^{ij}K\cdots S_{t-1}^{ij}KS_t^{ij}, \quad k=1,\cdots,p, \tag{20.18}$$

where

$$T_i = \binom{t}{i} = \frac{t!}{i!(t-i)!}.$$

Using Lemma 20.3 and equation (20.16), (20.18) can be expressed as

$$\begin{aligned} & c_k(A+BK)^tB \\ & = \sum_{i=0}^{t}\sum_{j=1}^{T_i} S_0^{ij} \ltimes (I_m \otimes V_r^T(S_1^{ij})) \ltimes Z \ltimes \cdots \ltimes (V_r^T(I_m \otimes S_t^{ij})) \ltimes Z \\ & = \sum_{i=0}^{t}\sum_{j=1}^{T_i} S_0^{ij} \ltimes (I_m \otimes V_r^T(S_1^{ij})) \ltimes (I_{m^2n} \otimes V_r^T(S_2^{ij})) \\ & \qquad \ltimes \cdots \ltimes (I_{m^tn^{t-1}} \otimes V_r^T(S_t^{ij})) \ltimes Z^t, \quad k=1,\cdots,p, \end{aligned} \tag{20.19}$$

where $Z = V_r(K)$.

Using (20.19), $W(K)$ and $T(K)$ can be expressed in canonical form as:

$$\begin{aligned} & W(K) = W_0 + W_1 \ltimes Z + \cdots + W_{l-1} \ltimes Z^{l-1} \in \mathcal{M}_{d\times m}, \quad d = \sum_{i=1}^{p}\rho_i - p, \\ & T(K) = T_0 + T_1 \ltimes Z + \cdots + T_l \ltimes Z^l \in \mathcal{M}_{p\times m}, \end{aligned} \tag{20.20}$$

where $l = \max\{\rho_i - 1 \mid i = 1,\cdots,p\}$.

Denote by $W^s = \mathrm{Row}_s(W(K))$, then the size of W^s is

$$|W^s| = \frac{d!}{s!(d-s)!}.$$

Now the Morgan's problem can be converted into an algebraic form as follows.

Proposition 20.1. *Morgan's problem is solvable, if and only if there exists an integer s with $1 \le s \le m-p+1$ such that*

$$R(Z) := \sum_{L\in W^s} \det\left(L(Z)L^T(Z)\right) = 0 \tag{20.21}$$

and

$$J(Z) := \sum_{L\in W^{s-1}} \det\left(\begin{bmatrix} T(Z) \\ L(Z) \end{bmatrix} (T^T(Z)\ L^T(Z))\right) > 0 \tag{20.22}$$

has solution Z.

Proof. If the system (20.21)−(20.22) has solution $Z = V_r(K)$, then from (20.21) we have $\mathrm{rank}(W(K)) < s$, and from (20.22) we have

$$\mathrm{rank}\begin{pmatrix} T(K) \\ W(K) \end{pmatrix} = p + s - 1.$$

According to Theorem 20.3, the conclusion follows. □

Now the Morgan's problem becomes a numerical problem: For each set of fixed $1 \le \rho_i \le n,\ i = 1, \cdots, p$ and $1 \le s \le m - p + 1$, solve system (20.21)−(20.22). Since there are only finite possible cases, the Morgan's problem is solvable as long as there is a case (a pair of $(\rho_i,\ s)$), under which system (20.21)−(20.22) has solution.

There are many numerical methods, which are applicable to solving this numerical problem.

For instance, we may convert it to the so called Wu's problem (Wu, 1995): Is the equation $R(Z) = 0$ implies $J(Z) = 0$? If for all cases the answer is "yes", then the Morgan's problem is not solvable. Otherwise, if at least there is one case, where the answer is "no", then the Morgan's problem is solvable.

An alternative approach is to convert it into an optimization problem:

$$\max_{R(Z)=0} J(Z).$$

If the maximum value is zero, then the Morgan's problem is not solvable. Otherwise, it is solvable.

Note that if every element of $A(Z)$ can be expressed as a polynomial of the form $a_0 + a_1 \ltimes Z + \cdots + a_L \ltimes Z^L$, the determinant $\det(A(Z))$ can be calculated directly. Hence, to get (20.21) and (20.22) we need to calculate the following product: Let

$$A = A_0 + A_1 \ltimes Z + \cdots + A_s \ltimes Z^s \in \mathcal{M}_{m\times n},$$

$$B = B_0 + B_1 \ltimes Z + \cdots + B_t \ltimes Z^t \in \mathcal{M}_{p\times n}.$$

Then

$$AB^T = \sum_{i=0}^{s}\sum_{j=0}^{t} A_i \ltimes Z^i \ltimes (Z^T)^j \ltimes B_j^T.$$

Using (20.16)–(20.17), we have

$$(Z^T)^j = (Z^T)^j I_{n^j} = V_r^T(I_{n^j}) \ltimes Z^j;$$

$$Z^i \ltimes V_r^T(I_{n^j}) = (I_{n^i} \ltimes V_r^T(I_{n^j})) \ltimes Z^i;$$

and

$$Z^{i+j} \ltimes B_j^T = (I_{n^{i+j}} \ltimes B_j^T) \ltimes Z^{i+j}.$$

Using them, we have

$$AB^T = \sum_{i=0}^{s} \sum_{j=0}^{t} A_i \ltimes \left(I_{n^i} \ltimes V_r^T(I_{n^j})\right) \ltimes \left(I_{n^{i+j}} \ltimes B_j^T\right) \ltimes Z^{i+j}. \quad (20.23)$$

An immediate consequence of the above argument is: when $1 \le \rho_i \le 2$, $i = 1, \cdots, p$, we have the following result.

Corollary 20.4. *Assume* $1 \le \rho_i \le 2$, $i = 1, \cdots, p$, *and* $A1$ *holds. Then Morgan's problem is solvable, if and only if*

$$J(Z) := \det\left(T(Z)T^T(Z)\right) > 0 \quad (20.24)$$

has solution Z.

In this case, the Morgan's problem converts to a free (i.e., without restriction) optimization problem:

$$\max J(Z).$$

If the maximum value is zero, the Morgan's problem is not solvable. Otherwise, it is solvable.

Summarizing the above argument, we present the numerical algorithm for solving Morgan's problem.

Algorithm 5.

Step 1. For $\rho_1, \cdots, \rho_p = 1, \cdots, n$, using (20.18)–(20.19) to express $T(K)$ and $W(K)$ into the standard polynomial form as (20.20).

Step 2. For $s = 1, \cdots, m-p+1$ and each $L \in W^s$, using (20.23) to calculate

$$L(Z)L^T(Z), \quad \begin{bmatrix} T(Z) \\ L(Z) \end{bmatrix} \left[T^T(Z)\ L^T(Z)\right].$$

Step 3. Using (20.21) to calculate $R(Z)$, and using (20.22) to calculate $J(Z)$ respectively.

Step 4. Solving the numerical problem

$$\begin{cases} R(Z) = 0, \\ J(Z) > 0, \end{cases} \quad (20.25)$$

where $R(Z)$ and $J(Z)$ are polynomials obtained from Step 3.

Finally, we give a numerical example to depict it.

Example 20.1. Consider a linear system

$$\begin{cases} \dot{x} = \begin{bmatrix} 0 & 6 & -2 & -4 & 0 \\ -1 & 0 & 0 & 0 & 1 \\ 0 & -2 & 1 & 1 & -1 \\ -1 & 1 & -1 & -1 & 1 \\ 0 & 2 & 0 & 0 & 1 \end{bmatrix} x + \begin{bmatrix} -2 & 1 & 0 \\ 0 & 0 & 0 \\ 2 & 0 & -2 \\ -1 & 0 & 1 \\ -1 & 1 & 1 \end{bmatrix} u, \\ y = \begin{bmatrix} 0 & 1 & -1 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 \end{bmatrix} x. \end{cases} \tag{20.26}$$

Note that $\rho_1 + \rho_2$ can not be great than the dimension 5, we have to verify the following possible cases: $\rho_1 = 1, \rho_2 = 1, 2, 3, 4; \rho_1 = 2, \rho_2 = 1, 2, 3; \rho_1 = 3, \rho_2 = 1, 2; \rho_1 = 4, \rho_2 = 1$. As an example we verify the case when $\rho_1 = 3, \rho_2 = 2$. In this case we have

$$W(K) = \begin{bmatrix} c_1 B \\ c_1(A + BK)B \\ c_2 B \end{bmatrix} = \begin{bmatrix} -1 & 0 & 1 \\ p_1(Z) & p_2(Z) & p_3(Z) \\ 1 & 0 & -1 \end{bmatrix},$$

where

$$Z = V_r(K) = (k_{11}, \ldots, k_{15}, \ \ldots \ , k_{31}, \ldots, k_{35})^T.$$

Using (4.33), we have

$$\begin{bmatrix} p_1(Z) \\ p_2(Z) \\ p_3(Z) \end{bmatrix} = V_r(c_1 AB + c_1 BKB) = \begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix} + ((c_1 B) \otimes B^T)Z.$$

Then we can calculate that

$$\begin{aligned} p_1(Z) &= -1 + 2k_{11} - 2k_{13} + k_{14} + k_{15} - 2k_{31} + 2k_{33} - k_{34} - k_{35}, \\ p_2(Z) &= 1 - k_{11} - k_{15} + k_{31} + k_{35}, \\ p_3(Z) &= 1 - k_{11} + 2k_{13} - k_{14} - 2k_{15} + k_{31} - 2k_{33} + k_{34} + 2k_{35}. \end{aligned}$$

Moreover,

$$\begin{aligned} T(K) &= \begin{bmatrix} c_1(A + BK)^2 B \\ c_2(A + BK)B \end{bmatrix} \\ &= \begin{bmatrix} c_1 A^2 B \\ c_2 AB \end{bmatrix} + \begin{bmatrix} c_1 ABKB + c_1 BKAB \\ c_2 BKB \end{bmatrix} + \begin{bmatrix} c_1 BKBKB \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & -1 \end{bmatrix} + T_1 Z + T_2 Z^2, \end{aligned}$$

where

$$T_1 = \left[\begin{matrix} 2 & -1 & 0 & -1 & 0 & -1 & -4 & 1 & 4 & 1 & 0 & -3 & 2 & -2 & -2 \\ -2 & 1 & 0 & 0 & 0 & 0 & 2 & 0 & -2 & -1 & 0 & 1 & -1 & 1 & 1 \end{matrix}\right.$$

$$\begin{matrix} -2 & 1 & 0 & 0 & 0 & 0 & 2 & 0 & -2 & -1 & 0 & 1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

$$\left.\begin{matrix} -2 & 1 & 0 & 1 & 0 & 1 & 4 & -1 & -4 & -1 & 0 & 3 & -2 & 2 & 2 \\ 2 & -1 & 0 & 0 & 0 & 0 & -2 & 0 & 2 & 1 & 0 & -1 & 1 & -1 & -1 \end{matrix}\right],$$

and T_2 is a 2×675 matrix, which is skipped here. (The reader can calculate it via computer easily.)

According to Proposition 20.1, we need to check it for two cases: $s = 1$ and $s = 2$. It is clear that when $s = 1$, $R(Z) > 0$. For $s = 2$, from (20.13) one sees that to make D nonsingular, $\text{rank}(H) \geq 2$. Then, from (20.12), $\text{rank}\, W(K) \leq 1$. Hence, we can assume

$$p_1(Z) = 0, \quad p_2(Z) = 0, \quad p_3(Z) = 0. \tag{20.27}$$

It is easy to find a set of solutions as

$$K = V_r^{-1}(Z) = \begin{bmatrix} 0 & 2 & -1 & -2 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 2 & -1 & -2 & -1 \end{bmatrix},$$

and

$$W(K) = \begin{bmatrix} -1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & -1 \end{bmatrix}, \qquad T(K) = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

It is obvious that $R(Z) = 0$. Moreover,

$$\begin{aligned} J(Z) &= \det\left(\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}^T \right) \\ &\quad + \det\left(\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}^T \right) \\ &\quad + \det\left(\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}^T \right) \\ &= 4 > 0. \end{aligned}$$

According to Proposition 20.1, we conclude that the Morgan's problem for system (20.26) is solvable.

In the following we look for the required feedback matrix H. Since $\mathrm{Span}\{\mathrm{Col}(H)\} \subset \ker(W(K))$, we can choose

$$H = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Using feedback control $u = Kx + Hv$, the closed-loop system becomes

$$\begin{cases} \dot{x} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 \\ 0 & -2 & 1 & 1 & 1 \\ -1 & 1 & -1 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} x + \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 0 & -2 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} v, \\ y = \begin{bmatrix} 0 & 1 & -1 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 \end{bmatrix} x. \end{cases} \tag{20.28}$$

It is ready to verify that

$$WH = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad D = TH = \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix}.$$

We then have $\rho_1 = 3$ and $\rho_2 = 2$. Moreover, the decoupling matrix D is nonsingular.

Remark 20.1.

(1) From the argument of the above example one sees that all the K, satisfying (20.27) are the solutions of the Morgan's problem.

(2) Using the K and H obtained in the above example, we can verify that the equality (20.9) in Lemma 20.1 holds. In addition, (20.4) is nonsingular. We can also check that the equality (20.14) in Lemma 20.2 holds, and $T(K)$ has full row rank.

(3) For Theorem 20.3, it is easy to check that

$$\mathrm{rank}\left(\begin{bmatrix} T(K_0) \\ W(K_0) \end{bmatrix}\right) = 3, \quad \mathrm{rank}(W(K_0)) = 1, \quad p = 2,$$

hence, (20.15) holds.

Exercises

20.1 Consider a linear system

$$\begin{cases} \dot{x}_1 = u \\ \dot{x}_2 = 2\sin(x_3) - u \\ \dot{x}_3 = \ln(1 + \frac{x_1 - x_2}{2}). \end{cases}$$

Find the relative degree with respect to the following different outputs: (i) $y = \frac{x_1 - x_2}{2}$; (ii) $y = x_3$; (iii) $y = \frac{x_1 + x_2}{2}$. Compare them to see that the relative degree depends on the outputs.

20.2 Consider a linear system

$$\begin{cases} \dot{x}_1 = x_1 + \frac{1}{2}u_2 \\ \dot{x}_2 = \frac{1}{2}(x_3 - x_1 - x_2) - \frac{1}{2}u_2 \\ \dot{x}_3 = x_1 - x_2 + u_1 \\ y_1 = x_1 + x_2 \\ y_2 = x_1 - x_2. \end{cases} \tag{20.29}$$

(i) Find the relative degree vector.

(ii) Construct the decoupling matrix and the decoupling vector.

(iii) Is the decoupling matrix non-singular? If "yes" solve the input-output decoupling problem.

20.3 Consider the attitude control of a missile. The dynamic equation of the attitude of missiles is (Qiao *et al.*, 2011)

$$\begin{cases} J_x \frac{d\varnothing_x}{dt} + (J_z - J_y)\varnothing_y \varnothing_z = u_1 \\ J_y \frac{d\varnothing_y}{dt} + (J_x - J_z)\varnothing_x \varnothing_z = u_2 \\ J_z \frac{d\varnothing_z}{dt} + (J_y - J_x)\varnothing_x \varnothing_y = u_3 \\ \frac{d\gamma}{dt} = \varnothing_x - (\varnothing_y \cos\gamma - \varnothing_z \sin\gamma)\tan\theta \\ \frac{d\phi}{dt} = \frac{1}{\cos\theta}(\varnothing_y \cos\gamma - \varnothing_z \sin\gamma) \\ \frac{d\theta}{dt} = \varnothing_y \sin\gamma + \varnothing_z \cos\gamma \\ y_1 = \gamma(t) \\ y_2 = \phi(t) \\ y_3 = \theta(t), \end{cases} \tag{20.30}$$

where J_x, J_y, J_z are the inertias with respect to x, y, and z axes of the missile (for simplicity, they are assumed to be constants), $\varnothing_x$, $\varnothing_y$, and $\varnothing_z$ are the angular velocities with respect to the axes respectively; γ, ϕ, and θ are attitude angles.

Denoting $x = (x_1, x_2, x_3, x_4, x_5, x_6)^T$, where

$$x_1 = \o_x,\ x_2 = \o_y,\ x_3 = \o_z,\ x_4 = \gamma,\ x_5 = \phi,\ x_6 = \theta,$$

then the system (20.30) can be converted into a canonical form as

$$\begin{cases} \dot{x} = f(x) + g_1(x)u_1 + g_2(x)u_2 + g_3(x)u_3 \\ y_1 = x_4 \\ y_2 = x_5 \\ y_3 = x_6. \end{cases} \tag{20.31}$$

(i) Prove that the decoupling matrix of (20.30) is

$$D(x) = \begin{bmatrix} \frac{1}{J_x} & -\frac{1}{J_y}\cos(x_4)\tan(x_6) & \frac{1}{J_z}\sin(x_4)\tan(x_6) \\ 0 & \frac{1}{J_y}\cos(x_4)\sec(x_6) & -\frac{1}{J_z}\sin(x_4)\sec(x_6) \\ 0 & \frac{1}{J_y}\sin(x_4) & \frac{1}{J_z}\cos(x_4) \end{bmatrix}. \tag{20.32}$$

(ii) When the local input-output decoupling problem is solvable?

20.4 Let

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Assume $K \in \mathcal{M}_{2\times 3}$, which is unknown. Express $(A+BK)^3$ as $W_0 + W_1 Z + W_2 Z^2 + W_3 Z^3$, where $Z = V_r(K)$.

20.5 Assume $A(x) \in \mathbb{R}^m$ with $x \in \mathbb{R}^n$ is expressed as

$$A(x) = A_0 + A_1 x + A_2 x^2 + \cdots.$$

Express $A(x)^T A(x)$ as

$$A(x)^T A(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots.$$

Find formulas for calculating α_i, $i = 0, 1, 2, \cdots$.

20.6 $A_i \in \mathcal{M}_{n\times n}$, $i = 0, 1, \cdots$ and $K \in \mathcal{M}_{n\times n}$.

(i) Express

$$A(x) = A_0 + A_1 K + A_2 K^2 + \cdots = C_0 + C_1 Z + C_2 Z^2 + \cdots,$$

where $Z = V_c(K)$.

(ii) Express

$$[A(x)]^T = D_0 + D_1 Z + D_2 Z^2 + \cdots.$$

(iii) Express

$$[A(x)]^T A(x) = E_0 + E_1 Z + E_2 Z^2 + \cdots.$$

20.7 Consider a linear control system and assume $p < m$. Prove that if

$$\mathrm{rank}(D(x_0) = p,$$

then the input-output decoupling problem is solvable.

20.8 Consider Lemma 20.2. To complete the proof, show that the following two are equivalent:

(i) $T(K)((W(K)^T)^\perp) = \mathbb{R}^p$;

(ii) $[(T^T(K))^{-1}(W(K)^T)]^\perp = \mathbb{R}^p$.

20.9 Let $M(z) \in \mathcal{M}_{m\times n}$. Show that

(i) there exists z_0 such that $\mathrm{rank}(M(z)) = m$, if and only if

$$\max_z M(z)M^T(z) > 0;$$

(ii) there exists z_0 such that $\mathrm{rank}(M(z)) = n$, if and only if

$$\max_z M^T(z)MT(z) > 0.$$

20.10 Consider Example 20.1.

(i) Find another K, satisfying (20.27). Show that the corresponding feedback control solves the decoupling problem.

(ii) Using both the K obtained in the above and the K from textbook to check items 2 and 3 of the Remark 20.1.

Chapter 21

Linearization of Nonlinear Control Systems

Since the dynamics of a linear system is much simpler than that of a nonlinear system, if a nonlinear (control) system can be converted into a linear (control) system, the analysis and design tools for linear (control) systems can then be used. Hence different kinds of linearizations become an interesting topic in physics and systems and control. In this chapter we consider some special linearizations, in which STP players an important role.

21.1 Carleman Linearization

Consider a dynamic system

$$\dot{x} = f(x), \quad x \in \mathbb{R}^n, \tag{21.1}$$

where $f(x)$ is an analytic vector field with $f(0) = 0$.

J. Carleman proposed a method to merge the system into an infinite-dimensional linear system. In this section its basic form and the realization are discussed. Its original form is rather complicated. STP makes it much simpler.

Choosing x, x^2, $\cdots$ as a set of basis, the system (21.1) can be expressed as

$$\dot{x} = F_1 x + F_2 x^2 + F_3 x^3 + \cdots, \tag{21.2}$$

where F_1 is an $n \times n$ matrix, and F_2 is an $n \times n^2$ matrix, and so on.

