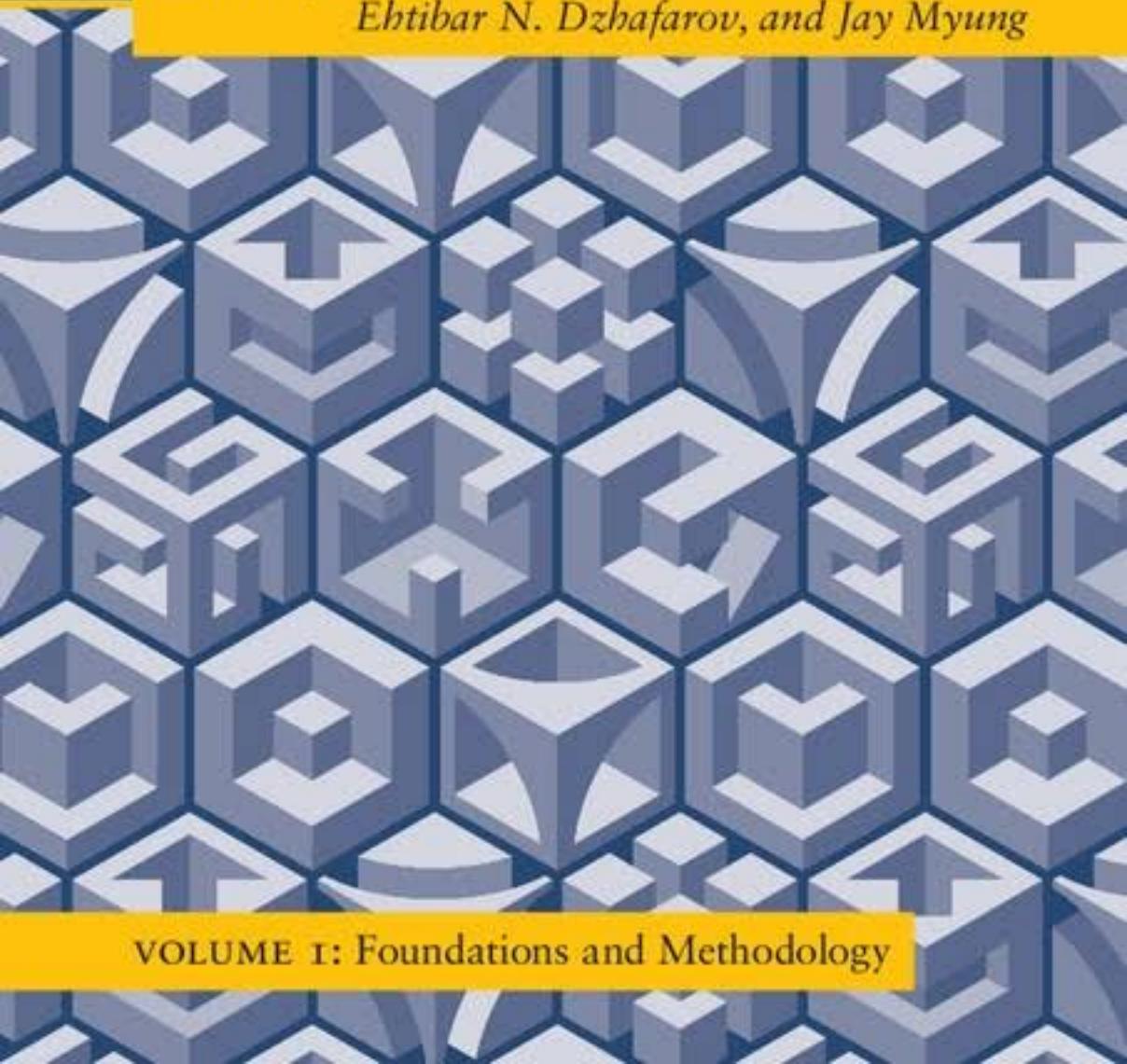


NEW HANDBOOK OF  
MATHEMATICAL  
PSYCHOLOGY

*Edited by William H. Batchelder, Hans Colonius,  
Ehtibar N. Dzhafarov, and Jay Myung*



VOLUME 1: Foundations and Methodology

## **New Handbook of Mathematical Psychology**

The field of mathematical psychology began in the 1950s and includes both psychological theorizing in which mathematics plays a key role, and applied mathematics motivated by substantive problems in psychology. Central to its success was the publication of the first *Handbook of Mathematical Psychology* in the 1960s. The psychological sciences have since expanded to include new areas of research, and significant advances have been made in both traditional psychological domains and in the applications of the computational sciences to psychology. Upholding the rigor of the original Handbook, the *New Handbook of Mathematical Psychology* reflects the current state of the field by exploring the mathematical and computational foundations of new developments over the last half century. The first volume focuses on select mathematical ideas, theories, and modeling approaches to form a foundational treatment of mathematical psychology.

WILLIAM H. BATCHELDER is Professor of Cognitive Sciences at the University of California Irvine.

HANS COLONIUS is Professor of Psychology at Oldenburg University, Germany.

EHTIBAR N. DZHAFAROV is Professor of Psychological Sciences at Purdue University.

JAY MYUNG is Professor of Psychology at Ohio State University.



# **New Handbook of Mathematical Psychology**

## **Volume 1. Foundations and Methodology**

Edited by

William H. Batchelder

Hans Colonius

Ehtibar N. Dzhafarov

Jay Myung



**CAMBRIDGE**  
UNIVERSITY PRESS

**CAMBRIDGE**  
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of  
education, learning and research at the highest international levels of excellence.

[www.cambridge.org](http://www.cambridge.org)

Information on this title: [www.cambridge.org/9781107029088](http://www.cambridge.org/9781107029088)

© Cambridge University Press 2017

This publication is in copyright. Subject to statutory exception  
and to the provisions of relevant collective licensing agreements,  
no reproduction of any part may take place without the written  
permission of Cambridge University Press.

First published 2017

Printed in the United Kingdom by TJ International Ltd. Padstow Cornwall

*A catalogue record for this publication is available from the British Library*

ISBN 978-1-107-02908-8 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy  
of URLs for external or third-party internet websites referred to in this publication,  
and does not guarantee that any content on such websites is, or will remain,  
accurate or appropriate.

# Contents

<i>List of contributors</i>	<i>page</i> vii
<i>Preface</i>	
WILLIAM H. BATCHELDER, HANS COLONIUS, EHTIBAR N. DZHAFAROV, AND JAY MYUNG	ix
<b>1 Selected concepts from probability</b>	
HANS COLONIUS	1
<b>2 Probability, random variables, and selectivity</b>	
EHTIBAR N. DZHAFAROV AND JANNE KUJALA	85
<b>3 Functional equations</b>	
CHE TAT NG	151
<b>4 Network analysis</b>	
JOHN P. BOYD AND WILLIAM H. BATCHELDER	194
<b>5 Knowledge spaces and learning spaces</b>	
JEAN-PAUL DOIGNON AND JEAN-CLAUDE FALMAGNE	274
<b>6 Evolutionary game theory</b>	
J. MCKENZIE ALEXANDER	322
<b>7 Choice, preference, and utility: probabilistic and deterministic representations</b>	
A. A. J. MARLEY AND MICHEL REGENWETTER	374
<b>8 Discrete state models of cognition</b>	
WILLIAM H. BATCHELDER	454
<b>9 Bayesian hierarchical models of cognition</b>	
JEFFREY N. ROUDER, RICHARD D. MOREY, AND MICHAEL S. PRATTE	504
<b>10 Model evaluation and selection</b>	
JAY MYUNG, DANIEL R. CAVAGNARO, AND MARK A. PITTS	552
<b><i>Index</i></b>	<b>599</b>



## Contributors

J. MCKENZIE ALEXANDER, London School of Economics (UK)

WILLIAM H. BATCHELDER, University of California at Irvine (USA)

JOHN P. BOYD, Institute for Mathematical Behavioral Sciences, University of California at Irvine (USA)

DANIEL R. CAVAGNARO, Mihaylo College of Business and Economics, California State University at Fullerton (USA)

HANS COLONIUS, Oldenburg University (Germany)

JEAN-PAUL DOIGNON, Département de Mathématique, Université Libre de Bruxelles (Belgium)

EHTIBAR N. DZHAFAROV, Purdue University (USA)

JEAN-CLAUDE FALMAGNE, Department of Cognitive Sciences, University of California at Irvine (USA)

JANNE V. KUJALA, University of Jyväskylä (Finland)

ANTHONY A. J. MARLEY, Department of Psychology, University of Victoria (Canada)

RICHARD D. MOREY, University of Groningen (The Netherlands)

JAY MYUNG, Ohio State University (USA)

CHE TAT NG, Department of Pure Mathematics, University of Waterloo (Canada)

MARK A. PITTS, Department of Psychology, Ohio State University (USA)

MICHAEL S. PRATTE, Department of Psychology, Vanderbilt University (USA)

MICHEL REGENWETTER, Department of Psychology, University of Illinois at Urbana-Champaign (USA)

JEFFREY N. ROUDER, Department of Psychological Sciences, University of Missouri (USA)



# Preface

## About mathematical psychology

There are three fuzzy and interrelated understandings of what mathematical psychology is: part of mathematics, part of psychology, and analytic methodology. We call them “fuzzy” because we do not offer a rigorous way of defining them. As a rule, a work in mathematical psychology, including the chapters of this handbook, can always be argued to conform to more than one if not all three of these understandings (hence our calling them “interrelated”). Therefore, it seems safer to think of them as three constituents of mathematical psychology that may be differently expressed in any given line of work.

### 1. Part of mathematics

Mathematical psychology can be understood as a collection of mathematical developments inspired and motivated by problems in psychology (or at least those traditionally related to psychology). A good example for this is the algebraic theory of semiorders proposed by R. Duncan Luce (1956). In algebra and unidimensional topology there are many structures that can be called orders. The simplest one is the total, or linear order ( $S, \leq$ ), characterized by the following properties: for any  $a, b, c \in S$ ,

- (O1)       $a \leq b$  or  $b \leq a$ ;
- (O2)      if  $a \leq b$  and  $b \leq c$  then  $a \leq c$ ;
- (O3)      if  $a \leq b$  and  $b \leq a$  then  $a = b$ .

The ordering relation here has the intuitive meaning of “not greater than.” One can, of course, think of many other kinds of order. For instance, if we replace the property (O1) with

$$(O4) \quad a \leq a,$$

we obtain a weaker (less restrictive) structure, called a partial order. If we add to the properties (O1–O3) the requirement that every nonempty subset  $X$  of  $S$  possesses an element  $a_X$  such that  $a_X \leq a$  for any  $a \in X$ , then we obtain a stronger (more restrictive) structure called a well-order. Clearly, one needs motivation for

introducing and studying various types of order, and for one of them, semiorders, it comes from psychology.<sup>1</sup>

Luce (1956) introduces the issue by the following example:

Find a subject who prefers a cup of coffee with one cube of sugar to one with five cubes (this should not be difficult). Now prepare 401 cups of coffee with  $(1 + \frac{i}{100})x$  grams of sugar,  $i = 0, 1, \dots, 400$ , where  $x$  is the weight of one cube of sugar. It is evident that he will be indifferent between cup  $i$  and cup  $i + 1$ , for any  $i$ , but by choice he is not indifferent between  $i = 0$  and  $i = 400$ . (p. 179)

This example involves several idealizations, e.g., Luce ignores here the probabilistic nature of a person's judgments of sweetness/bitterness, treating the issue as if a given pair of cups of coffee was always judged in one and the same way. However, this idealized example leads to the idea that there may be an interesting order such that  $a < b$  only if  $a$  and  $b$  are "sufficiently far apart"; otherwise  $a$  and  $b$  are "too similar" to be ordered ( $a \sim b$ ). Luce formalizes this idea by the following four properties of the structure  $(S, \prec, \sim)$ : for any  $a, b, c, d \in S$ ,

- (SO1)     exactly one of three possibilities obtains: either  $a \prec b$ , or  $b \prec a$  or else  $a \sim b$ ;
- (SO2)      $a \sim a$ ;
- (SO3)     if  $a \prec b$ ,  $b \sim c$  and  $c \prec d$  then  $a \prec d$ ;
- (SO4)     if  $a \prec b$ ,  $b \prec c$  and  $b \sim d$  then either  $a \sim d$  or  $c \sim d$  does not hold.

There are more compact ways of characterizing semiorders, but Luce's seems most intuitive.

Are there familiar mathematical entities that satisfy the requirements (SO1–SO4)? Consider a set of reals and suppose that  $A$  is a set of half-open intervals  $[x, y)$  with the following property: if  $[x_1, y_1)$  and  $[x_2, y_2)$  belong to  $A$ , then  $x_1 \leq x_2$  holds if and only if  $y_1 \leq y_2$ . Let's call the intervals in  $A$  monotonically ordered. Define  $[x, y) \prec [v, w)$  to mean  $y \leq v$ . Define  $[x, y) \sim [v, w)$  to mean that the two intervals overlap. It is easy to check then that the system of monotonically ordered intervals with the  $\prec$  and  $\sim$  relations just defined forms a semiorder.

As it turns out, under certain constraints imposed on  $S$ , the reverse of this statement is also true. To simplify the mathematics, let us assume that  $S$  can be one-to-one mapped onto an interval (finite or infinite) of real numbers. Thus, in Luce's example with cups of coffee we can assume that each cup is uniquely characterized by the weight of sugar in it. Then all possible cups of coffee form a set  $S$  that is bijectively mapped onto an interval of reals between 1 and 5 cubes of sugar (ignoring discreteness due to the molecular structure of sugar). Under this assumption, it follows from a theorem proved by Peter Fishburn (1973) that the semiorder  $(S, \prec, \sim)$  has a monotonically ordered representation. The latter means that there is a monotonically ordered set  $A$  of real intervals and a function  $f : S \rightarrow A$  such

<sup>1</sup> The real history, as often happens, is more complicated, and psychology was the main but not the only source of motivation here (see Fishburn and Monjardet, 1992).

that, for all  $a, b \in S$ ,

$$a \prec b \text{ if and only if } f(a) = [x_1, y_1], f(b) = [x_2, y_2], \text{ and } y_1 \leq x_2; \quad (0.1)$$

$$a \sim b \text{ if and only if } f(a) = [x_1, y_1], f(b) = [x_2, y_2], \text{ and } [x_1, y_1] \cap [x_2, y_2] \neq \emptyset. \quad (0.2)$$

Although as a source of inspiration for abstract mathematics psychology cannot compete with physics, it has motivated several lines of mathematical development. Thus, a highly sophisticated study of  $m$ -point-homogeneous and  $n$ -point-unique monotonic homeomorphisms (mappings that are continuous together with their inverses) of conventionally ordered real numbers launched by Louis Narens (1985) and Theodore M. Alper (1987) was motivated by a well-known classification of measurements by Stanley Smith Stevens (1946). In turn, this classification was inspired by the variety of measurement procedures used in psychology, some of them clearly different from those used in physics. Psychology has inspired and continues to inspire abstract foundational studies in the representational theory of measurement (essentially an area of abstract algebra with elements of topology), probability theory, geometries based on nonmetric dissimilarity measures, topological and pre-topological structures, etc. Finally and prominently, the modern developments in the area of functional equations, beginning with the highly influential work of János Aczél (1966), have been heavily influenced by problems in or closely related to psychology.

## 2. Part of psychology

According to this understanding, mathematical psychology is simply psychological theorizing and model-building in which mathematics plays a central role (but does not necessarily evolve into new mathematical developments). A classical example of work that falls within this category is Gustav Theodor Fechner's derivation of his celebrated logarithmic law in the *Elemente der Psychophysik* (1861, Ch. 17).<sup>2</sup> From this book (and this law) many date the beginnings of scientific psychology. The problem Fechner faced was how to relate "magnitude of physical stimulus" to "magnitude of psychological sensation," and he came up with a principle: *equal ratios of stimulus magnitudes correspond to equal differences of sensation magnitudes*. This means that for any stimulus values  $x_1, x_2$  (real numbers at or above some positive threshold value  $x_0$ ) we have

$$\psi(x_2) - \psi(x_1) = F\left(\frac{x_2}{x_1}\right), \quad (0.3)$$

where  $\psi$  is the hypothetical psychophysical function (mapping stimulus magnitudes into positive reals representing sensation magnitudes), and  $F$  is some unknown function.

<sup>2</sup> The account that follows is not a reconstruction but Fechner's factual derivation (pp. 34–36 of vol. 2 of the *Elemente*). It has been largely overlooked in favor of the less general and less clearly presented derivation of the logarithmic law in Chapter 16 (see Dzhafarov and Colonius, 2012).

Once the equation was written, Fechner investigated it as a purely mathematical object. First, he observed its consequence: for any three suprathreshold stimuli  $x_1, x_2, x_3$ ,

$$F\left(\frac{x_3}{x_1}\right) = F\left(\frac{x_3}{x_2}\right) + F\left(\frac{x_2}{x_1}\right). \quad (0.4)$$

Second, he observed that  $u = x_2/x_1$  and  $v = x_3/x_2$  can be any positive reals, and  $x_3/x_1$  is the product of the two. We have therefore, for any  $u > 0$  and  $v > 0$ ,

$$F(uv) = F(u) + F(v). \quad (0.5)$$

This is an example of a simple functional equation: the function is unknown, but it is constrained by an identity that holds over a certain domain (positive reals).

Functional equations were introduced in pure mathematics only 40 years before Fechner's publication, by Augustin-Louis Cauchy, in his famous *Cours d'analyse* (1821). Cauchy showed there that the only continuous solution for Equation (0.5) is the logarithmic function

$$F(x) \equiv k \log x, \quad x > 0, \quad (0.6)$$

where  $k$  is a constant. The functional equations of this kind were later called the Cauchy functional equations. We know now that one need not even assume that  $F$  is continuous. Thus, it is clear from (0.3) that  $F$  must be positive on at least some interval of values for  $x_2/x_1$ : if  $x_2$  is much larger than  $x_1$ , it is empirically plausible to assume that  $\psi(x_2) > \psi(x_1)$ . This alone is sufficient to derive (0.6) as the only possible solution for (0.5), and to conclude that  $k$  is a positive constant.

The rest of the work for Fechner was also purely mathematical, but more elementary. Putting in (0.1)  $x_2 = x$  (an arbitrary value) and  $x_1 = x_0$  (the threshold value), one obtains

$$\psi(x) - \psi(x_0) = \psi(x) = k \log\left(\frac{x}{x_0}\right), \quad (0.7)$$

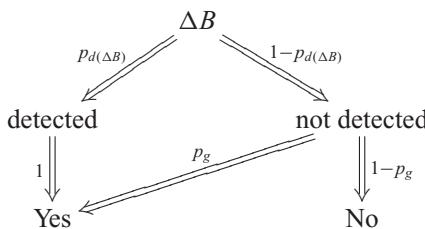
which is the logarithmic law of psychophysics. Fechner thus used sophisticated (by standards of his time) mathematical work by Cauchy to derive the first justifiable quantitative relation in the history of psychology. The value of Fechner's reasoning is entirely in psychology, bringing nothing new to mathematics, but the reasoning itself is entirely mathematical.

There are many other problems and areas in psychology whose analysis falls within the considered category because it essentially consists of purely mathematical reasoning. Thus, analysis of response times that involves distribution or quantile functions is one such area, and so are some areas of psychophysics (especially, theory of detection and discrimination), certain paradigms of decision making, memory and learning, etc.

### 3. Analytic methodology

A third way one can think of mathematical psychology is as an applied, or service field, a set methodological principles and techniques of experimental design, data analysis, and model assessment developed for use by psychologists. The spectrum of examples here extends from purely statistical research to methodology based on substantive theoretical constructs falling within the scope of the first of our three understandings of mathematical psychology.

A simple but representative example of the latter category is H. Richard Blackwell's (1953) correction-for-guessing formula and recommended experimental design. Blackwell considered a simple detection experiment: an observer is shown a stimulus that may have a certain property and asked whether she is aware of this property being present (Yes or No). Thus, the property may be an increment of intensity  $\Delta B$  in the center of a larger field of some fixed intensity  $B$ . Depending on the value of  $\Delta B$ , the observer responds Yes with some probability  $p$ . Blackwell found that this probability  $p(\Delta B)$  was not zero even at  $\Delta B = 0$ . It was obvious to Blackwell (but not to the detection theorists later on) that this indicated that the observer was "guessing" that the signal was there, with probability  $p_g = p(0)$ . It is clear, however, that the observer cannot distinguish the situation in which  $\Delta B = 0$  (and therefore, according to Blackwell, she could not perceive an intensity increment) from one in which  $\Delta B > 0$  but she failed to detect it. Assuming that  $\Delta B$  is detected with probability  $p_d(\Delta B)$ , we have the following tree of possibilities:



We can now express the probability  $p(\Delta B)$  of the observer responding Yes to  $\Delta B$  through the probability  $p_d(\Delta B)$  that she detects  $\Delta B$  and the probability  $p_g$  that she says Yes even though she has not detected  $\Delta B$ :

$$p(\Delta B) = p_d(\Delta B) + (1 - p_d(\Delta B))p_g. \quad (0.8)$$

The value of  $p_d(\Delta B)$  decreases with decreasing  $\Delta B$ , reaching zero at  $\Delta B = 0$ . At this value therefore the formula turns into

$$p(0) = p_g, \quad (0.9)$$

as it should. That is,  $p_g$  is directly observable (more precisely, can be estimated from data): it is the probability with which the observer says Yes to "catch" or "empty" stimuli, those with  $\Delta B = 0$ . Blackwell therefore should insist that catch trials be an integral part of experimental design in any Yes/No detection experiment. Once  $p_g = p(0)$  is known (estimated), one can "correct" the observed

---

(estimated) probability  $p(\Delta B)$  for any nonzero  $\Delta B$  into the true probability of detection:

$$p_d(\Delta B) = \frac{p(\Delta B) - p(0)}{1 - p(0)}. \quad (0.10)$$

Therefore, we end up with a strong recommendation on experimental design (which is universally followed by all experimenters) and a formula for finding true detection probabilities (which is by now all but abandoned). Therefore, Blackwell's work is an example of a methodological development to be used in experimental design and data analysis. At the same time, however, it is also a substantive model of sensory detection, and as such falls within the category of work in psychology in which mathematics plays a central role. The mathematics here is technically simple but ingeniously applied.

The list of methodological developments based on substantive psychological ideas is long. Other classical examples it includes are Louis Leon Thurstone's (1927) analysis of pairwise comparisons and Georg Rasch's analysis of item difficulty and responder aptitude (1960).

On the other pole of the spectrum we find methodological developments that have purely data-analytic character, and their relation to psychology is determined by historical tradition rather than internal logic of these developments. For instance, nowadays we see a rapid growth of sophisticated Bayesian data-analytic and model-comparison procedures, as well as those based on resampling and permutation techniques. Some psychologists prefer to consider all these applied-statistical developments part of psychometrics rather than mathematical psychology. The relationship between the two disciplines is complex, but they are traditionally separated, with different societies and journals.

## About this handbook

The *New Handbook of Mathematical Psychology* (NHMP) is about all three of the meanings of mathematical psychology outlined above. The title of the handbook stems from a very important series of three volumes called the *Handbook of Mathematical Psychology* (HMP), edited by R. Duncan Luce, Robert R. Bush, and Eugene Galanter (1963a; 1963b; 1965). These three volumes played an essential role in defining the character of a new field called mathematical psychology that had begun only 10 years earlier. The 21 chapters of the HMP, totalling 1800 pages, were written by scholars who had ingeniously employed serious mathematics in their work, such as information theory, automata theory, probability theory (including stochastic processes), logic, modern algebra, and set theory. The HMP sparked a great deal of research eventually leading, among other things, to the founding of the European Mathematical Psychology Group, the Society for Mathematical Psychology, the *Journal of Mathematical Psychology*, and a

number of special graduate programs within psychology departments in Europe and the USA. In our view, the main feature of the HMP was that it focused on foundational issues and emphasized mathematical ideas. These two foci were central to the philosophy of the editors of the HMP, who believed that the foundations of any serious science had to be mathematical. It is in this sense that our concept of the NHMP derives from the HMP. We realize, however, we are attempting to fill very big shoes. Also, we are facing more complex circumstances than were the editors and authors of the HMP. In the early 1960s there were fewer topics to cover, and there was less material to cover in each topic: the chapters therefore could very well be both conveyors of major mathematical themes and surveyors of empirical studies. We have to be more selective to make our task manageable.

One could see it as a success of mathematical psychology that almost every area of psychology nowadays employs a variety of formal models and analytic methods, some of them quite sophisticated. It seems also the case, however, that the task of constructing new formal models in an area has to some extent displaced mathematical foundational work. Thus, in our modern age of computation, it is possible to use formal probabilistic models and estimate them with standard statistical packages without a deep understanding of the probabilistic and mathematical underpinnings of the models' assumptions. We hope the NHMP will serve to counteract such tendencies.

Our goal in this and subsequent volumes of the NHMP is to focus on foundational issues, on mathematical themes, ideas, theories, and approaches rather than on empirical facts and specific competing models. Empirical studies are reviewed in the NHMP primarily to motivate or illustrate a class of models or a mathematical formulation. Rather than briefly touching on a large number of pertinent topics in an attempt to provide a comprehensive overview, each chapter discusses in depth and with relevant mathematical explanations just a few topics chosen for being fundamental or highly representative of the field.

In relation to our “three fuzzy and interrelated understandings” of mathematical psychology, the first four chapters of the present volume can be classed into the category “part of mathematics,” as they deal primarily with broad mathematical themes. Chapter 1, by Hans Colonius, discusses the important notions of probabilistic couplings and probabilistic copulas, as well as other foundational notions of probabilistic analysis, such as Fréchet–Hoeffding inequalities and different forms of stochastic dependence and stochastic ordering. The theme of foundations of probability with a prominent role of probabilistic couplings continues in Chapter 2, by Ehtibar Dzhafarov and Janne Kujala. It deals with systems of random variables recorded under variable conditions and adds the notion of selectiveness (in the dependence of the random variables on these conditions) to the conceptual framework of probability theory. Chapter 3, by Che Tat Ng, takes on the traditional topic of functional equations. As we have seen, their use in mathematical psychology dates back to Gustav Theodor Fechner. Chapter 4, by John Boyd and

William Batchelder, takes on the field of network analysis, focusing on discrete networks representable by graphs and digraphs. The chapter presents algebraic (matrix) methods of network analysis, as well as probabilistic networks, such as Markov random fields.

Chapters 5–8 can be classed into the category “part of psychology,” as they primarily deal with substantive theories and classes of models. In Chapter 5, Jean-Paul Doignon and Jean-Claude Falmagne describe a theory of knowledge and learning spaces, which are highly abstract pre- (or proto-) topological constructs that nevertheless have considerable applied value in assessment and guidance of knowledge acquisition. Chapter 6, by McKenzie Alexander, is about interdisciplinary applications of classical game theory to dynamic systems, such as behavior of animals, cultural norms, or linguistic conventions, and about how these systems evolve into evolutionary stable structures within a Darwinian concept of adaptability. The classical topic of choice, preference, and utility models is taken on in Chapter 7, by Anthony A. J. Marley and Michel Regenwetter. The chapter focuses primarily on probabilistic models, treating deterministic representations as their special case. Chapter 8, by William Batchelder, deals with another classical topic, that of modeling cognitive processes by discrete state models representable as a special class of parameterized full binary trees. Such models range from discrete state models of signal detection to Markov chain models of learning and memory to the large class of multinomial processing tree (MPT) models.

The last two chapters of the handbook deal primarily with the relation between psychological models and empirical data. They can therefore be classed into the category of “analytic methodology.” Chapter 9, by Jeffrey Rouder, Richard Morey, and Michael Pratte, deals with data structures where several participants each give responses to several classes of similar experimental items. The chapter describes how Bayesian hierarchical models can specify both subject and item parameter distributions. In Chapter 10, Jay Myung, Daniel Cavagnaro, and Mark Pitt discuss statistical techniques, both Bayesian and frequentist, of evaluating and comparing parametric probabilistic models applied to a given body of data, as well as ways to optimally select a sequence of experimental conditions in data gathering to maximally differentiate the competing models.

There is no particular order in which the chapters in the NHMP should be read: they are independent of each other. We strived to ensure that each chapter is self-contained, requiring no prior knowledge of the material except for a certain level of mathematical maturity (ability to read mathematics) and some knowledge of basic mathematics. The latter includes calculus, elementary probability theory, and elementary set theory, say, within the scope of one- or two-semester introductory courses at mathematics departments. The intended readership of the handbook are behavioral and social scientists, mathematicians, computer scientists, and analytic philosophers – ranging from graduate students, or even advanced undergraduates, to experts in one of these fields.

## References

- Aczél, J. (1966). *Lectures on Functional Equations and Their Applications*. (Mathematics in Science and Engineering 19.) New York, NY: Academic Press.
- Alper, T. M. (1987). A classification of all order-preserving homeomorphism groups of the real that satisfy finite uniqueness. *Journal of Mathematical Psychology*, **31**: 135–154.
- Blackwell, H. R. (1953). *Psychological Thresholds: Experimental Studies of Methods of Measurement* (Bulletin No. 36). Ann Arbor, MI: University of Michigan, Engineering Research Institute.
- Cauchy, A.-L. (1821). *Cours d'analyse de l'École royale polytechnique*. Paris: Imprimerie royale.
- Dzhafarov, E. N., and Colonius, H. (2012). The Fechnerian idea. *American Journal of Psychology*, **124**: 127–140.
- Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf & Härtel.
- Fishburn, P. (1973). Interval representations for interval orders and semiorders. *Journal of Mathematical Psychology*, **10**; 91–105.
- Fishburn, P., and Monjardet, B. (1992). Norbert Wiener on the theory of measurement (1914, 1915, 1921). *Journal of Mathematical Psychology*, **36**; 165–184.
- Luce, R. D. (1956). Semiorders and a theory of utility discrimination. *Econometrica*, **24**: 178–191.
- Luce, R. D. Bush, R. R. and Galanter, E. (1963a). *Handbook of Mathematical Psychology*, vol. 1. New York, NY: Wiley.
- Luce, R. D. Bush, R. R. and Galanter, E. (1963b). *Handbook of Mathematical Psychology*, vol. 2. New York, NY: Wiley.
- Luce, R. D. Bush, R. R. and Galanter, E. (1965). *Handbook of Mathematical Psychology*, vol. 3. New York, NY: Wiley.
- Narens, L. (1985). *Abstract Measurement Theory*. Cambridge, MA: MIT University Press.
- Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Paedagogiske Institut.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, **103**; 677–680
- Thurstone, L. L. (1927). Psychophysical analysis. *American Journal of Psychology*, **38**; 368–389.



# 1 Selected concepts from probability

Hans Colonius

1.1	Introduction	2
1.1.1	Goal of this chapter	2
1.1.2	Overview	6
1.2	Basics	7
1.2.1	$\sigma$ -Algebra, probability space, independence, random variable, and distribution function	7
1.2.2	Random vectors, marginal and conditional distribution	13
1.2.3	Expectation, other moments, and tail probabilities	16
1.2.4	Product spaces and convolution	19
1.2.5	Stochastic processes	21
	The Poisson process	22
	The non-homogeneous Poisson process	23
1.3	Specific topics	24
1.3.1	Exchangeability	24
1.3.2	Quantile functions	25
1.3.3	Survival analysis	27
	Survival function and hazard function	27
	Hazard quantile function	29
	Competing risks models: hazard-based approach	30
	Competing risks models: latent-failure times approach	32
	Some non-identifiability results	33
1.3.4	Order statistics, extreme values, and records	34
	Order statistics	34
	Extreme value statistics	37
	Record values	40
1.3.5	Coupling	44
	Coupling event inequality and maximal coupling	48
1.3.6	Fréchet–Hoeffding bounds and Fréchet distribution classes	50
	Fréchet–Hoeffding bounds for $n = 2$	50
	Fréchet–Hoeffding bounds for $n \geq 3$	52
	Fréchet distribution classes with given higher-order marginals	53
	Best bounds on the distribution of sums	55

---

1.3.7	Copula theory	56
	Definition, examples, and Sklar's theorem	56
	Copula density and pair copula constructions (vines)	58
	Survival copula, dual and co-copula	59
	Copulas with singular components	60
	Archimedean copulas	60
	Example: Clayton copula and the copula of max and min order statistics	62
	Operations on distributions not derivable from operations on random variables	62
1.3.8	Concepts of dependence	63
	Positive dependence	63
	Negative dependence	66
	Measuring dependence	67
1.3.9	Stochastic orders	68
	Univariate stochastic orders	68
	Univariate variability orders	71
	Multivariate stochastic orders	75
	Positive dependence orders	76
1.4	Bibliographic references	78
1.4.1	Monographs	78
1.4.2	Selected applications in mathematical psychology	78
1.5	Acknowledgments	79
	References	79

## 1.1 Introduction

### 1.1.1 Goal of this chapter

Since the early beginnings of mathematical psychology, concepts from probability theory have always played a major role in developing and testing formal models of behavior and in providing tools for data-analytic methods. Moreover, fundamental measurement theory, an area where such concepts have not been mainstream, has been diagnosed as wanting of a sound probabilistic base by founders of the field (see Luce, 1997). This chapter is neither a treatise on the role of probability in mathematical psychology nor does it give an overview of its most successful applications. The goal is to present, in a coherent fashion, a number of probabilistic concepts that, in my view, have not always found appropriate consideration in mathematical psychology. Most of these concepts have been around in mathematics for several decades, like coupling, order statistics, records, and copulas; some of them, like the latter, have seen a surge of interest in recent years, with copula theory providing a new means of modeling dependence in high-dimensional data

(see Joe, 2015). A brief description of the different concepts and their interrelations follows in the second part of this introduction.

The following three examples illustrate the type of concepts addressed in this chapter. It is no coincidence that they all relate, in different ways, to the measurement of reaction time (RT), which may be considered a prototypical example of a random variable in the field. Since the time of Dutch physiologist Franciscus C. Donders (Donders, 1868/1969), mathematical psychologists have developed increasingly sophisticated models and methods for the analysis of RTs.<sup>1</sup> Nevertheless, the probabilistic concepts selected for this chapter are, in principle, applicable in any context where some form of randomness has been defined.