We may consider x, x^2, x^3, $\cdots$ as a set of independent variables and then calculate their derivatives to get a linear form, called the Carleman linearization, as follows:

$$\begin{bmatrix} \dot{x} \\ \dot{x^2} \\ \dot{x^3} \\ \vdots \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} & \cdots \\ 0 & A_{22} & A_{23} & A_{24} & \cdots \\ 0 & 0 & A_{33} & A_{34} & \cdots \\ \vdots & & & & \end{bmatrix} \begin{bmatrix} x \\ x^2 \\ x^3 \\ \vdots \end{bmatrix}. \tag{21.3}$$

Theorem 21.1. *In Carleman linearization form (21.3) the coefficients A_{ij} are determined by the following equations.*

$$\begin{cases} A_{1i} = F_i, \quad i \geq 1, \\ A_{k,k+s} = \sum\limits_{i=0}^{k-1} I_{n^i} \otimes F_{s+1} \otimes I_{n^{k-1-i}}. \end{cases} \tag{21.4}$$

Proof. Using chair rule, we have

$$\frac{d}{dt}(x^k) = \sum_{i=0}^{k-1} x^i \dot{x} x^{k-i-1} = \sum_{s=0}^{\infty} \sum_{i=0}^{k-1} x^i F_{s+1} x^{k-i+s}.$$

Using (2.56), we have

$$x^i F_{s+1} x^{k-i+s} = (I_{n^i} \otimes F_{s+1}) \ltimes x^{k+s} = (I_{n^i} \otimes F_{s+1} \otimes I_{n^{k-i-1}}) x^{k+s}.$$

(21.4) follows. □

We can express (21.3) into the following linear form

$$\dot{X} = AX, \tag{21.5}$$

where A is an infinite-dimensional block upper triangular matrix.

The infinite-dimensional block upper triangular matrices have some special properties, which make (21.5) meaningful. We give a brief discussion about this.

Denote the kth left-upper block of A by A_k, i.e.,

$$A_k = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1k} \\ 0 & A_{22} & \cdots & A_{2k} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & A_{kk} \end{bmatrix}.$$

A square block upper triangular matrix is said to have a set of structure parameters $(k_1, k_2, \cdots)$, if the diagonal blocks have dimensions $\dim(A_{ii}) = k_i \times k_i$. For instance, consider (21.3), its structure parameters are $(n, n^2, n^3, \cdots)$. For statement ease, we identify A_k with its infinite-dimensional extension A_k^e, which is an infinite-dimensional matrix with A_k

as its left upper minor and all other elements are zero. Using this convention, the coefficient matrix A can be expressed as

$$A = \lim_{k\to\infty} A_k.$$

This limit is well defined in the following sense: Denote the (i,j)th element of A_k by a_{ij}^k. Then the sequence $\{a_{ij}^k,\ k = 1,2,\cdots\}$ has the following form

$$(a_{ij}^1, a_{ij}^2, \cdots, a_{ij}^k, \cdots) = (0, \cdots, 0, c_{ij}, c_{ij}, c_{ij}, \cdots).$$

That is, after first finite terms, it becomes a constant sequence. Hence this sequence converges. Based on the same reason, the following operations are also well defined.

Definition 21.1.

(1) Let A and B be two infinite-dimensional upper triangular matrices with same structure parameters. The product of A and B is defined as

$$AB := \lim_{k\to\infty} A_k B_k.$$

(2) Assume $A_{ii},\ i = 1,2,\cdots$ are invertible. Then its inverse is defined as

$$A^{-1} := \lim_{k\to\infty} A_k^{-1}.$$

(3)

$$e^A := \lim_{k\to\infty} e^{A_k}.$$

Now it is natural to use the solution of the linearized system (21.5)

$$X = e^{At} X_0$$

as a solution of (21.1). In fact, we can use only finite terms to approximate the real solution.

Denote by $E_{ij}^k(t)$ the (i,j)th block of $e^{A_k t}$. It follows from the structure of A that

$$E_{ij}^s(t) = E_{ij}^k(t), \quad s > k, \quad i,j \le k.$$

Hence we can define

$$X^n(t) = \sum_{k=1}^n E_{1k}^k(t) X_0^k.$$

Observing (21.3), one sees easily that if

$$X(t) = \lim_{n\to\infty} X^n(t)$$

exists, then it is the solution of (21.1) with initial value $X(0) = X_0$.

We are particularly interested in the upper triangular matrix generated by Carleman linearization. In Carleman linearized form (21.3). We assume $F_1 = A_{11}$ is stable (anti-stable), that is, all the eigenvalues of A_{11} have negative real part $\mathrm{Re}\,\sigma(A_{11}) < 0$ (correspondingly, $\mathrm{Re}\,\sigma(A_{11}) > 0$), then A is invertible. In fact, we have

Theorem 21.2. *Assume $F_1 = A_{11}$ has eigenvalues $\sigma(A_{11}) = \{\lambda_1, \cdots, \lambda_n\}$, then A_{ii}, $i \geq 2$ has eigenvalues*

$$\sigma(A_{ii}) = \{\lambda_{k_1} + \cdots + \lambda_{k_i} \,|\, k_1, \cdots, k_i = 1, \cdots, n\}.$$

Proof. First, we assume the eigenvalues of A_{11} are distinct, and their corresponding eigenvectors are

$$\{\xi_1, \cdots, \xi_n\}.$$

A straightforward computation shows

$$A_{ii}(\xi_{k_1} \ltimes \cdots \ltimes \xi_{k_i}) = (\lambda_{k_1} + \cdots + \lambda_{k_i})(\xi_{k_1} \ltimes \cdots \ltimes \xi_{k_i}).$$

To avoid notational confusion, we consider it only for the case of $i = 2$. Using (21.4), we have

$$A_{22} = I_n \otimes A_{11} + A_{11} \otimes I_n.$$

Hence,

$$\begin{aligned} A_{22}(\xi_i \ltimes \xi_j) &= (I_n \otimes A_{11} + A_{11} \otimes I_n)(\xi_i \ltimes \xi_j) \\ &= (I_n \otimes A_{11})(\xi_i \ltimes \xi_j) + (A_{11} \otimes I_n)(\xi_i \ltimes \xi_j) \\ &= \lambda_j \xi_i \ltimes \xi_j + \lambda_i \xi_i \ltimes \xi_j. \end{aligned}$$

It follows that $\lambda_i + \lambda_j$, $i, j = 1, \cdots, n$ are eigenvalues of A_{22}. To prove that they are the complete set of eigenvalues it is enough to prove that

$$\{\xi_i \ltimes \xi_j \,|\, i, j = 1, \cdots, n\}$$

are linearly independent. Note that since all λ_i, $i = 1, \cdots, n$, are distinct, it follows that all ξ_i $i = 1, \cdots, n$, are linearly independent. Assume

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_{i,j} \xi_i \xi_j = 0.$$

Rewrite it as

$$\sum_{i=1}^{n} \xi_i [\sum_{j=1}^{n} c_{ij} \xi_j] = 0.$$

Using Proposition 2.14, we have

$$\sum_{j=1}^{n} c_{ij}\xi_j = 0, \quad i = 1, \cdots, n.$$

Hence, it is clear that $c_{i,j} = 0$, $i = 1, \cdots, n$, $j = 1, \cdots, n$.

We conclude that the eigenvalues of A_{22} are

$$\sigma(A_{22}) = \{\lambda_i + \lambda_j \,|\, i, j = 1, \cdots, n\}. \tag{21.6}$$

Finally, by continuity we know that even multi-eigenvalues exist, the structure of eigenvalues of A_{22}, precisely, (21.6) remains true. □

21.2 First Integral

This section considers the first integral of a vector field.

Definition 21.2. Let $f(x)$, $x \in \mathbb{R}^n$, be a smooth vector field. A smooth time-varying function $\phi(t, x)$ is said to be the first integral of $f(x)$ if it satisfies

$$\frac{d}{dt}\phi(t, x) = \frac{\partial \phi}{\partial t} + \frac{\partial \phi}{\partial x} f(x) = 0.$$

Consider a polynomial system

$$\dot{x} = F_1 x + F_2 x^2 + \cdots + F_k x^k. \tag{21.7}$$

We search for the following type of first integral

$$H(t, x) = e^{-\xi t} P(x).$$

Carleman linearization technique can be used to investigate it.

Assume $P(x) = P_0 + P_1 x + \cdots + P_s x^s$, with its symmetric coefficients as $P_1, \cdots, P_s$. (Where the "symmetry" means the coefficients for same items with different factor orders are the same. For instance, $x_1^2 x_2$, $x_1 x_2 x_1$, and $x_2 x_1^2$ have the same coefficients.) It is easy to see that if $\xi \neq 0$ then $P_0 = 0$. Hence, we can simply assume $P_0 = 0$.

Setting $dH(t, x)/dt = 0$, we have

$$\begin{aligned} &\begin{bmatrix} P_1 & \cdots & P_s \end{bmatrix} \begin{bmatrix} A_{11} & \cdots & \cdots & A_{1k} & & & \\ & \ddots & & & \ddots & & \\ & & A_{ss} & \cdots & \cdots & A_{s,s+k-1} \end{bmatrix} \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^{s+k-1} \end{bmatrix} \\ &= \xi \begin{bmatrix} P_1 & \cdots & P_s \end{bmatrix} \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^s \end{bmatrix}. \end{aligned} \tag{21.8}$$

Since x^k is a redundant generator, the coefficients are not unique. Using this form to search first integral is conservative, and the result obtained is, in general, not necessity. Because under other equivalent coefficients may produce other first integrals.

To get necessary and sufficient condition we convert it into natural basis. Set

$$P_i = \tilde{P}_i T_B(n,i), \quad \tilde{A}_{ij} = T_B(n,i) A_{ij} T_N(n,j).$$

Plugging them into (21.8) yields

$$\begin{aligned}
&\left[\tilde{P}_1 \cdots \tilde{P}_s\right] \begin{bmatrix} \tilde{A}_{11} \cdots & \cdots \ \tilde{A}_{1k} & & \\ \ddots & & \ddots & \\ & \tilde{A}_{ss} & \cdots & \cdots \ \tilde{A}_{s,s+k-1} \end{bmatrix} \begin{bmatrix} x \\ x_{(2)} \\ \vdots \\ x_{(s+r-1)} \end{bmatrix} \\
&= \xi \left[\tilde{P}_1 \cdots \tilde{P}_s\right] \begin{bmatrix} x \\ x_{(2)} \\ \vdots \\ x_{(s)} \end{bmatrix}.
\end{aligned}$$

Theorem 21.3. *Denote $h_i = \tilde{P}_i^T$ and $B_{ij} = \tilde{A}_{ji}^T$. System (21.7) has first integral $H(t,x)$, if and only if there exists ξ such that the following system has nonzero solution $(h_1, \cdots, h_s)$.*

$$\begin{cases}
\begin{bmatrix} B_{11} & 0 & 0 \cdots & 0 \\ B_{21} & B_{22} & 0 \cdots & 0 \\ \vdots & & & \\ B_{s1} & B_{s2} & \cdots & B_{s,s} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_s \end{bmatrix} = \xi \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_s \end{bmatrix}, \\[2ex]
\begin{bmatrix} B_{s+1,1} & B_{s+1,2} & \cdots & B_{s+1,s} \\ \vdots & & & \\ B_{k,1} & B_{k,2} & \cdots & B_{k,s} \\ 0 & B_{k+1,2} & \cdots & B_{k+1,s} \\ \vdots & & & \\ 0 & \cdots & 0 & B_{s+k-1,s} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_s \end{bmatrix} = 0.
\end{cases} \tag{21.9}$$

Next, we consider the solution of (21.9).

The following lemma is itself interesting. Let $\xi = (i_1, \cdots, i_k)$. We use notation $\mathrm{id}(\xi; n^k)$ for $\mathrm{id}(i_1, \cdots, i_k; n, \cdots, n)$.

Lemma 21.1. *Assume the row vector* $h \in \mathbb{R}^{n^k}$ *is labeled by multi-index* $\mathrm{id}(\xi; n^k)$*, and it is symmetric with respect to* $\mathrm{id}(\xi; n^k)$*,* $F \in M_{n\times n}$*. Define*

$$A = F \otimes I_{n^{k-1}} + I_n \otimes F \otimes I_{n^{k-2}} + \cdots + I_{n^{k-1}} \otimes F.$$

Then hA *is also symmetric with respect to* $\mathrm{id}(\xi; n^k)$.

Proof. Since interchange any two indices can be realized by swapping two adjacent indices, we need only to prove that hA is invariant under the swap of two adjacent indices. Define

$$\Phi = I_{n^{j-1}} \otimes W_{[n]} \otimes I_{n^{k-j-1}}.$$

It is clear that exchanging the jth and $(j+1)$th indices yields a new vector $h\Phi$. Since h is symmetric with respect to $\mathrm{id}(\xi; n^k)$, then $h\Phi = h$, and $h\Phi A = hA$. To prove hA is symmetric, i.e., $hA\Phi = hA$, it suffices to prove

$$A\Phi = \Phi A.$$

Express A as

$$A = \sum_{i=1}^{k} A_i,$$

where

$$A_i = \underbrace{I \otimes \cdots \otimes I}_{i-1} \otimes F \otimes \underbrace{I \otimes \cdots \otimes I}_{k-1}.$$

If F does not lie on the jth or $(j+1)$th position, then it is obvious that Φ and $A_j + A_{j+1}$ are commutative. Hence we need only to consider the two related terms of $A_j + A_{j+1}$. This means

$$W_{[n]}(F \otimes I + I \otimes F) = (F \otimes I + I \otimes F)W_{[n]}.$$

Note that $W_{[n]}^{-1} = W_{[n]}$, the above equality is true. □

Proposition 21.1. *Assume the* F_1 *in (21.7) has eigenvalues* $\sigma = \{\lambda_1, \cdots, \lambda_n\}$*, then the eigenvalues of* B_{kk} *in (21.9) are*

$$\sigma_k := \sigma(B_{kk}) = \{\lambda_{i_1} + \cdots + \lambda_{i_k} \,|\, i_1, \cdots, i_k = 1, \cdots, n\}.$$

Proof. Since $B_{kk} = \tilde{A}_{kk}^T$, and the eigenvalues of A_{kk} are σ_k, it is enough to prove that the $\tilde{A}_{kk}$ and A_{kk} have the same eigenvalues. Assume μ is an eigenvalue of $\tilde{A}_{kk}$, then there exists a $\tilde{P} \neq 0$ such that

$$\tilde{P}\tilde{A}_{kk} = \mu\tilde{P}.$$

By definition we have $\tilde{A}_{kk} = T_B(n,k)A_{kk}T_N(n,k)$. Note that $T_B(n,k)T_N(n,k) = I$, hence

$$\tilde{P}T_B(n,k)A_{kk}T_N(n,k)\begin{bmatrix} x_{(1)} \\ x_{(2)} \\ \vdots \\ x_{(k)} \end{bmatrix} = \mu\tilde{P}T_B(n,k)T_N(n,k)\begin{bmatrix} x_{(1)} \\ x_{(2)} \\ \vdots \\ x_{(k)} \end{bmatrix}. \quad (21.10)$$

Let $P = \tilde{P}T_B(n,k)$. Then $P \neq 0$ is a symmetric set. For P, (21.10) becomes

$$PA_{kk}\begin{bmatrix} x^1 \\ x^2 \\ \vdots \\ x^k \end{bmatrix} = \mu P\begin{bmatrix} x^1 \\ x^2 \\ \vdots \\ x^k \end{bmatrix}. \quad (21.11)$$

According to Lemma 21.1, PA_{kk} is a symmetric set. Note that the symmetric coefficients are unique, we have

$$PA_{kk} = \mu P.$$

Hence, μ is also an eigenvalue of A_{kk}.

Conversely, assume μ is an eigenvalue of A_{kk}, then $\mu = \lambda_{i_1} + \cdots + \lambda_{i_k}$. Denote by Y_j the eigenvector corresponding to λ_{i_j} of F_1, then we construct

$$Y = \sum_{\sigma\in\mathbf{S}_k} Y_{\sigma(1)} \otimes \cdots \otimes Y_{\sigma(k)},$$

where $\mathbf{S}_k$ is the kth order symmetric group. Then we have

$$YA_{kk} = \mu Y.$$

Since Y is symmetric, there exists $\tilde{Y} \neq 0$ such that $Y = \tilde{Y}T_B(n,k)$. Hence, we have

$$\tilde{Y}T_B(n,k)A_{kk} = \mu\tilde{Y}T_B(n,k).$$

Right-multiplying both sides of the above equality by $T_N(n,k)$ yields

$$\tilde{Y}\tilde{A}_{kk} = \mu\tilde{Y}.$$

That is, μ is also an eigenvalue of $\tilde{A}_{kk}$. □

Proposition 21.2.

(1) If (21.9) has solution $h \neq 0$, then

$$\xi = c_1\lambda_{i_1} + \cdots + c_s\lambda_{i_s},$$

where $\lambda_{i_1}, \cdots, \lambda_{i_s} \in \sigma(F_1)$; $c_1, \cdots, c_s$ take value 1 or 0.

(2) If h has a component $h_j \neq 0$, then $\xi \in \sigma^j$. If h has t non-zero components, then σ^s has at least one t fold element, where $\sigma^t = \{c_1\lambda_{i_1} + \cdots + c_t\lambda_{i_t} \,|\, c_1, \cdots, c_t \in \{0,1\}\}$.

(3) If (21.7) has a linear first integral $H(t,x) = e^{-\xi t}h^T x$, then for arbitrary integer $j > 0$, $H_j(t,x) = e^{-j\xi t}(h^T)^j x^j$ is a first integral of (21.7).

Proof. (1) and (2) are the immediate consequences of Proposition 21.1. We prove (3). If (21.7) has a linear first integral $H(t,x) = e^{-\xi t}h^T X$, then

$$\begin{cases} F_1 h = \xi h, \\ F_i h = 0, \quad i = 2, \cdots, k. \end{cases}$$

Assume $p = (0_n, 0_{n^2}, \cdots, 0_{n^{j-1}}, h^j)$, where 0_k is the zero vector in $\mathbb{R}^k$. Since

$$A_{j,j+s-1} = I_{n^{j-1}} \otimes F_s + I_{n^{j-2}} \otimes F_s \otimes I + \cdots + F_s \otimes I_{n^{j-1}},$$

we have

$$\begin{cases} A_{jj} h^j = j\xi h^j, \\ A_{jt} h^j = 0, \quad t = j+1, \cdots, j+k-1, \end{cases}$$

which means p satisfies (21.8) with ξ being replaced by $j\xi$. □

Remark 21.1. Proposition 21.2 provides a convenient way to find first integer. In fact, after fixing ξ the problem becomes solving a linear algebraic system. For Lorenz system, (1) and (2) of Proposition 21.2 have been proved in Cairo and Feix (1992). Here the statement is a generalization of their result.

Example 21.1. Lotka-Volterra equation established the interactive relation in chemistry or for co-existence of spices. Lotka-Volterra equation is

$$\dot{x}_i = x_i\left(a_i + \sum_{j=1}^{n} b_{ij}x_j\right), \quad i = 1, \cdots, n. \tag{21.12}$$

Consider $n = 2$. Set

$$A_{11} = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix}, \quad A_{12} = \begin{bmatrix} b_{11} & b_{12} & 0 & 0 \\ 0 & 0 & b_{21} & b_{22} \end{bmatrix}.$$

Then

$$A_{22} = A_{11} \otimes I_2 + I_2 \otimes A_{11} = \begin{bmatrix} 2a_1 & 0 & 0 & 0 \\ 0 & a_1 + a_2 & 0 & 0 \\ 0 & 0 & a_1 + a_2 & 0 \\ 0 & 0 & 0 & 2a_2 \end{bmatrix},$$

$$\begin{aligned} A_{23} &= A_{12} \otimes I_2 + I_2 \otimes A_{12} \\ &= \begin{bmatrix} 2b_{11} & b_{12} & b_{12} & 0 & 0 & 0 & 0 & 0 \\ 0 & b_{11} & b_{21} & b_{12} + b_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & b_{11} + b_{21} & b_{12} & b_{22} & 0 \\ 0 & 0 & 0 & 0 & 0 & b_{21} & b_{21} & 2b_{22} \end{bmatrix}. \end{aligned}$$

Carleman linearized form becomes

$$\begin{bmatrix} \dot{x} \\ \dot{x^2} \\ \dot{x^3} \\ \vdots \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & 0 & 0 & \cdots \\ 0 & A_{22} & A_{23} & 0 & \cdots \\ 0 & 0 & A_{33} & A_{34} & \cdots \\ \vdots & & & & \end{bmatrix} \begin{bmatrix} x \\ x^2 \\ x^3 \\ \vdots \end{bmatrix}.$$

Assume we look for the first integral with the form as $H(t,x) = e^{-\xi t}(P_1 x + P_2 x^2)$. Then

$$T_B(2,2) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad T_N(2,2) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$T_N(2,3) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$B_{11} = A_{11}^T, \quad B_{21} = \tilde{A}_{12}^T = T_B(2,1) A_{12} T_N(2,2) = \begin{bmatrix} b_{11} & b_{12} & 0 \\ 0 & b_{21} & b_{22} \end{bmatrix},$$

$$B_{22}^T = T_B(2,2) A_{22} T_N(2,2) = \begin{bmatrix} 2a_1 & 0 & 0 \\ 0 & a_1 + a_2 & 0 \\ 0 & 0 & 2a_2 \end{bmatrix},$$

$$B_{32}^T = T_B(2,2)A_{23}T_N(2,3) = \begin{bmatrix} 2b_1 & 2b_{12} & 0 & 0 \\ 0 & b_{11}+b_{21} & b_{12}+b_{22} & 0 \\ 0 & 0 & 2b_{21} & 2b_{22} \end{bmatrix}.$$

We conclude that the second order first integral exists, if and only if the following equation has solution.

$$\begin{cases} \begin{bmatrix} B_{11} & 0 \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \xi \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}, \\ B_{32}h_2 = 0, \end{cases}$$

where $\xi \in \{a_1, a_2, 2a_1, 2a_2, a_1 + a_2\}$.

21.3 Invariance of Polynomial System

Definition 21.3. Consider a smooth dynamic system

$$\dot{x} = f(x), \quad x \in \mathbb{R}^n. \tag{21.13}$$

A smooth time-varying function $h(t,x)$ is said to be an invariance of (21.13) if

$$\left.\frac{dh}{dt}\right|_{(21.13)} = 0. \tag{21.14}$$

In this section we consider a polynomial system

$$\dot{x} = F_0 + F_1x + \cdots + F_kx^k, \quad x \in \mathbb{R}^2, \tag{21.15}$$

and look for an invariance with the following form:

$$H(x,t) = e^{\xi t}x_1^{\alpha}x_2^{\beta}(P_0 + P_1x + \cdots + P_lx^l). \tag{21.16}$$

This form of invariance has been investigated in Cairo and Feix (1992). Recently, Darboux method has also been used for this investigation (Cairo *et al.*, 1999).

Our purpose is to convert it into an algebraic system. Using formula (18.72), the time derivative of $H(x,t)$ can be calculated as

$$\begin{aligned} \tfrac{dH}{dt} &= \tfrac{\partial H}{\partial t} + DH \cdot \dot{x} \\ &= \xi e^{\xi t}x_1^{\alpha}x_2^{\beta}(P_0 + P_1x + \cdots + P_lx^l) \\ &\quad + \ e^{\xi t}(\alpha x_1^{\alpha-1}x_2^{\beta},\ \beta x_1^{\alpha}x_2^{\beta-1})(P_0 + P_1x + \cdots + P_lx^l) \\ &\quad \times (F_0 + F_1x + \cdots + F_kx^k) \\ &\quad + \ e^{\xi t}x_1^{\alpha}x_2^{\beta}(P_1 + P_2\Phi_1x + \cdots + P_l\Phi_{l-1}x^{l-1})(F_0 + F_1x + \cdots + F_kx^k). \end{aligned}$$

Setting

$$\frac{dH}{dt} = 0,$$

and observing that

$$x_1x_2 = (0\ 1\ 0\ 0)x^2, \quad (\alpha x_2,\ \beta x_1) = (0\ \beta\ \alpha\ 0)x,$$

we have

$$\begin{aligned}
&\xi(0\ 1\ 0\ 0)x^2(P_0 + P_1x + \cdots + P_lx^l)\\
&\quad +\ (0\ \beta\ \alpha\ 0)x(P_0 + P_1x + \cdots + P_lx^l)(F_0 + F_1x + \cdots + F_kx^k)\\
&\quad +\ (0\ 1\ 0\ 0)x^2(P_1 + P_2\Phi_1x + \cdots + P_l\Phi_{l-1}x^{l-1})(F_0 + F_1x + \cdots + F_kx^k)\\
&= 0.
\end{aligned}$$

Using (2.56), it can be expressed as

$$\begin{aligned}
&\xi(0\ 1\ 0\ 0)[(I_4 \otimes P_0)x^2 + (I_4 \otimes P_1)x^3 + \cdots + (I_4 \otimes P_l)x^{l+2}]\\
&\quad + (0\ \beta\ \alpha\ 0)[(I_2 \otimes P_0)x + (I_2 \otimes P_1)x^2\\
&\qquad + \cdots + (I_2 \otimes P_l)x^{l+1}](F_0 + F_1x + \cdots + F_kx^k)\\
&\quad + (0\ 1\ 0\ 0)[(I_4 \otimes P_1)x^2 + (I_4 \otimes P_2\Phi_1)x^3\\
&\qquad + \cdots + (I_4 \otimes P_l\Phi_{l-1})x^{l+1})(F_0 + F_1x + \cdots + F_kx^k)\\
&= 0.
\end{aligned}$$

Using (2.56) again, we can multiply the above form out as

$$\begin{aligned}
&\sum_{s=1}^{l+1} \xi(0\ 1\ 0\ 0)[I_4 \otimes P_{s-1}]x^{s+1}\\
&\quad +\ \sum_{s=0}^{k+l}(0\ \beta\ \alpha\ 0)\sum_{i=0,j=0}^{i+j=s}[(I_2 \otimes P_i)(I_{2^{i+1}} \otimes F_{s-i})]x^{s+1}\\
&\quad +\ \sum_{s=1}^{k+l}(0\ 1\ 0\ 0)\sum_{i=1,j=0}^{i+j=s}[(I_4 \otimes P_i\Phi_{i-1})(I_{2^{i+1}} \otimes F_{s-i})]x^{s+1}\\
&= 0.
\end{aligned}$$

Converting each terms into the forms with naturel basis, we have the following result.

Theorem 21.4. *System (21.15) has the invariance of the form (21.16), if and only if the following algebraic system has solution* $(\xi, \alpha, \beta, P_0, \cdots, P_l)$.

$$
\begin{aligned}
&(0\ \beta\ \alpha\ 0)[(I_2 \otimes P_0)(I_2 \otimes F_0)] = 0\\
&\Bigg\{\xi(0\ 1\ 0\ 0)(I_4 \otimes P_{s-1}) + (0\ \beta\ \alpha\ 0)\sum_{i=0,j=0}^{i+j=s}[(I_2 \otimes P_i)(I_{2^{i+1}} \otimes F_{s-i})]\\
&\quad + (0\ 1\ 0\ 0)\sum_{i=1,j=0}^{i+j=s}[(I_4 \otimes P_i\Phi_{i-1})(I_{2^{i+1}} \otimes F_{s-i})]\Bigg\} T_N(2, s+1) = 0,\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad s = 1, \cdots, l+1\\
&\Bigg\{(0\ \beta\ \alpha\ 0)\sum_{i=0,j=0}^{i+j=s}[(I_2 \otimes P_i)(I_{2^{i+1}} \otimes F_{s-i})]\\
&\quad + (0\ 1\ 0\ 0)\sum_{i=1,j=0}^{i+j=s}[(I_4 \otimes P_i\Phi_{i-1})(I_{2^{i+1}} \otimes F_{s-i})]\Bigg\} T_N(2, s+1) = 0,\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad s = l+2, \cdots, l+k.
\end{aligned}
\tag{21.17}
$$

Remark 21.2. The advantage of this approach lies on

(i) the solution is easily obtained via computer;
(ii) it can easily be generated into higher-dimensional cases.

21.4 Feedback Linearization of Nonlinear Control System

Consider an affine nonlinear system

$$\dot{x} = f(x) + \sum_{i=1}^{m} g_i(x)u_i, \quad f(0) = 0, \quad x \in \mathbb{R}^n, u \in \mathbb{R}^m, \tag{21.18}$$

where $f(x)$, $g_i(x)$, $i = 1, \cdots, m$ are analytic vector fields. The feedback linearization is defined as following.

Definition 21.4. System (21.18) is locally state feedback linearizable, if there exists a state feedback control

$$u = \alpha(x) + \beta(x)v, \tag{21.19}$$

and a local diffeomorphism $z = \xi(x)$ on a neighborhood U of the origin, such that under the new coordinate frame the closed-loop system becomes a completely controllable linear system

$$\dot{z} = Az + \sum_{i=1}^{m} b_i v_i, \quad z \in U, v \in \mathbb{R}^m. \tag{21.20}$$

If $\beta(x)$ in control (21.19) is an $m \times m$ nonsingular matrix, the linearization is called a regular feedback linearization; otherwise, say, $\beta(x)$ is an $m \times k$ matrix ($k < m$), it is called a non-regular feedback linearization. Particularly, when $k = 1$ it is called the single-input feedback linearization.

Heymann's Lemma (Heymann, 1968) says that for a completely controllable linear system there exists a linear feedback

$$u = Hv, \quad v \in \mathbb{R},$$

such that the closed-loop system becomes a single-input completely controllable system. The following lemma is an immediate consequence of the Heymann's Lemma.

Lemma 21.2. *System (21.18) is state feedback linearizable, if and only if it is single-input feedback linearizable. That is, it is linearizable via (21.19), where $\beta(x)$ is an $m \times 1$ vector.*

The following lemma is useful and easily verifiable.

Lemma 21.3 (Sun and Xia, 1997). *Set $A = J_f(0)$, which is the Jacobi matrix of f at the origin, and $B = g(0)$. If the system (21.18) is linearizable, then (A, B) is completely controllable.*

In the following investigation we need one more concept, called the normal form, which has been used to investigate many nonlinear (control) systems (Krener and Kang, 1990; Cheng, 2002; Devanathan, 2001). We first introduce it briefly (Guckenheimer and Holmes, 1983): Let H_n^k be the set of kth order homogeneous polynomial vector fields. Then

(1) H_n^k is an $\mathbb{R}$-linear vector space;
(2) Let $Ax \in H_n^1$ be a given linear vector field, where A is an $n \times n$ constant matrix. Then the derivative $\mathrm{ad}_{Ax} : H_n^k \to H_n^k$ is a linear mapping.

The following normal form expression (Arnold, 1983) and its application to linearization (Devanathan, 2001) are the starting point of our study.

Definition 21.5. Assume $A \in \mathcal{M}_{n\times n}$, and $\sigma(A) = \lambda = \{\lambda_1, \cdots, \lambda_n\}$ is the set of its eigenvalues. A is an resonant matrix, if there exist $m =$

$(m_1 \cdots, m_n) \in \mathbb{Z}_+^n$ and $|m| \geq 2$, i.e., $m_i \geq 0$ and $\sum\limits_{i=1}^{n} m_i \geq 2$, such that for an $1 \leq s \leq n$, $\lambda_s = \langle m, \lambda \rangle$. Otherwise, A is called a non-resonant matrix.

Theorem 21.5 (Poincaré Theorem (Arnold, 1983)). *Consider an analytic system*

$$\dot{x} = Ax + f_2(x) + f_3(x) + \cdots, \quad x \in \mathbb{R}^n, \tag{21.21}$$

where $f_i(x)$, $i \geq 2$ are ith order homogeneous polynomial vector fields. If A is non-resonant, then there exists a coordinate transformation

$$x = y + h(y), \tag{21.22}$$

where $h(y) = h_2(y) + h_3(y) + \cdots$ with $h_i(y)$ the ith homogeneous vectors, such that system (21.21) can be expressed into a linear system as

$$\dot{y} = Ay. \tag{21.23}$$

The following lemma gives a sufficient condition for non-resonant matrix.

Proposition 21.3 (Devanathan, 2001). *Let $\lambda = (\lambda_1, \cdots, \lambda_n)$ be the set of eigenvalues of a Hurwitz matrix A. A is non-resonant, if*

$$\max\{|\operatorname{Re}(\lambda_i)| \,|\, \lambda_i \in \sigma(A)\} \leq 2\min\{|\operatorname{Re}(\lambda_i)| \,|\, \lambda_i \in \sigma(A)\}. \tag{21.24}$$

21.5 Single Input Feedback Linearization

First, we give a normal form for non-regular feedback linearization.

A constant vector $b = (b_1, \cdots, b_n)^T \in \mathbb{R}^n$ is called a component nonzero vector, if $b_i \neq 0$, $\forall\, i$.

Proposition 21.4. *A linear control system*

$$\dot{x} = Ax + \sum_{i=1}^{m} b_i u_i := Ax + Bu, \quad x \in \mathbb{R}^n,\ u \in \mathbb{R}^m \tag{21.25}$$

is completely controllable, if and only if there exist two matrices F, G such that the closed-loop system

$$\dot{x} = (A + BF)x + BGv$$

can be converted, via a linear coordinate transformation, into the following form

$$\dot{z} = Az + bv := \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix} z + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} v, \tag{21.26}$$

where d_i, $i = 1, \cdots, n$ are distinct, and b is a nonzero component vector.

The proof is simple. The key is for such a system the controllability matrix C has its determinant as

$$\det(C) = \prod_{i=1}^{n} b_i \prod_{i<j} (d_j - d_i) \neq 0. \tag{21.27}$$

It is essentially a Vandermonde matrix. Because of Proposition 21.4 we call (21.26) the non-regular single input feedback A-diagonal (NRSIFAD) normal form. Moreover, we give the following assumption:

A1. A is diagonal with distinct diagonal elements, d_i, $i = 1, \cdots, n$, and A is non-resonant.

Lemma 21.4. *Assume A satisfies A1, g is a kth order homogeneous polynomial vector fields, $k \geq 2$. Then there exists a kth order homogeneous vector field η such that*

$$\mathrm{ad}_{Ax}\, \eta = g. \tag{21.28}$$

***Proof*.** For a given η, assume $f = ad_{Ax}\eta$. Then a straightforward computation shows that the ith component f_i of f depends only on the ith component η_i of η. Now assume $x_1^{r_1} \cdots x_n^{r_n}$ is a term of η_i, a simple computation shows that

$$\mathrm{ad}_{Ax}\, \eta = \left[\begin{bmatrix} d_1 x_1 \\ \vdots \\ d_i x_i \\ \vdots \\ d_n x_n \end{bmatrix}, \begin{bmatrix} \times \\ \vdots \\ x_1^{r_1} \cdot x_n^{r_n} \\ \vdots \\ \times \end{bmatrix} \right] = \begin{bmatrix} \times \\ \vdots \\ \mu_i x_1^{r_1} \cdot x_n^{r_n} \\ \vdots \\ \times \end{bmatrix}, \tag{21.29}$$

where

$$\mu_i = d_1 r_1 + \cdots + d_n r_n - d_i. \tag{21.30}$$

Since A is non-resonant and $\mu_i \neq 0$, then for every term of g_i, say, $x_1^{r_1} \cdots x_n^{r_n}$, we can construct a corresponding term of η_i, say, $\frac{1}{\mu_i} x_1^{r_1} \cdots x_n^{r_n}$, such that $\mathrm{ad}_{Ax}\, \eta = g$. □

Note that since all the vector fields and functions involved are analytic, then all the functions and their derivatives concerned have convergent Taylor expansions.

Note also that if A satisfies A1, then for vector field $g = g_k x^k + g_{k+1}x^{k+1} + \cdots \in O(\|x\|^k)$, we can apply Lemma 21.4 to its each components, then we can construct a vector field $\eta \in O(\|x\|^k)$, such that $\mathrm{ad}_{Ax}\,\eta = g$.

Go back to the linearization problem. We consider the following system:

$$\dot{x} = Ax + \xi(x) + \sum_{i=1}^{m} g_i(x)u_i, \tag{21.31}$$

where A satisfies A1, and $\xi(x) = O(\|x\|^2)$. An immediate conclusion from the above argument is

Proposition 21.5. *Consider system (21.31). It is non-regular feedback linearizable, if*

(i) $\xi(x) \in \mathrm{Span}\{g_1, \cdots, g_m\}$*;*
(ii) there exists a nonzero component vector b, such that

$$b \in \mathrm{Span}\{g_1, \cdots, g_m\}.$$

When conditions of Proposition 21.5 are not satisfied, we may use another normal form to investigate the linearization problem again.

According to Lemma 21.4, we can always find a vector field $\eta(x)$, such that

$$\mathrm{ad}_{Ax}\,\eta(x) = \xi(x). \tag{21.32}$$

Define a local diffeomorphism $z_1 = x - \eta(x)$. Then on coordinate chart z_1 the system (21.31) can be expressed as

$$\dot{z}_1 = Az_1 - J_0(x)\xi(x) + \sum_{i=1}^{m} g_i^1(x)u_i, \tag{21.33}$$

where $J_0(x)$ is the Jacobi matrix of $\eta(x)$. Moreover, $g_i^1(x) = (I - J_0(x))g_i(x)$.

For notational convenience, we denote by $x := z_0$, $\xi(x) := \xi_0(x)$, $\eta(x) := \eta_0(x)$, $g_i(x) := g_i^0(x)$. Hence, we can continue the previous procedure to define a new coordinate chart by

$$\mathrm{ad}_{Ax}(\eta_k) = \xi_k, \quad z_{k+1} = z_k - \eta_k(x), \quad k \geq 0,$$

and a new vector field as

$$g_i^{k+1}(x) = (I - J_k(x))\, g_i^k(x), \quad 1 \leq i \leq m, \quad k \geq 0,$$

where $J_k(x)$ is the Jacobi matrix of $\eta_k(x)$. Using it, one sees easily that under the coordinate chart z_k the system can be expressed as

$$\dot{z}_k = Az_k + \xi_k(x) + \sum_{i=1}^{m} g_i^k(x)u_i, \quad k \geq 1. \tag{21.34}$$

Summarizing the above argument, we have

Corollary 21.1. *System (21.31) is non-regular state feedback linearizable, if there exists $k \geq 0$, such that (21.34) verify the conditions (i) and (ii) of Proposition 21.5.*

From the recursive calculation one sees that

$$deg(\xi_i) = c_{i+1} + 1, \quad i = 0, 1, \cdots,$$

where $\{c_i\}$ is the Fibonacci series, i.e., $(c_1, c_2, \cdots) = (1, 1, 2, 3, 5, 8, \cdots)$. Hence when $k \to \infty$ we have $\xi_k(x) \to 0$, because we assume it converges. Hence, we have

Corollary 21.2. *System (21.31) is non-regular state feedback linearizable, if there exists a nonzero component of constant vector b, such that*

$$b \in \operatorname{Span}\left\{ \prod_{i=0}^{\infty} (I - J_i(x)) g_j(x),\ j = 1, \cdots, m \right\}.$$

21.6 Algorithm for Non-Regular Feedback Linearization

This section first provides a formula to realize the Poincaré coordinate transformation (21.22), then the necessary and sufficient conditions will be given for (approximate) linearization.

To begin with, applying Taylor expansion to $f(x)$, system (21.21) can be expressed as

$$\dot{x} = Ax + F_2x^2 + F_3x^3 + \cdots, \tag{21.35}$$

where F_k are $n \times n^k$ constant matrices.

Next, we assume

$$\mathrm{ad}_{Ax}\, \eta_k = F_k x^k.$$

Using Lemma 21.4, we can obtain that

$$\eta_k = (\Gamma_k^n \odot F_k)\, x^k, \quad x \in \mathbb{R}^n. \tag{21.36}$$

Here $\odot$ is the Hadamard product. (We refer to Chapter 1 for it.) According to (21.30), Γ_k^n can be constructed as

$$(\Gamma_k^n)_{ij} = \frac{1}{\left(\sum\limits_{s=1}^{n} \alpha_s^j \lambda_s\right) - \lambda_i}, \quad i = 1, \cdots, n; \; j = 1, \cdots, n^k \tag{21.37}$$

where $\alpha_1^j, \cdots, \alpha_n^j$ are the powers of $x_1, \cdots, x_n$ respectively in the jth component of x^k.

To make the definition of the parameters $\{\alpha_t^j\}$ clear, we give an example for this.

Example 21.2. Assume $n = 3$ and $k = 2$. Then we have

$$x^2 = [x_1^2 \; x_1x_2 \; x_1x_3 \; x_2x_1 \; x_2^2 \; x_2x_3 \; x_3x_1 \; x_3x_2 \; x_3^2]^T.$$

Let $j = 1$. That is, we consider the first component of x^2, which is x_1^2. Hence we have:

$$\alpha_1^1 = 2, \quad \alpha_2^1 = 0, \quad \alpha_3^1 = 0.$$

Similarly, we have

$$\begin{aligned}
&\alpha_1^2 = 1, \alpha_2^2 = 1, \alpha_3^2 = 0,\\
&\alpha_1^3 = 1, \alpha_2^3 = 0, \alpha_3^3 = 1,\\
&\alpha_1^4 = 1, \alpha_2^4 = 1, \alpha_3^4 = 0,\\
&\alpha_1^5 = 0, \alpha_2^5 = 2, \alpha_3^5 = 0,\\
&\alpha_1^6 = 0, \alpha_2^6 = 1, \alpha_3^6 = 1,\\
&\alpha_1^7 = 1, \alpha_2^7 = 0, \alpha_3^7 = 1,\\
&\alpha_1^8 = 0, \alpha_2^8 = 1, \alpha_3^8 = 1,\\
&\alpha_1^9 = 0, \alpha_2^9 = 0, \alpha_3^9 = 2.
\end{aligned}$$

Then we have the following main result.

Theorem 21.6. *Assume A satisfies A1, then system (21.35) can be transformed via the coordinate transformation*

$$z = x - \sum_{i=2}^{\infty} E_i x^i \tag{21.38}$$

to the linear form

$$\dot{z} = Az, \tag{21.39}$$

where E_i can be determined by the following recursive formula.

$$\begin{aligned}
&E_2 = \Gamma_2 \odot F_2,\\
&E_s = \Gamma_s \odot \left(F_s - \sum_{i=2}^{s-1} E_i \Phi_{i-1}(I_{n^{i-1}} \otimes F_{s+1-i})\right), \; s \geq 3.
\end{aligned} \tag{21.40}$$

(In the above formula Φ_i is defined in (18.73).)