**Example 1.1** (Random variables vs. distribution functions) Assume that the time to respond to a stimulus depends on the attentional state of the individual; the response may be the realization of a random variable with distribution function  $F_H$  in the high-attention state and  $F_L$  in the low-attention state. The distribution of observed RTs could then be modeled as a mixture distribution,

$$F(t) = pF_H(t) + (1 - p)F_L(t),$$

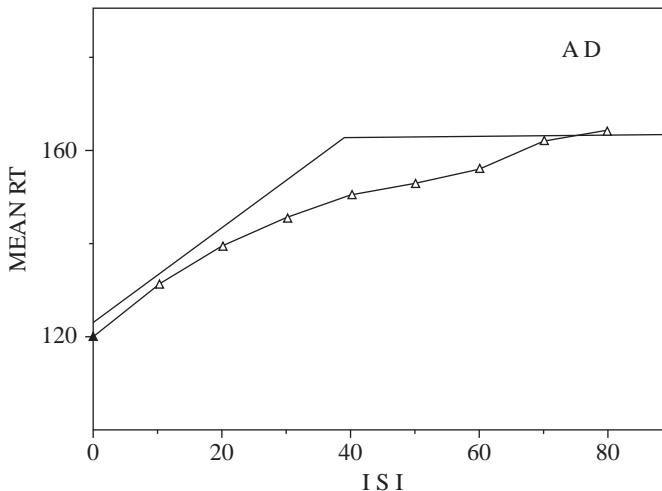
for all  $t \geq 0$  with  $0 \leq p \leq 1$  the probability of responding in a state of high attention.

Alternatively, models of RT are often defined directly in terms of operations on random variables. Consider, for example, Donders' *method of subtraction* in the detection task; if two experimental conditions differ by an additional decision stage,  $D$ , total response time may be conceived of as the sum of two random variables,  $D + R$ , where  $R$  is the time for responding to a high-intensity stimulus.

In the case of a mixture distribution, one may wonder whether it might also be possible to represent the observed RTs as the sum of two random variables  $H$  and  $L$ , say, or, more generally, if the observed RTs follow the distribution function of some  $Z(H, L)$ , where  $Z$  is a measurable two-place function of  $H$  and  $L$ . In fact, the answer is negative and follows as a classic result from the *theory of copulas* (Nelsen, 2006), to be treated later in this chapter.

**Example 1.2** (Coupling for audiovisual interaction) In a classic study of intersensory facilitation, Hershenson (1962) compared reaction time to a moderately intense visual or acoustic stimulus to the RT when both stimuli were presented more or less simultaneously. Mean RT of a well-practiced subject to the sound ( $RT_A$ , say) was approximately 120 ms, mean RT to the light ( $RT_V$ ) about 160 ms. When both stimuli were presented synchronously, mean RT was still about 120 ms. Hershenson reasoned that intersensory facilitation could only occur if the “neural events” triggered by the visual and acoustic stimuli occurred simultaneously somewhere in the processing. That is, “physiological synchrony,” rather than “physical (stimulus) synchrony” was required. Thus, he presented bimodal stimuli with light leading sound giving the slower system a kind of “head start.” In the absence of

<sup>1</sup> For monographs, see Townsend and Ashby (1983), Luce (1986), Schweickert *et al.* (2012).



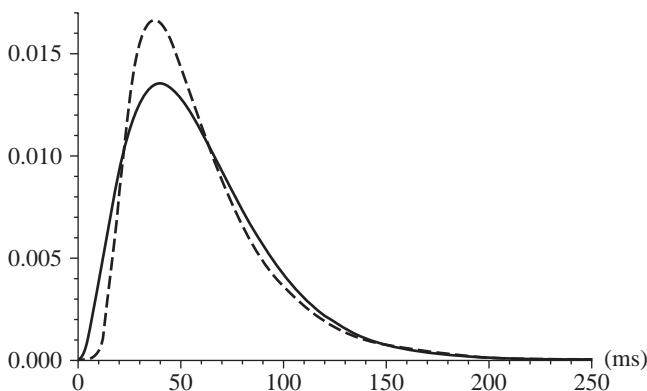
**Figure 1.1** Bimodal (mean) reaction time to light and sound with interstimulus interval (ISI) and sound following light,  $RT_V = 160$  ms,  $RT_A = 120$  ms. Upper graph: prediction in absence of interaction, lower graph: observed mean RTs; data from Diederich and Colonius (1987).

interaction, reaction time to the bimodal stimulus with presentation of the acoustic delayed by  $\tau$  ms, denoted as  $RT_{V\tau A}$ , is expected to increase linearly until the sound is delivered 40 ms after the light (the upper graph in Figure 1.1). Actual results, however, looked more like the lower graph in Figure 1.1, where maximal facilitation occurs at about physiological synchrony. Raab (1962) suggested an explanation in terms of a probability summation (or, *race*) mechanism: response time to the bimodal stimulus,  $RT_{V\tau A}$ , is considered to be the winner of a race between the processing times for the unimodal stimuli, i.e.,  $RT_{V\tau A} \equiv \min\{RT_V, RT_A + \tau\}$ . It then follows for the expected values (mean RTs):

$$E[RT_{V\tau A}] = E[\min\{RT_V, RT_A + \tau\}] \leq \min\{E[RT_V], E[RT_A + \tau]\},$$

a prediction that is consistent with the observed facilitation. It has later been shown that this prediction is not sufficient for explaining the observed amount of facilitation, and the discussion of how the effect should be modeled is ongoing, attracting a lot of attention in both psychology and neuroscience.

However, as already observed by Luce (1986, p. 130), the above inequality only makes sense if one adds the assumption that the three random variables  $RT_{V\tau A}$ ,  $RT_V$ , and  $RT_A$  are jointly distributed. The existence of a joint distribution is not automatic because each variable relates to a different underlying probability space defined by the experimental condition: visual, auditory, or bimodal stimulus presentation. From the *theory of coupling* (Thorisson, 2000), constructing such a joint distribution is always possible by assuming stochastic independence of the random variables. However – and this is the main point of this example – independence is not the only coupling possibility, and alternative assumptions yielding distributions



**Figure 1.2** Inverse gaussian (dashed line) and gamma densities with identical mean (60 ms) and standard deviation (35 ms).

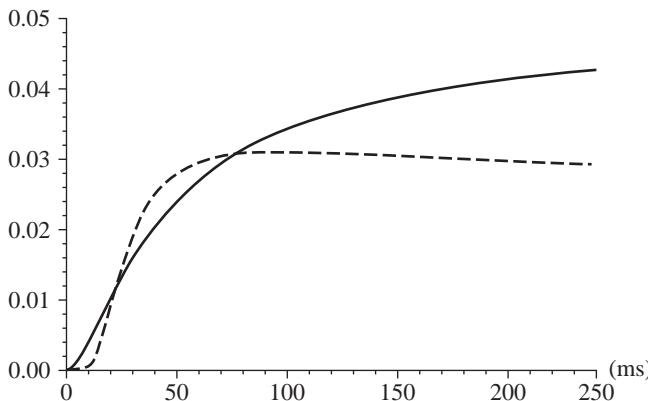
with certain dependency properties may be more appropriate to describe empirical data.

**Example 1.3** (Characterizing RT distributions: hazard function) Sometimes, a stochastic model can be shown to predict a specific parametric distribution, e.g., drawing on some asymptotic limit argument (central limit theorem or convergence to extreme-value distributions). It is often notoriously difficult to tell apart two densities when only a histogram estimate from a finite sample is available. Figure 1.2 provides an example of two theoretically important distributions, the gamma and the inverse gaussian densities with identical means and standard deviations, where the rather similar shapes make it difficult to distinguish them on the basis of a histogram.

An alternative, but equivalent, representation of these distributions is terms of their *hazard functions* (see Section 1.10). The hazard function  $h_X$  of random variable  $X$  with distribution function  $F_X(x)$  and density  $f_X(x)$  is defined as

$$h_X(x) = \frac{f_X(x)}{1 - F_X(x)}.$$

As Figure 1.3 illustrates, the gamma hazard function is increasing with decreasing slope, whereas the inverse gaussian is first increasing and then decreasing. Although estimating hazard functions also has its intricacies (Kalbfleisch and Prentice, 2002), especially at the right tail, there is a better chance to tell the distributions apart based on estimates of the hazard function than on the density or distribution function. Still other methods to distinguish classes of distribution functions are based on the concept of *quantile function* (see Section 1.3.2), among them the method of *delta plots*, which has recently drawn the attention of researchers in RT modeling (Schwarz and Miller, 2012). Moreover, an underlying theme of this chapter is to provide tools for a model builder that do not depend on committing oneself to a particular parametric distribution assumption.



**Figure 1.3** Hazard functions of the inverse gaussian (dashed line) and gamma distributions corresponding to the densities of Figure 1.2.

We hope to convey in this chapter that even seemingly simple situations, like the one described in Example 1.2, may require some careful consideration of the underlying probabilistic concepts.

### 1.1.2 Overview

In trying to keep the chapter somewhat self-contained, the first part presents basic concepts of probability and stochastic processes, including some elementary notions of measure theory. Because of space limitations, some relevant topics had to be omitted (e.g., random walks, Markov chains), or are only mentioned in passing (e.g., martingale theory). For the same reason, statistical aspects are considered only when suggested by the context.<sup>2</sup> Choosing what material to cover was guided by the specific requirements of the topics in the second, main part of the chapter.

The second part begins with a brief introduction to the notion of *exchangeability* (with a reference to an application in vision) and its role in the celebrated “theorem of de Finetti.” An up-to-date presentation of quantile (density) functions follows, a notion that emerges in many areas including survival analysis. The latter topic, while central to RT analysis, has also found applications in diverse areas, like decision making and memory, and is treated next at some length, covering an important non-identifiability result. Next follow three related topics: order statistics, extreme values, and the theory of records. Whereas the first and, to a lesser degree, the second of these topics have become frequent tools in modeling psychological processes, the third has not yet found the role that it arguably deserves.

The method of coupling, briefly mentioned in introductory Example 1.2, is a classic tool of probability theory concerned with the construction of a joint

<sup>2</sup> For statistical issues of reaction time analysis, see the competent treatments by Van Zandt (2000, 2002); and Ulrich and Miller (1994), for discussing effects of truncation.

probability space for previously unrelated random variables (or, more general random entities). Although it is used in many parts of probability, e.g., Poisson approximation, and in simulation, there are not many systematic treatises of coupling and it is not even mentioned in many standard monographs of probability theory. We can only present the theory at a very introductory level here, but the expectation is that coupling will have to play an important conceptual role in psychological theorizing. For example, its relevance in defining “selective influence/contextuality” has been demonstrated in the work by Dzhafarov and Kujala (see also Chapter 2 by Dzhafarov and Kujala in this volume).

While coupling strives to construct a joint probability space, the existence of a multivariate distribution is presumed in the next two sections. *Fréchet classes* are multivariate distributions that have certain of their marginal distributions fixed. The issues are (i) to characterize upper and lower bounds for all elements of a given class, and (ii) to determine conditions under which (bivariate or higher) margins with overlapping indices are compatible. Copula theory allows one to separate the dependency structure of a multivariate distribution from the specific univariate margins. This topic is pursued in the subsequent section presenting a brief overview of different types of multivariate dependence. Comparing uni- and multivariate distribution functions with respect to location and/or variability is the topic of the final section, stochastic orders.

A few examples of applications of these concepts to issues in mathematical psychology are interspersed in the main text. Moreover, the comments and reference section at the end gives a number of references to further pertinent applications.

## 1.2 Basics

Readers familiar with basic concepts of probability and stochastic processes, including some measure-theoretic terminology, may skip this first section of the chapter.

### 1.2.1 $\sigma$ -Algebra, probability space, independence, random variable, and distribution function

A fundamental assumption of practically all models and methods of response time analysis is that the response latency measured in a given trial of a reaction time task is the realization of a random variable. In order to discuss the consequences of treating response time as a random variable or, more generally, as a function of several random variables, some standard concepts of probability theory will first be introduced.<sup>3</sup>

<sup>3</sup> Limits of space do not permit a completely systematic development here, so only a few of the most relevant topics will be covered in detail. For a more comprehensive treatment see the references in the final section (and Chapter 2 for a more general approach).

Let  $\Omega$  be an arbitrary set, often referred to as the *sample space* or set of *elementary outcomes* of a random experiment, and  $\mathcal{F}$  a system of subsets of  $\Omega$  endowed with the properties of a  $\sigma$ -algebra (*of events*), i.e.,

- (i)  $\emptyset \in \mathcal{F}$  (“impossible” event  $\emptyset$ ).
- (ii) If  $A \in \mathcal{F}$  then also its complement:  $A^c \in \mathcal{F}$ .
- (iii) For a sequence of events  $\{A_n \in \mathcal{F}\}_{n \geq 1}$ , then also  $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ .

The pair  $(\Omega, \mathcal{F})$  is called *measurable space*. Let  $\mathcal{A}$  be any collection of subsets of  $\Omega$ . Because the power set,  $\mathfrak{P}(\Omega)$ , is a  $\sigma$ -algebra, it follows that there exists at least one  $\sigma$ -algebra containing  $\mathcal{A}$ . Moreover, the intersection of any number of  $\sigma$ -algebras is again a  $\sigma$ -algebra. Thus, there exists a unique *smallest*  $\sigma$ -algebra containing  $\mathcal{A}$ , defined as the intersection of all  $\sigma$ -algebras containing  $\mathcal{A}$ , called the  $\sigma$ -algebra *generated by*  $\mathcal{A}$  and denoted as  $\mathcal{S}(\mathcal{A})$ .

**Definition 1.1** (Probability space) The triple  $(\Omega, \mathcal{F}, P)$  is a *probability space* if  $\Omega$  is a sample space with  $\sigma$ -algebra  $\mathcal{F}$  such that  $P$  satisfies the following (*Kolmogorov*) axioms:

- (1) For any  $A \in \mathcal{F}$ , there exists a number  $P(A) \geq 0$ ; the probability of  $A$ .
- (2)  $P(\Omega) = 1$ .
- (3) For any sequence of mutually disjoint events  $\{A_n, n \geq 1\}$ ,

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

Then  $P$  is called the *probability measure*, the elements of  $\mathcal{F}$  are the *measurable* subsets of  $\Omega$ , and the probability space  $(\Omega, \mathcal{F}, P)$  is an example of *measure spaces* which may have measures other than  $P$ . Some easy to show consequences of the three axioms are, for measurable sets  $A, A_1, A_2$ ,

1.  $P(A^c) = 1 - P(A)$ ;
2.  $P(\emptyset) = 0$ ;
3.  $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$ ;
4.  $A_1 \subset A_2 \rightarrow P(A_1) \leq P(A_2)$ .

A set  $A \subset \Omega$  is called a *null set* if there exists  $B \in \mathcal{F}$ , such that  $B \supset A$  with  $P(B) = 0$ . In general, null sets need not be measurable. If they are, the probability space  $(\Omega, \mathcal{F}, P)$  is called *complete*.<sup>4</sup> A property that holds everywhere except for those  $\omega$  in a null set is said to hold ( $P$ )-*almost everywhere* (a.e.).

**Definition 1.2** (Independence) The events  $\{A_k, 1 \leq k \leq n\}$  are *independent* if, and only if,

$$P\left(\bigcap A_{i_k}\right) = \prod P(A_{i_k}),$$

<sup>4</sup> Any given  $\sigma$ -algebra can be enlarged and the probability measure can be uniquely extended to yield a complete probability space, so it will be assumed in the following without further explicit mentioning that a given probability space is complete.

where intersections and products, respectively, are to be taken over all subsets of  $\{1, 2, \dots, n\}$ . The events  $\{A_n, n \geq 1\}$  are independent if  $\{A_k, 1 \leq k \leq n\}$  are independent for all  $n$ .

**Definition 1.3** (Conditional probability) Let  $A$  and  $B$  be two events and suppose that  $P(A) > 0$ . The *conditional probability of  $B$  given  $A$*  is defined as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

**Remark 1.1** If  $A$  and  $B$  are independent, then  $P(B|A) = P(B)$ . Moreover,  $P(\cdot|A)$  with  $P(A) > 0$  is a probability measure. Because  $0 \leq P(A \cap B) \leq P(A) = 0$ , null sets are independent of “everything.”

The following statements about any subsets (events) of  $\Omega$ ,  $\{A_k, 1 \leq k \leq n\}$ , turn out to be very useful in many applications in response time analysis and are listed here for later reference.

**Remark 1.2** (Inclusion–exclusion formula)

$$\begin{aligned} P\left(\bigcup_{k=1}^n A_k\right) &= \sum_{k=1}^n P(A_k) - \sum_{1 \leq i \leq j \leq n} P(A_i \cap A_j) \\ &\quad + \sum_{1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k) \\ &\quad + \cdots - (-1)^n P(A_1 \cap A_2 \cap \cdots \cap A_n). \end{aligned}$$

If the events are independent, this reduces to

$$P\left(\bigcup_{k=1}^n A_k\right) = 1 - \prod_{k=1}^n (1 - P(A_k)).$$

**Definition 1.4** (Measurable function) Let  $(\Omega, \mathcal{F})$  and  $(\Omega', \mathcal{F}')$  be measure spaces and  $T : \Omega \rightarrow \Omega'$  a mapping from  $\Omega$  to  $\Omega'$ .  $T$  is called  $\mathcal{F}\text{-}\mathcal{F}'$ -measurable if

$$T^{-1}(A') \in \mathcal{F} \quad \text{for all } A' \in \mathcal{F}',$$

where

$$T^{-1}(A') = \{\omega \in \Omega \mid T(\omega) \in A'\}$$

is called the *inverse image* of  $A'$ .

For the introduction of (real-valued) random variables, we need a special  $\sigma$ -algebra. Let  $\Omega = \mathbb{R}$ , the set of real numbers. The  $\sigma$ -algebra of *Borel sets*, denoted as  $\mathcal{B}(\mathbb{R}) \equiv \mathcal{B}$ , is the  $\sigma$ -algebra generated by the set of open intervals<sup>5</sup> of  $\mathbb{R}$ . Importantly, two probability measures  $P$  and  $Q$  on  $(\mathbb{R}, \mathcal{B})$  that agree on all open intervals are identical,  $P = Q$ .

<sup>5</sup> It can be shown that  $\mathcal{B}$  can equivalently be generated by the sets of closed or half-open intervals of real numbers. In the latter case, this involves extending  $\mathcal{B}$  to a  $\sigma$ -algebra generated by the extended real line,  $\mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$ .

**Definition 1.5** (Random variable) Let  $(\Omega, \mathcal{F}, P)$  be a probability space. A (real-valued) *random variable*  $X$  is a  $\mathcal{F}$ - $\mathcal{B}$ -measurable function from the sample space  $\Omega$  to  $\mathbb{R}$ ; that is, the inverse image of any Borel set  $A$  is  $\mathcal{F}$ -measurable:

$$X^{-1}(A) = \{\omega \mid X(\omega) \in A\} \in \mathcal{F}, \quad \text{for all } A \in \mathcal{B}.$$

If  $X : \Omega \rightarrow [-\infty, +\infty]$ , we call  $X$  an *extended random variable*.

Random variables that differ only on a null set are called *equivalent* and for two random variables  $X$  and  $Y$  from the same equivalence class we write  $X \sim Y$ .

To each random variable  $X$  we associate an *induced probability measure*,  $\Pr$ , through the relation

$$\Pr(A) = P(X^{-1}(A)) = P(\{\omega \mid X(\omega) \in A\}), \quad \text{for all } A \in \mathcal{B}.$$

The induced space  $(\mathbb{R}, \mathcal{B}, \Pr)$  can be shown to be a probability space by simply checking the above (Kolmogorov) axioms.  $\Pr$  is also called the *distribution* of  $X$ .

**Remark 1.3** Most often one is only interested in the random variables and “forgets” the exact probability space behind them. Then no distinction is made between the probability measures  $P$  and  $\Pr$ , one omits the brackets {} emphasizing that  $\{X \in A\}$  actually is a set, and simply writes  $P(X \in A)$ , or  $P_X(A)$ , instead of  $\Pr(A)$ .

However, sometimes it is important to realize that two random variables are actually defined with respect to two different probability spaces. A case in point is our introductory Example 1.2, where the random variable representing reaction time to a visual stimulus and the one representing reaction time to the acoustic stimulus are not a-priori defined with respect to the same probability space. In such a case, for example, it is meaningless to ask whether two events are independent (see Section 1.3.5).

“The equality” of random variables can be interpreted in different ways.

**Remark 1.4** Random variables  $X$  and  $Y$  are *equal in distribution* iff they are governed by the same probability measure:

$$X =_d Y \iff P(X \in A) = P(Y \in A), \quad \text{for all } A \in \mathcal{B}.$$

$X$  and  $Y$  are *point-wise equal* iff they agree for almost all elementary events<sup>6</sup>:

$$X \stackrel{a.s.}{=} Y \iff P(\{\omega \mid X(\omega) = Y(\omega)\}) = 1,$$

i.e., iff  $X$  and  $Y$  are equivalent random variables,  $X \sim Y$ .

The following examples illustrate that two random variables may be equal in distribution, and at the same time there is no elementary event where they agree.

**Example 1.4** (Gut, 2013, p. 27) Toss a fair coin once and set

$$X = \begin{cases} 1, & \text{if the outcome is heads,} \\ 0, & \text{if the outcome is tails,} \end{cases} \quad \text{and } Y = \begin{cases} 1, & \text{if the outcome is tails,} \\ 0, & \text{if the outcome is heads.} \end{cases}$$

<sup>6</sup> Here, a.s. is for “almost sure”.

Clearly,  $P(X = 1) = P(X = 0) = P(Y = 1) = P(Y = 0) = 1/2$ , in particular,  $X =_d Y$ . Moreover,  $X(\omega)$  and  $Y(\omega)$  differ for every  $\omega$ .

**Remark 1.5** A function  $X : \Omega \rightarrow \mathfrak{N}$  is a random variable iff  $\{\omega \mid X(\omega) \leq x\} \in \mathcal{F}$ , for all  $x \in \mathfrak{N}$ . This important observation follows because the Borel  $\sigma$ -algebra can be generated from the set of half-open intervals of the form  $(-\infty, x]$  (see footnote 5).

**Remark 1.6** Let  $X_1, X_2$  be random variables. Then  $\max\{X_1, X_2\}$  and  $\min\{X_1, X_2\}$  are random variables as well. In general, it can be shown that a Borel measurable function of a random variable is a random variable.

**Definition 1.6** (Distribution function) Let  $X$  be a real-valued random variable. The *distribution function* of  $X$  is

$$F_X(x) = P(X \leq x), \quad x \in \mathfrak{N}.$$

From Remark 1.5 it is clear that the distribution function of  $X$ ,  $F_X(\cdot)$ , provides a complete description of the distribution  $P(\cdot)$  of  $X$  via the relation

$$F_X(b) - F_X(a) = P(X \in (a, b]), \quad \text{for all } a, b, \quad -\infty < a \leq b < \infty.$$

Thus, every probability measure on  $(\mathfrak{N}, \mathcal{B})$  corresponds uniquely to the distribution function of some random variable(s) but, as shown in Example 1.4, different random variables may have the same distribution function.

Next we list without proof the most important properties of any distribution function (Gut, 2013, p. 30). Let  $F$  be a distribution function. Then:

1.  $F$  is non-decreasing:  $x_1 < x_2 \implies F(x_1) \leq F(x_2)$ ;
2.  $F$  is right-continuous:  $\lim_{h \rightarrow 0^+} F(x + h) = F(x)$ ;
3.  $F$  has left-hand limits:  $\lim_{h \rightarrow 0^+} F(x - h) = F(x-) = P(X < x)$ ;
4.  $\lim_{x \rightarrow -\infty} F(x) = 0$  and  $\lim_{x \rightarrow +\infty} F(x) = 1$ ;
5.  $F$  has at most a countable number of discontinuities.

Thus, the jump in  $F$  at  $x$  is

$$F(x) - F(x-) = P(X = x).$$

Conversely, some of the above properties can be shown to characterize the distribution function:

**Proposition 1.1** If a function  $F : \mathfrak{N} \rightarrow [0, 1]$  is non-decreasing, right-continuous,  $\lim_{x \rightarrow -\infty} F(x) = 0$ , and  $\lim_{x \rightarrow \infty} F(x) = 1$ , then  $F$  is the distribution function of some random variable.

*Proof* See, e.g., Severini (2005, p. 11). □

**Example 1.5** (Weibull/exponential distribution) Define the *Weibull* distribution function for random variable  $T$  by

$$F(t) = \begin{cases} 0, & \text{if } t < 0, \\ 1 - \exp[-\lambda t^\alpha], & \text{if } t \geq 0, \end{cases}$$

with parameters  $\lambda$  and  $\alpha$ ,  $\lambda > 0, \alpha > 0$ . For  $\alpha = 1$ , this is the *exponential distribution*,  $\text{Exp}(\lambda)$ , for short. While the exponential is not descriptive for empirically observed reaction times, it plays a fundamental role in response time modeling as a building block for more realistic distributions. Its usefulness derives from the fact that exponentially distributed random variables *have no memory*, or are *memoryless*, in the following sense:

$$P(X > s + t | X > t) = P(X > s), \quad \text{for } s, t \geq 0.$$

In the last example, the distribution function is a continuous function. In order to more fully describe the different types of distribution functions that exist, the concept of *Lebesgue integral* is required.<sup>7</sup>

**Definition 1.7** (Distribution function types) A distribution function  $F$  is called

- (i) *discrete* iff for some countable set of numbers  $\{x_j\}$  and point masses  $\{p_j\}$ ,

$$F(x) = \sum_{x_j \leq x} p_j, \quad \text{for all } x \in \mathfrak{N}.$$

The function  $p$  is called *probability function*;

- (ii) *continuous* iff it is continuous for all  $x$ ;
- (iii) *absolutely continuous* iff there exists a non-negative, Lebesgue integrable function  $f$ , such that

$$F(b) - F(a) = \int_a^b f(x) dx, \quad \text{for all } a < x < b.$$

The function  $f$  is called *density* of  $f$ ;

- (iv) *singular* iff  $F \neq 0, F'$  exists and equals 0 a.e. (almost everywhere).

The following *decomposition theorem* (Gut, 2013, pp. 36–38) shows that any distribution can be described uniquely by three types:

**Theorem 1.1** Every distribution function can be decomposed uniquely into a convex combination of three pure types, a discrete one, an absolutely continuous one, and a continuous singular one. Thus, if  $F$  is a distribution function, then

$$F = \alpha F_{ac} + \beta F_d + \gamma F_{cs},$$

where  $\alpha, \beta, \gamma \geq 0$  and  $\alpha + \beta + \gamma = 1$ . This means that

<sup>7</sup> This topic is treated in introductory analysis books (see references in the final section) and some familiarity is assumed here.

- $F_{ac}(x) = \int_{-\infty}^x f(y) dy$ , where  $f(x) = F'_{ac}(x)$  a.e.;
- $F_d$  is a pure jump function with at most a countable number of jumps;
- $F_{cs}$  is continuous and  $F'_{cs} = 0$  a.e.

The exponential distribution is an example of an absolutely continuous distribution function, with density  $f(t) = \lambda \exp[-\lambda t]$ , for  $t \geq 0$ , and  $f(t) = 0$ , for  $t < 0$ .

### 1.2.2 Random vectors, marginal and conditional distribution

**Definition 1.8** (Random vector) An  $n$ -dimensional *random vector* (or, *vector-valued random variable*)  $\mathbf{X}$  on a probability space  $(\Omega, \mathcal{F}, P)$  is a measurable function from the sample space  $\Omega$  to  $\mathbb{R}^n$ ; thus, this requires that the inverse image of any Borel set is  $\mathcal{F}$ -measurable:

$$\mathbf{X}^{-1}(A) = \{\omega \mid \mathbf{X}(\omega) \in A\} \in \mathcal{F}, \quad \text{for all } A \in \mathcal{B}^n,$$

where  $\mathcal{B}^n$  denotes the  $\sigma$ -algebra of Borel sets of  $\mathbb{R}^n$  generated by the  $n$ -dimensional “rectangles” of  $\mathbb{R}^n$ . Random vectors are column vectors:

$$\mathbf{X} = (X_1, X_2, \dots, X_n)',$$

where  $'$  denotes transpose.

The *joint, or multivariate, distribution function* of  $\mathbf{X}$  is

$$F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n),$$

for  $x_k \in \mathbb{R}$ ,  $k = 1, 2, \dots, n$ . This is written more compactly as  $F_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x})$ , where the event  $\{\mathbf{X} \leq \mathbf{x}\}$  is to be interpreted component-wise. For discrete distributions, the joint probability function is defined by

$$p_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n,$$

and in the absolutely continuous case we have a *joint density*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^n F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \partial x_2 \cdots \partial x_n}, \quad \mathbf{x} \in \mathbb{R}^n.$$

Let  $(X, Y)$  be a two-dimensional random vector with joint probability distribution  $p_{X,Y}(x, y)$ . Then the *marginal probability function* is defined as

$$p_X(x) = P(X = x) = \sum_y p_{X,Y}(x, y),$$

and  $p_Y(y)$  is defined analogously. The *marginal distribution function* is obtained by

$$F_X(x) = \sum_{u \leq x} p_X(u) = P(X \leq x, Y < +\infty).$$

In the absolutely continuous case, we have

$$F_X(x) = P(X \leq x, Y < \infty) = \int_{-\infty}^x \int_{-\infty}^{\infty} f_{X,Y}(u, v) du dv,$$

and differentiate to obtain the *marginal density function*,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy.$$

The concept of conditional probability (Definition 1.3) extends to distributions:

**Definition 1.9** Let  $X$  and  $Y$  be discrete, jointly distributed random variables. For  $P(X = x) > 0$ , the *conditional probability function of  $Y$  given that  $X = x$*  equals

$$p_{Y|X=x}(y) = P(Y = y|X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)},$$

and the *conditional distribution function of  $Y$  given that  $X = x$*  is

$$F_{Y|X=x}(y) = \sum_{z \leq y} p_{Y|X=x}(z).$$

**Definition 1.10** Let  $X$  and  $Y$  have a joint absolutely continuous distribution. For  $f_X(x) > 0$ , the *conditional density function of  $Y$  given  $X = x$*  equals

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)},$$

and the *conditional distribution function of  $Y$  given that  $X = x$*  is

$$F_{Y|X=x}(y) = \int_{-\infty}^y f_{Y|X=x}(z) dz.$$

Analogous formulas hold in higher dimensions and for more general distributions.

Sometimes, it is necessary to characterize a multivariate distribution function by its properties. For the bivariate case, we have, in analogy to Proposition 1.1,

**Proposition 1.2** *Necessary and sufficient conditions for a bounded, nondecreasing, and right-continuous function  $F$  on  $\mathbb{R}^2$  to be a bivariate distribution function are:*

1.  $\lim_{x_1 \rightarrow -\infty} F(x_1, x_2) = 0$ , and  $\lim_{x_2 \rightarrow -\infty} F(x_1, x_2) = 0$ ;
2.  $\lim_{(x_1, x_2) \rightarrow (+\infty, +\infty)} F(x_1, x_2) = 1$ ;
3. (rectangle inequality or 2-increasing) for any  $(a_1, a_2), (b_1, b_2)$  with  $a_1 < b_1$ ,  $a_2 < b_2$ ,

$$F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2) \geq 0.$$

This result can be extended to the  $n$ -dimensional case. For a proof, see, e.g., Leadbetter *et al.* (2014, p. 197)

So far, independence has only been defined for events (Definition 1.2). It can be used to define independence of random variables:

**Definition 1.11** (Independence of random variables) The random variables  $X_1, X_2, \dots, X_n$  are *independent* iff, for arbitrary Borel measurable sets  $A_1, A_2, \dots, A_n$ ,

$$P\left(\bigcap_{k=1}^n \{X_k \in A_k\}\right) = \prod_{k=1}^n P(X_k \in A_k).$$

This can be shown to be equivalent to defining independence of the components  $X_1, X_2, \dots, X_n$  of a random vector  $\mathbf{X}$  by

$$F_{\mathbf{X}}(\mathbf{x}) = \prod_{k=1}^n F_{X_k}(x_k), \quad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

For discrete and absolutely continuous distributions, independence is equivalent to factorization of the joint probability function and density into the product of the marginals, respectively.

**Example 1.6** (Independent exponentials) Let  $X_1, X_2, \dots, X_n$  be independent random variables following the exponential distribution with parameters  $\lambda_i$  ( $\lambda_i > 0$ ),  $i = 1 \dots, n$ . Then, it is routine to show that

1.  $Z = \min\{X_1, \dots, X_n\}$  is also exponentially distributed, with parameter  $\lambda_Z = \sum_{i=1}^n \lambda_i$ ;
2. for any  $i$ ,

$$P(Z = X_i) = \frac{\lambda_i}{\lambda_Z}.$$

**Proposition 1.3** Let  $X_1, X_2, \dots, X_n$  be random variables and  $h_1, h_2, \dots, h_n$  be measurable functions. If  $X_1, X_2, \dots, X_n$  are independent, then so are  $h_1(X_1), h_2(X_2), \dots, h_n(X_n)$ .

This proposition easily follows from Definition 1.11. The following bivariate exponential distribution, often denoted as BVE, has generated much interest and will be referred to again in this chapter:

**Example 1.7** (Marshall–Olkin BVE) Let  $X_1 = \min\{Z_1, Z_{12}\}$ , and  $X_2 = \min\{Z_2, Z_{12}\}$  where  $Z_1, Z_2, Z_{12}$  are independent exponential random variables with parameters  $\lambda_1, \lambda_2, \lambda_{12}$ , respectively. Marshall and Olkin (1967) define a bivariate distribution by

$$\bar{F}(x_1, x_2) \equiv P(X_1 > x_1, X_2 > x_2) = \exp[-\lambda_1 x_1 - \lambda_2 x_2 - \lambda_{12} \max(x_1, x_2)].$$

Then,

$$\begin{aligned} P(X_1 = X_2) &= P(Z_{12} = \min\{Z_1, Z_2, Z_{12}\}) \\ &= \frac{\lambda_{12}}{\lambda} \\ &> 0, \end{aligned}$$

from Example 1.6, with  $\lambda = \lambda_1 + \lambda_2 + \lambda_{12}$ . Thus, the BVE distribution  $F(x_1, x_2)$  has a singular component. We list without proof (see, e.g., Barlow and Proschan, 1975) some further properties of BVE:

- (a) BVE has exponential marginals, with

$$\begin{aligned}\bar{F}_1(x_1) &= P(X_1 > x_1) = \exp[-(\lambda_1 + \lambda_{12})x_1], \quad \text{for } t \geq 0, \\ \bar{F}_2(x_2) &= P(X_2 > x_2) = \exp[-(\lambda_2 + \lambda_{12})x_1], \quad \text{for } t \geq 0;\end{aligned}$$

- (b) BVE has the lack-of-memory property:

$$P(X_1 > x_1 + t, X_2 > x_2 + t | X_1 > t, X_2 > t) = P(X_1 > x_1, X_2 > x_2),$$

for all  $x_1 \geq 0, x_2 \geq 0, t \geq 0$ , and BVE is the only bivariate distribution with exponential marginals possessing this property;

- (c) BVE distribution is a mixture of an absolutely continuous distribution  $F_{ac}(x_1, x_2)$  and a singular distribution  $F_{cs}(x_1, x_2)$ :

$$\bar{F}(x_1, x_2) = \frac{\lambda_1 + \lambda_2}{\lambda} \bar{F}_{ac}(x_1, x_2) + \frac{\lambda_{12}}{\lambda} \bar{F}_s(x_1, x_2),$$

where  $\bar{F}_s(x_1, x_2) = \exp[-\lambda \max(x_1, x_2)]$  and,

$$\begin{aligned}\bar{F}_{ac}(x_1, x_2) &= \frac{\lambda}{\lambda_1 + \lambda_2} \exp[-\lambda_1 x_1 - \lambda_2 x_2 - \lambda_{12} \max(x_1, x_2)] \\ &\quad - \frac{\lambda_{12}}{\lambda_1 + \lambda_2} \exp[-\lambda \max(x_1, x_2)].\end{aligned}$$

### 1.2.3 Expectation, other moments, and tail probabilities

The consistency of the following definitions is based on the theory of Lebesgue integration and the Riemann–Stieltjes integral (some knowledge of which is presupposed here).

**Definition 1.12** (Expected value) Let  $X$  be a real-valued random variable on a probability space  $(\Omega, \mathcal{F}, P)$ . The *expected value* of  $X$  (or *mean* of  $X$ ) is the integral of  $X$  with respect to measure  $P$ :

$$EX = \int_{\Omega} X(\omega) dP(\omega) = \int X dP.$$

For  $X \geq 0$ ,  $EX$  is always defined (it may be infinite); for the general  $X$ ,  $EX$  is defined if at least one of  $EX^+$  or  $EX^-$  is finite.<sup>8</sup> in which case

$$EX = EX^+ - EX^-;$$

if both values are finite, that is, if  $E|X| < \infty$ , we say that  $X$  is *integrable*.

Let  $X, Y$  be integrable random variables. The following are basic consequences of the (integral) definition of expected value:

<sup>8</sup>  $X^+ = \sup\{+X, 0\}$  and  $X^- = \sup\{-X, 0\}$ .

1. If  $X = 0$  a.s., then  $\mathbb{E}X = 0$ ;
2.  $|X| < \infty$  a.s., that is,  $P(|X| < \infty) = 1$ ;
3. If  $\mathbb{E}X > 0$ , then  $P(X > 0) > 0$ ;
4. Linearity:  $\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y$ , for any  $a, b \in \mathbb{R}$ ;
5.  $\mathbb{E}XI\{X \neq 0\} = \mathbb{E}X$ ;<sup>9</sup>
6. Equivalence: If  $X = Y$  a.s., then  $\mathbb{E}X = \mathbb{E}Y$ ;
7. Domination: If  $Y \leq X$  a.s., then  $\mathbb{E}X \leq \mathbb{E}Y$ ;
8. Domination: If  $|Y| \leq X$  a.s., then  $\mathbb{E}|Y| \leq \mathbb{E}|X|$ .

**Remark 1.7** Define an integral over measurable sets  $A \in \mathcal{F}$ :

$$\mu_X(A) = \mathbb{E}XI\{A\} = \int_A X \, dP = \int_{\Omega} XI\{A\} \, dP.$$

In other words,  $\mu_X(\cdot)$  is an “ordinary” expectation applied to the random variable  $XI\{\cdot\}$ . Note that  $\mu_{I\{B\}}(A) = P(B \cap A)$ , for  $B \in \mathcal{F}$ .

The expected value of random variables has so far been defined in terms of integrals over the sample space of the underlying probability space. Just as the probability space behind the random variables is “hidden” once they have been defined by the induced measure (see Remark 1.3), one can compute an integral on the real line rather than over the probability space invoking the Riemann–Stieltjes integral: suppose  $X$  is integrable, then

$$\mathbb{E}X = \int_{\Omega} X \, dP = \int_{\mathbb{R}} x \, dF_X(x).$$

**Example 1.8** (Example 1.5 cont’d) Expected value (mean) of an exponentially distributed random variable  $T$  is

$$\mathbb{E}T = \int_{\mathbb{R}} t \, dF_T(t) = \int_{\mathbb{R}} t f_T(t) \, dt = \int_0^{\infty} t \lambda \exp[-\lambda t] \, dt = 1/\lambda,$$

using the fact that  $T$  is absolutely continuous, that is, a density exists.

Because measurable functions of random variables are new random variables (Remark 1.6), one can compute the expected values of functions of random variables: let  $X$  be a random variable and suppose that  $g$  is a measurable function such that  $g(X)$  is an integrable random variable. Then

$$\mathbb{E}g(X) = \int_{\Omega} g(X) \, dP = \int_{\mathbb{R}} g(x) \, dF_X(x).$$

For proof of this and the previous observations, see Gut (2013, p. 60).

**Example 1.9** (Expected value of  $1/T$ ) Let  $T$  be a positive random variable with density  $f_T(t)$ . Then  $T^{-1}$  has density  $f_T(1/t)/t^2$ . This is useful in RT analysis if one wants to go from reaction time ( $t$ ) to “speed” ( $1/t$ ).

<sup>9</sup>  $I\{A\}$  is the indicator function for a set  $A$ .

An alternative way of computing expected values, which is often useful in RT analyses, is based on the following *tail probability formula*:

**Remark 1.8** For a real-valued random variable with distribution function  $F$ ,

$$EX = \int_0^\infty (1 - F(x)) dx - \int_{-\infty}^0 F(x) dx,$$

assuming  $E|X| < \infty$ . For nonnegative random variables this reduces to

$$EX = \int_0^\infty (1 - F(x)) dx,$$

where both sides of the equation are either finite or infinite. If  $X$  is a nonnegative, integer-valued random variable, then

$$EX = \sum_{n=1}^{+\infty} P(X \geq n).$$

The expected value measures the “center of gravity” of a distribution, it is a *location* measure. Other properties of a distribution, like dispersion (spread), skewness (asymmetry), and kurtosis (peakedness), are captured by different types of *moments*:

**Definition 1.13** Let  $X$  be a random variable. The

1. *moments* are  $EX^n$ ,  $n = 1, 2, \dots$ ;
2. *central moments* are  $E(X - EX)^n$ ,  $n = 1, 2, \dots$ ;
3. *absolute moments* are  $E|X|^n$ ,  $n = 1, 2, \dots$ ;
4. *absolute central moments* are  $E|X - E|^n$ ,  $n = 1, 2, \dots$ .

Clearly, the first moment is the mean,  $EX = \mu$ . The second central moment is called *variance*:

$$\text{Var}X = E(X - \mu)^2 \quad (= EX^2 - (EX)^2).$$

A (standardized) measure of skewness is obtained using the third central moment:

$$\gamma_1 = \frac{E(X - \mu)^3}{[\text{Var}X]^{3/2}}. \tag{1.1}$$

A classic, but not undisputed,<sup>10</sup> measure of kurtosis is based on the fourth (standardized) moment:

$$\beta_2 = \frac{E(X - \mu)^4}{(\text{Var}X)^2}. \tag{1.2}$$

Although moments often provide a convenient summary of the properties of a distribution function, they are not always easy to work with or may not be available

<sup>10</sup> See Balanda and MacGillivray (1990).

in simple, explicit form. Alternative ways of comparing distributions with respect to location, dispersion, skewness, and kurtosis will be discussed in Section 1.3.9.

Joint moments of two or more random variables, introduced next, are also commonly used.

**Definition 1.14** Let  $(X, Y)'$  be a random vector. The *covariance* of  $X$  and  $Y$  is

$$\text{Cov}(X, Y) = E(X - EX)(Y - EY) (= EXY - EXEY).$$

A bivariate version of the tail-probability formula in Remark 1.8, also known as *Hoeffding's identity* (Hoeffding, 1940), will be quite helpful later on in Section 1.3.7:

**Proposition 1.4** Let  $F_{XY}$ ,  $F_X$ , and  $F_Y$  denote the joint and marginal distribution functions for random variables  $X$  and  $Y$ , and suppose that  $E|XY|$ ,  $E|X|$ , and  $E|Y|$  are finite; then

$$\text{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [F_{XY}(x, y) - F_X(x)F_Y(y)] dx dy.$$

An instructive proof of this is found in Shea (1983). Finally, the (linear) *correlation coefficient* is defined as

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}X \text{Var}Y}}.$$

## 1.2.4 Product spaces and convolution

**Definition 1.15** (Product space) For any finite number of measurable spaces,

$$(\Omega_1, \mathcal{F}_1), (\Omega_2, \mathcal{F}_2), \dots, (\Omega_n, \mathcal{F}_n),$$

one can construct a *product measurable space* in a natural way. The set of all ordered  $n$ -tuples  $(\omega_1, \dots, \omega_n)$ , with  $\omega_i \in \Omega_i$  for each  $i$ , is denoted by  $\Omega_1 \times \dots \times \Omega_n$  is called the *product* of the  $\{\Omega_i\}$ . A set of the form

$$A_1 \times \dots \times A_n = \{(\omega_1, \dots, \omega_n) \in \Omega_1 \times \dots \times \Omega_n \mid \omega_i \in \Omega_i \text{ for each } i\},$$

with  $A_i \in \mathcal{F}_i$  for each  $i$ , is called a *measurable rectangle*. The *product  $\sigma$ -algebra*  $\mathcal{F}_1 \otimes \dots \otimes \mathcal{F}_n$  is defined to be the  $\sigma$ -algebra generated by the set of all measurable rectangles. Within the measure-theoretic approach to probability (Pollard, 2002), one can then construct a unique product measure  $\mathbb{P}$  on the product  $\sigma$ -algebra such that

$$\mathbb{P}(A_1 \times \dots \times A_n) = \prod_{k=1}^n P_k(A_k) \quad \text{for all } A_k \in \mathcal{F}_k, 1 \leq k \leq n.$$

Note that the product probability measure has built-in independence.

Given that models of response time typically make the assumption that observable RT is composed of different stages (e.g., Donders' method of subtraction)

being able to calculate the distribution function for the sum (or difference) of random variables is clearly essential. Consider the product space  $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, P_1 \times P_2)$ , and suppose that  $(X_1, X_2)$  is a two-dimensional random variable whose marginal distribution functions are  $F_1$  and  $F_2$ , respectively. The convolution formula provides the distribution of  $X_1 + X_2$ .

**Proposition 1.5** *In the above setting, for independent  $X_1$  and  $X_2$ ,*

$$F_{X_1+X_2}(u) = \int_{-\infty}^{\infty} F_1(u-y) dF_2(y).$$

*If, in addition,  $X_2$  is absolutely continuous with density  $f_2$ , then*

$$F_{X_1+X_2}(u) = \int_{-\infty}^{\infty} F_1(u-y) f_2(y) dy.$$

*If  $X_1$  is absolutely continuous with density  $f_1$ , the density of the sum equals*

$$f_{X_1+X_2}(u) = \int_{-\infty}^{\infty} f_1(u-y) dF_2(y).$$

*If both are absolutely continuous, then*

$$f_{X_1+X_2}(u) = \int_{-\infty}^{\infty} f_1(u-y) f_2(y) dy.$$

*Proof* See, e.g., Gut (2013, p. 68). □

**Example 1.10** (Schwarz, 2001) Let  $D$  be a random variable following an *inverse Gauss* (or *Wald*) distribution (see Example 1.3) with density<sup>11</sup>

$$f(w | \mu, \sigma, a) = \frac{a}{\sigma \sqrt{2\pi w^3}} \exp\left[-\frac{(a - \mu w)^2}{2\sigma^2 w}\right];$$

The distribution function is

$$F(w | \mu, \sigma, a) = \Phi\left(\frac{\mu w - a}{\sigma \sqrt{w}}\right) + \exp\left(\frac{2a\mu}{\sigma^2}\right) \Phi\left(-\frac{\mu w + a}{\sigma \sqrt{w}}\right),$$

with  $\Phi$  the standard normal distribution function. Let  $M$  be an exponentially distributed random variable with density

$$g(w | \gamma) = \gamma \exp(-\gamma w).$$

If  $D$  and  $M$  are independent, their sum has density  $h$

$$h(t | \mu, \sigma, \gamma) = \int_0^t f(w | \mu, \sigma, a) \times g(t-w | \gamma) dw.$$

Inserting for  $f$  and  $g$ , after some algebraic manipulations, this yields

$$h(t | \mu, \sigma, a, \gamma) = \gamma \exp\left[-\gamma t + \frac{a(\mu - k)}{\sigma^2}\right] \times F(t | k, \sigma, a),$$

<sup>11</sup> This parametrization expresses  $D$  as the first-passage time through a level  $a > 0$  in a Wiener diffusion process starting at  $x = 0$  with drift  $\mu > 0$  and variance  $\sigma^2 > 0$ .

with  $k \equiv \sqrt{\mu^2 - 2\gamma\sigma^2} \geq 0$ . The latter condition amounts to  $EM = 1/\gamma \geq 2\text{Var}D/ED$ ; if  $D$  is interpreted as time to detect and identify a signal and  $M$  the time to execute the appropriate response, this condition is typically satisfied. Otherwise, a more elaborate computation of the convolution is required (see Schwarz, 2002). In the above interpretation of  $D$ , often the Wald distribution is replaced by the normal distribution (Hohle, 1965; Palmer *et al.*, 2011).

### 1.2.5 Stochastic processes

A *stochastic process*  $X$  is a collection of random variables  $\{X_t : t \in T\}$ , all defined on the same probability space, say  $(\Omega, \mathcal{F}, P)$ . The *index set*  $T$  can be an arbitrary set but, in most cases,  $T$  represents time.  $X$  can be a *discrete-parameter* process,  $T = \{1, 2, \dots\}$ , or a *continuous-parameter* process, with  $T$  an interval of the real line. For a fixed  $\omega \in \Omega$ , the function  $X_t(\omega)$  of  $t$  is called a *realization* or *sample path* of  $X$  at  $\omega$ . For each  $n$ -tuple  $\mathbf{t} = (t_1, \dots, t_n)$  of distinct elements of  $T$ , the random vector  $(X_{t_1}, \dots, X_{t_n})$  has a joint distribution function

$$F_{\mathbf{t}}(\mathbf{x}) = P(X_{t_1} \leq x_1, \dots, X_{t_n} \leq x_n).$$

As  $\mathbf{t}$  ranges over all  $n$ -tuples of  $T$ , the collection  $\{F_{\mathbf{t}}\}$  is the *system of finite-dimension distributions* of  $X$  from which the distributional properties of the stochastic process can be studied.<sup>12</sup>

A stochastic process  $\{N(t), t \geq 0\}$  is called a *counting process* if  $N(t)$  represents the total number of “events” that have occurred up to time  $t$ . Hence, a counting process  $N(t)$  satisfies:

1.  $N(t) \geq 0$ ;
2.  $N(t)$  is integer-valued;
3. If  $s < t$ , then  $N(s) \leq N(t)$ ;
4. For  $s < t$ ,  $N(t) - N(s)$  equals the number of events that have occurred in the interval  $(s, t]$ .

A counting process has *independent increments* if the numbers of events that occur in disjoint time intervals are independent random variables. It has *stationary increments* if the distribution of the number of events in the interval  $(t_1 + s, t_2 + s]$ , that is,  $N(t_2 + s) - N(t_1 + s)$ , does not depend on  $s$ , for all  $t_1 < t_2$  and  $s > 0$ .

Let  $X_1$  be the time of the first event of a counting process, and  $X_n$  the time between the  $(n-1)$ th and the  $n$ th event. The sequence  $\{X_n, n \geq 1\}$  is called the *sequence of interarrival times*. Moreover,

$$S_n = \sum_{i=1}^n X_i, \quad n \geq 1,$$

is the *waiting time* until the  $n$ th event.

<sup>12</sup> For the general measure-theoretic approach to stochastic processes, including *Kolmogorov's existence theorem*, see e.g., Leadbetter *et al.* (2014).

### The Poisson process

The most simple counting process is the Poisson process, which is a basic building block for many RT models. It can be defined in several, equivalent ways (see, e.g., Gut, 2013). Here we introduce it as follows:

**Definition 1.16** The counting process  $\{N(t), t \geq 0\}$  is said to be a *Poisson process with rate  $\lambda$* ,  $\lambda > 0$ , if:

1.  $N(0) = 0$ ;
2.  $N(t)$  has independent increments;
3. the number of events in any interval of length  $t$  is Poisson distributed with mean  $\lambda t$ ; that is, for all  $s, t \geq 0$ ,

$$P(N(t+s) - N(s) = n) = \exp[-\lambda t] \frac{(\lambda t)^n}{n!}, \quad n = 0, 1, \dots$$

Obviously, a Poisson process has stationary increments. Moreover,  $E[N(t)] = \lambda t$ . For the interarrival times, one has

**Proposition 1.6** *The interarrival times  $\{X_n, n \geq 1\}$  of a Poisson process with rate  $\lambda$  are independent identically distributed (i.i.d.) exponential random variables with mean  $1/\lambda$ .*

*Proof* Obviously,

$$P(X_1 > t) = P(N(t) = 0) = \exp[-\lambda t],$$

so  $X_1$  has the exponential distribution. Then

$$\begin{aligned} P(X_2 > t | X_1 = s) &= P(\text{no events in } (s, s+t] | X_1 = s) \\ &= P(\text{no events in } (s, s+t]) \quad (\text{by independent increments}) \\ &= \exp[-\lambda t] \quad (\text{by stationary increments}). \end{aligned} \quad \square$$

Given the memoryless property of the exponential, the last result shows that the Poisson process *probabilistically* restarts itself at any point in time. For the waiting time until the  $n$ th event of the Poisson process, the density is

$$f(t) = \lambda \exp[-\lambda t] \frac{(\lambda t)^{n-1}}{(n-1)!}, \quad t \geq 0, \quad (1.3)$$

i.e., the gamma density,  $\text{Gamma}(n, \lambda)$ . This follows using convolution (or moment generating functions), but it can also be derived by appealing to the following logical equivalence:

$$N(t) \geq n \Leftrightarrow S_n \leq t.$$

Hence,

$$P(S_n \leq t) = P(N(t) \geq n) \quad (1.4)$$

$$= \sum_{j=n}^{\infty} \exp[-\lambda t] \frac{(\lambda t)^j}{j!}, \quad (1.5)$$

which upon differentiation yields the gamma density.

Note that a function  $f$  is said to be  $o(h)$  if

$$\lim_{h \rightarrow 0} \frac{f(h)}{h} = 0.$$

One alternative definition of the Poisson process is contained in the following:

**Proposition 1.7** *The counting process  $\{N(t), t \geq 0\}$  is a Poisson process with rate  $\lambda$ ,  $\lambda > 0$ , if and only if:*

1.  $N(0) = 0$ ;
2. *the process has stationary and independent increments*;
3.  $P(N(h) = 1) = \lambda h + o(h)$ ;
4.  $P(N(h) \geq 2) = o(h)$ .

*Proof* See Ross (1983, pp. 32–34). □

### The non-homogeneous Poisson process

Dropping the assumption of stationarity leads to a generalization of the Poisson process that also plays an important role in certain RT model classes, the *counter models*.

**Definition 1.17** The counting process  $\{N(t), t \geq 0\}$  is a *nonstationary, or non-homogeneous Poisson process with intensity function  $\lambda(t)$ ,  $t \geq 0$* , if:

1.  $N(0) = 0$ ;
2.  $\{N(t), t \geq 0\}$  has independent increments;
3.  $P(N(t+h) - N(t) = 1) = \lambda(t)h + o(h)$ ;
4.  $P(N(t+h) - N(t)) \geq 2) = o(h)$ .

Letting

$$m(t) = \int_0^t \lambda(s) ds,$$

it can be shown that

$$\begin{aligned} P(N(t+s) - N(t) = n) \\ = \exp[-(m(t+s) - m(t))] [m(t+s) - m(t)]^n / n!, \quad n = 0, 1, 2 \dots, \end{aligned} \tag{1.6}$$

that is,  $N(t+s) - N(t)$  is Poisson distributed with mean  $m(t+s) - m(t)$ .

A non-homogeneous Poisson process allows for the possibility that events may be more likely to occur at certain times than at other times. Such a process can be interpreted as being a random sample from a homogeneous Poisson process, if function  $\lambda(t)$  is bounded, in the following sense (p. 46 Ross, 1983):

Let  $\lambda(t) \leq \lambda$ , for all  $t \geq 0$ ; suppose that an event of the process is counted with probability  $\lambda(t)/\lambda$ , then the process of counted events is a non-homogeneous Poisson process with intensity function  $\lambda(t)$ . Properties 1, 2, and 3 of Definition 1.17 follow because they also hold for the homogeneous Poisson process. Property 4

follows because

$$\begin{aligned} P(\text{one counted event in } (t, t+h)) &= P(\text{one event in } (t, t+h)) \frac{\lambda(t)}{\lambda} + o(h) \\ &= \lambda h \frac{\lambda(t)}{\lambda} + o(h) \\ &= \lambda(t)h + o(h). \end{aligned}$$

An example of this process in the context of *record values* is presented at the end of Section 1.3.4.

## 1.3 Specific topics

### 1.3.1 Exchangeability

When stochastic independence does not hold, the multivariate distribution of a random vector can be a very complicated object that is difficult to analyze. If the distribution is *exchangeable*, however, the associated distribution function  $F$  can often be simplified and written in a quite compact form. Exchangeability intuitively means that the dependence structure between the components of a random vector is completely symmetric and does not depend on the ordering of the components (Mai and Scherer, 2012, p. 39).

**Definition 1.18** (Exchangeability) Random variables  $X_1, \dots, X_n$  are *exchangeable* if, for the random vector  $(X_1, \dots, X_n)$ ,

$$(X_1, \dots, X_n) =_d (X_{i_1}, \dots, X_{i_n})$$

for all permutations  $(i_1, \dots, i_n)$  of the subscripts  $(1, \dots, n)$ . Furthermore, an infinite sequence of random variables  $X_1, X_2, \dots$  is *exchangeable* if any finite subsequence of  $X_1, X_2, \dots$  is exchangeable.

Notice that a subset of exchangeable random variables can be shown to be exchangeable. Moreover, any collection of independent identically distributed random variables is exchangeable. The distribution function of a random vector of exchangeable random variables is invariant with respect to permutations of its arguments. The following proposition shows that the correlation structure of a finite collection of exchangeable random variables is limited:

**Proposition 1.8** *Let  $X_1, \dots, X_n$  be exchangeable with existing pairwise correlation coefficients  $\rho(X_i, X_j) = \rho$ ,  $i \neq j$ . Then  $\rho \geq -1/(n-1)$ .*

*Proof* Without loss of generality, we scale the variables such that  $EZ_i = 0$  and  $EZ_i^2 = 1$ . Then

$$0 \leq E \left( \sum_i Z_i \right)^2 = \sum_i EZ_i^2 + \sum_{i \neq j} EZ_i Z_j = n + n(n-1)\rho,$$

from which the inequality follows immediately.  $\square$

Note that this also implies that  $\rho \geq 0$ , for an infinite sequence of exchangeable random variables. The assumption that a finite number of random variables is exchangeable is significantly different from the assumption that they are a finite segment of an infinite sequence of exchangeable random variables (for a simple example, see Galambos, 1978, pp. 127–128).

The most important result for exchangeable random variables is known as “de Finetti’s theorem,” which is often stated in words such as “An infinite exchangeable sequence is a mixture of i.i.d. sequences.” The precise result, however, requires measure theoretic tools like *random measures*. First, let us consider the “Bayesian viewpoint” of sampling.

Let  $X_1, \dots, X_n$  be observations on a random variable with distribution function  $F(x, \theta)$ , where  $\theta$  is a random variable whose probability function  $p$  is assumed to be independent of the  $X_j$ . For a given value of  $\theta$ , the  $X_j$  are assumed to be i.i.d. Thus, the distribution of the sample is

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = \int \prod_{j=1}^n F(x_j, \theta) p(\theta) d\theta = E \left[ \prod_{j=1}^n F(x_j, \theta) \right]. \quad (1.7)$$

Obviously,  $X_1, \dots, X_n$  are exchangeable random variables. The following theorem addresses the converse of this statement.

**Theorem 1.2** (“de Finetti’s theorem”) *An infinite sequence  $X_1, X_2, \dots$  of random variables on a probability space  $(\Omega, \mathcal{F}, P)$  is exchangeable if, and only if, there is a sub- $\sigma$ -algebra  $\mathcal{G} \subset \mathcal{F}$  such that, given  $\mathcal{G}$ ,  $X_1, X_2, \dots$  are a.s. independent and identically distributed, that is,*

$$P(X_1 \leq x_1, \dots, X_n \leq x_n | \mathcal{G}) = \prod_{j=1}^n P(X_1 \leq x_j | \mathcal{G}) \text{ a.s.},$$

for all  $x_1, \dots, x_n \in \mathfrak{N}$ .

Consequently, the  $n$ -dimensional distribution of the  $X_j$  can always be expressed as in Equation (1.7). A more complete statement of the theorem and a proof can be found, for example, in Aldous (1985). The above theorem is considered a key result in the context of Bayesian (hierarchical) modeling, with the probability function of  $\theta$  taking the role of a *prior distribution*.

### 1.3.2 Quantile functions

While specific distributions are typically described by certain moments (mean and variance) and transforms thereof, in particular skewness ( $\gamma_1$  from Equation (1.1)) and kurtosis ( $\beta_2$  from Equation (1.2)), this approach has certain shortcomings. For example, the variance may not be finite; there exist asymmetric distributions

with  $\gamma_1 = 0$ ; and  $\beta_2$  does not reflect the sign of the difference between mean and median, which is a traditional basis for defining skewness. The concept of a *quantile function* is closely related to that of a distribution function and plays an increasingly important role in RT analysis.

The specification of a distribution through its quantile function takes away the need to describe a distribution through its moments. Alternative measures in terms of quantiles are available, e.g., the median as a measure of location, defined as  $M = Q(0.5)$ , or the interquartile range, as a measure of dispersion, defined as  $IQR = Q(0.75) - Q(0.25)$ . Another measure of dispersion in terms of quantiles, the *quantile spread*, will be discussed in Section 1.3.9.

**Definition 1.19** (Quantile function) Let  $X$  be a real-valued random variable with distribution function  $F(x)$ . Then the *quantile function* of  $X$  is defined as

$$Q(u) = F^{-1}(u) = \inf\{x \mid F(x) \geq u\}, \quad 0 \leq u \leq 1. \quad (1.8)$$

For every  $-\infty < x < +\infty$  and  $0 < u < 1$ , we have

$$F(x) \geq u \quad \text{if, and only if, } Q(u) \leq x.$$

Thus, if there exists  $x$  with  $F(x) = u$ , then  $F(Q(u)) = u$  and  $Q(u)$  is the smallest value of  $x$  satisfying  $F(x) = u$ . If  $F(x)$  is continuous and strictly increasing,  $Q(u)$  is the unique value  $x$  such that  $F(x) = u$ .

**Example 1.11** Let  $X$  be an  $\text{Exp}(\lambda)$  random variable,  $F(x) = 1 - \exp[-\lambda x]$ , for  $x > 0$  and  $\lambda > 0$ . Then  $Q(u) = \lambda^{-1}(-\log(1-u))$ .

If  $f(x)$  is the probability density function of  $X$ , then  $f(Q(u))$  is called the *density quantile function*. The derivative of  $Q(u)$ , i.e.,

$$q(u) = Q'(u),$$

is known as the *quantile density function* of  $X$ . Differentiating  $F(Q(u)) = u$ , we find

$$q(u)f(Q(u)) = 1. \quad (1.9)$$

We collect a number of properties of quantile functions:

1.  $Q(u)$  is non-decreasing on  $(0, 1)$  with  $Q(F(x)) \leq x$ , for all  $-\infty < x < +\infty$  for which  $0 < F(x) < 1$ ;
2.  $F(Q(u)) \geq u$ , for any  $0 < u < 1$ ;
3.  $Q(u)$  is continuous from the left, or  $Q(u-) = Q(u)$ ;
4.  $Q(u+) = \inf\{x : F(x) > u\}$  so that  $Q(u)$  has limits from above;
5. any jumps of  $F(x)$  are flat points of  $Q(u)$  and flat points of  $F(x)$  are jumps of  $Q(u)$ ;
6. a non-decreasing function of a quantile function is a quantile function;
7. the sum of two quantile functions is again a quantile function;

- 
8. two quantile density functions, when added, produce another quantile density function;
  9. if  $X$  has quantile function  $Q(u)$ , then  $1/X$  has quantile function  $1/Q(1 - u)$ .

The following simple but fundamental fact shows that any distribution function can be conceived as arising from the uniform distribution transformed by  $Q(u)$ :

**Proposition 1.9** *If  $U$  is a uniform random variable over  $[0, 1]$ , then  $X = Q(U)$  has its distribution function as  $F(x)$ .*

*Proof*

$$P(Q(U) \leq x) = P(U \leq F(x)) = F(x).$$

□

### 1.3.3 Survival analysis

#### Survival function and hazard function

Given that time is a central notion in survival analysis, actuarial, and reliability theory, it is not surprising that many concepts from these fields have an important role to play in the analysis of psychological processes as well. In those contexts, realization of a random variable is interpreted as a “critical event,” e.g., the death of an individual or the breakdown of a system of components.

Let  $F_X(x)$  be the distribution function of random variable  $X$ ; then  $S_X(x) = 1 - F_X(x) = P(X > x)$ , for  $x \in \mathbb{R}$ , is called the *survival function* (or, *tail function*) of  $X$ . From the properties of distribution functions, it follows that the survival function is non-increasing, right-continuous and such that

$$\lim_{x \rightarrow -\infty} S_X(x) = 1 \text{ and } \lim_{x \rightarrow +\infty} S_X(x) = 0.$$

While a more general framework for the following concepts exists, e.g., including discrete distribution functions, in the remainder of this section we assume all random variables or vectors to be nonnegative with absolutely continuous distribution functions.

**Definition 1.20** The hazard rate  $h_X$  of  $X$  with distribution function  $F_X(x)$  and density  $f_X(x)$  is defined as

$$h_X(x) = \frac{f_X(x)}{S_X(x)} = -\frac{d \log S_X(x)}{dx}. \quad (1.10)$$

In reliability theory, the hazard rate is often called *failure rate*. The following, equivalent definition helps in understanding its meaning.

#### Proposition 1.10

$$h_X(x) = \lim_{\Delta x \downarrow 0} \frac{P(x < X \leq x + \Delta x | X > x)}{\Delta x}. \quad (1.11)$$

*Proof* For  $P(X > x) > 0$ ,

$$\begin{aligned} P(x < X \leq x + \Delta x | X > x) &= \frac{P(x < X \leq x + \Delta x)}{1 - F_X(x)} \\ &= \frac{F_X(x + \Delta x) - F_X(x)}{1 - F_X(x)}, \end{aligned}$$

whence it follows that

$$\begin{aligned} \lim_{\Delta x \downarrow 0} \frac{P(x < X \leq x + \Delta x | X > x)}{\Delta x} &= \frac{1}{1 - F_X(x)} \lim_{\Delta x \downarrow 0} \frac{F_X(x + \Delta x) - F_X(x)}{\Delta x} \\ &= \frac{1}{S_X(x)} f_X(x). \quad \square \end{aligned}$$

Thus, for “sufficiently” small  $\Delta x$ , the product  $h_X(x) \cdot \Delta x$  can be interpreted as the probability of the “critical event” occurring in the next instant of time given it has not yet occurred. Or, written with the  $o()$  function:

$$P(x < X \leq x + \Delta x | X > x) = h_X(x)\Delta x + o(\Delta x).$$

The following conditions have been shown to be necessary and sufficient for a function  $h(x)$  to be the hazard function of some distribution (see Marshall and Olkin, 2007).

1.  $h(x) \geq 0$ ;
2.  $\int_0^x h(t) dt < +\infty$ , for some  $x > 0$ ;
3.  $\int_0^{+\infty} h(t) dt = +\infty$ ;
4.  $\int_0^x h(t) dt = +\infty$  implies  $h(y) = +\infty$ , for every  $y > x$ .

It is easy to see that the exponential distribution with parameter  $\lambda > 0$  corresponds to a constant hazard function,  $h_X(x) = \lambda$  expressing the “lack of memory” property of that distribution.

A related quantity is the *cumulative* (or, *integrated*) *hazard function*  $H(x)$  defined as

$$H(x) = \int_0^x h(t) dt = -\log S(x).$$

Thus,

$$S(x) = \exp \left[ - \int_0^x h(t) dt \right],$$

showing that the distribution function can be regained from the hazard function.