Proof. Applying (21.38) to the system yields

$$\begin{aligned}\dot{z} &= \left(Ax + \sum_{i=2}^{\infty} F_i x^i\right) - \sum_{i=2}^{\infty} \frac{\partial E_i x^i}{\partial x} \left(Ax + \sum_{i=2}^{\infty} F_i x^i\right)\\ &= Az + \sum_{i=2}^{\infty} F_i x^i + A\sum_{i=2}^{\infty} E_i x^i - \sum_{i=2}^{\infty} \frac{\partial E_i x^i}{\partial x} Ax\\ &\quad - \left(\sum_{i=2}^{\infty} \frac{\partial E_i x^i}{\partial x}\right)\left(\sum_{j=2}^{\infty} F_j x^j\right)\\ &= Az - \sum_{i=2}^{\infty} \mathrm{ad}_{Ax}(E_i x^i) + F_2 x^2 + \sum_{s=3}^{\infty} \left(F_s x^s - \right.\\ &\quad \left. \sum_{i=2}^{s-1} \frac{\partial E_i x^i}{\partial x} F_{s+1-i} x^{s+1-i}\right)\\ &:= Az - \sum_{i=2}^{\infty} \mathrm{ad}_{Ax}(E_i x^i) + \sum_{s=2}^{\infty} L_s,\end{aligned} \tag{21.41}$$

where

$$\begin{aligned}L_2 &= F_2 x^2\\ L_s &= F_s x^s - \sum_{i=2}^{s-1} \frac{\partial E_i x^i}{\partial x} \ltimes F_{s+1-i} x^{s+1-i}\\ &= \left(F_s - \sum_{i=2}^{s-1} E_i \Phi_{i-1}(I_{n^{i-1}} \otimes F_{s+1-i})\right) x^s,\ s \geq 3.\end{aligned} \tag{21.42}$$

Because of assumption A1 we can define

$$E_s x^s = \mathrm{ad}_{Ax}^{-1}(L_s), \quad s = 2, 3, \cdots.$$

Hence, (21.35) becomes (21.39). □

The advantage of this Taylor series expansion is that we can get the linear form directly, without calculating the infinite times of coordinate transformations z_i, $i = 1, 2, 3, \cdots$.

Next, we consider the linearization of system (21.18). Denote $A = \frac{\partial f}{\partial x}|_0$, $B = g(0)$, and assume (A, B) is completely controllable. Then we can find feedback coefficient matrix K and a linear coordinate transformation T, such that $\tilde{A} = T^{-1}(A + BK)T$ satisfies A1. For statement ease, the above transformation is called a non-resonant (NR) transformation.

Using the above notations and computations, the following result is obvious.

Theorem 21.7. *System (21.18) is single input feedback linearizable, if and only if there exists an NR transformation and a component nonzero constant* *vector b such that*

$$b \in \mathrm{Span}\left\{ \left(I - \sum_{i=2}^{\infty} E_i \Phi_{i-1} x^{i-1} \right) g_j \;\middle|\; j = 1, \cdots, m \right\}. \tag{21.43}$$

In the following we consider approximate linearization, which is practically useful.

Definition 21.6. System (21.18) is said to be kth order non-regular (NR-k) feedback approximately linearizable, if there exist a state feedback and a local coordinate chart z, such that within this chart the closed-loop system becomes

$$\dot{z} = Az + O(\|z\|^{k+1}) + (b + O(\|z\|^{k}))v, \tag{21.44}$$

where (A, b) is completely controllable.

For approximate linearization, the non-resonant requirement can be relaxed a little bit.

Definition 21.7. Let $\lambda = \{\lambda_1, \cdots, \lambda_n\}$ be the set of eigenvalues of A. A is kth order resonant, if there exists $m = (m_1 \cdots, m_n) \in \mathbb{Z}_+^n$ with $2 \leq |m| \leq k$, such that for certain $1 \leq s \leq n$, we have $\lambda_s = \langle m, \lambda \rangle$.

From equation (21.37) it is ready to verify the following result, which is a corollary of Poincaré's theorem.

Corollary 21.3. *Consider an analytic system (21.21). If A is kth order non-resonant, then there exists a coordinate transformation (21.22), which transforms the system (21.21) into an approximately linear system*

$$\dot{z} = Az + O(\|z\|^{k+1}). \tag{21.45}$$

If we consider the kth order approximate linearization of system (21.35), we need only to adjust (21.38) to

$$z = x - \sum_{i=2}^{k} E_i x^i. \tag{21.46}$$

Then the formulas in (21.40) remain available for $s \leq k$. Moreover, (21.39) becomes

$$\dot{z} = Az + O(||x||^{k+1}). \tag{21.47}$$

We call a transformation kth order non-resonant transformation, if it is the same as NR transformation except the condition of non-resonant is replaced by the one of kth order non-resonant.

Theorem 21.8. *System (21.4) is kth order single input feedback approximately linearizable, if and only if there exists an NR-k transformation and a component nonzero constant vector b, such that*

$$b \in \text{Span}\left\{ \left(I - \sum_{i=2}^{k} E_i \Phi_{i-1} x^{i-1} \right) g_j \,\middle|\, \forall j \right\} + O(\|x\|^k). \tag{21.48}$$

We use the following example to depict the linearization process.

Example 21.3. Consider the 4th order feedback approximate linearization of the following system

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} &= \begin{bmatrix} -4\sin x_1 - \frac{2}{3}x_1^3 + 5x_2^2 + 6x_2^3 \\ -5x_2 - 3x_3^2 \\ -6x_3 \end{bmatrix} \\ &\quad + \begin{bmatrix} 0 \\ 6(1+x_3) \\ 7 \end{bmatrix} u_1 + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u_2. \end{aligned} \tag{21.49}$$

Using Taylor series expansion, we can express system (21.49) as

$$\begin{aligned} \dot{x} &= \begin{bmatrix} -4 & 0 & 0 \\ 0 & -5 & 0 \\ 0 & 0 & -6 \end{bmatrix} x + \begin{bmatrix} 5x_2^2 \\ -3x_3^2 \\ 0 \end{bmatrix} + \begin{bmatrix} 6x_2^3 \\ 0 \\ 0 \end{bmatrix} \\ &\quad + O(\|x\|^5) + \begin{bmatrix} 0 \\ 6 + 6x_3 \\ 7 \end{bmatrix} u_1 + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u_2. \end{aligned} \tag{21.50}$$

It is easy to calculate that

$$\begin{aligned} L_2 &= (5x_2^2,\ -3x_3^2,\ 0)^T, \\ E_2 x^2 &= ad_{Ax}^{-1}(L_2) = (-\tfrac{5}{6}x_2^2,\ \tfrac{3}{7}x_3^2,\ 0)^T; \\ L_3 &= (6x_2^3 - 5x_2x_3^2,\ 0,\ 0)^T, \\ E_3 x^3 &= ad_{Ax}^{-1}(L_3) = \left(-\tfrac{6}{11}x_2^3 + \tfrac{5}{13}x_2x_3^2,\ 0,\ 0\right)^T. \end{aligned}$$

The expected coordinate transformation is

$$z = x - \begin{bmatrix} -\frac{5}{6}x_2^2 - \frac{6}{11}x_2^3 + \frac{5}{13}x_2x_3^2 \\ \frac{3}{7}x_3^2 \\ 0 \end{bmatrix}. \tag{21.51}$$

Under this new coordinate frame (21.49) can be expressed as

$$\dot{z} = \begin{bmatrix} -4 & 0 & 0 \\ 0 & -5 & 0 \\ 0 & 0 & -6 \end{bmatrix} z + O(||x||^5) + \begin{bmatrix} h(x) & 1 \\ 6 & 0 \\ 7 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \tag{21.52}$$

where $h(x) = (6 + 6x_3)(\frac{5}{3}x_2 + \frac{18}{11}x_2^2 - \frac{5}{13}x_3^2) - \frac{70}{13}x_2x_3$. Since

$$\begin{bmatrix} 1 \\ 6 \\ 7 \end{bmatrix} = \begin{bmatrix} h(x) \\ 6 \\ 7 \end{bmatrix} \times 1 + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \times (-h(x) + 1),$$

Theorem 21.8 assures that the system is 4th order single input feedback approximately linearizable.

Choosing state feedback control as

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -h(x) + 1 \end{bmatrix} v. \tag{21.53}$$

Plugging it into (21.52), we have

$$\dot{z} = \begin{bmatrix} -4 & 0 & 0 \\ 0 & -5 & 0 \\ 0 & 0 & -6 \end{bmatrix} z + O(||x||^5) + \begin{bmatrix} 1 \\ 6 \\ 7 \end{bmatrix} v, \tag{21.54}$$

which is the 4th order single input approximately linearized system of the system (21.49).

Exercises

21.1 Using chain rule, we have

$$\frac{d}{dt}x^k = D(x^k) \times \frac{d}{dt}x.$$

Using this equation and the formulas (18.72)–(18.73) to provide an alternative proof of formula (21.4).

21.2 Consider system (21.1).

$$\begin{bmatrix} \dot{x} \\ \dot{x}^2 \\ \vdots \\ \dot{x}^k \end{bmatrix} = A_k \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^k \end{bmatrix}, \tag{21.55}$$

where

$$A_k = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1k} \\ 0 & A_{22} & \cdots & A_{2k} \\ \vdots & & & \\ 0 & 0 & \cdots & A_{kk} \end{bmatrix}$$

is called the kth approximated Carleman linearization (CL) of (21.1). Consider the system

$$\begin{cases} \dot{x}_1 = x_1 + x_2^2 \\ \dot{x}_2 = -x_2 + x_1 x_2. \end{cases}$$

Find its (i) second approximated CL A_2; (ii) third approximated CL A_3; (iii) 4th approximated CL A_4.

21.3 As a linear system, (21.55) can easily be solved. Then we can find an approximate solution as

$$x(t) \approx B_1(t)x(0) + B_2(t)x^2(0) + \cdots + B_k(t)x^k(0),$$

where $[B_1(t), B_2(t), \cdots, B_k(t)]$ is the first row of $e^{A_k t}$.

Consider the following system

$$\begin{cases} \dot{x}_1 = x_2 + x_1 x_2 \\ \dot{x}_2 = -x_1 - 2x_2 - x_1^2. \end{cases} \tag{21.56}$$

(i) Find the 3rd approximated CL of (21.56).

(ii) Find an approximate solution using the 3rd approximated CL.

21.4 Given two infinite-dimensional upper triangular matrices

$$A = \begin{bmatrix} 1\,0\,1\,0\,1 \cdots \\ 0\,1\,0\,1\,0 \cdots \\ 0\,0\,1\,0\,1 \cdots \\ \ddots \end{bmatrix}, \quad B = \begin{bmatrix} 0\,1\,0\,1\,0 \cdots \\ 0\,0\,1\,0\,1 \cdots \\ 0\,0\,0\,1\,0 \cdots \\ \ddots \end{bmatrix}.$$

Calculate the following. (If you are not able to find precise form, find (i) $k = 3$, (ii) $k = 4$, (iii) $k = 5$, approximated forms.

(i) $A \times B = ?$

(ii) $A^{-1} = ?$

(iii) $e^B = ?$

21.5 When $i = 3$, give a direct proof for Theorem 21.2. Precisely, show that

$$\sigma(A_{33}) = \{\lambda_{k_1} + \lambda_{k_2} + \lambda_{k_3} \mid k_1, k_2, k_3 = 1, \cdots, n\}.$$

21.6 Consider the system

$$\begin{cases} \dot{x}_1 = -x_2 + x_1x_2 \\ \dot{x}_2 = x_1 - x_1x_2. \end{cases} \tag{21.57}$$

Check if it has a first integral of the form $H(x,t) = e^{-\xi t}P(x)$. If “yes”, solve it.

21.7 Consider the system (21.15) with $x \in \mathbb{R}^3$, and assume an invariance is of the form

$$H(x,t) = e^{\xi t}x_1^{\alpha}x_2^{\beta}x_3^{\gamma}(P_0 + P_1x + \cdots + P_{\ell}x^{\ell}).$$

Find the corresponding algebraic equations for $\{\alpha, \beta, \gamma, P_0, P_1, \cdots, P_{\ell}\}$.

21.8 When $h(t,x) = h(x)$ is time-invariant, then (21.13)–(21.14) becomes the Lie derivative

$$L_f(h(x)) = 0. \tag{21.58}$$

Assume

$$f(x) = \begin{bmatrix} e^{x_1+x_2+x_3} \\ x_2 \\ x_3 \end{bmatrix}.$$

(i) Find a time-invariant $h(x)$ satisfying (21.58).

(ii) Prove that there are two linearly independent $h_1(x)$ and $h_2(x)$, satisfying (21.58). (Hint: Linear independence of $h_1(x)$ and $h_2(x)$ means $dh_1(x)$ and $dh_2(x)$ are linearly independent.)

21.9 Complete the proof of Proposition 21.4.

21.10 Consider the following system

$$\dot{x} = f(x) + g_1(x)u_1 + g_2(x)u_2, \quad x \in \mathbb{R}^3 \tag{21.59}$$

where

$$f(x) = \begin{bmatrix} x_2e^{x_1} \\ x_3e^{x_2} \\ x_1x_2x_3 \end{bmatrix}, \; g_1(x) = \begin{bmatrix} x_3 \\ e^{x_2} \\ \cos(x_1) \end{bmatrix}, \; g_2(x) = \begin{bmatrix} x_3e^{x_1} \\ e^{x_1+x_2} \\ 0 \end{bmatrix}.$$

(i) Show that the system (21.59) is not regular feedback linearizable.

(ii) Show that the system is non-regular feedback linearizable.

(Hint: Consider $K = \begin{bmatrix} 1 \\ -e^{-x_1} \end{bmatrix}$.)

21.11 Check which one(s) of the following matrices is(are) non-resonant.

$$A = \begin{bmatrix} 12 & -6 & -8 \\ -4 & 3 & 3 \\ 19 & -10 & -13 \end{bmatrix}, \quad B = \begin{bmatrix} -11 & 5 & 10 \\ 19 & -8 & -15 \\ -27 & 12 & 23 \end{bmatrix}, \quad C = \begin{bmatrix} -4 & 1 & 4 \\ 12 & -4 & -9 \\ -13 & 4 & 11 \end{bmatrix}.$$

21.12 Consider the following system

$$\begin{cases} \dot{x}_1 = x_1 + x_2 - x_1^2 \\ \dot{x}_2 = \ln(1 - 2x_1). \end{cases} \tag{21.60}$$

(i) Prove that there exists a local coordinate transformation $x = y + h(y)$, such that under coordinates of y (21.60) can be expressed locally as

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = A \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

(ii) Find the coordinate transformation.

21.13 Let

$$\mathrm{ad}_{Ax}\,\eta = g(x),$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -3 & 3 \end{bmatrix}, \quad g(x) = \begin{bmatrix} x_1x_2 \\ -2x_3^2 \\ x_2x_3 \end{bmatrix}.$$

Solve η.

21.14 Consider the following system

$$\begin{cases} \dot{x}_1 = 4\ln(1 + x_3) - 5x_1 - 4x_2 \\ \dot{x}_2 = -4\ln(1 + x_3) + 6x_1 + 5x_2 \\ \dot{x}_3 = (1 + x_3)\ln(1 + x_3). \end{cases} \tag{21.61}$$

(i) Find A such that (21.61) can be expressed as

$$\dot{x} = Ax + HOT,$$

where $HOT = O(\|x^2\|)$ represents the higher order terms.

(ii) Check that A is resonant.

(iii) Find a (local) coordinate transformation to convert (21.61) into the form

$$\dot{y} = Ay.$$

(Hint: Consider the following transformation.)

$$\begin{cases} y_1 = \frac{x_1+x_2}{2} \\ y_2 = \frac{x_1-x_2}{2} \\ y_3 = \ln(1 + x_3) - x_1. \end{cases}$$

(Note that this example shows that the Poincaré Theorem is only a sufficient condition. The necessary and sufficient condition for linearization via coordinate transformation is still an open problem.)

21.15 Consider the following system

$$\begin{cases} \dot{x}_1 = \ln(1 + x_2) \\ \dot{x}_2 = 3\sin(x_2 - x_1) + x_1. \end{cases} \tag{21.62}$$

(i) Express (21.62) into Taylor expansion form as

$$\dot{x} = Ax + F_2 x^2 + F_3 x^3 + F_4 x^4 + O(\|x^5\|).$$

(ii) Calculate

$$\Gamma_k^2, \quad k = 1, 2, 3, 4.$$

(iii) Calculate

$$E_i, \quad i = 2, 3, 4.$$

(iv) Define a coordinate transformation as

$$z = x - \sum_{i=2}^{4} E_i x^i.$$

Using it to the original system (21.62) to see what we can get.

Note that this example shows that truncated form of formula (21.38) can be used for approximate linearization.

21.16 Consider the system (21.59) with

$$f(x) = \begin{bmatrix} -4x_1 - \frac{2}{3}x_1^3 + 5x_2^2 + 6x_2^3 \\ -5x_2 - 3x_3^2 \\ -6x_3 \end{bmatrix}, \ g_1(x) = \begin{bmatrix} 0 \\ 1 + x_3 \\ \frac{7}{6} \end{bmatrix}, \ g_2(x) = \begin{bmatrix} e^{x_1} \\ 0 \\ 0 \end{bmatrix}.$$

Check whether it is 4th order feedback approximate linearizable. If "yes", linearize it.

This page intentionally left blank

Chapter 22

Stability Region of Dynamic Systems

Consider a dynamic system. The stability region of a stable equilibrium plays very important role in practice, because the stability region is the allowed working area of an engineering system. Particularly, consider a power system, there are many working points (stable equilibriums), and investigating their stability region is fundamental for the safety of the system.

22.1 Stability Region

Consider the following nonlinear dynamic system,

$$\dot{x} = f(x), \quad x \in \mathbb{R}^n, \tag{22.1}$$

where $f(x)$ is an analytic field.

Definition 22.1. Let x_e be an equilibrium of (22.1).

(1) The stable and unstable submanifold of x_e, denoted by $W^s(x_e)$, is defined as

$$W^s(x_e) = \left\{ p \in \mathbb{R}^n \;\middle|\; \lim_{t \to \infty} x(t,p) \to x_e \right\}. \tag{22.2}$$

(2) The unstable submanifold of x_e, denoted by $W^u(x_e)$, is defined as

$$W^u(x_e) = \left\{ p \in \mathbb{R}^n \;\middle|\; \lim_{t \to -\infty} x(t,p) \to x_e \right\}. \tag{22.3}$$

Definition 22.2.

(1) Let x_s be a stable equilibrium of (22.1). The region of attraction of x_s is defined as

$$A(x_s) = \left\{ p \in \mathbb{R}^n \;\middle|\; \lim_{t \to \infty} x(t,p) \to x_s \right\}. \tag{22.4}$$

The boundary of $A(x_s)$ is denoted by $\partial A(x_s)$.

(2) An equilibrium x_e is said to be hyperbolic, if the Jacobi matrix of f at x_e, $J_f(x_e)$ has no zero real part eigenvalues.
(3) A hyperbolic equilibrium is said to be of type-k, if $J_f(x_e)$ has exactly k positive real part eigenvalues.

The following result is fundamental for our approach.

Theorem 22.1 (Zaborszky *et al.*, 1988; Chiang *et al.*, 1988). *Consider system (22.1). Assume x_s is a stable equilibrium, satisfying the following three assumptions*

(i) the equilibriums on $\partial A(x_s)$ are all hyperbolic;
(ii) the stable and unstable submanifolds of the equilibriums on $\partial A(x_s)$ are transversal;
(iii) each trajectory on $\partial A(x_s)$ converges to an equilibrium as $t \to \infty$.

Then the boundary of the stability region consists of the stable submanifolds of the unstable equilibriums on the boundary.

Fig. 22.1 illustrates this.

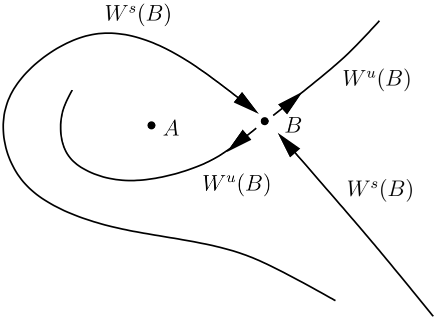


Fig. 22.1 Boundary of stability region

Note that two submanifolds N and S of a manifold M are said to be transversal, if for any $x \in N \cap S$, the union of the tangent spaces of these two submanifolds is the tangent space of M. Precisely,

$$T_x(N) \cup T_x(S) = T_x(M), \quad x \in N \cap S.$$

It is well known that (Chiang *et al.*, 1988) if the state manifold is of dimension n, then the boundary of the stability region is of dimension $n-1$. Hence, the boundary is basically generated by the stable submanifolds of type-1 equilibriums. Based on this consideration, the stable submanifolds of type-1 equilibriums are particularly important. There are many algorithms to calculate approximation of the stable sun-manifold of type-1 equilibrium.

The purpose of this chapter is to explore the Taylor expansion of the equation describing the type-1 unstable submanifold. Particularly, it can be used to obtain a best quadratic approximation, comparing previously existing results.

22.2 Stable Submanifold

In this section we search for the equation of the stable submanifold of type-1 equilibrium.