**Example 1.12** (Hazard function of the minimum) Let  $X_1, X_2, \dots, X_n$  be independent random variables, defined on a common probability space, and

$Z = \min(X_1, X_2, \dots, X_n)$ . Then

$$\begin{aligned} S_Z(x) &= P(Z > x) \\ &= P(X_1 > x, X_2 > x, \dots, X_n > x) \\ &= S_{X_1}(x) \dots S_{X_n}(x). \end{aligned}$$

Logarithmic differentiation leads to

$$h_Z(x) = h_{X_1}(x) + \dots + h_{X_n}(x).$$

In reliability theory, this model constitutes a *series system* with  $n$  independent components, each having possibly different life distributions: the system breaks down as soon as any of the components does, i.e., the “hazards” cumulate with the number of components involved. For example, the Marshall–Olkin BVE distribution (Example 1.7) can be interpreted to describe a series system exposed to three independent “fatal” breakdowns with a hazard function  $\lambda = \lambda_1 + \lambda_2 + \lambda_{12}$ .

### Hazard quantile function

For many distributions, like the normal, the corresponding quantile function cannot be obtained in closed form, and the solution of  $F(x) = u$  must be obtained numerically. Similarly, there are quantile functions for which no closed-form expressions for  $F(x)$  exists. In that case, analyzing the distribution via its hazard function is of limited use, and a translation in terms of quantile functions is more promising (Nair *et al.*, 2013).

We assume an absolutely continuous  $F(x)$  so that all quantile related functions are well defined. Setting  $x = Q(u)$  in Equation (1.10) and using the relationship  $f(Q(u)) = [q(u)]^{-1}$ , the *hazard quantile function* is defined as

$$H(u) = h(Q(u)) = [(1 - u)q(u)]^{-1}. \quad (1.12)$$

From  $q(u) = [(1 - u)H(u)]^{-1}$ , we have

$$Q(u) = \int_0^u \frac{dv}{(1 - v)H(v)}.$$

The following example illustrates these notions.

**Example 1.13** (Nair *et al.*, 2013, p. 47) Taking

$$Q(u) = u^{\theta+1}(1 + \theta(1 - u)), \quad \theta > 0,$$

we have

$$q(u) = u^\theta[1 + \theta(\theta + 1)(1 - u)],$$

and so

$$H(u) = [(1 - u)u^\theta(1 + \theta(\theta + 1)(1 - u))]^{-1}.$$

Note that there is no analytic solution for  $x = Q(u)$  that gives  $F(x)$  in terms of  $x$ .

### Competing risks models: hazard-based approach

An extension of the hazard function concept, relevant for response times involving, for example, a choice between different alternatives, is to assume that not only the time of breakdown but also its *cause*  $C$  are observable. We define a joint *sub-distribution function*  $F$  for  $(X, C)$ , with  $X$  the time of breakdown and  $C$  its cause, where  $C = 1, \dots, p$ , a (small) number of labels for the different causes:

$$F(x, j) = P(X \leq x, C = j),$$

and a *sub-survival distribution* defined as

$$S(x, j) = P(X > x, C = j).$$

Then,  $F(x, j) + S(x, j) = p_j = P(C = j)$ . Thus,  $F(x, j)$  is not a proper distribution function because it only reaches the values  $p_j$  instead of 1 when  $x \rightarrow +\infty$ . It is assumed implicitly that  $p_j > 0$  and  $\sum_1^p p_j = 1$ .

Note that  $S(x, j)$  is not, in general, equal to the probability that  $X > x$  for breakdowns of type  $j$ ; rather, that is the conditional probability

$$P(X > x | C = j) = S(x, j)/p_j.$$

For continuous  $X$ , the *sub-density function* is  $f(x, j) = -dS(x, j)/dx$ . The *marginal survival function* and *marginal density* of  $X$  are calculated from

$$S(x) = \sum_{j=1}^p S(x, j) \quad \text{and} \quad f(x) = -dS(x)/dx = \sum_{j=1}^p f(x, j). \quad (1.13)$$

Next, one defines the hazard function for breakdown from cause  $j$ , in the presence of all risks, in similar fashion as the overall hazard function:

**Definition 1.21** (Sub-hazard function) The sub-hazard function (for breakdown from cause  $j$  in the presence of all risks  $1, \dots, p$ ) is

$$\begin{aligned} h(x, j) &= \lim_{\Delta x \downarrow 0} \frac{P(x < X \leq x + \Delta x, C = j | X > x)}{\Delta x} \\ &= \lim_{\Delta x \downarrow 0} (\Delta x)^{-1} \frac{S(x, j) - S(x + \Delta x, j)}{S(x)} \\ &= \frac{f(x, j)}{S(x)}. \end{aligned} \quad (1.14)$$

Sometimes,  $h(x, j)$  is called *cause-specific* or *crude hazard function*. The overall hazard function<sup>13</sup> is  $h(x) = \sum_{j=1}^p h(x, j)$ .

**Example 1.14** (Weibull competing risks model) One can specify a Weibull *competing risks model* by assuming sub-hazard functions of the form

$$h(x, j) = \alpha_j \beta_j^{-\alpha_j} x^{\alpha_j - 1},$$

13 Dropping subscript  $X$  here for simplicity.

with  $\alpha_j, \beta_j > 0$ . The overall hazard function then is

$$h(x) = \sum_{j=1}^p \alpha_j \beta_j^{-\alpha_j} x^{\alpha_j - 1},$$

and the marginal survival function follows as

$$S(x) = \exp \left\{ - \int_0^x h(y) dy \right\} = \exp \left\{ - \sum_{j=1}^p (x/\beta_j)^{\alpha_j} \right\}.$$

In view of Example 1.12, this shows that  $X$  has the distribution of  $\min(X_1, \dots, X_p)$ , the  $X_j$  being independent Weibull variates with parameters  $(\alpha_j, \beta_j)$ . The sub-densities  $f(x, j)$  can be obtained as  $h(x, j)S(x)$ , but integration of this in order to calculate the sub-survival distributions is generally intractable.

A special type of competing risks model emerges when *proportionality of hazards* is assumed; that is, when  $h(x, j)/h(x)$  does not depend on  $x$  for each  $j$ . It means that, over time, the instantaneous risk of a “critical event” occurring may increase or decrease, but the relative risks of the various causes remain invariant. The meaning of the assumption is captured more precisely by the following lemma (for proof see, e.g., Kocherlakota and Proschan, 1991).

**Lemma 1.1** (Independence of time and cause) *The following conditions are equivalent:*

1.  $h(x, j)/h(x)$  does not depend on  $x$  for all  $j$  (proportionality of hazards);
2. the time and cause of breakdown are independent;
3.  $h(x, j)/h(x, k)$  is independent of  $x$ , for all  $j$  and  $k$ .

In this case,  $h(x, j) = p_j h(x)$  or, equivalently,  $f(x, j) = p_j f(x)$  or  $F(x, j) = p_j F(x)$ . The second condition in the lemma means that breakdown during some particular period does not make it any more or less likely to be from cause  $j$  than breakdown in some other period.

**Example 1.15** (Weibull sub-hazards) With  $h(x, j) = \alpha_j \beta_j^{-\alpha_j} x^{\alpha_j - 1}$  (see Example 1.14) proportional hazards only holds if the  $\alpha_j$  are all equal, say to  $\alpha$ . Let  $\beta^+ = \sum_{j=1}^p \beta_j^{-\alpha}$ ; in this case,

$$S(x) = \exp[-\beta^+ x^\alpha] \text{ and } f(x, j) = h(x, j)S(x) = \alpha \beta_j^{-\alpha} x^{\alpha-1} \exp[-\beta^+ x^\alpha].$$

With  $\pi_j = \beta_j^{-\alpha} / \beta^+$ , the sub-survival distribution is

$$S(x, j) = \pi_j \exp[-\beta^+ x^\alpha],$$

showing that this is a mixture model in which  $p_j = P(C = j) = \pi_j$  and

$$P(X > x | C = j) = P(X > x) = \exp[-\beta^+ x^\alpha],$$

i.e., time and cause of breakdown are stochastically independent.

### Competing risks models: latent-failure times approach

While the sub-hazard function introduced above has become the core concept of competing risks modeling, the more traditional approach was based on assuming the existence of a *latent failure time* associated with each of the  $p$  potential causes of breakdown. If  $X_j$  represents the time to system failure from cause  $j$  ( $j = 1, \dots, p$ ), the smallest  $X_j$  determines the time to overall system failure, and its index is the cause of failure  $C$ . This is the prototype of the concept of *parallel processing* and, as such, central to reaction time modeling as well.

For the latent failure times we assume a vector  $\mathbf{X} = (X_1, \dots, X_p)$  with absolutely continuous survival function  $G(\mathbf{x}) = P(\mathbf{X} > \mathbf{x})$  and marginal survival distributions  $G_j(x) = P(X_j > x)$ . The latent failure times are, of course, not observable, nor are the marginal hazard functions

$$h_j(x) = -d \log G_j(x) / dx.$$

However, we clearly must have an identity between  $G(\mathbf{x})$  and the marginal survival function  $S(x)$  introduced in Equation (1.13),

$$G(x\mathbf{1}_p) = S(x), \quad (1.15)$$

where  $\mathbf{1}_p' = (1, \dots, 1)$  of length  $p$ . It is then easy to show (see, e.g., Crowder, 2012, p. 236) that the sub-densities can be calculated from the joint survival distribution of the latent failure times as

$$f(x, j) = [-\partial G(\mathbf{x})/\partial x_j]_{x\mathbf{1}_p},$$

and the sub-hazards follow as

$$h(x, j) = [-\partial \log G(\mathbf{x})/\partial x_j]_{x\mathbf{1}_p},$$

where the notation  $[\dots]_{x\mathbf{1}_p}$  indicates that the enclosed function is to be evaluated at  $\mathbf{x} = (x, \dots, x)$ .

#### *Independent risks*

Historically, the main aim of competing risks analysis was to estimate the marginal survival distributions  $G_j(x)$  from the sub-survival distributions  $S(x, j)$ . While this is not possible in general, except by specifying a fully parametric model for  $G(\mathbf{x})$ , it can also be done by assuming that the risks act independently. In particular, it can be shown (see Crowder, 2012, p. 245) that the set of sub-hazard functions  $h(x, j)$  determines the set of marginals  $G_j(x)$  via

$$G_j(x) = \exp \left\{ - \int_0^x h(y, j) dy \right\}.$$

Without making such strong assumptions, it is still possible to derive some bounds for  $G_j(x)$  in terms of  $S(x, j)$ . Obviously,

$$S(x, j) = P(X_j > x, C = j) \leq P(X_j > x) = G_j(x).$$

More refined bounds are derived in Peterson (1976):

Let  $y = \max\{x_1, \dots, x_p\}$ . Then

$$\sum_{j=1}^p S(y, j) \leq G(\mathbf{x}) \leq \sum_{j=1}^p S(x_j, j). \quad (1.16)$$

Setting  $x_k = x$  and all other  $x_j = 0$  yields, for  $x > 0$  and each  $k$ ,

$$\sum_{j=1}^p S(x, j) \leq G_k(x) \leq S(x, k) + (1 - p_k). \quad (1.17)$$

These bounds cannot be improved because they are attainable by particular distributions, and they may be not very restrictive; that is, even knowing the  $S(x, j)$  quite well may not suffice to say much about  $G(\mathbf{x})$  or  $G_j(x)$ . This finding is echoed in some general non-identifiability results considered next.

### Some non-identifiability results

A basic result, much discussed in the competing risks literature, is (for proof, see Tsiatis, 1975):

**Proposition 1.11** (Tsiatis' Theorem) *For any arbitrary, absolutely continuous survival function  $G(\mathbf{x})$  with dependent risks and sub-survival distributions  $S(x, j)$ , there exists a unique “proxy model” with independent risks yielding identical  $S(x, j)$ . It is defined by*

$$G^*(\mathbf{x}) = \prod_{j=1}^p G_j^*(x_j), \text{ where } G_j^*(x_j) = \exp \left\{ - \int_0^x h(y, j) dy \right\} \quad (1.18)$$

and the sub-hazard function  $h(x, j)$  derives from the given  $S(x, j)$ .

The following example illustrates why this non-identifiability is seen as a problem.

**Example 1.16** (Gumbel's bivariate exponential distribution) The joint survival distribution is

$$G(\mathbf{x}) = \exp[-\lambda_1 x_1 - \lambda_2 x_2 - \nu x_1 x_2],$$

with  $\lambda_1 > 0$  and  $\lambda_2 > 0$  and  $0 \leq \nu \leq \lambda_1 \lambda_2$ . Parameter  $\nu$  controls dependence:  $\nu = 0$  implies stochastic independence between  $X_1$  and  $X_2$ . The univariate marginals are  $P(X_j > x_j) = G_j(x_j) = \exp[-\lambda_j x_j]$ . Then the marginal survival distribution is  $S(x) = \exp[-(\lambda_1 + \lambda_2)x - \nu x^2]$  and the sub-hazard functions are  $h(x, j) = \lambda_j + \nu x$ ,  $j = 1, 2$ . From the Tsiatis theorem, the proxy model has

$$G_j^*(x) = \exp \left\{ - \int_0^x (\lambda_j + \nu y) dy \right\} = \exp[-(\lambda_j x + \nu x^2/2)]$$

$$G^*(\mathbf{x}) = \exp \left[ - (\lambda_1 x_1 + \nu x_1^2/2) - (\lambda_2 x_2 + \nu x_2^2/2) \right].$$

Thus, to make predictions about  $X_j$ , one can either use  $G_j(x)$  or  $G_j^*(x)$ , and the results will not be the same (except if  $v = 0$ ). One cannot tell which is correct from just  $(X, C)$  data.

There are many refinements of Tsiatis' theorem and related results (see e.g., Crowder, 2012) and, ultimately, the latent-failure times approach in survival analysis has been superseded by the concept of sub-hazard function within the more general context of counting processes (Aalen *et al.*, 2008).

### 1.3.4 Order statistics, extreme values, and records

The concept of order statistics plays a major role in RT models with a parallel processing architecture. Moreover, several theories of learning, memory, and automaticity have drawn upon concepts from the asymptotic theory of specific order statistics, the extreme values. Finally, we briefly treat the theory of record values, a well-developed statistical area which, strangely enough, has not yet played a significant role in RT modeling.

#### Order statistics

Suppose that  $(X_1, X_2, \dots, X_n)$  are  $n$  jointly distributed random variables. The corresponding *order statistics* are the  $X_i$ 's arranged in ascending order of magnitude, denoted by  $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$ . In general, the  $X_i$  are assumed to be i.i.d. random variables with  $F(x) = P(X_i \leq x)$ ,  $i = 1, \dots, n$ , so that  $(X_1, X_2, \dots, X_n)$  can be considered as a random sample from a population with distribution function  $F(x)$ . The distribution function of  $X_{i:n}$  is obtained realizing that

$$\begin{aligned} F_{i:n}(x) &= P(X_{i:n} \leq x) \\ &= P(\text{at least } i \text{ of } X_1, X_2, \dots, X_n \text{ are less than or equal to } x) \\ &= \sum_{r=i}^n P(\text{exactly } r \text{ of } X_1, X_2, \dots, X_n \text{ are less than or equal to } x) \\ &= \sum_{r=i}^n \binom{n}{r} [F(x)]^r [1 - F(x)]^{n-r}, \quad -\infty < x < +\infty. \end{aligned} \quad (1.19)$$

For  $i = n$ , this gives

$$F_{n:n}(x) = P(\max(X_1, \dots, X_n) \leq x) = [F(x)]^n,$$

and for  $i = 1$ ,

$$F_{1:n}(x) = P(\min(X_1, \dots, X_n) \leq x) = 1 - [1 - F(x)]^n.$$

Assuming the existence of the probability density for  $(X_1, X_2, \dots, X_n)$ , the joint density for all  $n$ -order statistics – which, of course, are no longer stochastically independent – is derived as follows (Arnold *et al.*, 1992, p. 10). Given the realizations of the  $n$ -order statistics to be  $x_{1:n} < x_{2:n} < \dots < x_{n:n}$ , the original variables  $X_i$  are restrained to take on the values  $x_{i:n}$  ( $i = 1, 2, \dots, n$ ), which by symmetry

assigns equal probability for each of the  $n!$  permutations of  $(1, 2, \dots, n)$ . Hence, the joint density function of all  $n$  order statistics is

$$f_{1,2,\dots,n:n}(x_1, x_2, \dots, x_n) = n! \prod_{r=1}^n f(x_r), \quad -\infty < x_1 < x_2 < \dots < x_n < +\infty.$$

The marginal density of  $X_{i:n}$  is obtained by integrating out the remaining  $n - 1$  variables, yielding

$$f_{i:n}(x) = \frac{n!}{(i-1)!(n-i)!} [F(x)]^{i-1} [1-F(x)]^{n-i} f(x), \quad -\infty < x < +\infty. \quad (1.20)$$

It is relatively straightforward to compute the joint density and the distribution function of any two order statistics (see Arnold *et al.*, 1992).

Let  $U_1, U_2, \dots, U_n$  be a random sample from the standard uniform distribution (i.e., over  $[0, 1]$ ) and  $X_1, X_2, \dots, X_n$  be a random sample from a population with distribution function  $F(x)$ , and  $U_{1:n} \leq U_{2:n} \leq \dots \leq U_{n:n}$  and  $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$  the corresponding order statistics. For a continuous distribution function  $F(x)$ , the probability integral transformation  $U = F(X)$  produces a standard uniform distribution. Thus, for continuous  $F(x)$ ,

$$F(X_{i:n}) = {}_d U_{i:n}, \quad i = 1, 2, \dots, n.$$

Moreover, it is easy to verify that, for quantile function  $Q(u) \equiv F^{-1}(u)$ , one has

$$Q(U_i) = {}_d X_i, \quad i = 1, 2, \dots, n.$$

Because  $Q$  is non-decreasing, it immediately follows that

$$Q(U_{i:n}) = {}_d X_{i:n}, \quad i = 1, 2, \dots, n. \quad (1.21)$$

This implies simple forms for the quantile functions of the extreme order statistics,  $Q_n$  and  $Q_1$ ,

$$Q_n(u_n) = Q(u_n^{1/n}) \quad \text{and} \quad Q_1(u_1) = Q[1 - (1 - u_1)^{1/n}], \quad (1.22)$$

for the maximum and the minimum statistic, respectively.

Moments of order statistics can often be obtained from the marginal density (Equation (1.20)). Writing  $\mu_{i:n}^{(m)}$  for the  $m$ th moment of  $X_{i:n}$ ,

$$\begin{aligned} \mu_{i:n}^{(m)} &= \frac{n!}{(i-1)!(n-i)!} \int_{-\infty}^{+\infty} x^m [F(x)]^{i-1} [1-F(x)]^{n-i} f(x) dx, \\ &\quad 1 \leq i \leq n, \quad m = 1, 2, \dots. \end{aligned} \quad (1.23)$$

Using relation (1.21), this can be written more compactly as

$$\begin{aligned} \mu_{i:n}^{(m)} &= \frac{n!}{(i-1)!(n-i)!} \int_0^1 [F^{-1}(u)]^m u^{i-1} (1-u)^{n-i} du, \\ &\quad 1 \leq i \leq n, \quad m = 1, 2, \dots \end{aligned} \quad (1.24)$$

The computation of various moments of order statistics can be reduced drawing upon various identities and recurrence relations. Consider the obvious identity

$$\sum_{i=1}^n X_{i:n}^m = \sum_{i=1}^n X_i^m, \quad m \geq 1.$$

Taking expectation on both sides yields

$$\sum_{i=1}^n \mu_{i:n}^{(m)} = nEX^m = n\mu_{1:1}^{(m)}, \quad m \geq 1. \quad (1.25)$$

An interesting recurrence relation, termed triangle rule, is easily derived using Equation (1.24) (see Arnold *et al.*, 1992, p. 112):

For  $1 \leq i \leq n - 1$  and  $m = 1, 2, \dots$ ,

$$i\mu_{i+1:n}^{(m)} + (n - i)\mu_{i:n}^{(m)} = n\mu_{i:n-1}^{(m)}. \quad (1.26)$$

This relation shows that it is enough to evaluate the  $m$ th moment of a single-order statistic in a sample of size  $n$  if these moments in samples of size less than  $n$  are already available.

Many other recurrence relations, also including product moments and covariances of order statistics, are known and can be used not only to reduce the amount of computation (if these quantities have to be determined by numerical procedures), but also to provide simple checks of the accuracy of these computations (Arnold *et al.*, 1992).

#### *Characterization of distributions by order statistics*

Given that a distribution  $F$  is known, it is clear that the distributions  $F_{i:n}$  are completely determined, for every  $i$  and  $n$ . The interesting question is to what extent does knowledge of some properties of the distribution of an order statistic determine the parent distribution?

Clearly, the distribution of  $X_{n:n}$  determines  $F$  completely because  $F_{n:n}(x) = [F(x)]^n$  for every  $x$ , so that  $F(x) = [F_{n:n}(x)]^{1/n}$ . It turns out that, assuming the moments are finite, equality of the first moments of maxima,

$$EX_{n:n} = EX'_{n:n}, \quad n = 1, 2, \dots,$$

is sufficient to infer the equality of the underlying parent distributions  $F(x)$  and  $F'(x)$  (see Arnold *et al.*, 1992, p. 145).

Moreover, distributional relations between order statistics can be drawn upon in order to characterize the underlying distribution. For example, straightforward computation yields that for a sample of size  $n$  from an exponentially distributed population,

$$nX_{1:n} = {}_d X_{1:1}, \quad \text{for } n = 1, 2, \dots. \quad (1.27)$$

Notably, it has been shown (Desu, 1971) that the exponential is the *only* nondegenerate distribution satisfying the above relation. From this, a characterization of the Weibull is obtained as well:

**Example 1.17** (“Power Law of Practice”) Let, for  $\alpha > 0$ ,

$$F(x) = 1 - \exp[-x^\alpha];$$

then,  $W = X^\alpha$  is a unit exponential variable (i.e.,  $\lambda = 1$ ) and Equation (1.27) implies the following characterization of the Weibull:

$$n^{1/\alpha} X_{1:n} =_d X_{1:1}, \quad \text{for } n = 1, 2, \dots \quad (1.28)$$

At the level of expectations,

$$\mathbb{E}X_{1:n} = n^{-1/\alpha} \mathbb{E}X.$$

This has been termed the “power law of practice” (Newell and Rosenbloom, 1981): the time taken to perform a certain task decreases as a power function of practice (indexed by  $n$ ). This is the basis of G. Logan’s “theory of automaticity” (Logan, 1988, 1992, 1995) postulating Weibull-shaped RT distributions (see Colonius, 1995, for a comment on the theoretical foundations).

### Extreme value statistics

The behavior of extreme order statistics, that is, of the maximum and minimum of a sample when the sample size  $n$  goes to infinity, has attracted the attention of RT model builders, in particular in the context of investigating parallel processing architectures. Only a few results of this active area of statistical research will be considered here.

For the underlying distribution function  $F$ , its *right, or upper, endpoint* is defined as

$$\omega(F) = \sup\{x | F(x) < 1\}.$$

Note that  $\omega(F)$  is either  $+\infty$  or finite. Then the distribution of the maximum,

$$F_{n:n}(x) = P(X_{n:n} \leq x) = [F(x)]^n,$$

clearly converges to zero, for  $x < \omega(F)$ , and to 1, for  $x \geq \omega(F)$ . Hence, in order to obtain a nondegenerate limit distribution, a normalization is necessary. Suppose there exists sequences of constants  $b_n > 0$  and  $a_n$  real ( $n = 1, 2, \dots$ ) such that

$$\frac{X_{n:n} - a_n}{b_n}$$

has a nondegenerate *limit distribution*, or *extreme value distribution*, as  $n \rightarrow +\infty$ , i.e.,

$$\lim_{n \rightarrow +\infty} F^n(a_n + b_n x) = G(x), \quad (1.29)$$

for every point  $x$  where  $G$  is continuous, and  $G$  a nondegenerate distribution function.<sup>14</sup> A distribution function  $F$  satisfying Equation (1.29) is said to be in the *domain of maximal attraction* of a nondegenerate distribution function  $G$ , and we

14 Note that convergence here is convergence “in probability,” or “weak” convergence.

will write  $F \in \mathcal{D}(G)$ . Two distribution functions  $F_1$  and  $F_2$  are defined to be of the *same type* if there exist constants  $a_0$  and  $b_0 > 0$  such that  $F_1(a_0 + b_0 x) = F_2(x)$ .

It can be shown that (i) the choice of the norming constants is not unique, and (ii) a distribution function cannot be in the domain of maximal attraction of more than one type of distribution. The main theme of extreme value statistics is to identify all extreme value distributions and their domains of attraction and to characterize necessary and/or sufficient conditions for convergence.

Before a general result is presented, let us consider the asymptotic behavior for a special case.

**Example 1.18** Let  $X_{n:n}$  be the maximum of a sample from the exponential distribution with rate equal to  $\lambda$ . Then,

$$\begin{aligned} P(X_{n:n} - \lambda^{-1} \log n \leq x) &= \begin{cases} (1 - \exp[-(\lambda x + \log n)])^n, & \text{if } x > -\lambda^{-1} \log n \\ 0, & \text{if } x \leq -\lambda^{-1} \log n \end{cases} \\ &= (1 - \exp[-\lambda x]/n)^n, \quad \text{for } x > -\lambda^{-1} \log n \\ &\rightarrow \exp[-\exp[-\lambda x]], \quad -\infty < x < \infty, \end{aligned}$$

as  $n \rightarrow +\infty$ .

#### *Possible limiting distributions for the sample maximum*

The classic result on the limiting distributions, already described in Fréchet (1927) and Fisher and Tippett (1928), is the following:

**Theorem 1.3** *If Equation (1.29) holds, the limiting distribution function  $G$  of an appropriately normalized sample maximum is one of the following types:*

$$G_1(x; \alpha) = \begin{cases} 0, & x \leq 0 \\ \exp[-x^{-\alpha}], & x > 0; \alpha > 0 \end{cases} \quad (1.30)$$

$$G_2(x; \alpha) = \begin{cases} \exp[-(-x)^\alpha], & x < 0; \alpha > 0 \\ 1, & x \geq 0 \end{cases} \quad (1.31)$$

$$G_3(x; \alpha) = \exp[-\exp(-x)], \quad -\infty < x < +\infty. \quad (1.32)$$

Distribution  $G_1$  is referred to as the *Fréchet type*,  $G_2$  as the *Weibull type*, and  $G_3$  as the *extreme value* or *Gumbel distribution*. Parameter  $\alpha$  occurring in  $G_1$  and  $G_2$  is related to the tail behavior of the parent distribution  $F$ . It follows from the theorem that if  $F$  is in the maximal domain of attraction of some  $G$ ,  $F \in \mathcal{D}(G)$ , then  $G$  is either  $G_1$ ,  $G_2$ , or  $G_3$ .

Necessary and sufficient conditions on  $F$  to be in the domain of attraction of  $G_1$ ,  $G_2$ , or  $G_3$ , respectively, have been given in Gnedenko (1943). Because these are often difficult to verify (Arnold *et al.*, 1992, pp. 210–212), here we list sufficient conditions for weak convergence due to von Mises (1936):

**Proposition 1.12** *Let  $F$  be an absolutely continuous distribution function and let  $h(x)$  be the corresponding hazard function.*

1. If  $h(x) > 0$  for large  $x$  and for some  $\alpha > 0$ ,

$$\lim_{x \rightarrow +\infty} x h(x) = \alpha,$$

then  $F \in \mathcal{D}(G_1(x; \alpha))$ . Possible choices for the norming constants are  $a_n = 0$  and  $b_n = F^{-1}(1 - n^{-1})$ .

2. If  $F^{-1}(1) < \infty$  and for some  $\alpha > 0$ ,

$$\lim_{x \rightarrow F^{-1}(1)} (F^{-1}(1) - x) h(x) = \alpha,$$

then  $F \in \mathcal{D}(G_2(x; \alpha))$ . Possible choices for the norming constants are  $a_n = F^{-1}(1)$  and  $b_n = F^{-1}(1) - F^{-1}(1 - n^{-1})$ .

3. Suppose  $h(x)$  is nonzero and is differentiable for  $x$  close to  $F^{-1}(1)$  or, for large  $x$ , if  $F^{-1}(1) = +\infty$ . Then,  $F \in \mathcal{D}(G_3)$  if

$$\lim_{x \rightarrow F^{-1}(1)} \frac{d}{dx} \left( \frac{1}{h(x)} \right) = 0.$$

Possible choices for the norming constants are  $a_n = F^{-1}(1 - n^{-1})$  and  $b_n = [nf(a_n)]^{-1}$ .

#### *Limiting distributions for sample minima*

This case can be derived technically from the analysis of sample maxima, because the sample minimum from distribution  $F$  has the same distribution as the negative of the sample maximum from distribution  $F^*$ , where  $F^*(x) = 1 - F(-x)$ . We sketch some of the results for completeness here.

For parent distribution  $F$  one defines the *left*, or *lower endpoint* as

$$\alpha(F) = \inf\{x : F(x) > 0\}.$$

Note that  $\alpha(F)$  is either  $-\infty$  or finite. Then the distribution of the minimum,

$$F_{1:n}(x) = P(X_{1:n} \leq x) = 1 - [1 - F(x)]^n,$$

clearly converges to one, for  $x > \alpha(F)$ , and to zero, for  $x \leq \alpha(F)$ . The theorem paralleling Theorem 1.3 then is

**Theorem 1.4** Suppose there exist constants  $a_n^*$  and  $b_n^* > 0$  and a nondegenerate distribution function  $G^*$  such that

$$\lim_{n \rightarrow +\infty} 1 - [1 - F(a_n^* + b_n^* x)]^n = G^*(x)$$

for every point  $x$  where  $G^*$  is continuous, then  $G^*$  must be one of the following types:

1.  $G_1^*(x; \alpha) = 1 - G_1(-x; \alpha)$ ,
2.  $G_2^*(x; \alpha) = 1 - G_2(-x; \alpha)$ ,
3.  $G_3^*(x) = 1 - G_3(-x)$ ,

where  $G_1$ ,  $G_2$ , and  $G_3$  are the distributions given in Theorem 1.3.

Necessary and/or sufficient conditions on  $F$  to be in the domain of (minimal) attraction of  $G_1$ ,  $G_2$ , or  $G_3$ , respectively, can be found, e.g., in Arnold *et al.* (1992, pp. 213–214).

Let us state some important practical implications of these results on asymptotic behavior.

- Only three distributions (Fréchet, Weibull, and extreme value/Gumbel) can occur as limit distributions of independent sample maxima or minima.
- Although there exist parent distributions such that the limit does not exist, for most continuous distributions in practice one of them actually occurs.
- A parent distribution with non-finite endpoint in the tail of interest cannot lie in a Weibull-type domain of attraction.
- A parent distribution with finite endpoint in the tail of interest cannot lie in a Fréchet-type domain of attraction.

One important problem with practical implications is the speed of convergence to the limit distributions. It has been observed that, e.g., in the case of the maximum of standard normally distributed random variables, convergence to the extreme value distribution  $G_3$  is extremely slow and that  $F_n(a_n + b_n x)^n$  can be closer to a suitably chosen Weibull distribution  $G_2$  even for very large  $n$ . In determining the sample size required for the application of the asymptotic theory, estimates of the error of approximation do not depend on  $x$  because the convergence is uniform in  $x$ . They do depend, however, on the choice of the norming constants (for more details, see Galambos, 1978, pp. 111–116). Because of this so-called *penultimate approximation*, diagnosing the type of limit distribution even from rather large empirical samples may be problematic (see also the discussion in Colonius, 1995; Logan, 1995).

### **Record values**

The theory of records should not be seen as a special topic of order statistics, but their study parallels the study of order statistics in many ways. In particular, things that are relatively easy for order statistics are also feasible for records, and things that are difficult for order statistics are equally or more difficult for records. Compared to order statistics, there are not too many data sets for records available, especially for the asymptotic theory, the obvious reason being the low density of the  $n$ th record for large  $n$ . Nevertheless, the theory of records has become quite developed, representing an active area of research. Our hope in treating it here, although limiting exposition to the very basics, is that the theory may find some useful applications in the cognitive modeling arena.

Instead of a finite sample  $(X_1, X_2, \dots, X_n)$  of i.i.d. random variables, record theory starts with an infinite sequence  $X_1, X_2, \dots$  of i.i.d. random variables (with common distribution  $F$ ). For simplicity,  $F$  is assumed continuous so that ties are not possible. An observation  $X_j$  is called a *record* (more precisely, *upper record*) if it exceeds in value all preceding observations, that is,  $X_j > X_i$  for all  $i < j$ . *Lower records* are defined analogously.

Following the notation in Arnold *et al.* (1992, 1998), the sequence of *record times*  $\{T_n\}_{n=0}^{\infty}$  is defined by

$$T_0 = 1 \quad \text{with probability 1,}$$

and for  $n \geq 1$ ,

$$T_n = \min\{j \mid j > T_{n-1}, X_j > X_{T_{n-1}}\}.$$

The corresponding *record value sequence*  $\{R_n\}_{n=0}^{\infty}$  is defined as

$$R_n = X_{T_n}, \quad n = 0, 1, 2, \dots$$

*Interrecord times*,  $\Delta_n$ , are defined by

$$\Delta_n = T_n - T_{n-1}, \quad n = 1, 2, \dots$$

Finally, the *record counting process* is  $\{N_n\}_{n=1}^{+\infty}$ , where

$$N_n = \{\text{number of records among } X_1, \dots, X_n\}.$$

Interestingly, only the sequence of record values  $\{R_n\}$  depends on the specific distribution; monotone transformations of the  $X_i$ 's will not affect the values of  $\{T_n\}$ ,  $\{\Delta_n\}$ , and  $\{N_n\}$ . Specifically, if  $\{R_n\}$ ,  $\{T_n\}$ ,  $\{\Delta_n\}$ , and  $\{N_n\}$  are the record statistics associated with a sequence of i.i.d.  $X_i$ 's and if  $Y_i = \phi(X_i)$  (where  $\phi$ , increasing) has associated record statistics  $\{R'_n\}$ ,  $\{T'_n\}$ ,  $\{\Delta'_n\}$ , and  $\{N'_n\}$ , then

$$T'_n = {}_d T_n, \quad \Delta'_n = {}_d \Delta_n, \quad N'_n = {}_d N_n, \quad \text{and} \quad R'_n = {}_d \phi(R_n).$$

Thus, it is wise to select a convenient common distribution for the  $X_i$ 's which will make the computations simple. Because of its lack of memory property, the standard exponential distribution ( $\text{Exp}(1)$ ) turns out to be the most useful one.