Without loss of generality, we assume $x_u = 0$ is a type-1 equilibrium. Wright down the Taylor series expansion of the $f(x)$ in (22.1) as

$$f(x) = \sum_{i=1}^{\infty} F_i x^i = Jx + F_2 x^2 + \cdots , \tag{22.5}$$

where $F_1 = J = J_f(0)$, and $F_i = \frac{1}{i!} D^i f(0)$ are known $n \times n^i$ matrices.

We use A^{-T} for the inverse of A^T. Matrix A is said to be hyperbolic if it has no zero real part eigenvalue.

Lemma 22.1. *Let A be a hyperbolic matrix. Denote by V_s and V_u the stable and unstable submanifolds of A respectively, and by U_s and U_u the stable and unstable submanifolds of A^{-T}. Then*

$$V_s^{\perp} = U_u, \quad V_u^{\perp} = U_s. \tag{22.6}$$

Proof. Assume A is of the type-k, then we can convert A into a Jordan canonical form as

$$Q^{-1} A Q = \begin{bmatrix} J_s & 0 \\ 0 & J_u \end{bmatrix},$$

where J_s and J_u are stable and unstable blocks respectively. Splitting $Q = \begin{bmatrix} Q_1 \; Q_2 \end{bmatrix}$, where Q_1 and Q_2 are consisted by the first $n-k$ and the last k columns of Q. Then

$$V_s = \operatorname{Span} col\{Q_1\}, \quad V_u = \operatorname{Span} col\{Q_2\}.$$

It is easy to see that

$$Q^T A^{-T} Q^{-T} = \begin{bmatrix} J_s^{-T} & 0 \\ 0 & J_u^{-T} \end{bmatrix}.$$

Similarly, splitting $Q^{-T} = \left[\tilde{Q}_1\ \tilde{Q}_2\right]$, where $\tilde{Q}_1$ and $\tilde{Q}_2$ consist of the first $n-k$ and last k columns of Q^{-T} respectively, we have

$$U_s = \mathrm{Span}\, col\{\tilde{Q}_1\}, \quad U_u = \mathrm{Span}\, col\{\tilde{Q}_2\}.$$

The conclusion follows from $Q^{-1}Q = I$. □

The following corollary is an immediate consequence of the above lemma.

Corollary 22.1. *Let A be a matrix of type-1 with its unique unstable eigenvalue μ. Assume η is the eigenvector of A^T with respect to μ, then η is perpendicular to the stable subspace of A.*

***Proof*.** Since the only unstable eigenvalue of A^{-T} is $\frac{1}{\mu}$, denote by η the eigenvector of A^{-T} corresponding to this unstable eigenvalue. Then by Lemma 22.1 Span$\{\eta\} = U_u = V_s^{\perp}$. Hence we need only to prove that η is also the eigenvector of A^T with respect to μ. Since

$$A^{-T}\eta = \frac{1}{\mu}\eta \Rightarrow A^T\eta = \mu\eta,$$

the claim follows. □

Without loss of generality, we assume the unstable equilibrium $x_u = 0$ of the system (22.1), concerned in the sequel, is of type-1.

The following theorem provides a necessary and sufficient condition for the stable submanifold of a type-1 equilibrium.

Theorem 22.2. *Let $x_u = 0$ be an equilibrium of type-1 of the system (22.1).*

$$W^s(e_u) = \{x \mid h(x) = 0\}. \tag{22.7}$$

Then $h(x)$ is uniquely determined by the following equations (22.8)–(22.10).

$$h(0) = 0, \tag{22.8}$$

$$h(x) = \eta^T x + O(\|x\|^2), \tag{22.9}$$

$$L_f h(x) = \mu h(x), \tag{22.10}$$

where $L_f h(x)$ is the Lie derivative of $h(x)$ with respect to f, η is the eigenvector of $J_f^T(0)$ with respect to its unique positive eigenvalue μ.

Proof. (Necessity) The necessity of (22.8) and (22.9) are obvious. We need only to prove the necessity of (22.10). First, note that

$$\frac{\partial h}{\partial x} = \eta^T + O(\|x\|). \tag{22.11}$$

Hence, there exists a neighborhood U of the origin, such that

$$\text{rank}(h(x)) = 1, \quad x \in U. \tag{22.12}$$

Since $W^s(e_u)$ is f invariant, we have

$$\begin{cases} h(x) = 0, \\ L_f h(x) = 0, \quad x \in W^s(e_u). \end{cases} \tag{22.13}$$

Since $\dim(W^s(e_u)) = n - 1$, we have

$$\text{rank}\left(\begin{bmatrix} h(x) \\ L_f h(x) \end{bmatrix}\right) = 1,$$

this implies that $h(x)$ and $L_f h(x)$ are linearly dependent. A straightforward computation shows that

$$L_f h(x) = \eta^T J_f(0) x + O(\|x\|^2) = \mu \eta^T x + O(\|x\|^2).$$

Hence for $x \in U$, the linear dependence of $h(x)$ and $L_f h(x)$ yields (22.10). Finally, because of the analyticity of the system, we conclude that (22.10) is globally correct.

(Sufficiency) First, we prove that if $h(x)$ satisfies (22.8)$-$(22.10), then locally we have

$$\{x \in U \mid h(x) = 0\}$$

is the stable submanifold over U. According to the rank condition (22.12), we know that (refer to Boothby (1986), Theorem 5.8)

$$V := \{x \in U \mid h(x) = 0\}$$

is an $(n-1)$-dimensional submanifold.

Next, since $L_f h(x) = 0$, V is locally f invariant. Finally, (22.9) shows that zero is locally the asymptotically stable equilibrium of $f|_V$, which is the restriction of f on V. Hence, locally V is the stable submanifold of (22.1). But the stable submanifold is unique (Carr, 1981), it follows that locally $V = W^s(e_u)$.

Since the system is analytic, $\{x \mid h(x) = 0\}$ conincides globally with $W^s(e_u)$. □

22.3 Quadratic Approximation

In general, it is not easy to figure out the equation $h(x)$ of the stable submanifold. The quadratic approximation of the boundary of the stability region has been investigated by several authors (Venkatasubramanian and Ji, 1997; Saha *et al.*, 1997). This section provides a quadratic approximation of $h(x)$. The precise formula is provided, which is the unique approximation with the error of $O(\|x\|^3)$.

Denote the Taylor series expansion of $h(x)$ as

$$h(x) = H_1x + H_2x^2 + H_3x^3 + \cdots = H_1x + \frac{1}{2}x^T\Psi x + H_3x^3 + \cdots . \tag{22.14}$$

In the above we use two forms to express the quadratic terms: the semi-tensor product form H_2x^2 and the standard quadratic form $\frac{1}{2}x^T\Psi x$, where $\Psi = \mathrm{Hess}(h(0))$ is the Hessian matrix of $h(x)$ at $x = 0$, and $H_2 = V_c^T(\frac{1}{2}\Psi)$ is the row stacking form of $\frac{1}{2}\Psi$.

Note that for a real function $f(x, y) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ its Hessian matrix is

$$\mathrm{Hess}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial y_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial y_m} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial y_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial y_m} \end{bmatrix}.$$

Lemma 22.2. *Assume* 0 *is the type-1 equilibrium of (22.1). Then the quadratic terns of (22.14) satisfies*

$$\Psi\left(\frac{\mu}{2}I - J\right) + \left(\frac{\mu}{2}I - J^T\right)\Psi = \sum_{i=1}^{n} \eta_i \,\mathrm{Hess}(f_i(0)), \tag{22.15}$$

where μ *and* η *are as in Corollary 22.1,* $\mathrm{Hess}(f_i)$ *is the Hessian matrix of the* i*th component of* f*.*

Proof. First, the linear approximation of $h(x) = 0$ is

$$H_1x = 0,$$

which is the tangent space of the stable submanifold $W^s(x_u)$. Since η is perpendicular to $W^s(x_u)$ at x_u, we have $H_1 = \eta$.

According to Theorem 22.2, the Lie derivative satisfying

$$L_f h(x) = 0.$$

Using (18.72), we have

$$Dh(x) = H_1 + H_2\Phi_1 x + H_3\Phi_2 x^2 + \cdots = H_1 + x^T\Psi + H_3\Phi_2 x^2 + \cdots .$$

Note that the vector field f can be expressed as

$$f(x) = Jx + \frac{1}{2}\begin{bmatrix} x^T\,\mathrm{Hess}(f_1(0))x \\ \vdots \\ x^T\,\mathrm{Hess}(f_n(0))x \end{bmatrix} + O(\|x\|^3).$$

Calculating $L_f h$ out yields

$$\begin{aligned} L_f h &= \eta^T Jx + x^T\left(\tfrac{1}{2}\sum_{i=1}^n \eta_i\,\mathrm{Hess}(f_i(0)) + \Psi J\right)x + O(\|x\|^3) \\ &= \mu\eta^T x + x^T\left(\tfrac{1}{2}\sum_{i=1}^n \eta_i\,\mathrm{Hess}(f_i(0)) + \Psi J\right)x + O(\|x\|^3). \end{aligned} \tag{22.16}$$

Observing that as the invariant submanifold of f, we have

$$W^s(e_u) = \{x \mid h(x) = 0,\ L_f h(x) = 0\}. \tag{22.17}$$

Applying (22.14) and (22.17) to $W^s(e_u)$ yields

$$x^T\left(\frac{1}{2}\sum_{i=1}^n \eta_i\,\mathrm{Hess}(f_i(0)) + \Psi(J - \frac{\mu}{2}I)\right)x + O(\|x\|^3) = 0. \tag{22.18}$$

Converting the quadratic form into the symmetric form, we then have (22.15). □

Lemma 22.3. *Equation (22.15) has unique symmetric solution.*

***Proof*.** Express (22.15) into a linear system as

$$(A\otimes I_n + I_n\otimes A)V_c(\Psi) = V_c\left(\sum_{i=1}^n \eta_i\,\mathrm{Hess}(f_i(0))\right), \tag{22.19}$$

where

$$A = \frac{\mu}{2}I - J^T.$$

(22.19) is the linear form of Lyapunov mapping. Hence, let $\lambda_i \in \sigma(A)$, $i = 1, \cdots, n$ be the eigenvalues of A. Then the eigenvalues of $A\otimes I_n + I_n\otimes A$ are

$$\{\lambda_i + \lambda_j \,|\, 1 \le i, j \le n,\ \lambda_t \in \sigma(A)\}.$$

(We refer to Chapter 3 for Lyapunov mapping and its properties.)

To show $A \otimes I_n + I_n \otimes A^T$ is nonsingular, it suffices to show that all $\lambda_i + \lambda_j \neq 0$. Let $\xi_i \in \sigma(J)$, $i = 1, \cdots, n$ be the eigenvalues of J. Then

$$\lambda_i = \frac{\mu}{2} - \xi_i, \quad i = 1, \cdots, n.$$

Observing the eigenvalues of J, it is easy to see that the only negative eigenvalue of A is $-\frac{\mu}{2}$, and all other eigenvalues of A have positive real parts, which are greater than $\frac{\mu}{2}$. It follows that

$$\lambda_i + \lambda_j \neq 0, \quad 1 \leq i,\ j \leq n.$$

Hence (22.15) has unique solution. Finally, we prove the solution is symmetric. It is ready to verify that

$$(A \otimes I_n + I_n \otimes A)W_{[n]} = W_{[n]}(A \otimes I_n + I_n \otimes A). \tag{22.20}$$

Using (22.20), we have

$$\begin{aligned} &(A \otimes I_n + I_n \otimes A)V_r(\Psi) = (A \otimes I_n + I_n \otimes A)W_{[n]}V_c(\Psi) \\ &= W_{[n]}(A \otimes I_n + I_n \otimes A)V_c(\Psi) = W_{[n]}V_c\left(\sum_{i=1}^n \eta_i \operatorname{Hess}(f_i(0))\right) \\ &= V_r\left(\sum_{i=1}^n \eta_i \operatorname{Hess}(f_i(0))\right) = V_c\left(\sum_{i=1}^n \eta_i \operatorname{Hess}(f_i(0))\right). \end{aligned} \tag{22.21}$$

The last equality comes from the fact that $\sum_{i=1}^n \xi_i \operatorname{Hess}(f_i(0))$ is a symmetric matrix, hence its row and column stacking forms are the same. (22.21) shows that $V_r(\Psi)$ is another solution of (22.19). But the solution of (22.19) is unique, which leads to

$$V_r(\Psi) = V_c(\Psi).$$

That is, Ψ is symmetric. □

Denote by V_c^{-1} the inverse mapping of V_c, which retrieves A from its column stacking form $V_c(A)$.

Summarizing the Lemmas 22.1–22.3, we have the following result about the quadratic approximation of the stable submanifold.

Theorem 22.3. *Assume $x_u = 0$ is the type-1 equilibrium of system (22.1), and its stable submanifold is determined by $h(x) = 0$. Then*

$$h(x) = H_1 x + \frac{1}{2}x^T \Psi x + O(\|x\|^3), \tag{22.22}$$

where

$$\begin{cases} H_1 = \eta^T \\ \Psi = V_c^{-1}\left\{\left[(\frac{\mu}{2}I_n - J^T)\otimes I_n + I_n \otimes (\frac{\mu}{2}I_n - J^T)\right]^{-1} V_c \left(\sum_{i=1}^{n} \eta_i \operatorname{Hess}(f_i(0))\right)\right\}, \end{cases}$$

μ and η are defined as in Corollary 22.1 with respect to $J = F_1$, Hess(f_i) *is the Hessian matrix of the ith component f_i of f.*

Remark 22.1. If e_u is an equilibrium of type-$(n-1)$, μ is the unique negative eigenvalue, and its corresponding eigenvector is η, then all the above arguments remain available for describing the unstable submanifold. Particularly, (22.22) is the quadratic approximation of the equation of unstable submanifold.

Observing (22.18), the following corollary is an immediate consequence, which is sometimes useful for simplifying computations.

Corollary 22.2. *Assume*

$$\sum_{i=1}^{n} \eta_i \operatorname{Hess}(f_i(0)) \left(\frac{\mu}{2}I_n - J\right)^{-1}$$

is symmetric, then the quadratic approximation of the equation of stable submanifold is

$$h(x) = \eta^T x + \frac{1}{4}x^T \sum_{i=1}^{n} \eta_i \operatorname{Hess}(f_i(0)) \left(\frac{\mu}{2}I_n - J\right)^{-1} x = 0. \tag{22.23}$$

Example 22.1. Consider the system

$$\begin{cases} \dot{x}_1 = x_1, \\ \dot{x}_2 = -x_2 + x_1^2, \quad x \in \mathbb{R}^2. \end{cases} \tag{22.24}$$

Its stable and unstable submanifolds are respectively (reported in Saha *et al.* (1997))

$$W^s(0) = \{x \in \mathbb{R}^2 \mid x_1 = 0\},$$
$$W^u(0) = \{x \in \mathbb{R}^2 \mid x_2 = \tfrac{1}{3}x_1^2\}.$$

We use them to verify formula (22.23). For (22.24), we have

$$J = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

For stable submanifold $W^s(0)$, it is easy to verify that its stable eigenvalue is $\mu = 1$, its corresponding eigenvector is $\eta = (1\ 0)^T$. Moreover,

$$\mathrm{Hess}(f_1(0)) = 0, \quad \mathrm{Hess}(f_2(0)) = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}.$$

Hence

$$\frac{1}{4}\sum_{i=1}^{2} \eta_i \,\mathrm{Hess}(f_i(0)) \left(\frac{1}{2}I - J\right)^{-1} = 0,$$

that is,

$$h_s(x) = (1\ 0)x + 0 + O(\|x\|^3) = x_1 + O(\|x\|^3).$$

For unstable submanifold $W^u(0)$, it is easy to check that its unstable eigenvalue is $\mu = -1$, its corresponding eigenvector is $\eta = (0\ 1)^T$. Hence

$$\frac{1}{4}\sum_{i=1}^{2} \eta_i \,\mathrm{Hess}(f_i(0)) (\frac{-1}{2}I - J)^{-1} = \begin{bmatrix} -\frac{1}{3} & 0 \\ 0 & 0 \end{bmatrix}.$$

That is,

$$h_u(x) = (0\ 1)x + x^T \begin{bmatrix} -\frac{1}{3} & 0 \\ 0 & 0 \end{bmatrix} x + O(\|x\|^3) = x_2 - \frac{1}{3}x_1^2 + O(\|x\|^3).$$

In fact, if we use the conclusion in the next section, we can prove that the errors for the approximations $h_s(x)$ and $h_u(x)$ are both 0. Alternatively, we can also use Theorem 22.2 to verify this directly. For instance, we verify $h_u(x)$: Assume $h_u(x) = x_2 - \frac{1}{3}x_1^2$, then $W^u(e_u) = \{x \,|\, h_u(x) = 0\}$, if and only if $h_u(x) = 0$ implies $L_f h_u(x) = 0$. This is true, because

$$L_f(h_u(x)) = \begin{bmatrix} -\frac{2}{3}x_1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ -x_2 + x_1^2 \end{bmatrix} = -x_2 + \frac{1}{3}x_1^2 = -h_u(x).$$

22.4 Higher Order Approximation

This section considers the Taylor series expansion of the equation of stability submanifold. In the following calculation we need Φ_k. To calculate it, the following proposition is necessary.

Proposition 22.1.

$$W_{[n^s,n]} = \prod_{i=0}^{s-1} \left(I_{n^i} \otimes W_{[n,n]} \otimes I_{n^{s-i-1}}\right). \tag{22.25}$$

Proof. Using Proposition 2.12, we have

$$W_{[n^s,n]} = \left(W_{[n^{s-1},n]} \otimes I_n\right)\left(I_{n^{s-1}} \otimes W_{[n,n]}\right).$$

Using the first decomposition of Proposition 2.12 repeatedly, we finally obtain (22.25). Note that as a convention we have $I_{n^0} = 1$, $\Phi_0 = I_n$. □

Using (22.25), it is easy to calculate Φ_k. We use an example to depict it.

Example 22.2. Assume $n = 2$, then

$$\Phi_0 = I_n,$$

$$\Phi_1 = W_{[n,n]} + I \otimes W_{[1,n]} = W_{[n]} + I_{n^2} = \begin{bmatrix} 2&0&0&0\\ 0&1&1&0\\ 0&1&1&0\\ 0&0&0&2 \end{bmatrix},$$

$$\begin{aligned} \Phi_2 &= W_{[n^2,n]} + I_n \otimes W_{[n,n]} + I_{n^2} \otimes W_{[1,n]} \\ &= \left(W_{[n]} \otimes I_n\right)\left(I_n \otimes W_{[n]}\right) + I_n \otimes W_{[n]} + I_{n^3} \\ &= \begin{bmatrix} 3&0&0&0&0&0&0&0\\ 0&1&2&0&0&0&0&0\\ 0&1&1&0&1&0&0&0\\ 0&0&0&2&0&0&1&0\\ 0&1&0&0&2&0&0&0\\ 0&0&0&1&0&1&1&0\\ 0&0&0&0&0&2&1&0\\ 0&0&0&0&0&0&0&3 \end{bmatrix}, \quad \cdots. \end{aligned}$$

Next, we proceed to solve H_k from equations (22.8)–(22.10). First problem is: since x^k is a redundant generator of B_n^k, we are not able to get unique solution from (22.8)–(22.10). To overcome this difficulty we have to convert the equations to natural basis. Recall Chapter 15, let $S \in \mathbb{Z}_+^n$. The natural basis is defined as

$$N_n^k = \{x^S \mid S \in \mathbb{Z}_+^n,\ |S| = k\}.$$

We arrange the elements in N_n^k in the alphabetic order. That is, for $S^1 = (s_1^1, \cdots, s_n^1)$ and $S^2 = (s_1^2, \cdots, s_n^2)$ we use order $x^{S^1} \prec x^{S^2}$, if there exists a t, $1 \le t \le n-1$, such that

$$s_1^1 = s_1^2,\ \cdots,\ s_t^1 = s_t^2,\ s_{t+1}^1 > s_{t+2}^2.$$

In this way we arrange the elements of N_n^k as a column and denote it as $x_{(k)}$.

Example 22.3. Let $n = 3$ and $k = 2$. Then

$$x^2 = (x_1^2, x_1x_2, x_1x_3, x_2x_1, x_2^2, x_2x_3, x_3x_1, x_3x_2, x_3^2)^T,$$

and

$$x_{(2)} = (x_1^2, x_1x_2, x_1x_3, x_2^2, x_2x_3, x_3^2)^T.$$

In Chapter 16 it has been proved that the size of B_n^k is

$$|B_n^k| := d = \frac{(n+k-1)!}{k!(n-1)!}, \quad k \geq 0, \quad n \geq 1. \tag{22.26}$$

In Chapter 15 we have defined two matrices: $T_N(n,k) \in M_{n^k \times d}$ and $T_B(n,k) \in M_{d \times n^k}$, which can convert two generators x^k and $x_{(k)}$ back and forth. Precisely,

$$x^k = T_N(n,k)x_{(k)}, \quad x_{(k)} = T_B(n,k)x^k.$$

Moreover,

$$T_B(n,k)T_N(n,k) = I_d.$$

Recall (22.14), instead of solving H_k, we will try to solve G_k, which satisfies

$$H_kx^k = G_kx_{(k)}.$$

In addition, H_k is a symmetric coefficient matrix, if any two equal elements in x^k have the equal coefficients in H_kx^k. We use the following example to explain it.

Example 22.4. Let $n = 3$ and $k = 2$. Then x^2 is as in Example 22.3. For a given second order homogeneous polynomial $p(x) = x_1^2 + 2x_1x_2 - 3x_1x_3 + x_2^2 - x_3^2$, we can express it as

$$p(x) = H_1x^2 = (1,\ 2,\ -3,\ 0,\ 1,\ 0,\ 0,\ 0,\ -1)x^2.$$

Alternatively, we can also express it as

$$p(x) = H_2x^2 = \left(1,\ 1,\ -\frac{3}{2},\ 1,\ 1,\ 0,\ -\frac{3}{2},\ 0,\ -1\right)x^2.$$

It is easy to see that H_1 is not symmetric, while H_2 is.

We also know the following.

Proposition 22.2.

(1) The symmetric coefficient matrix H_k is unique.
(2)

$$H_k = G_k T_B(n,k), \quad G_k = H_k T_N(n,k). \tag{22.27}$$

Now we are ready to construct the higher terms of the equation of stable submanifold $h(x)$. Denote by

$$f(x) = F_1 x + F_2 x^2 + \cdots ;$$

and

$$h(x) = H_1 x + H_2 x^2 + \cdots .$$

Note that we already known that $F_1 = J_f(0) = J$, $H_1 = \eta^T$, and H_2 can be uniquely determined by (22.17).

Proposition 22.3. *The coefficients H_k, $k \geq 2$, of $h(x)$ satisfy the following equations*

$$\left[\sum_{i=1}^{k} H_i \Phi_{i-1}(I_{n^{i-1}} \otimes F_{k-i+1}) - \mu H_k\right] x^k = 0, \quad k \geq 2. \tag{22.28}$$

Proof. Note that $h(x) = 0$ is invariant with respect to vector field $f(x)$, that is, the Lie derivative

$$L_f h(x) = 0. \tag{22.29}$$

Using Proposition 18.4, we have

$$Dh(x) = H_1 + H_2\Phi_1 x + H_3\Phi_2 x^2 + \cdots = H_1 + 2x^T\Psi + H_3\Phi_2 x^2 + \cdots .$$

A straightforward computation shows

$$\begin{aligned} L_f h(x) = {} & \mu\eta^T x + [H_2\Phi_1(I_n \otimes F_1) + H_1 F_2]\, x^2 + \cdots \\ & + \left[\sum_{i=1}^{k} H_i \Phi_{i-1}(I_{n^{i-1}} \otimes F_{k+1-i})\right] x^k + \cdots . \end{aligned}$$

Note that $h(x)$ satisfies

$$\begin{cases} h(x) = 0, \\ L_f h(x) = 0. \end{cases} \tag{22.30}$$

Subtracting μ times the first equation of (22.30) from its second equation, we can prove, inductively on k, that

$$\left[\sum_{i=1}^{k} H_i \Phi_{i-1}(I_{n^{i-1}} \otimes F_{k-i+1}) - \mu H_k\right] x^k + O(\|x\|^{k+1}) = 0, \quad k \geq 2.$$

The conclusion follows. □

Observing (22.28), according to Proposition 22.2, it can be expressed as

$$\begin{aligned} & G_k \left[\mu I_d - T_B(n,k)\Phi_{k-1}(I_{n^{k-1}} \otimes F_1)T_N(n,k)\right] x_{(k)} \\ & = \left[\sum_{i=1}^{k-1} G_i T_B(n,i)\Phi_{i-1}(I_{n^{i-1}} \otimes F_{k-i+1})\right] T_N(n,k)x_{(k)}, \quad k \geq 3. \end{aligned} \tag{22.31}$$

The following theorem is a summarization of the above arguments, which can be used for general case.

Theorem 22.4. *Assume the matrix*

$$C_k := \mu I_d - T_B(n,k)\Phi_{k-1}(I_{n^{k-1}} \otimes F_1)T_N(n,k), \quad k \geq 3 \tag{22.32}$$

is nonsingular, then

$$G_k = \left[\sum_{i=1}^{k-1} G_i T_B(n,i)\Phi_{i-1}(I_{n^{i-1}} \otimes F_{k-i+1})\right] T_N(n,k)C_k^{-1}. \tag{22.33}$$

Remark 22.2. In fact, H_2 can also be solved in this way. (22.15) and (22.33) can produce the same result. In fact, when H_2 is solved from (22.15), since the symmetric quadratic equation is used, the symmetry of the coefficients has been automatically assured.

It is obvious that the efficiency of (22.33) depends on whether C_i is nonsingular. Unfortunately, unlike quadratic case, we are not able to assure it. The following example shows that (i) if $C_i \neq 0$ the above formula works well; (ii) $C_i \neq 0$ is not assured.

Example 22.5. Consider the following system

$$\begin{cases} \dot{x}_1 = -cx_1, \quad c > 0, \\ \dot{x}_2 = x_2 - 2x_1^2 + x_1^3, \end{cases} \tag{22.34}$$

where $c > 0$ is a parameter.

We calculate the equation of the stable submanifold. It is easy to calculate that $\mu = 1$, $\eta = (0\ 1)^T$,

$$J = \begin{bmatrix} -c & 0 \\ 0 & 1 \end{bmatrix},$$

and

$$\mathrm{Hess}(f_1(0)) = 0, \quad \mathrm{Hess}(f_2(0)) = \begin{bmatrix} -4 & 0 \\ 0 & 0 \end{bmatrix}.$$

Hence we can use (22.23) to calculate that

$$h(x) = (0\ 1)x + x^T \begin{bmatrix} -\frac{2}{2c+1} & 0 \\ 0 & 0 \end{bmatrix} x + O(\|x\|^3). \tag{22.35}$$

Using (18.58), (18.59), and the Φ_2 calculated in Example 22.2, we can calculate C_3 as

$$C_3 = \begin{bmatrix} 3c+1 & 0 & 0 & 0 \\ 0 & 2c & 0 & 0 \\ 0 & 0 & c-1 & 0 \\ 0 & 0 & 0 & -2 \end{bmatrix}. \tag{22.36}$$

Assume $c \neq 1$, then C_3 is invertible. Then we have

$$H_1 = (0,\ 1), \quad H_2 = \left(-\frac{2}{2c+1},\ 0,\ 0,\ 0\right),$$

$$F_2 = \begin{bmatrix} 0 & 0\,0\,0 \\ -2 & 0\,0\,0 \end{bmatrix}, \quad F_3 = \begin{bmatrix} 0\,0\,0\,0\,0\,0\,0\,0 \\ 1\,0\,0\,0\,0\,0\,0\,0 \end{bmatrix}.$$

Plugging them into (22.32) yields

$$G_3 = \left(\frac{1}{3c+1},\ 0,\ 0,\ 0\right).$$

Hence we have

$$h(x) = x_2 - \frac{2}{2c+1}x_1^2 + \frac{1}{3c+1}x_1^3 + O(\|x\|^4).$$

In fact, it is easy to verify that

$$h(x) = x_2 - \frac{2}{2c+1}x_1^2 + \frac{1}{3c+1}x_1^3 = 0,$$

and

$$L_f h(x) = h(x) = 0.$$

We conclude that

$$W^s(0) = \left\{ x \in \mathbb{R}^2 \;\middle|\; x_2 - \frac{2}{2c+1}x_1^2 + \frac{1}{3c+1}x_1^3 = 0 \right\}.$$

According to Theorem 22.4 and Example 22.5, we give the following algorithm:

Algorithm 6.

Step 1. If $C_3, \cdots, C_{k-1}$ are nonsingular, we continue to search H_k to approximate $h(x)$ till the accuracy is satisfied.

Step 2. If C_k is singular, we search for the least square solution G_k via

$$\begin{aligned} & G_k \left[\mu I_d - T_B(n,k)\Phi_{k-1}(I_{n^{k-1}} \otimes F_1)T_N(n,k)\right] \\ & = \left[\sum_{i=1}^{k-1} G_i T_B(n,i)\Phi_{i-1}(I_{n^{i-1}} \otimes F_{k-i+1})\right] T_N(n,k). \end{aligned} \tag{22.37}$$

Then use $G_3, \cdots, G_k$ to construct a kth order approximation of $h(x)$.

Step 3. (possible further improvement) If the least square solution is a real number solution for (22.37), solve the following system:

$$\begin{cases} G_k \left[\mu I_d - T_B(n,k)\Phi_{k-1}(I_{n^{k-1}} \otimes F_1)T_N(n,k)\right] \\ \quad = \left[\sum_{i=1}^{k-1} G_i T_B(n,i)\Phi_{i-1}(I_{n^{i-1}} \otimes F_{k-i+1})\right] T_N(n,k), \\ 0 = \left[\sum_{i=1}^{k} G_i T_B(n,i)\Phi_{i-1}(I_{n^{i-1}} \otimes F_{k-i+1})\right] T_N(n,k+1). \end{cases} \tag{22.38}$$

In fact, considering the kth and $(k+1)$th order terms leads to (22.38). Recall Example 22.5. When $c = 1$, the least square solution is

$$G_3 = \left(\frac{1}{3c+1},\ 0,\ t,\ 0\right),$$

where t is an arbitrary parameter. It is ready to verify that G_3 is a real number solution of (22.37). Hence, we can try to solve (22.38). A careful calculation shows that (22.38) has a solution $G_3 = \left(\frac{1}{3c+1},\ 0,\ 0,\ 0\right)$. It is easy to check that this G_3 is a real number solution.

In the following we consider another more general example.

Example 22.6. Consider the following system

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -x_1 - 2x_2, \\ \dot{x}_3 = 2x_3 - x_2(e^{x_1} - 1). \end{cases} \tag{22.39}$$

It is easy to check that $\mu = 2$, $\eta = (0\ 0\ 1)^T$,

$$J = \begin{bmatrix} 0 & 1 & 0 \\ -1 & -2 & 0 \\ 0 & 0 & 2 \end{bmatrix},$$

$$A = \frac{\mu}{2} I_3 - J^T = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 3 & 0 \\ 0 & 0 & -1 \end{bmatrix},$$

$$\text{Hess}(f_1(0)) = \text{Hess}(f_2(0)) = 0,$$

$$\text{Hess}(f_3(0)) = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Using formula (22.22), we have

$$\begin{aligned} h(x) &\approx \eta^T x + x^T(\tfrac{1}{2}\Psi)x \\ &= (0\ 0\ 1)x + x^T \begin{bmatrix} 0.09375 & -0.09375 & 0 \\ -0.09375 & -0.03125 & 0 \\ 0 & 0 & 0 \end{bmatrix} x \qquad (22.40) \\ &= x_3 + 0.09375x_1^2 - 0.1875x_1x_2 - 0.03125x_2^2. \end{aligned}$$

To calculate the third order terms, we verify C_3. Using (22.32), we have

$$C_3 = \begin{bmatrix} 2 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 6 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 2 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -4 \end{bmatrix}.$$

It can be calculated and verified to be invertible via computer. From the quadratic part of $h(x)$ we have

$$H_1 = \eta^T = (0,\ 0,\ 1),$$

$$H_2 = (0.09375,\ -0.09375,\ 0,\ -0.09375,\ -0.03125,\ 0,\ 0,\ 0,\ 0),$$

$F_2 \in M_{3\times 9}$ has all zero components except $F_2(3,2)$ and $F_2(3,4)$ which are

$$F_2(3,2) = F_2(3,4) = -\frac{1}{2},$$

$F_3 \in M_{3\times 29}$ has all zero components except 3 elements: $F_3(3,2)$, $F_3(3,4)$ $F_3(3,10)$, which are

$$F_3(3,2) = F_3(3,4) = F_3(3,10) = -\frac{1}{6}.$$

Plugging them into (22.33) yields

$$G_3 = (0.0408,\ -0.0816,\ 0,\ -0.0256,\ 0,\ 0,\ -0.0032,\ 0,\ 0,\ 0).$$

Hence the equation of the stable submanifold, approximated to third order terms, is

$$\begin{aligned} h(x) \approx\ & x_3 + 0.09375x_1^2 - 0.1875x_1x_2 - 0.03125x_2^2 + 0.0408x_1^3 \\ & - 0.0816x_1^2x_2 - 0.0256x_1x_2^2 - 0.0032x_2^3. \end{aligned} \tag{22.41}$$

Continuing this process, we can calculate the even higher order terms of $h(x)$.

In fact, for this special system the stable submanifold can be obtained by a suitable coordinate transformation. Hence the above result can be precisely verified.

22.5 Differential-Algebraic System

This section considers the stability region of a differential-algebraic system. Such systems exist widely. For instance, the power network is of this type. Consider the following system

$$\begin{cases} \dot{x} = f(x,y), & x \in \mathbb{R}^n,\ y \in \mathbb{R}^m, \\ \Phi(x,y) = 0, & \Phi(x,y) \in \mathbb{R}^m, \end{cases} \tag{22.42}$$

where $f(0,0) = 0$, $\Phi(0,0) = 0$. Moreover, the dynamics determined by this set of equations is assumed to be unique, hence we require that

$$\operatorname{rank}\left(\frac{\partial \Phi}{\partial y}(0,0)\right) = m. \tag{22.43}$$

Based on the aforementioned reason, we assume $(0,0)$ is a type-1 unstable equilibrium. We will use the result obtained in the previous sections to deduce the equation of the stable submanifold. For convenience, we consider only the quadratic approximation of the (22.42). Higher order terms can be calculated in a similar way.

According to the implicit function theorem, (22.43) implies that y can be solved from the second equation of (22.42) as $y = y(x)$. Substituting it into the first equation of (22.42) yields an equation of the form of (22.1) as

$$\dot{x} = f(x, y(x)). \tag{22.44}$$

Of course, equation (22.44) exists locally. But the Taylor series expansion requires only local information, hence local expression is enough. Now the only obstacle is solving $y = y(x)$, which is, in general, impossible. Recall (22.15), what do we need is only

$$\begin{aligned} & J := \tfrac{\partial f(x,y(x))}{\partial x}, \\ & H_i := Hess(f_i(0,0)), \quad i = 1, \cdots, n. \end{aligned} \tag{22.45}$$

Hence instead of solving y, we can calculate J and H_i, and then the formula (22.15) can be used to find the quadratic approximation.

Since

$$\frac{\partial \Phi}{\partial x} + \frac{\partial \Phi}{\partial y}\frac{\partial y}{\partial x} = 0,$$

then

$$\frac{\partial y}{\partial x} = -\left(\frac{\partial \Phi}{\partial y}\right)^{-1}\frac{\partial \Phi}{\partial x}. \tag{22.46}$$

Using chain rule, we have

$$J = \frac{\partial f}{\partial x}(0,0) - \frac{\partial f}{\partial y}(0,0)\left(\frac{\partial \Phi}{\partial y}(0,0)\right)^{-1}\frac{\partial \Phi}{\partial x}(0,0). \tag{22.47}$$

Recall Corollary 18.1, let $A(x)$ and $B(x)$ be $p \times q$ and $q \times r$ functional matrices. Then

$$DA(x)B(x) = DA(x)B(x) + A(x)DB(x). \tag{22.48}$$

Moreover, according to the chain rule, we have

$$DA(x, y(x)) = D_x A(x,y) + D_y A(x,y)\left(I_n \otimes \frac{\partial y}{\partial x}\right). \tag{22.49}$$

Here we use D_x (D_y) to express the differential with respect to x (y) only.

Now we calculate H_i. First, the gradient of f_i can be expressed as

$$\begin{aligned}\nabla f_i(x, y(x)) &= \nabla_x f_i(x,y) + \left(d_y f_i(x,y)\tfrac{\partial y}{\partial x}\right)^T \\ &= \nabla_x f_i(x,y) - \left(\tfrac{\partial \Phi}{\partial x}\right)^T\left(\tfrac{\partial \Phi}{\partial y}\right)^{-T}\nabla_y f_i(x,y).\end{aligned} \tag{22.50}$$

Since $y = y(x)$ is a function of x , we use $\nabla_x f_i(x,y)$ and $\nabla_y f_i(x,y)$ for the gradients with respect to x and y respectively.

Then, by definition we have

$$H_i = D(\nabla f_i)\big|_{(0,0)}, \quad i = 1, \cdots, n. \tag{22.51}$$

Applying (22.49) to the first term of (22.50), we have

$$D\left(\nabla_x f_i\right) = \frac{\partial^2 f_i}{\partial x \partial x} + \frac{\partial^2 f_i}{\partial x \partial y}\left(\frac{\partial y}{\partial x}\right). \tag{22.52}$$

Similarly, we have

$$D\left(\nabla_y f_i\right) = \frac{\partial^2 f_i}{\partial y \partial x} + \frac{\partial^2 f_i}{\partial y \partial y}\left(\frac{\partial y}{\partial x}\right). \tag{22.53}$$

Note that hereafter for any function $\xi(x,y)$, we use $\frac{\partial^2\xi(x,y)}{\partial x\partial y}$ to represent an $n\times m$ matrix, which has its (i,j)th element as $\frac{\partial^2\xi}{\partial x_i\partial y_j}$. Hence in general,

$$\frac{\partial^2\xi(x,y)}{\partial x\partial y}=\left[\frac{\partial^2\xi(x,y)}{\partial y\partial x}\right]^T.$$

Applying (22.49) to the second term of (22.50), we have

$$\begin{aligned}D\left[d_yf_i(x,y)\tfrac{\partial y}{\partial x}\right]^T &= D\left\{\left(\tfrac{\partial y}{\partial x}\right)^T(\nabla_yf_i(x,y))\right\}\\ &= D\left(\tfrac{\partial y}{\partial x}\right)^T(\nabla_yf_i(x,y)\otimes I_n)+\left(\tfrac{\partial y}{\partial x}\right)^T D\left(\nabla_yf_i(x,y)\right).\end{aligned}\tag{22.54}$$

Next, we calculate (22.54) term by term. Using (22.46), we have

$$\left(\frac{\partial y}{\partial x}\right)^T\left(\frac{\partial\Phi}{\partial y}\right)^T+\left(\frac{\partial\Phi}{\partial x}\right)^T=0.$$

Differentiating both sides of the above equation and applying (22.48) yield

$$D\left(\frac{\partial y}{\partial x}\right)^T\left[\left(\frac{\partial\Phi}{\partial y}\right)^T\otimes I_n\right]+\left(\frac{\partial y}{\partial x}\right)^T D\left(\frac{\partial\Phi}{\partial y}\right)^T+D\left(\frac{\partial\Phi}{\partial x}\right)^T=0.\tag{22.55}$$

Each terms in (22.55) are calculated as follows:

$$\begin{aligned}X &:= D\left(\tfrac{\partial\Phi}{\partial x}\right)^T\Big|_{(0,0)}\\ &= \left[\left(\tfrac{\partial^2\Phi_1}{\partial x\partial x}+\tfrac{\partial^2\Phi_1}{\partial x\partial y}\cdot\tfrac{\partial y}{\partial x}\right),\cdots,\left(\tfrac{\partial^2\Phi_m}{\partial x\partial x}\quad+\tfrac{\partial^2\Phi_m}{\partial x\partial y}\cdot\tfrac{\partial y}{\partial x}\right)\right]\Big|_{(0,0)}.\end{aligned}\tag{22.56}$$

$$\begin{aligned}Y &:= D\left(\tfrac{\partial\Phi}{\partial y}\right)^T\Big|_{(0,0)}\\ &= \left[\left(\tfrac{\partial^2\Phi_1}{\partial y\partial x}+\tfrac{\partial^2\Phi_1}{\partial y\partial y}\cdot\tfrac{\partial y}{\partial x}\right),\cdots,\left(\tfrac{\partial^2\Phi_m}{\partial y\partial x}\ +\ \tfrac{\partial^2\Phi_m}{\partial y\partial y}\cdot\tfrac{\partial y}{\partial x}\right)\right]\Big|_{(0,0)}.\end{aligned}\tag{22.57}$$

Substituting (22.56) and (22.57) into (22.55) yields

$$D\left(\frac{\partial y}{\partial x}\right)^T\Bigg|_{(0,0)}=-\left[\left(\frac{\partial y}{\partial x}(0,0)\right)^TY+X\right]\left[\left(\frac{\partial\Phi}{\partial y}(0,0)\right)^{-T}\otimes I_n\right].\tag{22.58}$$

Finally, substituting (22.52), (22.53), and (22.58) into (22.51), we can get the expression of H_i as follows:

$$\begin{aligned}H_i=\tfrac{\partial^2f_i}{\partial x\partial x}+\tfrac{\partial^2f_i}{\partial x\partial y}\left(\tfrac{\partial y}{\partial x}\right)-\left[\left(\tfrac{\partial y}{\partial x}\right)^TY+X\right]\left[\left(\tfrac{\partial\Phi}{\partial y}\right)^{-T}\otimes I_n\right](\nabla_yf_i\otimes I_n)\\ +\left(\tfrac{\partial y}{\partial x}\right)^T\left[\tfrac{\partial^2f_i}{\partial y\partial x}+\tfrac{\partial^2f_i}{\partial y\partial y}\left(\tfrac{\partial y}{\partial x}\right)\right],\end{aligned}\tag{22.59}$$

where X, Y and the detailed expression of $\frac{\partial y}{\partial x}$ are (22.56), (22.57), and (22.46) respectively.