#### *Distribution of the nth upper record $R_n$*

Let  $\{X_i^*\}$  denote a sequence of i.i.d.  $\text{Exp}(1)$  random variables. Because of the lack of memory property,  $\{R_n^* - R_{n-1}^*, n \geq 1\}$ , the differences between successive records will again be i.i.d.  $\text{Exp}(1)$  random variables. It follows that

$$R_n^* \text{ is distributed as } \text{Gamma}(n+1, 1), \quad n = 0, 1, 2, \dots. \quad (1.33)$$

From this we obtain the distribution of  $R_n$  corresponding to a sequence of i.i.d.  $X_i$ 's with common *continuous* distribution function  $F$  as follows: note that if  $X$  has distribution function  $F$ , then  $-\log[1 - F(X)]$  is distributed as  $\text{Exp}(1)$  (see Proposition 1.9), so that

$$X = {}_d F^{-1}(1 - \exp^{-X_n^*}),$$

where  $X^*$  is  $\text{Exp}(1)$ . Because  $X$  is a monotone function of  $X^*$  and, consequently, the  $n$ th record of the  $\{X_n\}$  sequence,  $R_n$ , is related to the  $n$ th record  $R_n^*$  of the

exponential sequence by

$$R_n = {}_d F^{-1}(1 - \exp^{-R_n^*}).$$

From Equation (1.5), the survival of a Gamma( $n + 1, 1$ ) random variable is

$$P(R_n^* > r) = e^{-r^*} \sum_{k=0}^n (r^*)^k / k!$$

implying, for the survival function of the  $n$ th record corresponding to an i.i.d.  $F$  sequence,

$$P(R_n > r) = [1 - F(r)] \sum_{k=0}^n [-\log(1 - F(r))]^k / k!.$$

If  $F$  is absolutely continuous with density  $f$ , differentiating and simplifying yields the density for  $R_n$

$$f_{R_n}(r) = f(r) [-\log(1 - F(r))]^n / n!.$$

The asymptotic behavior of the distribution of  $R_n$  has also been investigated. For example, when the common distribution of the  $X_i$ 's is Weibull, the distribution of  $R_n$  is asymptotically normal.

*Distribution of the nth record time  $T_n$*

Because  $T_0 = 1$ , we consider first the distribution of  $T_1$ .

For any  $n \geq 1$ , clearly

$$P(T_1 > n) = P(X_1 \text{ is largest among } X_1, X_2, \dots, X_n).$$

Without loss of generality, we assume that  $X_1, X_2, \dots$  are independent random variables with a uniform distribution on  $(0, 1)$ . Then, by the total probability rule, for  $n \geq 2$

$$\begin{aligned} P(T_1 = n) &= \int_0^1 P(T_1 = n | X_1 = x) dx \\ &= \int_0^1 x^{n-2} (1-x) dx \\ &= \frac{1}{n-1} - \frac{1}{n} = \frac{1}{n(n-1)}. \end{aligned} \tag{1.34}$$

Note that the median of  $T_1$  equals 2, but the expected value

$$ET_1 = \sum_{n=2}^{+\infty} n P(T_1 = n) = +\infty.$$

Thus, after a maximum has been recorded by the value of  $X_1$ , the value of  $X_2$  that will be even larger has a probability of  $1/2$ , but the expected “waiting time” to a larger value is infinity.

Introducing the concept of record indicator random variables,  $\{I_n\}_{n=1}^{\infty}$ , defined by  $I_1 = 1$  with probability 1 and, for  $n > 1$

$$I_n = \begin{cases} 1, & \text{if } X_n > X_1, \dots, X_{n-1}, \\ 0, & \text{otherwise,} \end{cases}$$

it can be shown that the  $I_n$ 's are independent Bernoulli random variables with

$$P(I_n = 1) = 1/n$$

for  $n = 1, 2, \dots$ , and the joint density of the record times becomes

$$\begin{aligned} P(T_1 = n_1, T_2 = n_2, \dots, T_k = n_k) \\ = P(I_2 = 0, \dots, I_{n_1-1} = 0, I_{n_1} = 1, I_{n_1+1} = 0, \dots, I_{n_k} = 1) \\ = \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} \cdots \frac{n_1 - 2}{n_1 - 1} \cdot \frac{1}{n_1} \frac{n_1}{n_1 + 1} \cdots \frac{1}{n_k} \\ = [(n_1 - 1)(n_2 - 1) \cdots (n_k - 1)n_k]^{-1}. \end{aligned}$$

The marginal distributions can be calculated to be

$$P(T_k = n) = S_{n-1}^k / n!,$$

where  $S_{n-1}^k$  are the *Stirling numbers of the first kind*, i.e., the coefficients of  $s^k$  in

$$\prod_{j=1}^{n-1} (s + j - 1).$$

One can verify that  $T_k$  grows rapidly with  $k$  and that it has in fact an infinitely large expectation. Moreover, about half of the waiting time for the  $k$ th record is spent between the  $k - 1$ th and the  $k$ th record.

### Record counting process $N_n$

Using the indicator variables  $I_n$ , the record counting process  $\{N_n, n \geq 1\}$  is just

$$N_n = \sum_{j=1}^n I_j,$$

so that

$$\begin{aligned} E(N_n) &= \sum_{j=1}^n \frac{1}{j} \\ &\approx \log n + \gamma, \end{aligned} \tag{1.35}$$

and

$$\begin{aligned} \text{Var}(N_n) &= \sum_{j=1}^n \frac{1}{j} \left(1 - \frac{1}{j}\right) \\ &\approx \log n + \gamma - \frac{\pi^2}{6}, \end{aligned} \tag{1.36}$$

where  $\gamma$  is *Euler's constant*  $0.5772\dots$ . This last result makes it clear that records are not very common: in a sequence of 1000 observations one can expect to observe only about 7 records!<sup>15</sup>

When the parent  $F$  is  $\text{Exp}(1)$  distributed, the record values  $R_n^*$  were shown to be distributed as  $\text{Gamma}(n+1, 1)$  random variables (Equation (1.33)). Thus, the associated record counting process is identical to a homogeneous Poisson process with unit intensity ( $\lambda = 1$ ). When  $F$  is an arbitrary continuous distribution with  $F^{-1}(0) = 0$  and hazard function  $\lambda(t)$ , then the associated record counts become a nonhomogeneous Poisson process with intensity function  $\lambda(t)$ : to see this, note that there will be a record value between  $t$  and  $t+h$  if, and only if, the first  $X_i$  whose value is greater than  $t$  lies between  $t$  and  $t+h$ . But we have (conditioning on which  $X_i$  this is, say  $i = n$ ), by definition of the hazard function,

$$P(X_n \in (t, t+h) | X_n > t) = \lambda(t)h + o(h),$$

which proves the claim.

### 1.3.5 Coupling

The concept of *coupling* involves the joint construction of two or more random variables (or stochastic processes) on a common probability space. The theory of probabilistic coupling has found applications throughout probability theory, in particular in characterizations, approximations, asymptotics, and simulation. Before we give a formal definition of coupling, one may wonder why a modeler should be interested in this topic.

In the introduction, we discussed an example of coupling in the context of audio-visual interaction (Example 1.2). Considering a more general context, data are typically collected under various experimental conditions, e.g., different stimulus properties, different number of targets in a search task, etc. Any particular condition generates a set of data considered to represent realizations of some random variable (e.g., RT), given a stochastic model has been specified. However, there is no principled way of relating, for example, an observed time to find a target in a condition with  $n$  targets,  $T_n$ , say, to that for a condition with  $n+1$  targets,  $T_{n+1}$ , simply because  $T_n$  and  $T_{n+1}$  are not defined on a common probability space. Note that this would not prevent one from numerically comparing average data under both conditions, or even the entire distributions functions; any statistical hypothesis about the two random variables, e.g., about the correlation or stochastic (in)dependence in general, would be void, however. Beyond this cautionary note, which is prompted by some apparent confusion about this issue in the RT modeling area, the method of coupling can also play an important role as a modeling tool. For example, a coupling argument allows one to turn a statement about an ordering of two RT distributions, a-priori not defined on a common probability space,

<sup>15</sup> This number is clearly larger than typical empirical data sets and may be one reason why record theory has not seen much application in mathematical psychology (but see the remarks on time series analysis in Section 1.4).

Table 1.1 *Joint distribution of two Bernoulli random variables.*

		$\hat{X}_q$		
		0	1	
$\hat{X}_p$	0	$1 - q$	$q - p$	$1 - p$
	1	0	p	p
		$1 - q$	q	1

into a statement about a pointwise ordering of corresponding random variables on a common probability space. Moreover, coupling turns out to be a key concept in a general theory of “contextuality” being developed by Dzhafarov and Kujala (see Dzhafarov and Kujala, 2013, and also Chapter 2 in this volume).

We begin with a definition of “coupling” for random variables without referring to the measure-theoretic details:

**Definition 1.22** A *coupling* of a collection of random variables  $\{X_i, i \in I\}$ , with  $I$  denoting some index set, is a family of jointly distributed random variables

$$(\hat{X}_i : i \in I) \text{ such that } \hat{X}_i =_d X_i, \quad i \in I.$$

Note that the joint distribution of the  $\hat{X}_i$  need not be the same as that of  $X_i$ ; in fact, the  $X_i$  may not even have a joint distribution because they need not be defined on a common probability space. However, the family  $(\hat{X}_i : i \in I)$  has a joint distribution with the property that its marginals are equal to the distributions of the individual  $X_i$  variables. The individual  $\hat{X}_i$  is also called a *copy* of  $X_i$ .

Before we mention a more general definition of coupling, let us consider a simple example.

**Example 1.19** (Coupling two Bernoulli random variables) Let  $X_p, X_q$  be Bernoulli random variables, i.e.,

$$P(X_p = 1) = p \text{ and } P(X_p = 0) = 1 - p,$$

and  $X_q$  defined analogously. Assume  $p < q$ ; we can couple  $X_p$  and  $X_q$  as follows: Let  $U$  be a uniform random variable on  $[0, 1]$ , i.e., for  $0 \leq a < b \leq 1$ ,

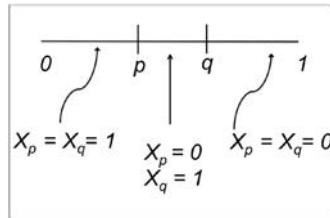
$$P(a < U \leq b) = b - a.$$

Define

$$\hat{X}_p = \begin{cases} 1, & \text{if } 0 < U \leq p; \\ 0, & \text{if } p < U \leq 1 \end{cases}; \quad \hat{X}_q = \begin{cases} 1, & \text{if } 0 < U \leq q; \\ 0, & \text{if } q < U \leq 1. \end{cases}$$

Then  $U$  serves as a common source of randomness for both  $\hat{X}_p$  and  $\hat{X}_q$ . Moreover,  $\hat{X}_p =_d X_p$  and  $\hat{X}_q =_d X_q$ , and  $\text{Cov}(\hat{X}_p, \hat{X}_q) = p(1 - p)$ . The joint distribution of  $(\hat{X}_p, \hat{X}_q)$  is presented in Table 1.1 and illustrated in Figure 1.4.

In this example, the two coupled random variables are obviously not independent. Note, however, that for any collection of random variables  $\{X_i, i \in I\}$  there



**Figure 1.4** Joint distribution of two Bernoulli random variables.

always exists a *trivial* coupling, consisting of independent copies of the  $\{X_i\}$ . This follows from construction of the product probability measure (see Definition 1.15).

Before we consider further examples of coupling, coupling of probability measures is briefly introduced. The definition is formulated for the binary case only, for simplicity; let  $(E, \mathcal{E})$  be a measurable space:

**Definition 1.23** A *random element* in the space  $(E, \mathcal{E})$  is the quadruple

$$(\Omega_1, \mathcal{F}_1, P_1, X_1),$$

where  $(\Omega_1, \mathcal{F}_1, P_1)$  is the underlying probability space (the sample space) and  $X_1$  is a  $\mathcal{F}_1$ - $\mathcal{E}$ -measurable mapping from  $\Omega_1$  to  $E$ . A *coupling of the random elements*  $(\Omega_i, \mathcal{F}_i, P_i, X_i)$ ,  $i = 1, 2$ , in  $(E, \mathcal{E})$  is a random element  $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P}, (\hat{X}_1, \hat{X}_2))$  in  $(E^2, \mathcal{E} \otimes \mathcal{E})$  such that

$$X_1 =_d \hat{X}_1 \quad \text{and} \quad X_2 =_d \hat{X}_2,$$

with  $\mathcal{E} \otimes \mathcal{E}$  the product  $\sigma$ -algebra on  $E^2$ .

This general definition is the starting point of a general theory of coupling for stochastic processes, but it is beyond the scope of this chapter (see Thorisson, 2000).

A *self-coupling* of a random variable  $X$  is a family  $(\hat{X}_i : i \in I)$  where each  $\hat{X}_i$  is a copy of  $X$ . A trivial, and not so useful, self-coupling is the one with all the  $\hat{X}_i$  identical. Another example is the *i.i.d. coupling* consisting of independent copies of  $X$ . This can be used to prove

**Lemma 1.2** For every random variable  $X$  and non-decreasing bounded functions  $f$  and  $g$ , the random variables  $f(X)$  and  $g(X)$  are positively correlated, i.e.,

$$\text{Cov}[f(X), g(X)] \geq 0.$$

The lemma also follows directly from a general concept of dependence (association) to be treated in Section 1.3.8. A simple, although somewhat fundamental, coupling is the following:

**Example 1.20** (Quantile coupling) Let  $X$  be a random variable with distribution function  $F$ , that is,

$$P(X \leq x) = F(x), \quad x \in \mathbb{R}.$$

Let  $U$  be a uniform random variable on  $[0, 1]$ . Then, for random variable  $\hat{X} = F^{-1}(U)$  (see Proposition 1.9),

$$P(\hat{X} \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x), \quad x \in \mathfrak{N},$$

that is,  $\hat{X}$  is a copy of  $X$ ,  $\hat{X} =_d X$ . Thus, letting  $F$  run over the class of all distribution functions (using the same  $U$ ) yields a coupling of all differently distributed random variables, the *quantile coupling*. Because  $F^{-1}$  is non-decreasing, it follows from Lemma 1.2 that the quantile coupling consists of positively correlated random variables.

An important application of quantile coupling is in reducing a stochastic ordering between random variables to a pointwise (a.s.) ordering: let  $X$  and  $X'$  be two random variables with distribution functions  $F$  and  $G$ , respectively. If there is a coupling  $(\hat{X}, \hat{X}')$  of  $X$  and  $X'$  such that  $\hat{X}$  is *pointwise dominated* by  $\hat{X}'$ , that is

$$\hat{X} \leq \hat{X}' \quad (\text{almost surely}),$$

then  $\{\hat{X} \leq x\} \supseteq \{\hat{X}' \leq x\}$ , which implies

$$P(\hat{X} \leq x) \geq P(\hat{X}' \leq x),$$

and thus

$$F(x) \geq G(x), \quad x \in \mathfrak{N}.$$

Then  $X$  is said to be *stochastically dominated* (or *dominated in distribution*) by  $X'$ :

$$X \leq_d X'.$$

However, the other direction also holds: *stochastic domination* implies *pointwise domination*: from  $F(x) \geq G(x), \quad x \in \mathfrak{N}$ , it follows that

$$G(x) \geq u \text{ implies } F(x) \geq u,$$

and thus,

$$\{x \in \mathfrak{N} \mid G(x) \geq u\} \subseteq \{x \in \mathfrak{N} \mid F(x) \geq u\}, \text{ thus,}$$

$$F^{-1}(u) \equiv \inf\{x \mid F(x) \geq u\} \leq \inf\{x \mid G(x) \geq u\} \equiv G^{-1}(u),$$

implying

$$F^{-1}(U) \leq G^{-1}(U), \text{ that is,}$$

$$\hat{X} \leq \hat{X}' \quad \text{pointwise (quantile coupling).}$$

Thus, we have proved a simple version of *Strassen's theorem*:

**Theorem 1.5** (Strassen, 1965) *Let  $X$  and  $X'$  be random variables. Then*

$$X \leq_d X'$$

if, and only if, there is a coupling  $(\hat{X}, \hat{X}')$  of  $X$  and  $X'$  such that a.s.

$$\hat{X} \leq \hat{X}'.$$

Note that it is often easier to carry out arguments using pointwise domination than stochastic domination. See the following example.

**Example 1.21** (Additivity of stochastic domination) Let  $X_1, X_2, X'_1$ , and  $X'_2$  be random variables such that  $X_1$  and  $X_2$  are independent,  $X'_1$  and  $X'_2$  are independent, and

$$X_1 \leq_d X'_1 \quad \text{and} \quad X_2 \leq_d X'_2.$$

For  $i = 1, 2$ , let  $(\hat{X}_i, \hat{X}'_i)$  be a coupling of  $X_i$  and  $X'_i$  such that  $\hat{X}_i \leq \hat{X}'_i$ . Let  $(\hat{X}_1, \hat{X}'_1)$  and  $(\hat{X}_2, \hat{X}'_2)$  be independent. Then  $(\hat{X}_1 + \hat{X}_2, \hat{X}'_1 + \hat{X}'_2)$  is a coupling of  $X_1 + X_2$  and  $X'_1 + X'_2$ , and

$$\hat{X}_1 + \hat{X}_2 \leq \hat{X}'_1 + \hat{X}'_2$$

implying

$$X_1 + X_2 \leq_d X'_1 + X'_2.$$

### Coupling event inequality and maximal coupling

The following question is the starting point of many convergence and approximation results obtained from coupling arguments. Let  $X$  and  $X'$  be two random variables with non-identical distributions. How can one construct a coupling of  $X$  and  $X'$ ,  $(\hat{X}, \hat{X}')$ , such that  $P(\hat{X} = \hat{X}')$  is maximal across all possible couplings? Here we follow the slightly more general formulation in Thorisson (2000), but limit presentation to the case of discrete random variables (the continuous case being completely analogous).

**Definition 1.24** Suppose  $(\hat{X}_i : i \in I)$  is a coupling of  $X_i, i \in I$ , and let  $C$  be an event such that if it occurs, then all the  $\hat{X}_i$  coincide, that is,

$$C \subseteq \{\hat{X}_i = \hat{X}_j, \quad \text{for all } i, j \in I\}.$$

Such an event is called a *coupling event*.

Assume all the  $X_i$  take values in a finite or countable set  $E$  with  $P(X_i = x) = p_i(x)$ , for  $x \in E$ . For all  $i, j \in I$  and  $x \in E$ , we clearly have

$$P(\hat{X}_i = x, C) = P(\hat{X}_j = x, C) \leq p_j(x),$$

and thus for all  $i \in I$  and  $x \in E$ ,

$$P(\hat{X}_i = x, C) \leq \inf_{j \in I} p_j(x).$$

Summing over  $x \in E$  yields the basic *coupling event inequality*:

$$P(C) \leq \sum_{x \in E} \inf_{j \in I} p_j(x). \tag{1.37}$$

As an example, consider again the case of two discrete random variables  $X$  and  $X'$  with coupling  $(\hat{X}, \hat{X}')$ , and set  $C = \{\hat{X} = \hat{X}'\}$ . Then

$$P(\hat{X} = \hat{X}') \leq \sum_x \min\{P(X = x), P(X' = x)\}. \quad (1.38)$$

Interestingly, it turns out that, at least in principle, one can always construct a coupling such that the above coupling event inequality (1.37) holds with identity. Such a coupling is called *maximal* and  $C$  a *maximal coupling event*.

**Proposition 1.13** (Maximal coupling) *Suppose  $X_i, i \in I$ , are discrete random variables taking values in a finite or countable set  $E$ . Then there exists a maximal coupling, that is, a coupling with coupling event  $C$  such that*

$$P(C) = \sum_{x \in E} \inf_{i \in I} p_i(x).$$

*Proof* Put

$$c := \sum_{x \in E} \inf_{i \in I} p_i(x) \quad (\text{the maximal coupling probability}).$$

If  $c = 0$ , take the  $\hat{X}_i$  independent and  $C = \emptyset$ . If  $c = 1$ , take the  $\hat{X}_i$  identical and  $C = \Omega$ , the sample space. For  $0 < c < 1$ , these couplings are mixed as follows: let  $J$ ,  $V$ , and  $W_i, i \in I$ , be independent random variables such that  $J$  is Bernoulli-distributed with  $P(J = 1) = c$  and, for  $x \in E$ ,

$$\begin{aligned} P(V = x) &= \frac{\inf_{i \in I} p_i(x)}{c} \\ P(W_i = x) &= \frac{p_i(x) - c P(V = x)}{1 - c}. \end{aligned}$$

Define, for each  $i \in I$ ,

$$\hat{X}_i = \begin{cases} V, & \text{if } J = 1, \\ W_i, & \text{if } J = 0. \end{cases} \quad (1.39)$$

Then

$$\begin{aligned} P(\hat{X}_i = x) &= P(V = x)P(J = 1) + P(W_i = x)P(J = 0) \\ &= P(X_i = x). \end{aligned}$$

Thus, the  $\hat{X}_i$  are a coupling of the  $X_i$ ,  $C = \{J = 1\}$  is a coupling event, and it has the desired value  $c$ . The representation (1.39) of the  $X_i$  is known as *splitting representation*.  $\square$

### Total variation distance and coupling

We conclude the treatment of coupling with a variation on the theme of maximal coupling. Given two random variables, what is a measure of closeness between

them when an appropriate coupling is applied to make them as close to being identical as possible?

The *total variation distance* between two probability distributions  $\mu$  and  $\nu$  on  $\Omega$  is defined as

$$\|\mu - \nu\|_{TV} := \max_{A \subseteq \Omega} |\mu(A) - \nu(A)|,$$

for all Borel sets  $A$ . Thus, the distance between  $\mu$  and  $\nu$  is the maximum difference between the probabilities assigned to a single event by the two distributions. Using the coupling inequality, it can be shown that

$$\|\mu - \nu\|_{TV} = \inf\{P(X \neq Y) \mid (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}. \quad (1.40)$$

A splitting representation analogous to the one in the previous proof then assures that a coupling can be constructed so that the infimum is obtained (see Levin *et al.*, 2008, pp. 50–52).

### 1.3.6 Fréchet–Hoeffding bounds and Fréchet distribution classes

We consider general classes of multivariate distributions with given marginals. In contrast to the previous section on coupling, here the existence of a multivariate distribution is taken for granted. Apart from studying the class of multivariate distributions  $\mathcal{F}(F_1, \dots, F_n)$  with fixed univariate marginals  $F_1, \dots, F_n$ , classes of distributions with fixed bivariate or higher-order marginals have been studied as well. For example, the class  $\mathcal{F}(F_{12}, F_{13})$  refers to trivariate distributions with fixed bivariate marginals  $F_{12}$  and  $F_{13}$ , where it is assumed that the first univariate margin  $F_1$  is the same. General questions to ask are (cf. Joe, 1997):

1. What are the bounds for the multivariate distributions in a specified class  $\mathcal{F}$ ?
2. Do the bounds correspond to proper multivariate distributions, and if so, is there a stochastic representation or interpretation of the bounds?
3. Are the bounds sharp if they do not correspond to proper multivariate distributions?
4. What are simple members of  $\mathcal{F}$ ?
5. Can one construct parametric subfamilies in  $\mathcal{F}$  with desirable properties?

The case of  $n = 2$  is considered first.

#### Fréchet–Hoeffding bounds for $n = 2$

For two events  $A_1, A_2$  in a probability space, the following sequence of inequalities is quite obvious:

$$\begin{aligned} \max\{0, P(A_1) + P(A_2) - 1\} &\leq P(A_1) + P(A_2) - P(A_1 \cup A_2) \\ &= P(A_1 \cap A_2) \leq \min\{P(A_1), P(A_2)\}. \end{aligned} \quad (1.41)$$

Defining  $A_1 = \{X_1 \leq x_1\}$  and  $A_2 = \{X_2 \leq x_2\}$  for random variables  $X_1, X_2$  with joint distribution  $F(x_1, x_2)$  and marginals  $F_1(x_1)$  and  $F_2(x_2)$  results in the Fréchet–Hoeffding bounds  $F^-$  and  $F^+$ :

**Proposition 1.14** *Let  $F(x_1, x_2)$  be a bivariate distribution function with marginals  $F_1$  and  $F_2$ . Then for all  $x_1, x_2$ ,*

$$\begin{aligned} F^-(x_1, x_2) &= \max\{0, F_1(x_1) + F_2(x_2) - 1\} \leq F(x_1, x_2) \\ &\leq \min\{F_1(x_1), F_2(x_2)\} = F^+(x_1, x_2), \end{aligned} \quad (1.42)$$

and both the lower bound  $F^-$  and the upper bound  $F^+$  are distribution functions.

It is routine to check that the conditions of Proposition 1.2 are satisfied, i.e., that both  $F^-$  and  $F^+$  are distribution functions as well. Moreover, it can be shown, for continuous marginals  $F_1, F_2$ , that  $F^+$  and  $F^-$  represent perfect positive and negative dependence, respectively, in the following sense:  $X_1$  is (almost surely) an increasing, respectively decreasing, function of  $X_2$ . This has also been referred to a *comonotonicity* and *countermonotonicity*, respectively.<sup>16</sup>

Rearranging the inequalities in (1.41) yields

$$\max\{P(A_1), P(A_2)\} \leq P(A_1 \cup A_2) \leq \min\{P(A_1) + P(A_2), 1\},$$

and, for the distributions,

$$\max\{F_1(x_1), F_2(x_2)\} \leq F^*(x_1, x_2) \leq \min\{F_1(x_1) + F_2(x_2), 1\}, \quad (1.43)$$

with  $F^*(x_1, x_2) = P(X_1 \leq x_1 \cup X_2 \leq x_2)$ .

**Example 1.22** (Race model inequality) This example relates to the audiovisual interaction context described in introductory Example 1.2, where visual ( $RT_V$ ), auditory ( $RT_A$ ), and visual–audio ( $RT_{VA}$ ) reaction times are observed. According to the *race model*,  $RT_{VA}$  is considered a random variable identical in distribution to  $\min\{RT_V, RT_A\}$ , with  $RT_V$  and  $RT_A$  random variables coupled with respect to a common probability space.

Rewriting Equation (1.43) restricted to the diagonal, i.e., to all  $(x_1, x_2)$  with  $x_1 = x_2 \equiv x$ , one obtains for the corresponding distribution functions

$$\max\{F_V(x), F_A(x)\} \leq F_{VA}(x) \leq \min\{F_V(x) + F_A(x), 1\}, \quad x \geq 0. \quad (1.44)$$

The upper bound was suggested as a test of the race model in the reaction time literature (Miller, 1982); its testing has become routine procedure in empirical studies for discriminating “genuine” multisensory integration, that is, integration based the action of multisensory neurons, from mere probability summation (race model) (Colonius and Diederich, 2006).

<sup>16</sup> Note that perfect dependence does not mean that the correlation equals +1 or −1.

### Fréchet–Hoeffding bounds for $n \geq 3$

Using<sup>17</sup> the simple generalization of the Inequalities (1.41) to  $n \geq 3$ ,

$$\max \left\{ 0, \sum_j P(A_j) - (n-1) \right\} \leq P(A_1 \cap \dots \cap A_n) \leq \min_j P(A_j), \quad (1.45)$$

yields the general *Fréchet bounds* (short for Fréchet–Hoeffding b)

$$\max\{0, F_1(x_1) + \dots + F_n(x_{n-1})\} \leq F(x_1, \dots, x_n) \leq \min_j F_j(x_j), \quad (1.46)$$

for all  $x_j \in \mathfrak{N}$ ,  $j = 1, \dots, n$  and with  $F \in \mathcal{F}(F_1, \dots, F_n)$ , the class of all  $n$ -variate distribution functions with fixed marginals  $F_1, \dots, F_n$ . Moreover, it can be shown that the bounds are sharp, i.e., they cannot be improved in general.

The Fréchet upper bound,  $\min_j F_j(x_j) \equiv F^+(\mathbf{x})$ ,  $\mathbf{x} \in \mathfrak{N}^n$ , is a distribution function. Indeed, if one of the univariate distributions is continuous, say  $F_1$ , then  $F^+(\mathbf{x})$  is the joint distribution of  $\mathbf{X}$  defined by quantile coupling via  $U = {}_d F_1(X_1)$  (cf. Example 1.20):

$$\mathbf{X} = {}_d (U, F_2^{-1}(U), \dots, F_n^{-1}(U)).$$

If all marginals are continuous, the random variables show perfect positive dependence (*comonotonicity*). For proof in the general case, see Joe (1997, p. 58).

Note that the lower Fréchet bound is not in general a distribution function for  $n \geq 3$ . This is easy to see in the case of  $n = 3$  with  $F_1$ ,  $F_2$ , and  $F_3$  non-degenerate. If the lower Fréchet bound  $F^-(\mathbf{x})$  is a distribution for  $(X_1, X_2, X_3)$ , then all three bivariate margins must also be (two-dimensional) lower Fréchet bounds. If one of the three distributions is continuous, say  $F_1$ , then, by using the probability transform and to satisfy countermonotonicity,

$$X_j = F_j^{-1}(1 - F_1(X_1)), \quad j = 2, 3.$$

However, in this case  $X_2, X_3$  are positively associated and this is a contradiction.

Nevertheless, a necessary and sufficient condition for the lower Fréchet bound of  $\mathcal{F}(F_1, \dots, F_n)$  to be a distribution can be stated for all  $n \geq 3$ :

$$\begin{aligned} & \text{either } \sum_j F_j(x_j) \leq 1, \quad \text{whenever } 0 < F_j(x_j) < 1, \quad j = 1, \dots, n; \\ & \text{or } \sum_j F_j(x_j) \geq n - 1, \quad \text{whenever } 0 < F_j(x_j) < 1, \quad j = 1, \dots, n. \end{aligned} \quad (1.47)$$

It turns out that these conditions imply that each of the  $F_j$  must have a discrete component.

<sup>17</sup> All results in this and the subsequent section are from Joe (1997, 2015), where a wealth of results are gathered together.

### Fréchet distribution classes with given higher-order marginals

We concentrate here on some examples for the case of  $n = 3$ .

#### (A) Class $\mathcal{F}(F_{12}, F_{13})$

We assume that the conditional distributions  $F_{2|1}$  and  $F_{3|1}$  are well defined.  $\mathcal{F}(F_{12}, F_{13})$  is always nonempty because it contains the trivariate distribution with the second and third variables conditionally independent given the first, defined by

$$F_l(\mathbf{x}) = \int_{-\infty}^{x_1} F_{2|1}(x_2|y) F_{3|1}(x_3|y) dF_1(y). \quad (1.48)$$

Now let  $F \in \mathcal{F}(F_{12}, F_{13})$  and write

$$F(\mathbf{x}) = \int_{-\infty}^{x_1} F_{23|1}(x_2, x_3|y) dF_1(y),$$

where  $F_{23|1}$  is the bivariate conditional distribution of the second and third variable given the first. Applying the bivariate bounds from Proposition 1.14 yields the Fréchet upper bound

$$F^+(\mathbf{x}) = \int_{-\infty}^{x_1} \min\{F_{2|1}(x_2|y), F_{3|1}(x_3|y)\} dF_1(y)$$

and the Fréchet lower bound

$$F^-(\mathbf{x}) = \int_{-\infty}^{x_1} \max\{0, F_{2|1}(x_2|y) + F_{3|1}(x_3|y) - 1\} dF_1(y),$$

and both of these bounds are proper distributions.

One application of this result yields a generalization of the race model inequality (cf. Example 1.22) by conditioning on a third random variable, for example, an indicator variable.

#### (B) Class $\mathcal{F}(F_{12}, F_{13}, F_{23})$

This class is more difficult to analyze. There is clearly a need to check for compatibility of the three bivariate margins. For example, if  $F_{12}$ ,  $F_{13}$  and  $F_{23}$  are bivariate standard normal distributions with corresponding correlations  $\rho_{12}$ ,  $\rho_{13}$ , and  $\rho_{23}$ , then compatibility would mean that the correlation matrix with these values is nonnegative definite.

It is straightforward, however, to obtain upper and lower bounds. For  $F \in \mathcal{F}(F_{12}, F_{13}, F_{23})$ , and  $\mathbf{x} \in \mathbb{R}^3$ :

$$\max\{0, b_1(\mathbf{x}), b_2(\mathbf{x}), b_3(\mathbf{x})\} \leq F(\mathbf{x}) \leq \min\{a_{12}(\mathbf{x}), a_{13}(\mathbf{x}), a_{23}(\mathbf{x}), a_{123}(\mathbf{x})\}, \quad (1.49)$$

where

$$\begin{aligned} a_{12}(\mathbf{x}) &= F_{12}(x_1, x_2), \quad a_{13}(\mathbf{x}) = F_{13}(x_1, x_3), \quad a_{23}(\mathbf{x}) = F_{23}(x_2, x_3), \\ a_{123}(\mathbf{x}) &= 1 - F_1(x_1) - F_2(x_2) - F_3(x_3) \\ &\quad + F_{12}(x_1, x_2) + F_{13}(x_1, x_3) + F_{23}(x_2, x_3), \\ b_1(\mathbf{x}) &= F_{12}(x_1, x_2) + F_{13}(x_1, x_3) - F_1(x_1), \\ b_2(\mathbf{x}) &= F_{12}(x_1, x_2) + F_{23}(x_2, x_3) - F_2(x_2), \\ b_3(\mathbf{x}) &= F_{13}(x_1, x_3) + F_{23}(x_2, x_3) - F_3(x_3). \end{aligned}$$

For  $F_{12}$ ,  $F_{13}$ , and  $F_{23}$  to be compatible bivariate margins, the upper bound must be greater or equal to the lower bound in the above inequality for all  $\mathbf{x}$ . Some versions of these bounds play a role in describing properties of general parallel processing models.

Several methods for obtaining compatibility conditions are presented in Joe (1997, 2015). Some distribution-independent results exist, but no solution for the general case has been suggested. For continuous distributions, a result is based on considering *Kendall's tau* ( $\tau$ ), a measure of dependence defined as follows.