Exercises

22.1 Consider a linear system

$$\dot{x} = Ax, \quad x \in \mathbb{R}^n. \tag{22.60}$$

(i) When $x = 0$ is a hyperbolic equilibrium?

(ii) Check which of the following matrices is hyperbolic?

$$A = \begin{bmatrix} 0 & 1 & -1 \\ 1 & 2 & 1 \\ 3 & -1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & -1 & -1 \\ 2 & 2 & 1 \\ 0 & -1 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & -1 \\ 4 & 2 & 1 \\ -3 & -1 & 0 \end{bmatrix}.$$

22.2 Consider the following system

$$\dot{x} = \begin{bmatrix} -1 & -1 & -1 \\ 4 & 4 & 3 \\ -4 & -5 & -4 \end{bmatrix} x.$$

Show that zero is a hyperbolic equilibrium. Find the stable and unstable submanifolds, (which are degenerated to subspaces).

22.3 Check whether 0 is the hyperbolic equilibrium of the following systems. If "yes", find type k for the system.

(i)

$$\begin{cases} \dot{x}_1 = \sin(-x_1 + x_3 - x_4) \\ \dot{x}_2 = -2\ln(1 + x_1 + x_2) + 3x_3 - \tan(x_4) \\ \dot{x}_3 = 2e^{x_3 + x_4 - x_1 - x_2} + x_4 - 2 \\ \dot{x}_4 = x_3 + x_4 - x_1 - x_2(1 + x_2). \end{cases}$$

(ii)

$$\begin{cases} \dot{x}_1 = e^{x_2 + x_3} - 4x_4 - 1 \\ \dot{x}_2 = \sin(3x_3 - 7x_4) \\ \dot{x}_3 = \ln(1 + 2x_3) - 3x_4 \\ \dot{x}_4 = \tan(x_3 - 2x_4). \end{cases}$$

(iii)

$$\begin{cases} \dot{x}_1 = \sin(x_1 - x_3) + 3x_4 \\ \dot{x}_2 = 2(x_1 - x_2) - \sin(x_3) + 7\tan(x_4) \\ \dot{x}_3 = 2\ln(x_1 - x_2 - x_3 + 5x_4) + x_4 \\ \dot{x}_4 = \tan(x_1 - x_2 - x_3) + 5x_4. \end{cases}$$

22.4 Assume 0 is a type-1 hyperbolic equilibrium and the stable submanifold of 0 is determined by $h(x) = 0$. Then the linear part of $h(x)$ is $\eta^T x$, where η is an eigenvector of A^T with respect to the unique unstable eigenvalue of A (equivalently, A^T). But the eigenvector is not unique. Does the choice of different eigenvectors affects the result? Explain why?

22.5 Consider the following system

$$\begin{cases} \dot{x}_1 = -\sin(x_1 + x_2 + x_3) \\ \dot{x}_2 = 4\ln(1 + x_1 + x_2 + x_3) - x_3 \\ \dot{x}_3 = x_1^2 - 4x_1 - 5x_2 - 4x_3. \end{cases} \tag{22.61}$$

(i) Show that 0 is a type-1 hyperbolic equilibrium.

(ii) Assume the stable submanifold of 0 is $h(x) = 0$, find its linear part $\eta^T x$.

(iii) Find its quadratic part $x^T \Psi x$.

22.6 The dynamics of a pendulum is described as (Khalil, 1996)

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -\sin(x_1) - 0.5x_2. \end{cases} \tag{22.62}$$

Prove that

(i) $(0, 0)$ is an stable equilibrium;

(ii) $(-\pi, 0)$ and $(0, \pi)$ are two unstable equilibriums on the stability boundary.

(iii) Calculate the stable submanifolds of the stable submanifolds of the two unstable equilibriums (1) up to quadratic degree; (2) up to cubic degree.

22.7 Consider the following system

$$\begin{cases} \dot{x}_1 = -4x_2 - x_3 - (x_1 + x_2)(x_3 + \frac{x_3^2}{2} + \frac{x_3^3}{6}) \\ \dot{x}_2 = -4x_1 - x_3 + (x_1 + x_2)(x_3 + \frac{x_3^2}{2} + \frac{x_3^3}{6}) \\ \dot{x}_3 = x_1 + x_2. \end{cases} \tag{22.63}$$

Prove that

(i) $(0, 0)$ is a type-1 hyperbolic equilibrium;

(ii) Let $h(x) = 0$ be its stable submanifold, calculate $h(x)$ up to cubic term. That is, express it as

$$h(x) = C_1 x + C_2 x^2 + C_3 x^3 + O(\|x^4\|).$$

22.8 Consider a linear system

$$Ax = b, \quad x \in \mathbb{R}^n,\ b \in \mathbb{R}^m. \tag{22.64}$$

When (22.64) has no solution? In this case, give the least square solution of (22.64). (Hint: Use pseudo-inverse (refer to Chapter 5).)

22.9 Consider the following differential-algebraic system

$$\begin{cases} \dot{x}_1 = \sin(x_2) \\ \dot{x}_2 = \ln(1 + x_1 + x_3) \\ x_2^3 + \sin(x_3) = 0. \end{cases} \tag{22.65}$$

(i) Show that $(x_1, x_2) = (0, 0)$ is a type-1 hyperbolic equilibrium;

(ii) Let $h(x_1, x_2) = 0$ be its stable submanifold, give the quadratic approximation of $h(x_1, x_2)$.

22.10 Consider system (22.1). Let 0 be a type-k hyperbolic equilibrium of the system. Show that the stable submanifold of zero can be expressed locally as $\{x \in U | h(x) = 0\}$, where U is a proper neighborhood of 0, and $h(x) : U \to \mathbb{R}^k$ can be expressed as

$$\begin{cases} h_1(x) = 0 \\ \vdots \\ h_k(x) = 0, \end{cases}$$

with its Jacobian matrix J has full rank. That is, $\operatorname{rank}(J) = k$, where

$$J = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \cdots & \frac{\partial h_1}{\partial x_n} \\ \vdots & & \\ \frac{\partial h_k}{\partial x_1} & \cdots & \frac{\partial h_k}{\partial x_n} \end{bmatrix}.$$

22.11 Consider system (22.1). Let 0 be a type-k hyperbolic equilibrium of the system. Denote by J_f the Jacobian matrix of F at 0. Express it into Jordan canonical form as

$$J_f = \operatorname{diag}(\Lambda_1, \Lambda_2),$$

where Λ_1 and Λ_2 correspond to its stable and unstable eigenvalue sets respectively.

Show that $h(x)$ is uniquely determined by the following condition

$$\begin{cases} L_f h(x) = \Lambda_1 h(x), \quad h(x) \in \mathbb{R}^k \\ \operatorname{rank}\left(\frac{\partial h}{\partial x}\right) = k, \quad h(0) = 0. \end{cases} \tag{22.66}$$

(Hint: Please refer to Xue *et al.* (2006).)

22.12 Consider system (22.1) with

$$f(x) = F_1 x + F_2 x^2 + \cdots.$$

Let 0 be a type-k hyperbolic equilibrium of the system and split the Jacobian matrix of f into the stable and unstable part as

$$J_f = F_1 = \operatorname{diag}(\Lambda_1, \Lambda_2).$$

Prove that the quadratic approximation of h is

$$h(x) = H_1 x + H_2 x^2 + O(\|x^3\|), \tag{22.67}$$

where H_1 satisfies

$$H_1 F_1 = \Lambda_1 H_1;$$

H_2 satisfies

$$(\Lambda_1 \otimes I_n) H_2 = H_2 \Phi_1 (I_n \otimes F_1) + H_1 F_2.$$

(Hint: (1) Φ_1 is defined in (18.73); (2) refer to Xue *et al.* (2006) for details.)

22.13 Consider the following system

$$\begin{cases} \dot{x}_1 = \sin(x_2) \\ \dot{x}_2 = 2x_1 + \ln(1 - x_2) \\ \dot{x}_3 = -3x_1 + \tan(x_3 - x_2). \end{cases} \tag{22.68}$$

(i) Show that 0 is a type-2 hyperbolic equilibrium;

(ii) Give the quadratic approximation of the stable submanifold, described as

$$\begin{cases} h_1(x) = 0 \\ h_2(x) = 0. \end{cases}$$

Appendix A

Numerical Algorithms

Since the semi-tensor product is different from the conventional matrix product, it is not convenient to use common tools, such as Matlab, to calculate the problems in this book. We, therefore, developed a toolbox to perform the numerical computations in the book. It can be found at `http://lsc.amss.ac.cn/~dcheng/STP/stp.zip`.

In this appendix, we will (i) print some basic functions to make the reader have a rough view to the toolbox; (ii) give some typical examples to show the usage of the toolbox. Please refer to the manual of the toolbox for more details.

A.1 Basic Functions

1. Calculate the STP of two matrices:

```
function c = sp(a,b)
% SP    (Left) Semi-Tensor Product of Matrices using ...
    Kronecker product
%
%   SP(A,B) is to calculate the (left) semi-tensor product ...
    of A and B.
%   The two matrices must meet the multiple dimension ...
    matching condition,
%   i.e., the number of columns of the fisrt matrix must be ...
    the divisor
%   or multiple of the number of rows of the last matrix.

if ~(isa(a,'sym') | isa(a,'double'))
    a = double(a);
end
if ~(isa(b,'sym') | isa(b,'double'))
```

```
    b = double(b);
end

if ndims(a) > 2 | ndims(b) > 2
    error('Input arguments must be 2-D.');
end

n = size(a,2);
p = size(b,1);
if n == p
    c = a*b;
elseif mod(n,p) == 0
    z = n/p;
    c = zeros(m,z*q);
    c = a*kron(b,eye(z));
elseif mod(p,n) == 0
    z = p/n;
    c = zeros(m*z,q);
    c = kron(a,eye(z))*b;
else
    error('The input arguments do not meet the multiple ...
        dimension matching condition.');
end
```

2. Calculate the STP of n (≥ 2) matrices:

```
function r = spn(varargin)
% SPN    (Left) Semi-tensor product of matrices with ...
    arbitrary number of matrices
%
%   SPN(A,B,C,...) calculates the (left) semi-tensor product ...
    of arbitrary
%   number of matrices which have the proper dimensions.

ni = nargin;
switch ni
    case 0
        error('No input arguments.')
    case 1
        r = varargin{1};
        return
    case 2
        r = sp(varargin{1},varargin{2});
        return
    otherwise
        r = sp(varargin{1},varargin{2});
        for i = 3:ni
            r = sp(r,varargin{i});
        end
```

```
end
```

3. Calculate the swap matrix $W_{[m,n]}$:

```
function w = wij(m,n)
% WIJ    Produces swap matrix
%
%   A = WIJ(N) produces an N^2-by-N^2 swap matrix.
%   A = WIJ(M,N) produces an MN-by-MN swap matrix.

if nargin == 1
    n=m;
end

d = m*n;
w = zeros(d);
for k = 1:d
    j = mod(k,n);
    if j == 0
        j = n;
    end
    i = (k-j)/n+1;
    w((j-1)*m+i,k) = 1;
end;
```

4. Calculate $V_c(A)$:

```
function v = vc(A)
% VC    produces column stacking form of a matrix
%
%   V = VC(A) produces a vector from the matrix A column by ...
    column.

v = A(:);
```

5. Calculate $V_r(A)$:

```
function v = vr(A)
% VR    Produces row stacking form of a matrix
%
%   V = VR(A) produces a vector from the matrix A row by row.

A = A';
v = A(:);
```

In the toolbox we define two classes *stp* and *lm* for performing the semi-tensor product and calculating the logical matrices respectively.

6. Create an *stp* object:

```
function m = stp(a)
% STP/STP    Semi-tensor product (STP) class constructor.
%  m = stp(A)    creates an STP object from the matrix A.
```

7. Create an *lm* object:

```
function m = lm(varargin)
% LM/LM    Logical matrix (LM) class constructor
%
%  M = LM(A)  creates an LM object from the matrix A.
%  Example: m = lm(eye(3))
%
%  M = LM(V,N)  creates an LM object from a vector V and a ...
     postive integer N
%  Example: m = lm([1,2,2,3],4)
```

A.2 Some Examples

1. Calculate the semi-tensor product:

```
% This example is to show how to perform semi-tensor product

%% using m function
x = [1 2 1 1;
     2 3 1 2;
     3 2 1 0];
y = [1 -2;
     2 -1];
r1 = sp(x,y)
r2 = spn(x,y,y)


%% using stp class
a = stp(x); % convert a double object to an stp object
b = stp(y);
c = a*b    % please compare the result with r1
c1 = a*b*b  % please compare the result with r2

% What will happen if we use (a double object multiplying an ...
    stp object)
r3 = a*y
r4 = x*b
```

```
%% The following shows more usage of stp class

% Convert an stp object to double
x1 = double(c), class(x1)

% size method for stp class
size(c)

% length method for stp class
length(c)

% subsref method for stp class
c(1,:)

% subsasgn method for stp class
c(2,1) = 3
```

2. Consider Example 15.3:

```
% Initialize
k = 2;
MN = lmn(k); % produce logical matrix for negation
MI = lmi(k); % produce logical matrix for implicaiton
MC = lmc(k); % produce logical matrix for conjunction
MD = lmd(k); % produce logical matrix for disjunction
ME = lme(k); % produce logical matrix for equivalence
MR = lmr(k); % produce logical matrix for power−reducing matrix
MU = lmu(k); % produce logical matrix for dummy matrix
options = [];

% Dynamics of Boolean network
    % A(t+1) = MC*B(t)*C(t)
    % B(t+1) = MN*A(t)
    % C(t+1) = MD*B(t)*C(t)
% Set X(t)=A(t)B(t)C(t), then
eqn = {'MC B C',
       'MN A',
       'MD B C'};

% Set the variables' order, otherwise they will be sorted in ...
    the alphabetic order
options = lmset('vars',{'A','B','C'});

% Convert the logical equations to its canonical form
[expr,vars] = stdform(strjoin(eqn),options,k);

% Calculate the network transition matrix
L = eval(expr)
```

```
% Analyze the dynamics of the Boolean network
[n,l,c,r0,T] = bn(L,k);

% Print the result
fprintf('Number of attractors: %d\n\n',n);
fprintf('Lengths of attractors:\n');
disp(l);
fprintf('\nAll attractors are displayed as follows:\n\n');
for i=1:length(c)
    fprintf('No. %d (length %d)\n\n',i,l(i));
    disp(c{i});
end
fprintf('Transient time: [T_t, T] = [%d %d]\n\n',r0,T);
```

Bibliography

Abraham, R. and Marsden, J. (1978). *Foundations of Mechanics*, 2nd edn. (Benjamin/Cummings Pub. Com. Inc.).

Agaian, S., Astola, J. and Egiazarian, K. (1995). *Binary Polynomial Transforms and Nonlinear Digital Filters* (Marcel Dekker, New York).

Agaian, S., Panetta, K., Nercessian, S. and Danahy, E. (2010). Boolean derivatives with application to edge detection for imaging systems, *IEEE Trans. Syst., Man., Man, Cybern. B, Cubern.* **40**, 2, pp. 371–382.

Akers, S. (1959). On a theory of Boolean functions, *Journal of the Society for Industrial and Applied Mathematics* **7**, 4, pp. 487–498.

Akutsu, T., Hayashida, M., Ching, W. and Ng, M. (2007). Control of Boolean networks: Hardness results and algorithms for tree structured networks, *J. Theoretical Biology* **244**, 4, pp. 670–679.

Akutsu, T., Miyano, S. and Kuhara, S. (2000). Inferring qualitative relations in genetic networks and metabolic pathways, *Bioinformatics* **16**, pp. 727–734.

Albert, R. and Barabási, A. (2000). Dynamics of complex systems: Scaling laws for the period of Boolean networks, *Physical Review Letters* **84**, 24, pp. 5660–5663.

Aldana, M. (2003). Boolean dynamics of networks with scale-free topology, *Physica D: Nonlinear Phenomena* **185**, 1, pp. 45–66.

Arnold, V. (1983). *Geometrical Methods on the Theory of Ordinary Differential Equations* (Springer-Verlag, New York).

Ashenhurst, R. (1957). The decomposition of switching functions, in *Proceedings of an International Symposium on the Theory of Switching*, pp. 74–116.

Azariadis, P. and Aspragathos, N. (2001). Computer graphics representation and transformation of geometric entities using dual unit vectors and line transformations, *Computers & Graphiscs* **25**, pp. 195–205.

Barnes, D. and Mack, J. (1975). *An Algebraic Introduction to Mathematical Logic* (Springer-Verlag, Texas).

Bates, D. and Watts, D. (1980). Relative curvature measures of nonlinearity, *Journal of the Royal Statistical Society. Series B (Methodological)* **42**, pp. 1–25.

Bates, D. and Watts, D. (1981). Parameter transformations for improved approximate confidence regions in nonlinear least squares, *The Annals of Statistics* **9**, pp. 1152–1167.

Bochmann, D. (1978). *Boolean Differential Calculus* (Larl Marx Stadt, German Democratic Pepublic).

Boothby, W. (1986). *An Introduction to Differentiable Manifolds and Riemannian Geometry*, 2nd edn. (Academic Press, Inc., Orlando).

Borota, N., Flores, E. and Osler, T. (2000). Spacetime numbers the easy way, *Mathematics and Computer Education* **34**, 2, pp. 159–168.

Brayton, R. and Khatri, S. (1999). Multi-valued logic synthesis, in *International Conference on VLSI Design* (Goa, India), pp. 196–205.

Brown, F. (2003). *Boolean Reasoning: The Logic of Boolean Equations*, 2nd edn. (Dover Pub., New York).

Burris, S. and Sankappanavar, H. (1981). *A Course in Universal Algebra*, Number 78 in Graduate Texts in Mathematics (Springer-Verlag).

Cairo, L. and Feix, M. (1992). Families of invariants of the motion for Lotka-Volterra 1st family, *J. Math. Phys.* **33**, 7, pp. 2240–2455.

Cairo, L., Feix, M. and Llibre, J. (1999). Darboux method and search of invariants for the Lotka-Volterra and complex quadratic systems, *J. Math. Phys.* **40**, 4, pp. 2074–2091.

Carlet, C. (2010). *Boolean Methods and Models in Mathematics, Computer Science, and Engineering*, chap. Boolean functions for cryptography and error-correcting codes (Cambridge Univ. Press, Cambridge), pp. 257–397.

Carr, J. (1981). *Applications of Center Manifold Theory* (Springer-Verlag).

Chartrand, G. and Zhang, P. (2005). *Introduction to Graph Theory* (McGraw-Hill Inc., New York).

Cheng, D. (2001). On Lyapunov mapping and its applications, *Commu. on Inform. Sys.* **1**, 3, pp. 255–272.

Cheng, D. (2002). *Matrix and Polynomial Approach to Dynamic Control Systems* (Science Press, Beijing).

Cheng, D. (2007). Sime-tensor product of matrices and its applications: A survey, in *Proc. 4th International Congress of Chinese Mathematicians* (Higher Edu. Press, Int. Press, Hangzhou), pp. 641–668.

Cheng, D. (2009). Input-state approach to Boolean networks, *IEEE Trans. Neural Networks* **20**, 3, pp. 512–521.

Cheng, D. (2011). Disturbance decoupling of Boolean control networks, *IEEE Trans. Aut. Contr.* **56**, 1, pp. 2–10.

Cheng, D., Feng, J. and Lv, H. (2011a). Solving fuzzy relational equations via semi-tensor product, IEEE Trans. Fuzzy System, (to appear).

Cheng, D., Guo, L. and Huang, J. (2003). On quadratic Lyapunov functions, *IEEE Trans. Aut. Contr.* **48**, 5, pp. 885–890.

Cheng, D. and Qi, H. (2007). *Semi-tensor Product of Matrices — Theory and Applications* (Science Press, Beijing), in Chinese.

Cheng, D. and Qi, H. (2009). Controllability and observability of Boolean control networks, *Automatica* **45**, 7, pp. 1659–1667.

Cheng, D. and Qi, H. (2010a). A lienar representation of dynamics of Boolean networks, *IEEE Trans. Aut. Contr.* **55**, 10, pp. 2251–2258.

Cheng, D. and Qi, H. (2010b). State-space analysis of Boolean networks, *IEEE Trans. Neural Networks* **21**, 4, pp. 584–594.

Cheng, D., Qi, H. and Li, Z. (2011b). *Analysis and Control of Boolean Networks: A Semi-tensor Product Approach* (Springer, London).

Cheng, D., Qi, H., Li, Z. and Liu, J. B. (2011c). Stability and stabilization of Boolean networks, *Int. J. Robust Nonlinear Contr.* **21**, 2, pp. 134–156.

Cheng, D., Xi, Z., Lu, Q. and Mei, S. (2000). Geometric structure of generalized controlled Hamiltonian systems and its application, *Science in China, Series E* **43**, 4, pp. 365–379.

Cheng, D. and Xu, X. (2011). Bi-decomposition of logical mappings via semi-tensor product of matrices, Preprint.

Cheng, D., Xue, W. and Huang, J. (1998). On general Hamiltonian systems, in *Proc. ICARCV'98* (Singapore), pp. 185–189.