**Definition 1.25** (Kendall's tau) Let  $F$  be a continuous bivariate distribution function and let  $(X_1, X_2)$ ,  $(X'_1, X'_2)$  be independent pairs with distribution  $F$ . Then (the population version of) *Kendall's tau* equals the probability of concordant pairs minus the probability of discordant pairs, i.e.,

$$\begin{aligned} \tau &= P[(X_1 - X'_1)(X_2 - X'_2) > 0] - P[(X_1 - X'_1)(X_2 - X'_2) < 0] \\ &= 2P[(X_1 - X'_1)(X_2 - X'_2) > 0] - 1 \\ &= 4P[X_1 > X'_1, X_2 > X'_2] - 1 = 4 \int_{I^2} F \, dF - 1. \end{aligned}$$

The following gives a necessary condition for compatibility:

**Proposition 1.15** (Joe, 1997, p. 76) *Let  $F \in \mathcal{F}(F_{12}, F_{13}, F_{23})$  and suppose  $F_{jk}$ ,  $j < k$ , are continuous. Let  $\tau_{jk} = \tau_{kj}$  be the value of Kendall's tau for  $F_{jk}$ ,  $j \neq k$ . Then the inequality*

$$-1 + |\tau_{ij} + \tau_{jk}| \leq \tau_{ik} \leq 1 - |\tau_{ij} - \tau_{jk}|$$

*holds for all permutations  $(i, j, k)$  of  $(1, 2, 3)$  and the bounds are sharp.*

Thus, if the above inequality does not hold for some  $(i, j, k)$ , then the three bivariate margins are not compatible. Sharpness follows from the special trivariate normal case: Kendall's tau for the bivariate normal is  $\tau = (2/\pi) \arcsin(\rho)$ , so that the inequality becomes

$$-\cos\left(\frac{1}{2}\pi(\tau_{12} + \tau_{23})\right) \leq \sin\left(\frac{1}{2}\pi\tau_{13}\right) \leq \cos\left(\frac{1}{2}\pi(\tau_{12} - \tau_{23})\right),$$

with  $(i, j, k) = (1, 2, 3)$ .

### Best bounds on the distribution of sums

Given that processing time in the psychological context is often assumed to be additively composed of certain sub-processing times, it is clearly of interest to establish bounds on the distribution of sums of random variables with arbitrary dependency structure. In keeping with the assumptions of the previous sections, existence of a multivariate distribution with given marginal distributions is presumed.

We start with the case of  $n = 2$ ; consider the sum  $S = X_1 + X_2$  with possibly dependent random variables  $X_1$  and  $X_2$ . We are interested in finding random variables  $S_{\min}$  and  $S_{\max}$  such that

$$P(S_{\min} \leq s) \leq P(S \leq s) \leq P(S_{\max} \leq s), \quad s \in \mathbb{R},$$

and these bounds should be the best possible (in the stochastic order sense).

Let  $F_i$ ,  $i = 1, 2$ , be the (marginal) distribution for  $X_i$ , and let  $F_i^-(s) = P(X_i < s)$  be the left limit, defined for all  $s \in \mathbb{R}$ , with  $i = 1, 2$ . We introduce

$$F_{\min}(s) = \sup_{x \in \mathbb{R}} \max \{F_1^-(x) + F_2^-(s - x) - 1, 0\} \quad (1.50)$$

and

$$F_{\max}(s) = \inf_{x \in \mathbb{R}} \min \{F_1(x) + F_2(s - x), 1\}. \quad (1.51)$$

It is routine to show that there exist random variables  $S_{\min}$  and  $S_{\max}$  such that  $F_{\min}(s) = P(S_{\min} < s)$  and  $F_{\max}(s) = P(S_{\max} \leq s)$  (see Denuit *et al.*, 2005, p. 364).

**Proposition 1.16** *If  $S = X_1 + X_2$  and  $F_{\min}(s)$  and  $F_{\max}(s)$  are defined by Equation (1.50) and Equation (1.51), then*

$$F_{\min}(s) \leq F_S(s) \leq F_{\max}(s),$$

for all  $s \in \mathbb{R}$ .

*Proof* (Denuit *et al.*, 2005, p. 364) For arbitrary  $s$  and  $x$  in  $\mathbb{R}$ , it is clear that  $X_1 > x$  and  $X_2 > s - x$  together imply  $S > s$ , so that

$$F_S(s) \leq P(X_1 \leq x \text{ or } X_2 \leq s - x) \leq F_1(x) + F_2(s - x),$$

obviously implying  $F_S(s) \leq F_{\max}(s)$  for all  $s$ . Moreover, note that

$$P(X_1 < x) + P(X_2 < s - x) - P(X_1 < x, X_2 < s - x) \leq 1,$$

from which it follows that

$$\max \{F_1^-(x) + F_2^-(s - x) - 1, 0\} \leq P(X_1 < x, X_2 < s - x) \leq F_S(s),$$

which implies the desired result.  $\square$

Note that the distribution functions of  $S_{\min}$  and  $S_{\max}$  provide the best possible bounds on the distribution of  $S$ , but these random variables are no longer expressible as sums. Here is an example allowing an explicit computation of  $F_{\min}$  and  $F_{\max}$ .

**Example 1.23** Let  $\text{Exp}(1/\lambda, \theta)$  denote a shifted exponential distribution  $F(x) = 1 - \exp[-(x - \theta)/\lambda]$  for  $\lambda > 0$  and  $x \geq \theta$ .

If  $X_i$  is distributed as  $\text{Exp}(1/\lambda_i, \theta_i)$ ,  $i = 1, 2$ , then, using the method of Lagrange multipliers to determine the extrema of  $F_S(s)$  under the constraint  $x_1 + x_2 = s$ , the distribution function of  $S = X_1 + X_2$  is bounded below by

$$F_{\min}(s) = 1 - \exp[-(s - \theta^*)/(\lambda_1 + \lambda_2)],$$

with  $\theta^* = \theta_1 + \theta_2 + (\lambda_1 + \lambda_2) \log(\lambda_1 + \lambda_2) - \lambda_1 \log(\lambda_1) - \lambda_2 \log(\lambda_2)$ . It is also bounded above by

$$F_{\max}(s) = 1 - \exp[-(s - (\theta_1 + \theta_2))/\max(\lambda_1, \lambda_2)].$$

Best bounds can also be obtained for sums of  $n \geq 3$  random variables and for non-decreasing continuous functions of the random variables. Moreover, assuming positive dependence among the random variables, the bounds in Proposition 1.16 can be sharpened.

### 1.3.7 Copula theory

There is a close connection between studying multivariate distributions with given marginals, i.e., the Fréchet distribution classes, and the theory of copulas to be introduced next.

#### Definition, examples, and Sklar's theorem

Let  $(X, Y)$  be a pair of random variables with joint distribution function  $F(x, y)$  and marginal distributions  $F_X(x)$  and  $F_Y(y)$ . To each pair of real numbers  $(x, y)$ , we can associate three numbers:  $F_X(x)$ ,  $F_Y(y)$ , and  $F(x, y)$ . Note that each of these numbers lies in the interval  $[0, 1]$ . In other words, each pair  $(x, y)$  of real numbers leads to a point  $(F_X(x), F_Y(y))$  in the unit square  $[0, 1] \times [0, 1]$ , and this ordered pair in turn corresponds to a number  $F(x, y)$  in  $[0, 1]$ .

$$(x, y) \mapsto (F_X(x), F_Y(y)) \mapsto F(x, y) = C(F_X(x), F_Y(y)),$$

$$\mathfrak{N} \times \mathfrak{N} \longrightarrow [0, 1] \times [0, 1] \xrightarrow{C} [0, 1].$$

The mapping  $C$  is an example of a copula (it “couples” the bivariate distribution with its marginals). Using the probability integral transformation, a straightforward definition for any finite dimension  $n$  is the following:

**Definition 1.26** A function  $C : [0, 1]^n \longrightarrow [0, 1]$  is called  $n$ -dimensional *copula* if there is a probability space  $(\Omega, \mathcal{F}, P)$  supporting a vector of standard uniform random variables  $(U_1, \dots, U_n)$  such that

$$C(u_1, \dots, u_n) = P(U_1 \leq u_1, \dots, U_n \leq u_n), \quad u_1, \dots, u_n \in [0, 1].$$

The concept of copula has stirred a lot of interest in recent years in several areas of statistics, including finance, mainly for the following reasons (see e.g.,

Joe, 2015): it allows (i) study of the structure of stochastic dependency in a “scale-free” manner, i.e., independent of the specific marginal distributions, and (ii) the construction of families of multivariate distributions with specified properties. The following important theorem laid the foundation of these studies (for a proof, see Nelsen, 2006, Theorem 2.10.9).

**Theorem 1.6** (Sklar’s Theorem, 1959) *Let  $F(x_1, \dots, x_n)$  be an  $n$ -variate distribution function with margins  $F_1(x_1), \dots, F_n(x_n)$ ; then there exists an  $n$ -copula  $C : [0, 1]^n \rightarrow [0, 1]$  that satisfies*

$$F(x_1, \dots, x_n) = C(F_1(x_1), \dots, F_n(x_n)), \quad (x_1, \dots, x_n) \in \mathbb{R}^n.$$

*If all univariate margins  $F_1, \dots, F_n$  are continuous, then the copula is unique. Otherwise,  $C$  is uniquely determined on  $\text{Ran}F_1 \times \text{Ran}F_2 \times \dots \times \text{Ran}F_n$ .*

*If  $F_1^{-1}, \dots, F_n^{-1}$  are the quantile functions of the margins, then for any  $(u_1, \dots, u_n) \in [0, 1]^n$*

$$C(u_1, \dots, u_n) = F(F_1^{-1}(u_1), \dots, F_n^{-1}(u_n)).$$

Sklar’s Theorem shows that copulas remain invariant under strictly increasing transformations of the underlying random variables. It is possible to construct a wide range of multivariate distributions by choosing the marginal distributions and a suitable copula.

**Example 1.24** (Bivariate exponential) For  $\delta > 0$ , the distribution

$$F(x, y) = \exp\{-[e^{-x} + e^{-y} - (e^{\delta x} + e^{\delta y})^{-1/\delta}]\}, \quad -\infty < x, y < +\infty,$$

with margins  $F_1(x) = \exp\{-e^{-x}\}$  and  $F_2(y) = \exp\{-e^{-y}\}$  corresponds to the copula

$$C(u, v) = uv \exp\{[(-\log u)^{-\delta} + (-\log v)^{-\delta}]^{-1/\delta}\},$$

an example of the class of *bivariate extreme value* copulas characterized by  $C(u^t, v^t) = C^t(u, v)$ , for all  $t > 0$ .

There is an alternative, analytical definition of copula based on the fact that distribution functions can be characterized as functions satisfying certain conditions, without reference to a probability space (cf. Proposition 1.2).

**Definition 1.27** An  $n$ -dimensional copula  $C$  is a function on the unit  $n$ -cube  $[0, 1]^n$  that satisfies the following properties:

1. the range of  $C$  is the unit interval  $[0, 1]$ ;
2.  $C(\mathbf{u})$  is zero for all  $\mathbf{u}$  in  $[0, 1]^n$  for which at least one coordinate is zero (groundedness);
3.  $C(\mathbf{u}) = u_k$  if all coordinates of  $\mathbf{u}$  are 1 except the  $k$ -the one;
4.  $C$  is  *$n$ -increasing*, that is, for every  $\mathbf{a} \leq \mathbf{b}$  in  $[0, 1]^n$  ( $\leq$  defined component-wise) the volume assigned by  $C$  to the  $n$ -box  $[\mathbf{a}, \mathbf{b}] = [a_1, b_1] \times \dots \times [a_n, b_n]$  is nonnegative.

One can show that groundedness and the  $n$ -increasing property are sufficient to define a proper distribution function (for  $n = 2$ , see Proposition 1.2). Moreover, from the results in Section 1.3.6 it follows that every copula is bounded by the Fréchet–Hoeffding bounds,

$$\max(u_1, + \cdots + u_n - n + 1, 0) \leq C(u_1, \dots, u_n) \leq \min(u_1, \dots, u_n).$$

For  $n = 2$ , both of the bounds are copulas themselves, but for  $n \geq 3$  the lower bound is no longer  $n$ -increasing. Another well-known copula is the *independence* copula,

$$C(u_1, \dots, u_n) = \prod_{i=1}^n u_i.$$

Moreover, copulas are uniformly continuous and all their partial derivatives exist almost everywhere, which is a useful property especially for computer simulations. Copulas with discrete margins have also been defined, but their treatment is less straightforward (for a review, see Pfeifer and Nešlehová, 2004).

### Copula density and pair copula constructions (vines)

If the probability measure associated with a copula  $C$  is absolutely continuous (with respect to the Lebesgue measure on  $[0, 1]^n$ ), then there exists a *copula density*  $c : [0, 1]^n \rightarrow [0, \infty]$  almost everywhere unique such that

$$C(u_1, \dots, u_n) = \int_0^{u_1} \cdots \int_0^{u_n} c(v_1, \dots, v_n) dv_n \dots dv_1, \quad u_1, \dots, u_n \in [0, 1].$$

Such an absolutely continuous copula is  $n$ -times differentiable and

$$c(u_1, \dots, u_n) = \frac{\partial}{\partial u_1} \cdots \frac{\partial}{\partial u_n} C(u_1, \dots, u_n), \quad u_1, \dots, u_n \in [0, 1].$$

For example, the independence copula is absolutely continuous with density equal to 1:

$$\Pi(u_1, \dots, u_n) = \prod_{k=1}^n u_k = \int_0^{u_1} \cdots \int_0^{u_n} 1 dv_n \dots dv_1.$$

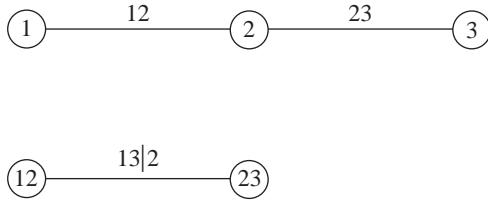
When the density of a distribution  $F_{12}(x_1, x_2)$  exists, differentiating yields

$$f_{12}(x_1, x_2) = f_1(x_1)f_2(x_2)c_{12}(F_1(x_1), F_2(x_2)).$$

This equation shows how independence is “distorted” by copula density  $c$  whenever  $c$  is different from 1. Moreover, this yields an expression for the conditional density of  $X_1$  given  $X_2 = x_2$ :

$$f_{1|2}(x_1|x_2) = c_{12}(F_1(x_1), F_2(x_2))f_1(x_1). \tag{1.52}$$

This is the starting point of a recent, important approach to constructing high-dimensional dependency structures from pairwise dependencies (“vine copulas”).



**Figure 1.5** Graphical illustration of the decomposition in Equation (1.53).  
A line in the graph corresponds to the indices of a copula linking two distributions, unconditional in the upper graph, conditional in the lower graph.

Note that a multivariate density of dimension  $n$  can be decomposed as follows, here taking the case for  $n = 3$ :

$$f(x_1, x_2, x_3) = f_{3|12}(x_3|x_1, x_2)f_{2|1}(x_2|x_1)f_1(x_1).$$

Applying the decomposition in Equation (1.52) to each of these terms yields,

$$\begin{aligned} f_{2|1}(x_2|x_1) &= c_{12}(F_1(x_1), F_2(x_2))f_2(x_2) \\ f_{3|12}(x_3|x_1, x_2) &= c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2))f_{3|2}(x_3|x_2) \\ f_{3|2}(x_3|x_2) &= c_{23}(F_2(x_2), F_3(x_3))f_3(x_3), \end{aligned}$$

resulting in the “regular vine tree” representation

$$\begin{aligned} f(x_1, x_2, x_3) &= f_3(x_3)f_2(x_2)f_1(x_1) \quad (\text{marginals}) \\ &\quad \times c_{12}(F_1(x_1), F_2(x_2)) \cdot c_{23}(F_2(x_2), F_3(x_3)) \quad (\text{unconditional pairs}) \\ &\quad \times c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) \quad (\text{conditional pair}). \end{aligned} \tag{1.53}$$

In order to visualize this structure, in particular for larger  $n$ , one defines a sequence of trees (acyclic undirected graphs), a simple version of which is depicted in Figure 1.5.

### Survival copula, dual and co-copula

Whenever it is more convenient to describe the multivariate distribution of a random vector  $(X_1, \dots, X_n)$  by means of its survival distribution, i.e.,

$$\hat{F}(x_1, \dots, x_n) = P(X_1 > x_1, \dots, X_n > x_n),$$

its *survival copula* can be introduced such that the analog of Sklar’s theorem holds, with

$$\hat{F}(x_1, \dots, x_n) = \hat{C}(\hat{F}_1(x_1), \dots, \hat{F}_n(x_n)),$$

where the  $\hat{F}_i$ ,  $i = 1, \dots, n$ , are the marginal survival distributions. In the continuous case there is a one-to-one correspondence between the copula and its survival copula. For  $n = 2$ , this is

$$C(u_1, u_2) = \hat{C}(1 - u_1, 1 - u_2) + u_1 + u_2 - 1.$$

For the general case, we refer to Mai and Scherer (2012, pp. 20–21).

Two other functions closely related to copulas and survival copulas are useful in the response time modeling context. The *dual of a copula*  $C$  is the function  $\tilde{C}$  defined by  $\tilde{C}(u, v) = u + v - C(u, v)$  and the *co-copula* is the function  $C^*$  defined by  $C^*(u, v) = 1 - C(1 - u, 1 - v)$ . Neither of these is a copula, but when  $C$  is the (continuous) copula of a pair of random variables  $X$  and  $Y$ , the dual of the copula and the co-copula each express a probability of an event involving  $X$  and  $Y$ :

$$\tilde{C}(F(x), G(y)) = P(X \leq x \text{ or } Y \leq y)$$

and

$$C^*(1 - F(x), 1 - G(y)) = P(X > x \text{ or } Y > y).$$

### Copulas with singular components

If the probability measure associated with a copula  $C$  has a singular component, then the copula also has a singular component which can often be detected by finding points  $(u_1, \dots, u_n) \in [0, 1]^n$ , where some (existing) partial derivative of the copula has a point of discontinuity. A standard example is the upper *Fréchet bound copula*

$$M(u_1, \dots, u_n) = \min(u_1, \dots, u_n),$$

where the partial derivatives have a point of discontinuity;

$$\frac{\partial}{\partial u_k} M(u_1, \dots, u_n) = \begin{cases} 1, & u_k < \min(u_1, \dots, u_{k-1}, u_{k+1}, \dots, u_n), \\ 0, & u_k > \min(u_1, \dots, u_{k-1}, u_{k+1}, \dots, u_n). \end{cases}.$$

The probability measure associated with  $M(u_1, \dots, u_n)$  assigns all mass to the diagonal of the unit  $n$ -cube  $[0, 1]^n$  (“perfect positive dependence”).

### Archimedean copulas

The class of *Archimedean copulas* plays a major role in constructing multivariate dependency. We start by motivating this class via a mixture model interpretation, following the presentation in Mai and Scherer (2012). Consider a sequence of i.i.d. exponential random variables  $E_1, \dots, E_n$  defined on a probability space  $(\Omega, \mathcal{F}, P)$  with mean  $E[E_1] = 1$ . Independent of the  $E_i$ , let  $M$  be a positive random variable and define the vector of random variables  $(X_1, \dots, X_n)$  by

$$(X_1, \dots, X_n) := \left( \frac{E_1}{M}, \dots, \frac{E_n}{M} \right).$$

This can be interpreted as a two-step experiment. First, the intensity  $M$  is drawn. In a second step, an  $n$ -dimensional vector of i.i.d.  $\text{Exp}(M)$  variables is drawn. Thus, for each margin  $k \in \{1, \dots, n\}$ , one has

$$\begin{aligned} \bar{F}_k(x) &= P(X_k > x) = P(E_k > xM) \\ &= E \left[ E[\mathbb{1}_{\{E_k > xM\}} | M] \right] = E[e^{-xM}] := \varphi(x), \quad x \geq 0, \end{aligned}$$

where  $\varphi$  is the *Laplace transform* of  $M$ . Thus, the margins are related to the mixing variable  $M$  via its Laplace transform. As long as  $M$  is not deterministic, the  $X_k$ 's are dependent, because they are all affected by  $M$ : whenever the realization of  $M$  is large (small), all realizations of  $X_k$  are more likely to be small (large). The resulting copula is parametrized by the Laplace transform of  $M$ :

**Proposition 1.17** *The survival copula of random vector  $(X_1, \dots, X_n) \equiv (\frac{E_1}{M}, \dots, \frac{E_n}{M})$  is given by*

$$\varphi(\varphi^{-1}(u_1) + \dots + \varphi^{-1}(u_n)) := C_\varphi(u_1, \dots, u_n), \quad (1.54)$$

where  $u_1, \dots, u_n \in [0, 1]$  and  $\varphi$  is the Laplace transform of  $M$ .

Copulas of this structure are called *Archimedean copulas*. Written differently,

$$P(X_1 > x_1, \dots, X_n > x_n) = C_\varphi(\varphi(x_1), \dots, \varphi(x_n)),$$

where  $x_1, \dots, x_n \geq 0$ . From Equation (1.54) it clearly follows that the copula  $C_\varphi$  is symmetric, i.e., the vector of  $X_k$ 's has an *exchangeable* distribution.

*Proof of Proposition 1.17.* The joint survival distribution of  $(X_1, \dots, X_n)$  is given by

$$\begin{aligned} \bar{F}(x_1, \dots, x_n) &= P(X_1 > x_1, \dots, X_n > x_n) \\ &= E[E[\mathbb{1}_{\{E_1 > x_1 M\}} \cdots \mathbb{1}_{\{E_n > x_n M\}} | M]] \\ &= E[e^{-M \sum_{k=1}^n x_k}] = \varphi\left(\sum_{k=1}^n x_k\right), \end{aligned}$$

where  $(x_1, \dots, x_n) \in [0, \infty]^n$ . The proof is established by transforming the marginals to  $U[0, 1]$  and by using the survival version of Sklar's theorem.  $\square$

Not all Archimedean copulas are obtained via the above probabilistic construction. The usual definition of Archimedean copulas is analytical: the function defined in Equation (1.54) is an *Archimedean copula generator* of the general form

$$\varphi(\varphi^{-1}(u_1) + \dots + \varphi^{-1}(u_n)) = C_\varphi(u_1, \dots, u_n),$$

if  $\varphi$  satisfies the following properties:

- (i)  $\varphi : [0, \infty] \rightarrow [0, 1]$  is strictly increasing and continuously differentiable (of all orders);
- (ii)  $\varphi(0) = 1$  and  $\varphi(\infty) = 0$ ;
- (iii)  $(-1)^k \varphi^{(k)} \geq 0$ ,  $k = 1, \dots, n$ , i.e., the derivatives are alternating in sign up to order  $n$ .

If Equation (1.54) is a copula for all  $n$ , then  $\varphi$  is a *completely monotone function* and turns out to be the Laplace transform of a positive random variable  $M$ , as in the previous probabilistic construction (cf. Mai and Scherer, 2012, p. 64).

### Example: Clayton copula and the copula of max and min order statistics

The generator of the Clayton family is given by

$$\varphi(x) = (1 + x)^{-1/\vartheta}, \quad \varphi^{-1}(x) = x^{-\vartheta} - 1, \quad \vartheta \in (0, \infty).$$

It can be shown that this corresponds to a  $\text{Gamma}(1/\vartheta, 1)$  distributed mixing variable  $M$ . The Clayton copula is hidden in the following application: let  $X_{(1)}$  and  $X_{(n)}$  be the min and max order statistics of a sample  $X_1, \dots, X_n$ ; one can compute the copula of  $(-X_{(1)}, X_{(n)})$  via

$$P(-X_{(1)} \leq x, X_{(n)} \leq y) = P(\cap_{i=1}^n \{X_i \in [-x, y]\}) = (F(y) - F(-x))^n,$$

for  $-x \leq y$  and zero otherwise. Marginal distributions of  $-X_{(1)}$  and  $X_{(n)}$  are, respectively,

$$F_{X_{(n)}}(y) = F(y)^n \quad \text{and} \quad F_{-X_{(1)}}(x) = (1 - F(-x))^n.$$

Sklar's theorem yields the copula

$$C_{-X_{(1)}, X_{(n)}}(u, v) = \max\{0, (u^{1/n} + v^{1/n} - 1)^n\},$$

which is the bivariate Clayton copula with parameter  $\vartheta = 1/n$ . Because  $C_{X_{(1)}, X_{(n)}}(u, v) = v - C_{-X_{(1)}, X_{(n)}}(1 - u, v)$ , it can be shown with elementary computations (see Schmitz, 2004) that

$$\lim_{n \rightarrow +\infty} C_{X_{(1)}, X_{(n)}}(u, v) = v - \lim_{n \rightarrow +\infty} C_{-X_{(1)}, X_{(n)}}(1 - u, v) = v - (1 - u)v = uv,$$

for all  $u, v \in (0, 1)$ , i.e.,  $X_{(1)}$  and  $X_{(n)}$  are *stochastically independent* asymptotically.

**Operations on distributions not derivable from operations on random variables**  
In the introductory section, Example 1.1 claimed that the mixture of two distributions  $F$  and  $G$ , say,  $\phi(F, G) = pF + (1 - p)G$ ,  $p \in (0, 1)$ , was not derivable by an operation defined directly on the corresponding random variables, in contrast to, for example, the convolution of two distributions which is derivable from addition of the random variables. Here we first explicate the definition of “derivability,” give a proof of that claim, and conclude with a caveat.

**Definition 1.28** A binary operation  $\phi$  on the space of univariate distribution functions  $\Delta$  is said to be *derivable* from a function on random variables if there exists a measurable two-place function  $V$  satisfying the following condition: for any distribution functions  $F, G \in \Delta$  there exist random variables  $X$  for  $F$  and  $Y$  for  $G$ , defined on a common probability space, such that  $\phi(F, G)$  is the distribution function of  $V(X, Y)$ .

**Proposition 1.18** *The mixture  $\phi(F, G) = pF + (1 - p)G$ ,  $p \in (0, 1)$  is not derivable.*

*Proof* (Alsina and Schweizer, 1988) Assume  $\phi$  is derivable, i.e., that a suitable function  $V$  exists. For any  $a \in \mathfrak{N}$ , define the unit step function  $\epsilon_a$  in  $\Delta$  by

$$\epsilon_a(t) = \begin{cases} 0, & \text{if } t \leq a, \\ 1, & \text{if } t > a; \end{cases}$$

and, for any given  $x$  and  $y$  ( $x \neq y$ ), let  $F = \epsilon_x$  and  $G = \epsilon_y$ . Then  $F$  and  $G$  are, respectively, the distribution functions of  $X$  and  $Y$ , defined on a common probability space and, respectively, equal to  $x$  and  $y$  a.s. It follows that  $V(X, Y)$  is a random variable defined on the same probability space as  $X$  and  $Y$  which is equal to  $V(x, y)$  a.s. The distribution of  $V(X, Y)$  is  $\epsilon_{V(x,y)}$ . However, because  $\phi$  is derivable from  $V$ , the distribution of  $V(X, Y)$  must be  $\phi(\epsilon_x, \epsilon_y)$ , which is  $p\epsilon_x + (1 - p)\epsilon_y$ . Because  $p\epsilon_x + (1 - p)\epsilon_y \neq \epsilon_{V(x,y)}$ , we have a contradiction.  $\square$

In Alsina and Schweizer (1988) it is shown that forming the geometric mean of two distributions is also not derivable. These examples demonstrate that the distinction between modeling with distribution functions, rather than with random variables, “... is intrinsic and not just a matter of taste” (Schweizer and Sklar, 1974), and that “the classical model for probability theory – which is based on random variables defined on a common probability space – has its limitations” (Alsina *et al.*, 1993).

It should be noted, however, that by a slight extension of the copula concept, i.e., allowing for randomized operations, the mixture operation on  $\Delta$  can be recovered by operations on random variables. Indeed, assume there is a Bernoulli random variable  $I$  independent of  $(X, Y)$  with  $P(I = 1) = p$ ; then the random variable  $Z$  defined by  $Z = IX + (1 - I)Y$  has a mixture distribution  $\phi(F, G)$ . Such a representation of a mixture is often called a *latent variable representation* of  $\phi$  and is easily generalized to a mixture of more than two components (see Faugeras, 2012, for an extensive discussion of this approach).

### 1.3.8 Concepts of dependence

It is obvious that copulas can represent any type of dependence among random variables. Moreover, we have seen that the upper and lower Fréchet–Hoeffding bounds correspond to maximal positive and negative dependence, respectively. There exist many, more or less strong, concepts of dependence that are often useful for model building. We can consider only a small subset of these.

#### Positive dependence

Here it is of particular interest to find conditions that guarantee multivariate positive dependence in the following sense. For a random vector  $(X_1, \dots, X_n)$

$$P(X_1 > x_1, \dots, X_n > x_n) \geq \prod_{i=1}^n P(X_i > x_i), \quad (1.55)$$

for all real  $x$ , called *positive upper orthant dependence (PUOD)*. Replacing “ $>$ ” by “ $\leq$ ” in the probabilities above yields the concept of *positive lower orthant dependence (PLOD)*. An example for *PLOD* are archimedean copulas with completely monotone generators (see Proposition 1.17), this property following naturally from the construction where independent components are made positively dependent by the random variable  $M$ :

$$(X_1, \dots, X_n) = \left( \frac{E_1}{M}, \dots, \frac{E_n}{M} \right).$$

Both *PUOD* and *PLOD* are implied by a stronger concept.

**Definition 1.29** A random vector  $\mathbf{X} = (X_1, \dots, X_n)$  is said to be *associated* if

$$\text{Cov}[f(\mathbf{X}), g(\mathbf{X})] \geq 0$$

for all non-decreasing real-valued functions for which the covariance exists.

Association implies the following properties:

1. Any subset of associated random variables is associated.
2. If two sets of associated random variables are independent of one another, then their union is a set of associated random variables.
3. The set consisting of a single random variable is associated.
4. Non-decreasing functions of associated random variables are associated.
5. If a sequence  $\mathbf{X}^{(k)}$  of random vectors, associated for each  $k$ , converges to  $\mathbf{X}$  in distribution, then  $\mathbf{X}$  is also associated.

Examples for applying the concept include the following.

1. Let  $X_1, \dots, X_n$  be independent. If  $S_k = \sum_{i=1}^k X_i$ ,  $k = 1, \dots, n$ , then  $(S_1, \dots, S_n)$  is associated.
2. The order statistics  $X_{(1)}, \dots, X_{(n)}$  are associated.
3. Let  $\Lambda$  be a random variable and suppose that, conditional on  $\Lambda = \lambda$ , the random variables  $X_1, \dots, X_n$  are independent with common distribution  $F_\lambda$ . If  $F_\lambda(x)$  is decreasing in  $\lambda$  for all  $x$ , then  $X_1, \dots, X_n$  are associated (without conditioning on  $\lambda$ ). To see this let  $U_1, \dots, U_n$  be independent standard uniform random variables that are independent of  $\Lambda$ . Define  $X_1, \dots, X_n$  by  $X_i = F_\Lambda^{-1}(U_i)$ ,  $i = 1, \dots, n$ . As the  $X_i$  are increasing functions of the independent random variables  $\Lambda, U_1, \dots, U_n$  they are associated (Ross, 1996, p. 449).

An important property of association is that it allows inferring independence from uncorrelatedness. This can be shown using *PUOD/PLOD* and Hoeffding’s Lemma (Proposition 1.4). Proposition 1.19 below gives a more general result.

Whenever it is difficult to directly test for association, a number of stronger dependency concepts are available.

**Definition 1.30** The bivariate density  $f(x, y)$  is of *total positivity of order 2* ( $TP_2$ ) if

$$f(x, y)f(x', y') = f(x, y')f(x', y),$$

for all  $x < x'$  and  $y < y'$  in the domain of  $f(x, y)$ .

This property has also been called the “positive likelihood ratio dependence” and can be shown to imply association. A multivariate density

$$f(x_1, \dots, x_i, \dots, x_j, \dots, x_n)$$

is called  $TP_2$  in pairs if the density, considered as a function of  $x_i$  and  $x_j$ , with all other arguments fixed, is  $TP_2$ . In  $\mathbb{R}^n$ , the concept of  $TP_2$  in pairs is equivalent to the following.

**Definition 1.31** A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is *multivariate total positive of order two*,  $MTP_2$ , if

$$f(\mathbf{x} \vee \mathbf{y})f(\mathbf{x} \wedge \mathbf{y}) \geq f(\mathbf{x})f(\mathbf{y}),$$

for all  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ , where  $\mathbf{x} \vee \mathbf{y}$  denotes the vector of maxima,  $\max(x_i, y_i)$ , and  $\mathbf{x} \wedge \mathbf{y}$  the vector of minima,  $\min(x_i, y_i)$ ,  $i = 1, \dots, n$ . If  $f$  is a density for random vector  $\mathbf{X}$  satisfying  $MTP_2$ , then  $\mathbf{x}$  is said to be  $MTP_2$ .

The class  $MTP_2$  possesses a number of nice properties, among them:

1. If  $f, g$  are  $MTP_2$ , then  $fg$  is  $MTP_2$ ;
2. A vector  $\mathbf{X}$  of independent random variables is  $MTP_2$ ;
3. the marginals of  $f$  which is  $MTP_2$  are also  $MTP_2$ ;
4. If  $f(\mathbf{x}, \mathbf{y})$  is  $MTP_2$  on  $\mathbb{R}^n \times \mathbb{R}^m$ ,  $g(\mathbf{y}, \mathbf{z})$  is  $MTP_2$  on  $\mathbb{R}^m \times \mathbb{R}^k$ , then

$$h(\mathbf{x}, \mathbf{z}) = \int f(\mathbf{x}, \mathbf{y})g(\mathbf{y}, \mathbf{z}) d\mathbf{y}$$

is  $MTP_2$  on  $\mathbb{R}^n \times \mathbb{R}^k$ ;

5. If  $f$  is  $MTP_2$ , then  $f(\phi_1(x_1), \dots, \phi_n(x_n))$  is  $MTP_2$ , where  $\phi_i$  are non-decreasing.

Moreover, it has been shown that  $MTP_2$  implies association. Examples of  $MTP_2$  random variables abound. For example, a multivariate normal density is  $MTP_2$  if, and only if, the elements of the inverse of the covariance matrix,  $B = \Sigma^{-1}$ ,  $b_{ij} \leq 0$  for  $i \neq j$ . The vector of order statistics is  $MTP_2$  as well, and so are negative multinomial variables. The following observation is of interest from a model-building point of view.

Let  $n$ -dimensional vectors  $\mathbf{X}$  and  $\mathbf{Y}$  be independent. If both  $\mathbf{X}$  and  $\mathbf{Y}$  are  $MTP_2$ , then the convolution  $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$  is no longer  $MTP_2$ , but  $\mathbf{Z}$  is still associated.

### Negative dependence

A concept of positive dependence for two random variables can easily be transformed into one of negative dependence. Specifically, if  $(X, Y)$  has positive dependence, then  $(X, -Y)$  on  $\mathbb{R}^2$ , or  $(X, 1 - Y)$  on the unit square, have negative dependence. However, with more than two variables, negative dependence is restricted. For example, if  $X_1$  has perfect negative dependence with  $X_2$  and with  $X_3$ , then  $X_2$  and  $X_3$  have perfect positive dependence, and the larger the number of variables, the weaker is negative dependence.

Reversing the Inequality (1.55) in the definition of *PUOD* turns this multivariate positive dependence concept into *negative upper orthant dependence (NUOD)* and, similarly, *PLOD* becomes *NLOD* by reversing the inequality. Both *NUOD* and *NLOD* clearly capture the idea of multivariate negative dependence but the challenge has been to find conditions under which these inequalities are fulfilled. The most influential concept was introduced by Joag-Dev and Proschan (1983) (where all results of this section can be found), based on the simple idea that negative dependence means that splitting the set of random variables into two parts leads to a negative covariance between the two sets.

**Definition 1.32** Random vector  $\mathbf{X}$  is *negatively associated (NA)* if, for every subset  $A \subseteq \{1, \dots, n\}$ ,

$$\text{Cov}(f(X_i, i \in A), g(X_j, j \in A^c)) \leq 0,$$

whenever  $f, g$  are non-decreasing.

*NA* may also refer to the set of random variables  $\{X_1, \dots, X_n\}$  or to the underlying distribution of  $\mathbf{X}$ . It has the following properties:

1. For a pair of random variables  $(X, Y)$ , *NA* and *negative quadrant dependence*,

$$P(X \leq x, Y \leq y) \leq P(X \leq x)P(Y \leq y)$$

are equivalent.

2. For disjoint subsets  $A_1, \dots, A_m$  of  $\{1, \dots, n\}$ , and non-decreasing functions  $f_1, \dots, f_n$ ,  $\mathbf{X}$  being *NA* implies

$$E \prod_{i=1}^m f_i(\mathbf{X}_{A_i}) \leq \prod_{i=1}^m E f_i(\mathbf{X}_{A_i}),$$

where  $\mathbf{X}_{A_i} = (X_j, j \in A_i)$ .

3. If  $\mathbf{X}$  is *NA*, then it is both *NUOD* and *NLOD*.
4. A subset of *NA* random variables is *NA*.
5. If  $\mathbf{X}$  has independent components, then it is *NA*.
6. Increasing functions defined on disjoint subsets of a set of *NA* variables are *NA*.
7. If  $\mathbf{X}$  is *NA* and  $\mathbf{Y}$  is *NA* and  $\mathbf{X}$  is independent of  $\mathbf{Y}$ , then  $(\mathbf{X}, \mathbf{Y})$  is *NA*.

Negative association is often created by conditioning. For example, let  $X_1, \dots, X_n$  be independent random variables. Then the joint conditional distribution of  $X_1, \dots, X_n$ , given  $\sum_{i=1}^n X_i = s$ , is *NA* for almost all  $s$ .

A result on independence via uncorrelatedness, generalizing the result cited in the previous section on positive dependence, is the following.

**Proposition 1.19** *Suppose  $\mathbf{X}$  is (positively) associated or negatively associated. Then*

- (a)  $\mathbf{X}_A$  is independent of  $\mathbf{X}_B$  if, and only if,  $\text{Cov}(X_i, X_j) = 0$ , for  $i \in A, j \in B, A \cap B = \emptyset$ ;
- (b)  $X_1, \dots, X_n$  are mutually independent if, and only if,  $\text{Cov}(X_i, X_j) = 0$  for all  $i \neq j$ .

For example, a multivariate normal density has negative (positive) association if and only if all its pairwise correlation coefficients are negative (positive).

### Measuring dependence

The dependence properties discussed so far are characterizable as being *local*. For example, the inequality defining *PQD* holds for all pairs  $(x, y)$  of the random variables  $X, Y$ . From this point of view, covariance and correlation are *global* measures of dependency. Because most of the (local) dependence concepts introduced above are invariant with respect to increasing transformations of the variables, two bivariate measures of monotone dependence for continuous variables are of special interest. The first, *Kendall's tau*, was introduced in Definition 1.25. The other is the following.

**Definition 1.33** (Spearman's rho) Let  $F$  be a continuous bivariate distribution function with univariate margins  $F_1, F_2$  and let  $(X_1, X_2)$  have distribution  $F$ . Then the (population version) of *Spearman's rho*,  $\rho$ , is defined as the correlation of  $F_1(X_1)$  and  $F_2(X_2)$ .

Because  $F_1(X_1)$  and  $F_2(X_2)$  are standard uniform random variables  $(U, V)$  with distribution function (copula)  $C$ , their expectations are  $1/2$ , their variances are  $1/12$ , and Spearman's rho becomes

$$\begin{aligned}\rho &= \frac{\text{EUV} - \text{EUEV}}{\sqrt{\text{Var}U\text{Var}V}} = \frac{\text{EUV} - 1/4}{1/12} = 12\text{EUV} - 3 \\ &= 12 \int_{I^2} uv \, dC(u, v) - 3.\end{aligned}\tag{1.56}$$

This parallels the copula version of Kendall's tau,

$$\tau = 4 \int_{I^2} C \, dC - 1,$$

which can be interpreted as the expected value of the function  $C(U, V)$  of standard uniform random variables,  $4\text{EC}(U, V) - 1$ . Besides being invariant with respect to increasing transformations, both  $\tau$  and  $\rho$  are equal to 1 for the bivariate Fréchet

upper bound (one variable is an increasing function of the other) and to  $-1$  for the Fréchet lower bound (one variable is a decreasing function of the other). There are numerous results for both measures in the literature, let us just mention an early one by Kruskal (1958) relating both measures. For the population versions of Kendall's tau and Spearman's rho,

$$-1 \leq 3\tau - 2\rho \leq 1.$$

### 1.3.9 Stochastic orders

The simplest way of comparing two distribution functions is by comparison of the associated means (or medians). However, such a comparison is often not very informative, being based on only two single numbers.<sup>18</sup> In studies of reaction time, for example, analyses of empirical data have long been limited to simply comparing mean reaction times measured under different experimental conditions. However, it is now generally acknowledged that much deeper information about the underlying mental processes can be uncovered by examining the entire distribution function (e.g., Palmer *et al.*, 2011).

Several orders of distribution functions that compare the “location” or the “magnitude” of random variables will be considered here. Moreover, when one is interested in comparing two distributions that have the same mean, one usually considers the standard deviations or some other measure of variability of these distributions. However, such a comparison is again based on only two single numbers. Several orders of distributions taking into account various forms of knowledge about the underlying distributions will be considered here. Orders associated with different notions of positive dependence as well as multivariate orders are briefly mentioned. There is a wealth of results on the various order concepts (see Shaked and Shantikumar, 2007, for a comprehensive compilation), but, due to space limitations, our presentation must be very selective.

#### Univariate stochastic orders

Let  $X$  and  $Y$  be two random variables with distribution functions  $F$  and  $G$ , respectively, not necessarily defined on the same probability space. In Section 1.3.5, the usual stochastic domination order was defined in the context of quantile coupling:

**Definition 1.34**  $X$  is said to be *stochastically dominated* (or *dominated in distribution*) by  $Y$ ,

$$X \leq_d Y, \text{ if } F(x) \geq G(x), \quad x \in \mathfrak{N}.$$

By Strassen's theorem (Theorem 1.5), there is a coupling  $(\hat{X}, \hat{Y})$  of  $X$  and  $Y$  such that  $\hat{X} \leq \hat{Y}$  pointwise. This can be used to prove an equivalent condition for the above stochastic order which easily generalizes to multivariate orders (see below).

<sup>18</sup> Sometimes, means may not even exist for a distribution.

**Proposition 1.20**

$$X \leq_d Y \quad \text{if, and only if,} \quad (*) \quad E[g(X)] \leq E[g(Y)]$$

for all bounded nondecreasing functions  $g$ .

*Proof* Suppose  $X \leq_d Y$  and obtain a coupling  $(\hat{X}, \hat{Y})$  such that  $\hat{X} \leq \hat{Y}$  pointwise. Then, for any bounded nondecreasing function  $g$ ,  $g(\hat{X}) \leq g(\hat{Y})$ , implying  $E[g(X)] \leq E[g(Y)]$ . Conversely, fix an  $x \in \mathfrak{N}$  and take  $g = 1_{(x, +\infty)}$  in  $(*)$ : then

$$P(X > x) \leq P(Y > x), \quad x \in \mathfrak{N},$$

implying  $X \leq_d Y$ .  $\square$

Stochastic domination is closed under non-decreasing transformations, in the following sense. Let  $X_1, \dots, X_m$  be a set of independent random variables and let  $Y_1, \dots, Y_m$  be another set of independent random variables. If  $X_i \leq_d Y_i$  for  $i = 1, \dots, m$ , then, for any non-decreasing function  $\psi : \mathfrak{N}^m \rightarrow \mathfrak{N}$ , one has

$$\psi(X_1, \dots, X_m) \leq_d \psi(Y_1, \dots, Y_m).$$

In particular,

$$\sum_{j=1}^m X_j \leq_d \sum_{j=1}^m Y_j,$$

that is, the order is closed under convolution.

Obviously, if both  $X \leq_d Y$  and  $Y \leq_d X$ , then  $X$  and  $Y$  are equal in distribution,  $X =_d Y$ . However, sometimes a weaker condition already implies equality in distribution. Clearly, if  $X \leq_d Y$  then  $EX \leq EY$ . However, we also have:

**Proposition 1.21** *If  $X \leq_d Y$  and if  $Eh(X) = Eh(Y)$  for some strictly increasing function  $h$ , then  $X =_d Y$ .*

*Proof* First, we assume  $h(x) = x$ ; let  $\hat{X}$  and  $\hat{Y}$  be the variables of the corresponding coupling, thus  $P(\hat{X} \leq \hat{Y}) = 1$ . If  $P(\hat{X} < \hat{Y}) > 0$ , then  $EX = E\hat{X} < EY = E\hat{Y}$ , a contradiction to the assumption. Therefore,  $X =_d \hat{X} = \hat{Y} =_d Y$ . Now let  $h$  be some strictly increasing function. Observe that if  $X \leq_d Y$ , then  $h(X) \leq_d h(Y)$  and, therefore, from the above result, we have that  $h(X) =_d h(Y)$ . The strict monotonicity of  $h$  yields  $X =_d Y$ .  $\square$

The next order is based on the concept of hazard rate function (see Definition 1.20).

**Definition 1.35** Let  $X$  and  $Y$  be two nonnegative random variables with absolutely continuous distribution functions  $F$  and  $G$  and with hazard rate functions  $r$  and  $q$ , respectively, such that

$$r(t) \geq q(t), \quad t \in \mathfrak{N}.$$

The  $X$  is said to be *smaller than  $Y$  in the hazard rate order*, denoted as  $X \leq_{hr} Y$ .

It turns out that  $\leq_{hr}$  can be defined for more general random variables, not necessarily nonnegative and even without requiring absolute continuity. The condition is

$$\frac{\bar{G}(t)}{\bar{F}(t)} \text{ increases in } t \in (-\infty, \max(u_X, u_Y))$$

( $a/0$  is taken to be equal to  $\infty$  whenever  $a > 0$ ). Here,  $u_X$  and  $u_Y$  denote the right endpoints of the supports of  $X$  and  $Y$ . The above can be written equivalently as

$$\bar{F}(x)\bar{G}(y) \geq \bar{F}(y)\bar{G}(x) \quad \text{for all } x \leq y, \quad (1.57)$$

which, in the absolutely continuous case, is equivalent to

$$\frac{f(x)}{\bar{F}(y)} \geq \frac{g(x)}{\bar{G}(y)} \quad \text{for all } x \leq y.$$

Setting  $x = -\infty$  in (1.57), one can show that

$$X \leq_{hr} Y \implies X \leq_d Y.$$

Many more results on  $\leq_{hr}$  are available, e.g., order statistics satisfy

$$X_{k:n} \leq_{hr} X_{k+1:n}$$

for  $k = 1, 2, \dots, n$ . Moreover, if  $W$  is a random variable with (mixture) distribution function  $pF + (1-p)G$  for some  $p \in (0, 1)$ , then  $X \leq_{hr} Y$  implies  $X \leq_{hr} W \leq_{hr} Y$ .

**Definition 1.36** Let  $X$  and  $Y$  have densities  $f$  and  $g$ , respectively, such that

$$\frac{g(t)}{f(t)} \text{ increases in } t \text{ over the union of the supports of } X \text{ and } Y$$

(here  $a/0$  is taken to be equal to  $\infty$  whenever  $a > 0$ ), or, equivalently,

$$f(x)g(y) \geq f(y)g(x) \quad \text{for all } x \leq y.$$

Then  $X$  is said to be *smaller than  $Y$  in the likelihood ratio order*, denoted by  $X \leq_{lr} Y$ .

An analogous definition can be given for discrete random variables, and even for mixed distributions. Moreover, it can be shown that, for distributions functions  $F$  and  $G$  for  $X$  and  $Y$ , respectively,

$$X \leq_{lr} Y \implies GF^{-1} \text{ is convex.}$$

The following (strict) hierarchy exists:

$$X \leq_{lr} Y \implies X \leq_{hr} Y \implies X \leq_d Y.$$

Results on order statistics and binary mixtures analogous to those for  $\leq_{hr}$  cited above also hold for  $\leq_{lr}$ .

## Univariate variability orders

### The convex order

The first variability order considered here requires the definition of a *convex function*  $\phi : \mathfrak{N} \rightarrow \mathfrak{N}$ , i.e.,

$$\phi(\alpha x + (1 - \alpha)y) \leq \alpha\phi(x) + (1 - \alpha)\phi(y),$$

for all  $x, y \in \mathfrak{N}$  and  $\alpha \in [0, 1]$ .

**Definition 1.37** Let  $X$  and  $Y$  be two random variables such that

$$E\phi(X) \leq E\phi(Y) \quad \text{for all convex functions } \phi : \mathfrak{N} \rightarrow \mathfrak{N},$$

provided the expectations exist. Then  $X$  is said to be *smaller than  $Y$  in the convex order*, denoted as

$$X \leq_{cx} Y.$$

Roughly speaking, convex functions take on their (relatively) larger values over regions of the form  $(-\infty, a) \cup (b, +\infty)$  for  $a < b$ . Therefore, when the condition in the above definition holds,  $Y$  is more likely to take on “extreme” values than  $X$ ; that is,  $Y$  is “more variable” than  $X$ .<sup>19</sup>

The convex order has some strong implications. Using the convexity of functions  $\phi(x) = x$ ,  $\phi(x) = -x$  and  $\phi(x) = x^2$ , it is easily shown that

$$X \leq_{cx} Y \quad \text{implies } EX = EY \quad \text{and } \text{Var}X \leq \text{Var}Y.$$

Moreover, when  $EX = EY$ , then a condition equivalent to  $X \leq_{cx} Y$  is

$$E[\max(X, a)] \leq E[\max(Y, a)], \quad \text{for all } a \in \mathfrak{N}.$$

A deeper result, an extension of Strassen’s theorem (Theorem 1.5) is the following.

**Theorem 1.7** (Müller and Rüscherdorf, 2001) *The random variables  $X$  and  $Y$  satisfy  $X \leq_{cx} Y$  if, and only if, there exist two random variables  $\hat{X}$  and  $\hat{Y}$ , defined on the same probability space, such that*

$$\hat{X} =_d X \quad \text{and} \quad \hat{Y} =_d Y,$$

*and  $\{\hat{X}, \hat{Y}\}$  is a martingale, i.e.,  $E[\hat{Y}|\hat{X}] = \hat{X}$  a.s. Furthermore, the random variables  $\hat{X}$  and  $\hat{Y}$  can be selected such that  $[\hat{Y}|\hat{X} = x]$  is increasing in  $x$  in the stochastic order  $\leq_d$ .*

As noted above, the convex order only compares random variables that have the same means. One way to drop this requirement is to introduce a so-called *dilation order* by defining

$$X \leq_{dil} Y, \quad \text{if} \quad [X - EX] \leq_{cx} [Y - EY].$$

<sup>19</sup> Clearly, it is sufficient to consider only functions  $\phi$  that are convex on the union of the supports of  $X$  and  $Y$  rather than over the whole real line.

For nonnegative random variables  $X$  and  $Y$  with finite means, one can alternatively define the *Lorenz order* by

$$X \leq_{\text{Lorenz}} Y, \quad \text{if} \quad \frac{X}{\mathbb{E}X} \leq_{\text{cx}} \frac{Y}{\mathbb{E}Y},$$

which can be used to order random variables with respect to the *Lorenz curve* used, for example, in economics to measure the inequality of incomes.

We conclude with a result that nicely connects the convex order with the multivariate dependence concepts from Section 1.3.8.

**Proposition 1.22** *Let  $X_1, \dots, X_n$  be positively associated (cf. Definition 1.29) random variables, and let  $Y_1, \dots, Y_n$  be independent random variables such that  $X_i =_d Y_i$ ,  $i = 1, \dots, n$ . Then*

$$\sum_{i=1}^n X_i \geq_{\text{cx}} \sum_{i=1}^n Y_i.$$

If “positively” is replaced by “negatively” above (cf. Definition 1.32), then the convex order sign is reversed.

This makes precise the intuition that, if random variables  $X_1, \dots, X_n$  are “more positively [negatively] associated” than random variables  $Y_1, \dots, Y_n$  in some sense, but otherwise have identical margins for each  $i = 1, \dots, n$ , then the sum of the  $X_i$  should be larger (smaller) in the convex order than the  $Y_i$ .

### The dispersive order

This order is based on comparing difference between quantiles of the distribution functions.

**Definition 1.38** Let  $X$  and  $Y$  be random variables with quantile functions  $F^{-1}$  and  $G^{-1}$ , respectively. If

$$F^{-1}(\beta) - F^{-1}(\alpha) \leq G^{-1}(\beta) - G^{-1}(\alpha), \quad \text{whenever } 0 < \alpha \leq \beta < 1,$$

then  $X$  is said to be *smaller than  $Y$  in the dispersive order*, denoted by  $X \leq_{\text{disp}} Y$ .

In contrast to the convex order, the dispersive order is clearly *location-free*:

$$X \leq_{\text{disp}} Y \Leftrightarrow X + c \leq_{\text{disp}} Y, \quad \text{for any real } c.$$

The dispersive order is also *dilative*, i.e.,  $X \leq_{\text{disp}} aX$  whenever  $a \geq 1$  and, moreover,

$$X \leq_{\text{disp}} Y \Leftrightarrow -X \leq_{\text{disp}} -Y.$$

However, it is not closed under convolutions. Its characterization requires two more definitions.

First, a random variable  $Z$  is called *dispersive* if  $X + Z \leq_{\text{disp}} Y + Z$  whenever  $X \leq_{\text{disp}} Y$  and  $Z$  is independent of  $X$  and  $Y$ . Second, a density (more general, any nonnegative function)  $g$  is called *logconcave* if  $\log g$  is concave, or, equivalently,

$$g[\alpha x + (1 - \alpha)y] \geq [g(x)]^\alpha [g(y)]^{1-\alpha}.$$

**Proposition 1.23** *The random variable  $X$  is dispersive if, and only if,  $X$  has a logconcave density.<sup>20</sup>*

It can be shown that the dispersion order implies the dilation order, that is, for random variables with finite means,

$$X \leq_{\text{disp}} Y \implies X \leq_{\text{dil}},$$

from which it immediately follows that also  $\text{Var}X \leq \text{Var}Y$ .

### *The quantile spread order*

All variability orders considered so far are strong enough to imply a corresponding ordering of the variances. Moreover, distributions with the same finite support are not related in terms of the dispersive order unless they are identical. Hence, a weaker concept is desirable. In fact, Townsend and Colonius (2005) developed a weaker concept of variability order in an attempt to describe the effect of sample size on the shape of the distribution of the order statistics  $X_{1:n}$  and  $X_{n:n}$ . It is based on the notion of *quantile spread*.

**Definition 1.39** The *quantile spread* of random variable  $X$  with distribution  $F$  is

$$QS_X(p) = F^{-1}(p) - F^{-1}(1 - p),$$

for  $0.5 < p < 1$ .

With  $S = 1 - F$  (the survival function),  $F^{-1}(p) = S^{-1}(1 - p)$  implies

$$QS_X(p) = S^{-1}(1 - p) - S^{-1}(p).$$

The quantile spread of a distribution<sup>21</sup> describes how probability mass is placed symmetrically about its median and hence can be used to formalize concepts such as peakedness and tailweight traditionally associated with kurtosis. This way it allows separation of the concepts of kurtosis and peakedness for asymmetric distributions.

For the extreme order statistics (cf. Equation (1.22)) one gets simple expressions for the quantile spread, making it easy to describe their behavior as a function of  $n$ :

$$QS_{\min}(p) = S^{-1}[(1 - p)^{1/n}] - S^{-1}(p^{1/n})$$

<sup>20</sup> Note that a logconcave density follows from its distribution function  $G$  being *strongly unimodal*, i.e., if the convolution  $G * F$  is unimodal for every unimodal  $F$ .

<sup>21</sup> It turns out that the concept of *spread function* defined by Balanda and MacGillivray (1990) is equivalent to the quantile spread.

and

$$QS_{\max}(p) = F^{-1}(p^{1/n}) - F^{-1}[(1-p)^{1/n}].$$

**Definition 1.40** Let  $X$  and  $Y$  be random variables with quantile spreads  $QS_X$  and  $QS_Y$ , respectively. Then  $X$  is called *smaller than  $Y$  in quantile spread order*, denoted as  $X \leq_{QS} Y$ , if

$$QS_X(p) \leq QS_Y(p), \quad \text{for all } p \in (0.5, 1).$$

The following properties of the quantile spread order are easily collected.

1. The order  $\leq_{QS}$  is *location-free*, i.e.,

$$X \leq_{QS} Y \quad \text{iff } X + c \leq_{QS} Y, \quad \text{for any real } c.$$

2. The order  $\leq_{QS}$  is *dilative*, i.e.,  $X \leq_{QS} aX$ , whenever  $a \geq 1$ .
3.  $X \leq_{QS} Y$  if, and only if,  $-X \leq_{QS} -Y$ .
4. Assume  $F_X$  and  $F_Y$  are symmetric, then

$$X \leq_{QS} Y \quad \text{if, and only if, } F_X^{-1}(p) \leq F_Y^{-1}(p) \quad \text{for } p \in (0.5, 1).$$

5.  $\leq_{QS}$  implies ordering of the *mean absolute deviation around the median*, MAD,

$$\text{MAD}(X) = E[|X - M(X)|]$$

( $M$  the median), i.e.,

$$X \leq_{QS} Y \quad \text{implies} \quad \text{MAD}(X) \leq \text{MAD}(Y).$$

The last point follows from writing

$$\text{MAD}(X) = \int_{0.5}^1 [F^{-1}(p) - F^{-1}(1-p)] dp.$$

**Example 1.25** (Cauchy distribution) The Cauchy distribution with density

$$f(x) = \frac{\gamma}{\pi(x^2 + \gamma^2)}, \quad \gamma > 0,$$

has no finite variance; the quantile function is

$$F^{-1}(p) = \gamma \tan[\pi(p - 0.5)],$$

yielding quantile spread

$$QS(p) = \gamma \{\tan[\pi(p - 0.5)] - \tan[\pi(0.5 - p)]\}.$$

Thus, parameter  $\gamma$  clearly defines the  $QS$  order for the Cauchy distribution.

The next example demonstrates how  $QS$  order can be used to describe the effect of sample size on the minimum order statistic.

**Example 1.26** The quantile spread for the Weibull minimum (cf. Example 1.5) is

$$QS_{\min}(p; n) = (n\lambda)^{-1/\alpha} \left[ \left( \ln \frac{1}{1-p} \right)^{1/\alpha} - \left( \ln \frac{1}{p} \right)^{1/\alpha} \right].$$

For  $p \in (0.5, 1)$ , this is easily seen to decrease in  $n$ . Thus,

$$X_{1:n} \leq_{QS} X_{1:n-1},$$

i.e., the quantile spread for the Weibull minimum decreases as a function of sample size  $n$ .

The quantile spread order has been used by Mitnik and Baek (2013) in ordering the *Kumaraswamy distribution*, an alternative to the Beta distribution, which possesses the advantage of having an invertible closed form (cumulative) distribution function.

### Multivariate stochastic orders

Various extensions of both univariate stochastic orders and univariate variability orders to the multivariate case exist. We only sketch two multivariate versions of the usual stochastic order  $\leq_d$  here.

For real vectors  $\mathbf{x} = (x_1, \dots, x_n), \mathbf{y} = (y_1, \dots, y_n)$  we define the coordinatewise ordering  $\mathbf{x} \leq \mathbf{y}$  by  $x_i \leq y_i, i = 1, \dots, n$ . A real-valued function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is *non-decreasing* if  $\mathbf{x} \leq \mathbf{y}$  implies  $f(\mathbf{x}) \leq f(\mathbf{y})$  for all  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ .

The following definition is a straightforward generalization of the condition stated in Proposition 1.20.

**Definition 1.41** Random vectors  $\mathbf{X} = (X_1, \dots, X_n)$  and  $\mathbf{Y} = (Y_1, \dots, Y_n)$  are *strongly stochastically ordered*, denoted as  $\mathbf{X} \leq_{st} \mathbf{Y}$ , if

$$\mathbb{E}[g(\mathbf{X})] \leq \mathbb{E}[g(\mathbf{Y})]$$

for all non-decreasing functions  $g$  for which the expectation exists.<sup>22</sup>

A multivariate version of Strassen's theorem (Theorem 1.5) holds:

**Theorem 1.8** *The random vectors  $\mathbf{X}$  and  $\mathbf{Y}$  satisfy  $\mathbf{X} \leq_{st} \mathbf{Y}$  if, and only if, there exist two random vectors  $\hat{\mathbf{X}}$  and  $\hat{\mathbf{Y}}$ , defined on the same probability space, such that*

$$\hat{\mathbf{X}} = \mathbf{X}, \quad \hat{\mathbf{Y}} = \mathbf{Y}, \quad \text{and} \quad P(\hat{\mathbf{X}} \leq \hat{\mathbf{Y}}) = 1.$$

The actual construction of  $\mathbf{X}$  and  $\mathbf{Y}$  is achieved by an iterative coupling argument (see, e.g., Szekli, 1995, pp. 51–52). Moreover, there exists a straightforward generalization of Proposition 1.21 also allowing the implication  $\mathbf{X} =_{st} \mathbf{Y}$ .

Another, straightforward generalization of the univariate stochastic order  $\leq_d$  is based on the multivariate distribution and survival functions. For a random vector

<sup>22</sup> Note that, for  $n = 1$ , the orders  $\leq_d$  and  $\leq_{st}$  are the same.

$\mathbf{X} = (X_1, \dots, X_n)$  with distribution function  $F$ , let  $\bar{F}$  be the multivariate survival function of  $\mathbf{X}$ , that is,

$$\bar{F}(x_1, \dots, x_n) \equiv P(X_1 > x_1, \dots, X_n > x_n)$$

for all  $\mathbf{x} = (x_1, \dots, x_n)$ . Let  $\mathbf{Y}$  be another  $n$ -dimensional random vector with distribution function  $G$  and survival function  $\bar{G}$ .

**Definition 1.42**  $\mathbf{X}$  is said to be *smaller than  $\mathbf{Y}$  in the upper orthant order*, denoted as  $\mathbf{X} \leq_{uo} \mathbf{Y}$ , if

$$\bar{F}(x_1, \dots, x_n) \leq \bar{G}(x_1, \dots, x_n), \quad \text{for all } \mathbf{x};$$

moreover,  $\mathbf{X}$  is said to be *smaller than  $\mathbf{Y}$  in the lower orthant order*, denoted as  $\mathbf{X} \leq_{lo} \mathbf{Y}$ , if

$$F(x_1, \dots, x_n) \leq G(x_1, \dots, x_n) \quad \text{for all } \mathbf{x}.$$

The following characterization of  $\leq_{uo}$  and  $\leq_{lo}$  demonstrates that these orders are actually, in general, weaker than  $\leq_{st}$ :

**Proposition 1.24** *Let  $\mathbf{X}$  and  $\mathbf{Y}$  be two  $n$ -dimensional random vectors. Then  $\mathbf{X} \leq_{uo} \mathbf{Y}$  if, and only if,*

$$E[\psi(\mathbf{X})] \leq E[\psi(\mathbf{Y})] \quad \text{for every distribution function } \psi;$$

*moreover,  $\mathbf{X} \leq_{lo} \mathbf{Y}$  if, and only if,*

$$E[\psi(\mathbf{X})] \leq E[\psi(\mathbf{Y})] \quad \text{for every survival function } \psi.$$

Clearly, we have the implication

$$\mathbf{X} \leq_{st} \mathbf{Y} \implies (\mathbf{X} \leq_{uo} \mathbf{Y} \quad \text{and} \quad \mathbf{X} \leq_{lo} \mathbf{Y}).$$

Finally, it should be mentioned that multivariate versions of the hazard rate order  $\leq_{hr}$  and of the likelihood ratio order  $\leq_{lr}$  have been developed, as well as generalizations of the variability orders  $\leq_{cx}$  and  $\leq_{disp}$ .

### Positive dependence orders

Whenever a concept of dependence has been defined, one can compare a given distribution with one obtained under stochastic independence. More generally, it may be interesting to compare two distributions with respect to their strength of dependence. However, comparing two distributions seems meaningful only if their marginals are the same. Therefore, comparisons will always be performed within a given *Fréchet* distribution class.

From Section 1.3.6, we know that, in the bivariate case, the upper and lower Fréchet bounds  $F^+$  and  $F^-$  represent distributions with perfect positive and negative dependence, respectively, providing natural upper and lower bounds in the

class  $\mathcal{F}(F_1, F_1)$ . The most basic dependence order is based on the concept of *positive quadrant dependence (PQD)*,

$$P(X_1 \leq x_1, X_2 \leq x_2) \geq P(X_1 \leq x_1)P(X_2 \leq x_2),$$

for all  $x_1, x_2$ . *PQD* is the  $n = 2$  version of the multivariate *PLOD* order introduced in Section 1.3.8.

**Definition 1.43** Let random vector  $(X_1, X_2)$  with distribution function  $F$  and random vector  $(Y_1, Y_2)$  with distribution function  $G$  be members of the Fréchet class  $\mathcal{F}(F_1, F_1)$ . Then,  $(X_1, X_2)$  is said to be *smaller than*  $(Y_1, Y_2)$  in the *PQD order*, denoted by  $(X_1, X_2) \leq_{PQD} (Y_1, Y_2)$ , if

$$F(x_1, x_2) \leq G(x_1, x_2), \quad \text{for all } x_1, x_2.$$

Sometimes it will be useful to write this as  $F \leq_{PQD} G$ . From Hoeffding's identity (Proposition 1.4), it follows readily that, if  $(X_1, X_2) \leq_{PQD} (Y_1, Y_2)$ , then

$$\text{Cov}(X_1, X_2) \leq \text{Cov}(Y_1, Y_2),$$

and, because  $\text{Var}X_i = \text{Var}Y_i$ ,  $i = 1, 2$ , we have

$$\text{Cor}(X_1, X_2) \leq \text{Cor}(Y_1, Y_2).$$

It can also be shown that Kendall's tau and Spearman's rho are preserved under the *PQD order*. Moreover, for every distribution  $F \in \mathcal{F}(F_1, F_1)$ , we have

$$F^- \leq_{PQD} F \leq_{PQD} F^+.$$

For random vectors  $(X_1, X_2)$  and  $(Y_1, Y_2)$  with distribution functions in  $\mathcal{F}(F_1, F_1)$ , we have

$$(X_1, X_2) \leq_{PQD} (Y_1, Y_2) \iff (X_1, X_2) \leq_{uo} (Y_1, Y_2)$$

and

$$(X_1, X_2) \leq_{PQD} (Y_1, Y_2) \iff (X_1, X_2) \geq_{lo} (Y_1, Y_2);$$

however, for the right-hand order relations it is not required that  $(X_1, X_2)$  and  $(Y_1, Y_2)$  have the same marginals. Therefore, whereas the upper and lower orthant orders measure the size (or location) of the underlying random vectors, the *PQD order* measures the degree of positive dependence of the underlying random vectors.

We conclude with an example of *PQD* relating to Archimedean copulas (cf. Section 1.3.7).

**Example 1.27** (Genest and MacKay, 1986) Let  $\phi$  and  $\psi$  be two Laplace transforms of nonnegative random variables. Then  $F$  and  $G$ , defined by

$$F(x_1, x_2) = \phi(\phi^{-1}(x_1) + \phi^{-1}(x_2)), \quad (x_1, x_2) \in [0, 1]^2,$$

and

$$G(y_1, y_2) = \psi(\psi^{-1}(y_1) + \psi^{-1}(y_2)), \quad (y_1, y_2) \in [0, 1]^2,$$

are bivariate distribution functions with standard uniform marginals, i.e.,  $F$  and  $G$  are Archimedean copulas. Then it can be shown that  $(X_1, X_2) \leq_{PQD} (Y_1, Y_2)$  if, and only if  $\psi^{-1}\phi$  is *superadditive*, i.e.,

$$\psi^{-1}\phi(x+y) \geq \psi^{-1}\phi(x) + \psi^{-1}\phi(y),$$

for all  $x, y \geq 0$ .