Cheng, D. and Zhao, Y. (2011). Identification of Boolean control networks, *Automatica* **47**, 4, pp. 702–710.

Cheng, D., Zhao, Y. and Mu, Y. (2010). Strategy optimization with its application to dynamic games, in *Proc. IEEE CDC'2010*, pp. 5822–5827, doi:10.1109/CDC.2010.5717060.

Cheng, D., Zhao, Y. and Xu, X. (2011d). Matrix approach to Boolean calculus, Preprint.

Cheng, D., Zhu, Y. and Qi, H. (2009). A conjecture on the norm of Lyapunov mapping, *J. Control Theory & Applications* **7**, 1, pp. 48–50.

Cheng, H. and Thompson, S. (1997). Dual iterative displacement analysis of spatial mechanisms using the C programming language, *Mechanism and Machine Theory* **32**, 2, pp. 193–207.

Chiang, H., Hirsch, M. and Wu, F. (1988). Stability regions of nonlinear autonomous dynamical systems, *IEEE Trans. Aut. Contr.* **33**, 1, pp. 16–27.

Choudhury, M. and Mohanram, K. (2010). Bi-decomposition of large Boolean functions using blocking edge graphs, in *Proc. 2010 IEEE/ACM Int. Conf. Comput.-aided Design* (San Jose), pp. 586–591.

Clifford, W. (1873). Preliminary sketch of biquaternions, *Proc. London Mathematical Society* **4**, 64, pp. 381–395.

Cohen, D. (1989). *An Introduction to Hilbert Space and Quantum Logic* (Springer-Verlag, New York).

Curtis, H. (1962). *A New Approach to the Design of Switching Circuits* (Van Nostrand, Princeton, N.J.).

Daniell, P. (1917). The modular difference of classes, *Bull. Amer. Math. Soc.* **23**, pp. 446–450.

Datta, A., Choudhary, A., Bittner, M. and Dougherty, E. (2003). External control in Markovian genetic regulatory networks, *Machine Learning* **52**, pp. 169–191.

Datta, A., Choudhary, A., Bittner, M. and Dougherty, E. (2004). External control in Markovian genetic regulatory networks: The imperfect information case, *Bioinformatics* **20**, pp. 924–930.

Davio, M., Deschamps, J. and Thayse, A. (1978). *Discrete and Switching Functions* (McGraw-Hill, New York).

Descusse, J. and Morg, C. (1985). Decoupling with dynamic compensation for strong invertible affine nonlinear systems, *Int. J. Contr.* **43**, pp. 1385–1398.

Devanathan, R. (2001). Linearization condition through state feedback, *IEEE Trans. Aut. Contr.* **46**, 8, pp. 1257–1260.

Di Benedetto, M., Glumineau, A. and Moog, C. (1994). The nonlnear interactor and its application to imput-output decoupling, *IEEE Trans. Aut. Contr.* **39**, 9, pp. 1240–1250.

Dixon, J. and Mortimer, B. (1996). *Permutation Groups* (Springer-Verlag, London).

Drossel, B., Mihaljev, T. and Greil, F. (2005). Number and length of attractors in a critical Kauffman model with connectivity one, *Physical Review Letters* **94**, 8, p. 88701.

Dubois, D. and Prade, H. (eds.) (2000). *Fundamentals of Fuzzy Sets* (Kluwer Acad. Pub., Boston).

Duchet, P. (1995). Hypergraphs, in R. Graham, M. Grötschel and L. Lovász (eds.), *Handbook of Combinatorics 1* (MIT Press, North-Holland), pp. 381–432.

Emmott, S. (2006). *Towards 2020 Science* (Microsoft Prsearch Ltd., Cambridge).

Falb, P. and Wolovich, W. (1967). Decoupling and synthesis of multivariable control systems, *IEEE Trans. Aut. Contr.* **12**, 5, pp. 651–669.

Farrow, C., Heidel, J., Maloney, J. and Rogers, J. (2004). Scalar equations for synchronous Boolean networks with biological applications, *IEEE Trans. Neural Networks* **15**, 2, pp. 348–354.

Feng, J., Lu, H. and Cheng, D. (2011). Matrix-based approach to fuzzy logic and fuzzy control, (Submitted for pub.).

Foster, J. and Nightngale, J. (1995). *A Short Course in General Relativity* (Springer-Verlag).

Freund, E. (1975). The structure of decoupled nonlinear systems, *Int. J. Contr.* **21**, pp. 443–450.

Fudenberg, D. and Tirole, J. (1991). *Game Theory* (MIT Press, Cambridge).

Gerla, G. (2001). *Fuzzy Logic, Mathematical Tools for Approximate Reasoning* (Kluwer).

Gibbons, R. (1992). *A Primer in Game Theory* (Prentice Hall).

Glumineau, A. and Moog, C. (1992). Nonlinear Morgan's problem: case of $(p+1)$ inputs and p outputs, *IEEE Trans. Aut. Contr.* **37**, 7, pp. 1067–1072.

Goodwin, B. (1963). *Temporal Organization in Cells* (Acad. Press, New York).

Greub, W. (1978). *Multilinear Algebra*, 2nd edn. (Springer-Verlag, New York).

Greub, W. (1981). *Linear Algebra*, 4th edn. (Springer-Verlag).

Guckenheimer, J. and Holmes, P. (1983). *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields* (Springer, Berlin).

Hachtel, G. and Somenzi, F. (2000). *Logic Synthesis and Verification* (Kluwer Academic Pub.).

Hamilton, A. (1988). *Logic for Mathematicians*, revised edn. (Cambridge Univ Press, Cambridge).

Han, D., Wei, Q., Li, Z. and Sun, W. (2008). Control of oriented mechanical systems: A method based on dual quaternion, in *Proc. 17th IFAC World*

Congress (Seoul), pp. 3836–3841.

Harris, S., Sawhill, B., Wuensche, A. and Kauffman, S. (2002). A model of transcriptional regulatory networks based on biases in the observed regulation rules, *Complexity* **7**, 4, pp. 23–40.

Hartshorne, R. (1977). *Algebraic Geometry* (Springer-Verlag, New York).

Heidel, J., Maloney, J., Farrow, C. and Rogers, J. (2003). Finding cycles in synchronous Boolean networks with applications to biochemical systems, *Int. J. Bifurcation & Chaos* **13**, 3, pp. 535–552.

Heymann, M. (1968). Pole assignment in multi-input linear systems, *IEEE Trans. Aut. Contr.* **13**, 6, pp. 748–749.

Horn, R. and Johnson, C. (1991). *Topics in Matrix Analysis*, Vol. 1, 2 (Cambridge Univ. Press, Cambridge).

Hu, B. (2010). *Foundation of Fuzzy Theory*, 2nd edn. (Wuhan University Press, Wuhan), in Chinese.

Huang, J. (2004). *Nonlinear Output Regulation, Theory and Applications* (Siam, Philadelphia).

Hungerford, T. (1974). *Algebra* (Springer-Verlag).

Ideker, T., Galitski, T. and Hood, L. (2001). A new approach to decoding life: Systems biology, *Annu. Rev. Genomics Hum. Genet.* **2**, pp. 343–372.

Isidori, A. (1995). *Nonlinear Control Systems*, 3rd edn. (Springer).

Kauffman, S. (1993). *The Origins of Order: Self-organization and Selection in Evolution* (Oxford University Press, New York).

Keren, O. (2008). Reduction of the average path length in binary decision diagrams by spectral methods, *IEEE Trans. Computers* **57**, 4, pp. 520–531.

Kerre, E., Huang, C. and Ruan, D. (2004). *Fuzzy Set Theory and Approximate Reasoning* (Wuhan University Press, Wuhan).

Khalil, H. (1996). *Nonlinear Systems*, 3rd edn. (Prentice Hall).

Knobloch, H., Isidori, A. and Flockerzi, D. (1993). *Topics in Control Theory* (Birkhäuser Verlag, Basel).

Krener, A. and Kang, W. (1990). Extended normal forms of quadratic systems, in *Proc. 29th IEEE Conference on Decision and Control*, pp. 2091–2096.

Lee, K. (2005). *First Course on Fuzzy theory and Applications* (Springer-Verlag, Berlin).

Lewis, F., Campos, J. and Selmic, R. (2002). *Neuro-Fuzzy Control of Industrial Systems with Actuator Nonlinearities* (SIAM, Philadelphia).

Li, D., Zhao, J. and Song, C. (2008). On the positive definiteness of the left semi-tensor product of matrices, in E. Jiang, T. Huang and C. Yang (eds.), *Advances in Matrix Theory and Its Applications*, Vol. 1 (World Academic Union, London), pp. 119–122.

Li, H. and Wang, Y. (2010). Boolean derivative calculation with application to fault detection of combinational circuits via the semi-tensor product method, Preprint.

Li, T., Tong, S. and Feng, G. (2010). A novel robust adaptive-fuzzy-tracking control for a class of nonlinear multi-input/multi-output systems, *IEEE Trans. Fuzzy Systems* **18**, 1, pp. 150–160.

Li, Z., Zhao, Y. and Cheng, D. (2011). Structure of higher order Boolean networks, *Journal of the Graduate School of the Chinese Academy of Sciences* **28**, 4, pp. 431–447, in Chinese.

Liu, X., Wang, H. and Zhao, J. (2008). The properties of bounded tridiagonal matrices, in E. Jiang, T. Huang and C. Yang (eds.), *Advances in Matrix Theory and Its Applications*, Vol. 2 (World Academic Union, London), pp. 104–109.

Liu, Z. and Liu, Y. (1996). *Fuzzy Logic and Neural Network* (BUAA Press, Beijing), in Chinese.

Ljung, L. and Söderström, T. (1982). *Theory and Practice of Recursive Identification* (MIT Press).

Mamdani, E. (1974). Applications of fuzzy algorithms for control of simple dynamic plant, *Proceedings of the Institution of Electrical Engineers* **121**, 12, pp. 1585–1588.

Mei, S., Liu, F. and Xue, A. (2010). *Semi-tensor Product Approach to Transient Analysis of Power Systems* (Tsinghua Univ. Press, Beijing), in Chinese.

Mishchenko, A., Steinbach, B. and Perkowski, M. (2001). An algorithm for bi-decomposition of logic functions, in *Proc. 2001 IEEE/ACM 38th Design Automation Conference* (Las Vegas), pp. 103–108.

Ooba, T. and Funahashi, Y. (1997). Two conditions concerning common quadratic Lyapunov functions for linear systems, *IEEE Trans. Aut. Contr.* **42**, 5, pp. 719–721.

Passino, K. and Yurkovich, S. (1998). *Fuzzy Control* (Addison Wesley Longman).

Passino, K. and Yurkovich, S. (2002). *Fuzzy Control* (Tsinghua Univ. Press & Addison-Wesley).

Pennestri, E. and Valentini, P. (2009). Dual quaternions as a tool for rigid body motion analysis: A tutorial with an application to biomechanics, in K. Arczewski, J. Fraczek and M. Wojtyra (eds.), *Multibody Dynamics 2009, ECCOMAS Thematic Conference* (Warsaw), pp. 1–17.

Posthoff, C. and Steinbach, B. (2004). *Logic Functions and Equations: Binary Models for Computer Science* (Springer, Dordrecht, The Netherlands).

Qiao, Y., Yuan, Y. and Cheng, D. (2011). Finite Fliess functional expansion and its application to flight control of missiles, *Int. J. Robot. Autom.* **26**, 2, pp. 173–181.

Rade, L. and Westergren, B. (1998). *Mathematics Handbook for Science and Engineering*, 4th edn. (Studentlitteratur, Sweden).

Reed, I. (1954). A class of multiple error-correction and the decoding scheme, *IRE Trans. Inform. Theory* **IT-4**, pp. 38–49.

Rohde, C. (1966). Some results on generalized inverses, *SIAM Review* **8**, 2, pp. 201–205.

Roth, J. and Karp, R. (1962). Minimization over Boolean graphs, *IBM Journal* , pp. 227–238.

Runkler, T. (1996). Extended defuzzification methods and their properties, in *Proc. 9th IEEE Int. Conf. Fuzzy Sys.*, pp. 694–700.

Saade, J. and Diab, H. (2000). Defuzzification techniques for fuzzy controllers, *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics* **30**, 1, pp. 223–229.

Sachs, R. and Wu, H. (1977). *General Relativity for Mathematicians* (Springer-Verlag).

Saha, S., Fouad, A., Kliemann, W. and Vittal, V. (1997). Stability boundary approximatin of a power system using the real normal form of vector fields, *IEEE Trans. Power Systems* **12**, 2, pp. 797–802.

Samuelsson, B. and Troein, C. (2003). Superpolynomial growth in the number of attractors in Kauffman networks, *Physical Review Letters* **90**, 9, p. 98701.

Sanchez, E. (1979). Inverses of fuzzy relations. Application to possibility distributions and medical diagnosis, *Fuzzy Sets and Systems* **2**, 1, pp. 75–86.

Sanchez, E. (1996). Truth-qualification and fuzzy relations in matural languages, application to medical diagnosis, *Fuzzy Sets and Systems* **84**, pp. 155–167.

Sanchez, E. (2002). Functional relations and fuzzy relational equations, in *Fuzzy Information Processing Society, 2002. Proceedings. NAFIPS. 2002 Annual Meeting of the North American* (IEEE), ISBN 0780374614, pp. 451–456.

Sasao, T. (1999). *Switching Theory for Logic Synthesis* (Kluwer Academic Publishers).

Sasao, T. and Butler, J. (1997). On bi-decompositions of logic functions, in *Proc. Int. Worksop on Logic Synthesis*, Vol. 2 (Lake Tahoe), pp. 1–6.

Sasao, T. and Fujita, M. (eds.) (1996). *Representations of Discrete Functions* (Kluwer Academic Publishers).

Schneeweiss, W. (1989). *Boolean Functions, with Engineering Applications and Computer Programs* (Springer).

Shmulevich, I., Dougherty, E., Kim, S. and Zhang, W. (2002). Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks, *Bioinformatics* **18**, 2, pp. 261–274.

Sidi, M. (1997). *Spacecraft Dynamics and Control, A Practical Engineering Approach* (Cambridge Univ. Press, Cambridge).

Song, C. (2009). *Some Properties and Applications of Left Semi-tensor Product of Matrices*, Master degree thesis, Liaocheng University.

Song, C., Zhao, J. and Li, D. (2008). The application of generalized permutation matrix in the commutation of tensor products of matrices, in E. Jiang, T. Huang and C. Yang (eds.), *Advances in Matrix Theory and Its Applications*, Vol. 1 (World Academic Union, London), pp. 16–18.

Soriano, J., Olarte, A. and Melgarejo, M. (2005). Fuzzy controller for MIMO sysems using defuzzification based on Boolean relations (DBR), in *Proc. 14th IEEE Int. Conf. Fuzzy Sys.*, pp. 271–275.

Sun, Z. and Xia, X. (1997). On nonregular feedback linearization, *Automatica* **33**, 7, pp. 1739–1744.

Svozil, K. (1998). *Quantum Logic* (Springer, New York).

Thayse, A. (1984). *P-functions and Boolean Matrix Factorization* (Springer, Berlin).

Truemper, K. (2004). *Design of Logic-based Intelligent Systems* (Wiley & Sons, New Jersey).

Tsai, C. (1983). *Contributions to the Design and Analysis of Nonlinear Models*, Ph.D. thesis, Univ. of Minisota.

Tsukamoto, Y. (1979). An approach to fuzzy reasoning method, in M. Gupta,

R. Ragade and R. Yager (eds.), *Advances in Fuzzy Set Theory and Application* (North-Holland, Amsterdam).

Tucker, J., Tapia, M. and Wayne Bennett, A. (1988). Boolean integral calculus, *Applied Mathematics and Computation* **26**, 3, pp. 201–236.

van der Schaft, A. (2000). *L_2-Gain and Passivity Techniques in Nonlinear Control* (Springer-Verlag).

Varadarajan, V. (1984). *Lie Groups, Lie Algebras, and Their Representations* (Springer-Verlag, New York).

Venkatasubramanian, V. and Ji, W. (1997). Numerical approximation of $(n-1)$-dimensional stable manifolds in large systems such as the power system, *Automatica* **33**, 10, pp. 1877–1883.

Verbruggen, H. and Babuška, R. (eds.) (1999). *Fuzzy Logic Control: Advances in Applications* (World Scientific Pub. Co. Inc., Singapore).

Vichniac, G. (1990). Boolean derivatives on cellular automata, *Physica D: Nonlinear Phenomena* **45**, 1-3, pp. 63–74.

Wang, G., Wei, Y. and Qiao, S. (2004). *Generalized Inverses: Theory and Computations* (Science Press, Beijing).

Wang, G. and Xu, Z. (2006). The reverse order law for the W-weighted Drazin inverse of multiple matrices product, *Journal of Applied Mathematics and Computing* **21**, 1, pp. 239–248.

Wang, H. (2010). *Some Advances of Left Semi-tensor Product of Matrice*, Master degree thesis, Liaocheng University.

Wang, H., Liu, X. and Zhao, J. (2009a). The estimates for singular values of Schur complements of the left semi-tensor product of matrices, in *Proc. 3rd Int. Workshop Matrix Analy. Appl.*, Vol. 1, pp. 286–289.

Wang, H., Si, C. and Niu, L. (2009b). On the metapositive of the left semi-tensor product of complex matrices, in *Proc. 3rd Int. Workshop Matrix Analy. Appl.*, Vol. 1, pp. 290–293.

Wang, L. (1996). *A Course in Fuzzy Systems and Control* (Prentice-Hall, Inc., Upper Saddle River, NJ, USA).

Wang, X. (2002). *Parameter Estimate of Nonlinear Models: Theory and Applications* (Wuhan Univ. Press, Wuhan), in Chinese.

Wei, B. (1986). Second moments of LS estimate of nonlinear regressive model, *Applied Math., J. Chinese Universities* **1**, 2, pp. 279–285, in Chinese.

Wen, Q., Niu, X. and Yang, Y. (2000). *Boolean Functions in Modern Cryptography* (Science Press, Beijing), in Chinese.

Willems, J. (1970). *Stability Theory of Dynamical Systems* (John Wiley & Sons Inc.).

Wilson, R. (1996). *Introduction to Graph Theory*, 4th edn. (Prentice Hall, London).

Wolfram, S. (1986). *Theory and Applications of Cellular Automata* (World Scientific, Singapore).

Wonham, W. (1979). *Linear Multivariable Control: A Geometric Aproach*, 2nd edn. (Springer, Berlin).

Wonham, W. and Morse, A. (1970). Decoupling and pole assignment in linear multivariable systems: A geometric approach, *SIAM J. Control* **8**, 1, pp.

1–18.

Wu, W. (1995). *On Mechanization of Mathematics* (Shandong Educational Press, Ji'nan).

Xue, A., Toe, K., Lu, Q. and Mei, S. (2006). Polynomial approximations for the stable and unstable manifolds of hyperbolic equilibrium point using semi-tensor product, *Int. J. Innov. Comp. Inform. Contr.* **2**, 3, pp. 1–16.

Yanushkevich, S. (1998). Logic differential calculus in multi-valued logic design, *Prace Naukowe Politechniki Szczecińskiej. Instytut Informatyki* **537**, 1, pp. 7–326.

Zaborszky, J., Huang, J., Zheng, B. and Leung, T. (1988). On the phase protraits of a class of large nonlinear dynamic systems such as the power systems, *IEEE Trans. Aut. Contr.* **33**, 1, pp. 4–15.

Zadeh, L. (1965). Fuzzy Sets, *Information and Control* **8**, pp. 338–353.

Zadeh, L. (1972). Outline of a new approach to the analysis of complex systems and decision processes, *IEEE Trans. Systems, Man and Cybernetics* **3**, 1, pp. 28–44.

Zadeh, L. and Kacprzyk, J. (eds.) (1992). *Fuzzy Logic for the Management of Uncertainty* (John Wiley & Sons, Inc. New York, NY, USA).

Zhang, X. (2004). *Matrix Analysis and Applications* (Tsinghua Univ. Press & Springer, Beijing).

Zhang, Y. (1993). *Theory of Multi-Edge Matrix* (Chinese Statistics Press, Henan), in Chinese.

Zhang, Y., Liu, Z. and Wang, Y. (2009). A three-dimensional probabilistic fuzzy control systems for network queue management, *J. Contr. Theory & Appl.* **7**, 1, pp. 29–34.

Zhao, Q. (2005). A remark on 'Scalar equations for synchronous Boolean networks with biologicapplications' by C. Farrow, J. Heidel, J. Maloney, and J. Rogers, *IEEE Trans. Neural Networks* **16**, 6, pp. 1715–1716.

Zhao, Y., Gao, X. and Cheng, D. (2010a). Semi-tensor product approach to Boolean functions, Preprint.

Zhao, Y., Li, Z. and Cheng, D. (2010b). Optimal control of logical control networks, Preprint.

Zhao, Y., Qi, H. and Cheng, D. (2010c). Input-state incidence matrix of Boolean control networks and its applications, *Sys. Contr. Lett.* **59**, 12, pp. 767–774.

Zhu, J. (2005). *Fuzzy System and Control Theory* (China Machine Press, Beijing), in Chinese.

This page intentionally left blank

Index