Several other positive dependence orders, derived from multivariate dependence concepts, like association, have been introduced in various contexts.

## 1.4 Bibliographic references

### 1.4.1 Monographs

The material in Section 1.2 is presented at the level of Durrett (2010) or Gut (2013); some of the measure-theoretic concepts are found in Leadbetter *et al.* (2014) and Pollard (2002); Severini (2005) develops distribution theory without measure theory. An extended discussion of the bivariate exponential distribution (Example 1.7) is found in Galambos and Kotz (1978). Aldous (1985) is a standard reference on exchangeability; see also the review of Kallenberg (2005) by Diaconis (2009); more recent presentations from a Bayesian viewpoint are Goldstein (2013) and Dawid (2013) (see also O'Hagen and Forster, 2004, pp. 108–114). A modern treatment of quantile functions is the quantile-based presentation of reliability analysis in Nair *et al.* (2013). A comprehensive discussion of multivariate survival analysis, including competing risks theory, is given in Crowder (2012). The order statistics section follows the presentation in Arnold *et al.* (1992) and, for the extreme value statistics and asymptotic results, Galambos (1978) was consulted. The introduction to the theory of records follows Arnold *et al.* (1998). The basics of coupling theory are found in the first chapter of Thorisson (2000); see also Lindvall (2002) and Levin *et al.* (2008). The results on Fréchet distribution classes are from Joe (1997) (see also Denuit *et al.*, 2005). A good introduction to copula theory is Nelsen (2006), recent results are collected in Joe (2015); see Mai and Scherer (2012) for simulation issues and Rüschenhoff (2013) for copulas and risk analysis. Concepts of dependence are treated in Mari and Kotz (2004), Joe (1997, 2015), and Barlow and Proschan (1975). The standard reference for stochastic orders is Shaked and Shantikumar (2007) and also Müller and Stoyan (2002) and Szekli (1995); for a compilation of papers on recent order results, see Li and Li (2013).

### 1.4.2 Selected applications in mathematical psychology

This final section presents a list, clearly not representative and far from being exhaustive, of mathematical psychology applications relating to the probabilistic concepts treated in this chapter.

An application of the concept of exchangeability is found in the study by Diaconis and Freedman (1981) disproving a conjecture of Bela Julesz concerning the discriminability of a class of random patterns. Quantile function, survival function, and hazard rate have become standard tools in RT analysis (e.g., Colonius, 1988; Townsend and Eidels, 2011; Houpt and Townsend, 2012; Schwarz and Miller, 2012). Chechile (2011) characterizes properties of the “reverse hazard rate.” The non-identifiability results in competing risks theory have found an interpretation in random utility theory (Marley and Colonius, 1992) and RT modeling (Townsend, 1976; Dzhafarov, 1993; Jones and Dzhafarov, 2014). Order statistics and (generalized) Poisson processes are essential parts of the RT models in Smith and Van Zandt (2000), Miller and Ulrich (2003), and Schwarz (2003). Issues of generalizing extreme-value statistics results have been addressed in Cousineau *et al.* (2002). The theory of coupling seems not to have played a role in mathematical psychology until the work by Dzhafarov and Kujala (2013), although the requirement of a coupling assumption had already been recognized explicitly in earlier work (cf. Ashby and Townsend, 1986; Luce, 1986; Colonius and Vorberg, 1994). Fréchet–Hoeffding bounds have been introduced in the context of audiovisual interaction models of RT (see Colonius, 1990; Diederich, 1992) and were further discussed in Townsend and Nozawa (1995), Townsend and Wenger (2004), and Colonius (2015). Copula theory has been applied in psychometric modeling (see Braeken *et al.*, 2013, and further references therein). The multivariate dependence concept of association was applied in a random utility context by Colonius (1983). Several of the stochastic order concepts presented here have been discussed in Townsend (1990). Stochastic orders related to the *delta plot method* have been characterized in Speckman *et al.* (2008) (see also Schwarz and Miller, 2012).

## 1.5 Acknowledgments

I am grateful to Maike Tahden for pointing out numerous typos and imprecisions in the original draft of the chapter.

## References

- Aalen, O. O., Borgan, Ø. and Gjessing, H. K. (2008). *Survival and Event History Analysis*. New York, NY: Springer Verlag.
- Aldous, D. J. (1985). *École d’Été de Probabilités de Saint-Flour XIII-1983. Lecture Notes in Mathematics*. Vol. 1117. Berlin: Springer, pp. 1–198.
- Alsina, C. and Schweizer, B. (1988). Mixtures are not derivable. *Foundations of Physics Letters*, **1**, 171–174.
- Alsina, C., Nelson, R. B. and Schweizer, B. (1993). On the characterization of class of binary operations on distribution functions. *Statistics and Probability Letters*, **17**, 85–89.

- Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1992). *A First Course in Order Statistics*. New York, NY: John Wiley & Sons.
- Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1998). *Records*. New York, NY: John Wiley & Sons.
- Ashby, F. G. and Townsend, J. T. (1986). Varities of perceptual independence. *Psychological Review*, **93**, 154–179.
- Balandia, K. P. and MacGillivray, H. L. (1990). Kurtosis and spread. *The Canadian Journal of Statistics*, **18**(1), 17–30.
- Barlow, R. E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing*. New York, NY: Holt, Rinehart and Winston.
- Braeken, J., Kuppens, P., De Boeck, P. and Tuerlincks, F. (2013). Contextualized personality questionnaires: a case for copulas in structural equation models for categorical data. *Multivariate Behavioral Research*, **48**, 845–870.
- Chechile, R. A. (2011). Properties of reverse hazard function. *Journal of Mathematical Psychology*, **55**, 203–222.
- Colonius, H. (1983). A characterization of stochastic independence, with an application to random utility theory. *Journal of Mathematical Psychology*, **27**, 103–106.
- Colonius, H. (1988). Modeling the redundant signals effect by specifying the hazard function. *Perception & Psychophysics*, **43**, 605–607.
- Colonius, H. (1990). Possibly dependent probability summation of reaction time. *Journal of Mathematical Psychology*, **34**, 253–275.
- Colonius, H. (1995). The instance theory of automaticity: why the Weibull? *Psychological Review*, **102**(4), 744–750.
- Colonius, H. (2015). Behavioral measures of multisensory integration: bounds on bimodal detection probability. *Brain Topography*, **28**(1), 1–4.
- Colonius, H. and Diederich, A. (2006). The race model inequality: interpreting a geometric measure of the amount of violation. *Psychological Review*, **113**(1), 148–154.
- Colonius, H. and Vorberg, D. (1994). Distribution inequalities for parallel models with unlimited capacity. *Journal of Mathematical Psychology*, **38**, 35–58.
- Cousineau, D., Goodman, V. W. and Shiffrin, R. M. (2002). Extending statistics of extremes to distributions varying on position and scale, and implications for race models. *Journal of Mathematical Psychology*, **46**(4), 431–454.
- Crowder, M. (2012). *Multivariate Survival Analysis and Competing Risks*. Boca Raton, FL: CRC Press.
- Dawid, A. P. (2013). Exchangeability and its ramifications. In Damien, P., Dellaportas, P., Polson, N. G. and Stephens, D. A. (eds), *Bayesian Theory and Applications* (pp. 19–29). Oxford: Oxford University Press.
- Denuit, M., Dhaene, J., Goovaerts, M. and Kaas, R. (2005). *Actuarial Theory for Dependent Risks*. Chichester: John Wiley and Sons, Ltd.
- Desu, M. M. (1971). A characterization of the exponential distribution by order statistics. *Annals of Mathematical Statistics*, **42**, 837–838.
- Diaconis, P. (2009). Review of “Probabilistic symmetries and invariance principles”, by Olav Kallenberg. *Bulletin of the American Mathematical Society*, **46**(4), 691–696.
- Diaconis, P. and Freedman, D. (1981). On the statistics of vision: the Julesz conjecture. *Journal of Mathematical Psychology*, **124**(2), 112–138.
- Diederich, A. (1992). Probability inequalities for testing separate activation models of divided attention. *Perception & Psychophysics*, **52**(6), 714–716.

- Diederich, A. and Colonius, H. (1987). Intersensory facilitation in the motor component? *Psychological Research*, **49**, 23–29.
- Donders, F. C. (1868/1969). Over de snelheid van psychische processen [On the speed of mental processes]. *Onderzoeken gedaan in het Physiologisch Laboratorium der Utrechtsche Hoogeschool, 1868–1869, Tweede reeks, II*, 92–120. Transl. by Koster, W. G. (1969). In *Attention and performance II*, Koster, W. G., Ed., *Acta Psychologica*, **30**, 412–431.
- Durrett, R. (2010). *Probability: Theory and Examples*. 4th edn. Cambridge: Cambridge University Press.
- Dzhafarov, E. N. (1993). Grice-representability of response time distribution families. *Psychometrika*, **58**(2), 281–314.
- Dzhafarov, E. N. and Kujala, J. V. (2013). All-possible-couplings approach to measuring probabilistic context. *PLoS ONE*, **8**(5), e61712.
- Faugeras, O. P. (2012). Probabilistic constructions of discrete copulas. Unpublished manuscript, hal-00751393.
- Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, **24**, 180–190.
- Frechét, M. (1927). Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Polonaise Math.*, **6**, 92–116.
- Galambos, J. (1978). *The Asymptotic Theory of Extreme Order Statistics*. New York, NY: John Wiley & Sons.
- Galambos, J. and Kotz, S. (1978). *Characterizations of Probability Distributions*. Lecture Notes in Mathematics No. 675. New York, NY: Springer.
- Genest, C. and MacKay, J. (1986). Copules archimédiennes et familles de loi bidimensionnelles dont les marges sont données. *The Canadian Journal of Statistics*, **14**(2), 145–159.
- Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, **44**, 423–453.
- Goldstein, M. (2013). Observables and models: exchangeability and the inductive argument. In Damien, P., Dellaportas, P., Polson, N. G., and Stephens, D. A. (eds), *Bayesian Theory and Applications*. Oxford: Oxford University Press, pp. 3–18.
- Gut, A. (2013). *Probability: A Graduate Course*. 2nd edn. New York, NY: Springer.
- Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. *Journal of Experimental Psychology*, **63**, 289–293.
- Hoeffding, W. (1940). Maßstabinvariante Korrelationstheorie. *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin*, **5**, 181–233.
- Hohle, R. H. (1965). Inferred components of reaction times as a function of foreperiod duration. *Journal of Experimental Psychology*, **69**, 382–386.
- Houpt, J. W. and Townsend, J. T. (2012). Statistical measures for workload capacity analysis. *Journal of Mathematical Psychology*, **56**, 341–355.
- Joag-Dev, K. and Proschan, F. (1983). Negative association of random variables with applications. *Annals of Statistics*, **11**, 286–295.
- Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman & Hall.
- Joe, H. (2015). *Dependence Modeling with Copulas*. Boca Raton, FL: CRC Press, Taylor & Francis.

- Jones, M. and Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review*, **121**, 1–32.
- Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd edn. Hoboken, NJ: Wiley-Interscience.
- Kallenberg, O. (2005). *Probabilistic Symmetries and Invariance Principles*. New York, NY: Springer-Verlag.
- Kocher, S. C. and Proschan, F. (1991). Independence of time and cause of failure in the multiple dependent competing risks model. *Statistica Sinica*, **1**(1), 295–299.
- Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, **53**, 814–861.
- Leadbetter, R., Cambanis, S. and Pipiras, V. (2014). *A Basic Course in Measure and Probability*. New York, NY: Cambridge University Press.
- Levin, D. A., Peres, Y. and Wilmer, E. L. (2008). *Markov Chains and Mixing Times*. Providence, RI: American Mathematical Society.
- Li, H. and Li, X. (2013). *Stochastic Orders in Reliability and Risk*. Lecture Notes in Statistics, vol. 208. New York, NY: Springer-Verlag.
- Lindvall, T. (2002). *Lectures on the Coupling Method*. Reprint of Wiley 1992 edn. Mineola, NY: Dover.
- Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, **95**, 492–527.
- Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: a test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**, 883–914.
- Logan, G. D. (1995). The Weibull distribution, the power law, and the instance theory of automaticity. *Psychological Review*, **102**(4), 751–756.
- Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. New York, NY: Oxford University Press.
- Luce, R. D. (1997). Several unresolved conceptual problems of mathematical psychology. *Journal of Mathematical Psychology*, **41**(1), 79–87.
- Mai, J. F. and Scherer, M. (2012). *Simulating Copulas: Stochastic Models, Sampling Algorithms and Applications*. Series in Quantitative Finance, Vol. 4. London: Imperial College Press.
- Mari, D. D. and Kotz, S. (2004). *Correlation and Dependence*. Reprint of 2001 edn. London: Imperial College Press.
- Marley, A. A. J. and Colonius, H. (1992). The “horse race” random utility model for choice probabilities and reaction times, and its competing risks interpretation. *Journal of Mathematical Psychology*, **36**, 1–20.
- Marshall, A. W. and Olkin, I. (1967). A generalized bivariate exponential distribution. *Journal of Applied Probability*, **4**, 291–302.
- Marshall, A. W. and Olkin, I. (2007). *Life Distributions*. New York, NY: Springer-Verlag.
- Miller, J. and Ulrich, R. (2003). Simple reaction time and statistical facilitation: a parallel grains model. *Cognitive Psychology*, **46**, 101–151.
- Miller, J. O. (1982). Divided attention: evidence for coactivation with redundant signals. *Cognitive Psychology*, **14**, 247–279.
- Mitnik, P. A. and Baek, S. (2013). The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation. *Statistical Papers*, **54**, 177–192.

- Müller, A. and Rüsendorf, L. (2001). On the optimal stopping values induced by general dependence structures. *Journal of Applied Probability*, **38**(3), 672–684.
- Müller, A. and Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*. Vol. 389. Chichester: Wiley.
- Nair, N. U., Sankaran, P. G. and Balakrishnan, N. (2013). *Quantile-based Reliability Analysis*. New York, NY: Springer Verlag.
- Nelsen, R. B. (2006). *An Introduction to Copulas*. 2nd edn. Lecture Notes in Statistics, vol. 139. New York, NY: Springer-Verlag.
- Newell, A. and Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In pages 1–55 of: Anderson, J. R. (ed.), *Cognitive Skills and Their Acquisition*. Mahwah, NJ: Erlbaum, pp. 1–55.
- O'Hagen, A. and Forster, J. (2004). *Bayesian Inference*. 2nd edn. Kendall's Advanced Theory of Statistics Vol. 2B. London: Arnold.
- Palmer, E. M., Horowitz, T. S., Torralba, A. and Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, **37**(1), 58–71.
- Peterson, A. V. (1976). Bounds for a joint distribution function with fixed subdistribution functions: application to competing risks. *Proceedings of the National Academy of Sciences USA*, **73**, 11–13.
- Pfeifer, D. and Nešlehová, J. (2004). Modeling and generating dependent risk processes for IRM and DFA. *ASTIN Bulletin*, **34**(2), 333–360.
- Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability*. New York, NY: Cambridge University Press.
- Raab, D. H. (1962). Statistical facilitation of simple reaction time. *Transactions of the New York Academy of Sciences*, **24**, 574–590.
- Ross, S. M. (1983). *Stochastic Processes*. New York, NY: John Wiley & Sons.
- Ross, S. M. (1996). *Stochastic processes*. 2nd edn. New York, NY: John Wiley & Sons.
- Rüsendorf, L. (2013). *Mathematical Risk Analysis*. New York, NY: Springer-Verlag.
- Schmitz, V. (2004). Revealing the dependence structure between  $X_{(1)}$  and  $X_{(n)}$ . *Journal of Statistical Planning and Inference*, **123**, 41–47.
- Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments, & Computers*, **33**(4), 457–469.
- Schwarz, W. (2002). On the convolution of inverse Gaussian and exponential random variables. *Communications in Statistics, Theory and Methods*, **31**(12), 2113–2121.
- Schwarz, W. (2003). Stochastic cascade processes as a model of multi-stage concurrent information processing. *Acta Psychologica*, **113**, 231–261.
- Schwarz, W. and Miller, J. (2012). Response time models of delta plots with negative-going slopes. *Psychomimic Bulletin & Review*, **19**, 555–574.
- Schweickert, R., Fisher, D. L. and Sung, K. (2012). *Discovering Cognitive Architecture by Selectively Influencing Mental Processes*. Advanced Series on Mathematical Psychology, vol. 4. Singapore: World Scientific.
- Schweizer, B. and Sklar, A. (1974). Operations of distribution functions not derivable from operations on random variables. *Studia Mathematica*, **52**, 43–52.
- Severini, T. A. (2005). *Elements of Distribution Theory*. New York, NY: Cambridge University Press.
- Shaked, M. and Shantikumar, J. G. (2007). *Stochastic Orders*. New York, NY: Springer Science+Business Media.

- Shea, G. A. (1983). Hoeffding's lemma. In Kotz, S., Johnson, N. L. and Read, C. B. (eds), *Encyclopedia of Statistical Science*, vol. 3. New York, NY: Wiley, pp. 648–649.
- Smith, P. L. and Van Zandt, T. (2000). Time-dependent Poisson counter models of response latency in simple judgment. *British Journal of Mathematical and Statistical Psychology*, **53**, 293–315.
- Speckman, P. L., Rouder, J. N., Morey, R. D. and Pratte, M. S. (2008). Delta plots and coherent distribution ordering. *The American Statistician*, **62**(3), 262–266.
- Strassen, V. (1965). The existence of probability measures with given marginals. *Annals of Mathematical Statistics*, **36**, 423–439.
- Szekli, R. (1995). *Stochastic Ordering and Dependence in Applied Probability*. New York, NY: Springer Verlag.
- Thorisson, H. (2000). *Coupling, Stationarity, and Regeneration*. New York, NY: Springer Verlag.
- Townsend, J. T. (1976). Serial and within-stage independent parallel model equivalence on the minimum completion time. *Journal of Mathematical Psychology*, **14**, 219–238.
- Townsend, J. T. (1990). Truth and consequences of ordinal differences in statistical distributions: toward a theory of hierarchical inference. *Journal of Mathematical Psychology*, **108**(551–567).
- Townsend, J. T. and Ashby, F. G. (1983). *The Stochastic Modeling of Elementary Psychological processes*. Cambridge: Cambridge University Press.
- Townsend, J. T. and Colonius, H. (2005). Variability of the max and min statistic: a theory of the quantile spread as a function of sample size. *Psychometrika*, **70**(4), 759–772.
- Townsend, J. T. and Eidels, A. (2011). Workload capacity spaces: a unified methodology for response time measures of efficiency as workload is varied. *Psychonomic Bulletin and Review*, **18**, 659–681.
- Townsend, J. T. and Nozawa, G. (1995). Spatio-temporal properties of elementary perception: an investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, **39**(4), 321–359.
- Townsend, J. T. and Wenger, M. J. (2004). A theory of interactive parallel processing: new capacity measures and predictions for a response time inequality series. *Psychological Review*, **111**(4), 1003–1035.
- Tsiatis, A. A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences USA*, **72**, 20–22.
- Ulrich, R. and Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, **123**(1), 34–80.
- Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin and Review*, **7**, 424–465.
- Van Zandt, T. (2002). *Stevens' Handbook of Experimental Psychology*. Vol. 4. New York, NY: Wiley, pp. 461–516.
- von Mises, R. (1936). La distribution de la plus grande de  $n$  valeurs. *Rev. Math. Union Inter-balkanique*, **1**, 141–160. Reproduced in Selected Papers of Richard von Mises, *Am. Math. Soc. II* (1964), 271–294.

# 2 Probability, random variables, and selectivity

Ehtibar N. Dzhafarov and Janne Kujala

2.1	What is it about?	85
2.2	What is a random variable?	92
2.3	Jointly distributed random variables	95
2.4	Random variables in the narrow sense	100
2.5	Functions of random variables	102
2.6	Random variables as measurable functions	105
2.7	Unrelated random variables and coupling schemes	107
2.8	On sameness, equality, and equal distributions	109
2.9	Random outputs depending on inputs	110
2.10	Selectiveness in the dependence of outputs on inputs	112
2.11	Selective influences in a canonical form	114
2.12	Joint distribution criterion	117
2.13	Properties of selective influences and tests	124
2.14	Linear feasibility test	129
2.15	Distance tests	135
2.16	(Non)Invariance of tests with respect to transformations	139
2.17	Conditional determinism and conditional independence of outcomes	144
2.18	Related literature	147
2.19	Acknowledgments	148
	References	148

## 2.1 What is it about?

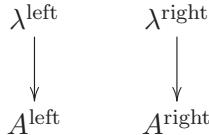
This chapter is about systems with several random outputs whose joint distribution depends on several inputs. More specifically, it is about selectiveness in the dependence of random outputs on inputs. That is, we are concerned with the question of which of the several outputs are influenced by which of the several inputs. A system can be anything: a person, animal, group of people, neural network, technical gadget, two entangled electrons running away from each other. Outputs are responses of the system or outcomes of measurements performed on it. Inputs are entities upon whose values the outputs of the system are conditioned.

Even if inputs are random variables in their own right, the outputs are being conditioned upon every specific stimulus. Inputs, therefore, are always deterministic (not random) entities insofar as their relationship to random outputs is concerned.

**Example 2.1** In a double-detection experiment, the stimulus presented in each trial may consist of two flashes, say, right one and left one, separated by some distance in visual field. Suppose that each flash can have one of two contrast levels, one zero and one (slightly) above zero. These contrasts play the role of two binary inputs, that we can call  $\lambda^{\text{left}}$  and  $\lambda^{\text{right}}$  (each one with values present/absent). The inputs are used in a completely crossed experimental design; that is, the stimulus in each trial is described by one of four combinations of the two inputs: ( $\lambda^{\text{left}} = \text{present}$ ,  $\lambda^{\text{right}} = \text{present}$ ), ( $\lambda^{\text{left}} = \text{present}$ ,  $\lambda^{\text{right}} = \text{absent}$ ), etc. In response to each such combination (called a treatment), the participant is asked to say whether the left flash was present (yes/no) and whether the right flash was present (yes/no). These are the two binary outputs, which we can denote  $A^{\text{left}}$  and  $A^{\text{right}}$  (each with two possible values, yes/no). The outputs are random variables. Theoretically, they are characterized by joint distributions tied to each of four treatments:

$(\lambda^{\text{left}} = i, \lambda^{\text{right}} = j)$		$A^{\text{right}} = \text{yes}$	$A^{\text{right}} = \text{no}$
$A^{\text{left}}$	$A^{\text{left}} = \text{yes}$	$p_{\text{yes},\text{yes}}$	$p_{\text{yes},\text{no}}$
	$A^{\text{left}} = \text{no}$	$p_{\text{no},\text{yes}}$	$p_{\text{no},\text{no}}$

where  $i, j$  each stand for “present” or “absent.” Suppose now that the experimenter hypothesizes that the response to the left stimulus depends only on the contrast of the left stimulus, and the response to the right stimulus depends only on the contrast of the right stimulus,



This hypothesis can be justified, for example, by one’s knowledge that the separation between the locations of the flashes is too large to allow for interference, and that subjectively, nothing seems to change in the appearance of the left stimulus as the right one is switched on and off, and vice versa. The meaning of this hypothesis is easy to understand if the two random outputs are known to be stochastically independent, which in this case means that, for every one of the four treatments,

$$p_{\text{yes},\text{yes}} = \Pr(A^{\text{left}} = \text{yes}, A^{\text{right}} = \text{yes}) = \Pr(A^{\text{left}} = \text{yes}) \Pr(A^{\text{right}} = \text{yes}).$$

In this case, the test of the selectiveness consists in finding out if the distribution of  $A^{\text{left}}$ , in this case defined by  $\Pr(A^{\text{left}} = \text{yes})$ , remains unchanged as one changes the value of  $\lambda^{\text{right}}$  while keeping  $\lambda^{\text{left}}$  fixed, and analogously for  $A^{\text{right}}$ . The experimenter, however, is likely to discover that stochastic independence in such an

experiment does not hold: for some, if not all of the four treatments,

$$p_{\text{yes},\text{yes}} \neq \Pr(A^{\text{left}} = \text{yes}) \Pr(A^{\text{right}} = \text{yes}).$$

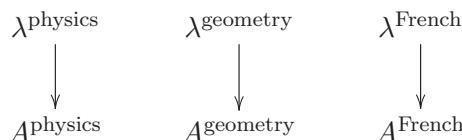
Now the conceptual clarity may be lost. Does the lack of stochastic independence invalidate the hypothesis that the outputs are selectively influenced by the corresponding inputs? Indeed, one might reason that it does, because if  $A^{\text{left}}$  and  $A^{\text{right}}$  are not independent, then  $A^{\text{left}}$  certainly “depends on”  $A^{\text{right}}$ , whence  $A^{\text{left}}$  should also depend on anything  $A^{\text{right}}$  depends on (and this includes  $\lambda^{\text{right}}$ ). However, one might also reason that the stochastic relationship between the two outputs can be ignored altogether. Cannot one declare that the hypothesis in question holds if one establishes that the marginal distributions (i.e.,  $\Pr(A^{\text{left}} = \text{yes})$  and  $\Pr(A^{\text{right}} = \text{yes})$ , taken separately) are invariant with respect to changes in the non-corresponding inputs (here,  $\lambda^{\text{right}}$  and  $\lambda^{\text{left}}$ , respectively)? We will see in this chapter that the stochastic relationship must not be ignored, but that the lack of stochastic independence does not by itself rule out selectiveness in the dependence of random outputs on inputs.  $\square$

It is easy to generate formally equivalent examples by trivial modifications. For instance, one can replace the two responses of a participant with activity levels of two neurons, determining whether each of them is above or below its background level. The two locations can be replaced with two stimulus features (say, orientation and spatial frequency of a grating pattern) that are hypothesized to selectively trigger the responses from the two neurons.

One can also easily modify any of such examples by increasing the number of inputs and outputs involved, or increasing the number of possible values per input or output. Thus, in the example with double-detection, one can think of several levels of contrast for each of the flashes. Or one can think of responses being a multilevel confidence rating instead of the binary yes/no.

Let us consider a few more examples, however, to appreciate the variety in the nature of inputs and outputs falling within the scope of our analysis.

**Example 2.2** Let a very large group of students have to take three exams, in physics, geometry, and French. Each student prepares for each of the exams, and the preparation times are classified as “short” or “long” by some criteria (which may be different for different exams). The three preparation times serve as the inputs in this example. We denote them by  $\lambda^{\text{physics}}$ ,  $\lambda^{\text{geometry}}$ , and  $\lambda^{\text{French}}$  (each with possible values short/long). The outputs are scores the students eventually receive:  $A^{\text{physics}}$ ,  $A^{\text{geometry}}$ , and  $A^{\text{French}}$  (say, from 0% to 100% each). The hypothesis to be tested is that preparation time for a given subject selectively affects the score in that subject,

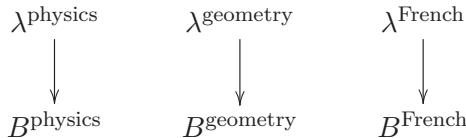


To see if this is the case, we subdivide the group of students into eight subgroups, corresponding to the eight combinations of the three preparation times,

$$(\lambda^{\text{physics}} = \text{short/long}, \lambda^{\text{geometry}} = \text{short/long}, \lambda^{\text{French}} = \text{short/long}).$$

Assuming each group is very large, we look at the joint distribution of scores within each of them. The conceptual difficulty here stems from the fact that, for any given treatment, test scores are typically positively correlated rather than stochastically independent.  $\square$

**Example 2.3** Let us modify the previous example by assigning to each student in each subject a binary grade, “high” or “low,” according to whether the student is, respectively, above or below the median score in this subject received by all student in the same preparation group. Thus, in the preparation group ( $\lambda^{\text{physics}} = \text{long}$ ,  $\lambda^{\text{geometry}} = \text{short}$ ,  $\lambda^{\text{French}} = \text{short}$ ), if the median score in physics is  $m$ , a student gets the grade “high” if her score is above  $m$  and “low” if it is not. This defines three outputs that we can call  $B^{\text{physics}}$ ,  $B^{\text{geometry}}$ ,  $B^{\text{French}}$ . The hypothesis represented by the diagram



is more subtle than in the previous example. It says that if one factors out the possible dependence of the median score in physics on all three preparation times (with no selectiveness assumed in this dependence), then whether a student’s physics score will or will not fall above the median may only depend on the preparation time for physics, and not on the preparation times for the two other subjects, and analogously for geometry and French. Because the grades assigned to students are binary, their theoretical distribution for each of the eight treatments is given by eight joint probabilities

$$\Pr(B^{\text{physics}} = \text{high/low}, B^{\text{geometry}} = \text{high/low}, B^{\text{French}} = \text{high/low}).$$

Again, the conceptual difficulty is in that this probability is not typically equal to  $1/8$  for all combinations of the high/low values, as it would have to be if the three random variables were independent. Indeed, the marginal (separately taken) probabilities here are, by the definition of median,

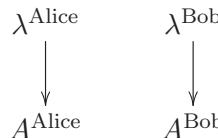
$$\Pr(B^{\text{physics}} = \text{high}) = \Pr(B^{\text{geometry}} = \text{high}) = \Pr(B^{\text{French}} = \text{high}) = \frac{1}{2}.$$

This example also shows why it is not wise to ignore the joint distributions and look at the marginal ones only. If we did this, none of the random outputs  $B^{\text{physics}}$ ,  $B^{\text{geometry}}$ ,  $B^{\text{French}}$  would be viewed as influenced by any of the inputs  $\lambda^{\text{physics}}$ ,  $\lambda^{\text{geometry}}$ ,  $\lambda^{\text{French}}$ . However, this view would clash with the fact that in different preparation groups the corresponding joint probabilities will typically be different.  $\square$

**Example 2.4** This example is not from behavioral sciences but from quantum physics. It is not as strange as it may appear to the reader. The fact is, the mathematical formalisms independently developed to study selective influences in psychology turn out to be identical to those developed in quantum physics to study the types of determinism involved in the behavior of so-called entangled particles. Two entangled particles can be thought of as being created as a single particle and then split into two mirror-images running away from each other. Particles possess a property called spin, something that can be measured along differently oriented spatial axes. In the case of so-called spin- $1/2$  particles, such as electrons, once an axis is chosen the spin can attain one of only two possible values, referred to as “spin-up” and “spin-down.” Suppose that two entangled electrons run away from each other towards two observers, Alice and Bob (a traditional way of referring to them in quantum physics), with previously synchronized clocks. At one and the same moment by these clocks Alice and Bob measure spins of their respective electrons along axes they previously chose. The nature of the entanglement is such that if the axes chosen by the two observers are precisely the same, then the spin values recorded will necessarily have opposite values: if Bob records spin-down, Alice will record spin-up. Suppose that Bob always chooses one of two axes, which we will denote  $\lambda^{\text{Bob}} = \beta_1$  and  $\lambda^{\text{Bob}} = \beta_2$ . We view  $\lambda^{\text{Bob}}$ , therefore, as one of the two inputs of the system. The other input is the axis chosen by Alice,  $\lambda^{\text{Alice}}$ . Let it also have two possible values,  $\lambda^{\text{Alice}} = \alpha_1$  and  $\lambda^{\text{Alice}} = \alpha_2$ . The outcome of Bob’s measurement is the first of two outputs of the system. We denote it by  $A^{\text{Bob}}$ , with the possible values “spin-up” and “spin-down.” The random output  $A^{\text{Alice}}$ , with the same two values, is defined analogously. The theoretical representation of this situation is given by the joint probabilities

$(\lambda^{\text{Alice}} = \alpha_i, \lambda^{\text{Bob}} = \beta_j)$	$A^{\text{Bob}} = \uparrow$	$A^{\text{Bob}} = \downarrow$
$A^{\text{Alice}} = \uparrow$	$p_{\uparrow\uparrow}$	$p_{\uparrow\downarrow}$
$A^{\text{Alice}} = \downarrow$	$p_{\downarrow\uparrow}$	$p_{\downarrow\downarrow}$

where  $i$  and  $j$  stand for 1 or 2 each. It is reasonable to hypothesize that



In other words, the spin recorded by Alice may depend on which axes she chose, but not on the axis chosen by Bob, and vice versa. However, the two outcomes here, for any of the four possible combinations of Alice’s and Bob’s axes, are not stochastically independent. This makes this situation formally identical to that described in the example with double-detection, except that in the entanglement paradigm the invariance of the marginal distributions is guaranteed:  $\Pr(A^{\text{Bob}} = \uparrow)$  is the same no matter which axis was chosen by Alice, and vice versa. In fact, it may very well be the case that these probabilities always remain equal to  $1/2$ , as in the second example with the three exams.  $\square$

Behavioral sciences abound with cases when selective influences are assumed with respect to random variables whose realizations are not directly observable. Rather, these random variables are hypothetical entities from which random variables with observable realizations can be derived theoretically. Thus, one may posit the existence of certain unobservable processes selectively influenced by certain experimental manipulations and manifested by their contribution to observable response times. For instance, one may assume the existence of processes called perception and response choice with respective durations  $A^{\text{percept}}$  and  $A^{\text{response}}$ , and assume that the observed response time is  $A^{\text{percept}} + A^{\text{response}}$ . One can further assume that stimulus characteristics selectively influence  $A^{\text{percept}}$  and instruction versions (such as speed emphasis versus accuracy emphasis) selectively influence  $A^{\text{response}}$ . The conceptual problem mentioned in the previous examples arises here if the two durations are not assumed to be stochastically independent.

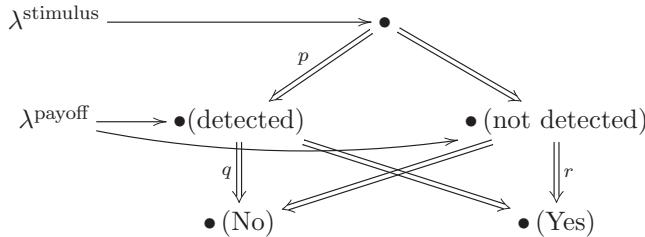
In analyzing “same–different” judgments for pairs of sounds, the observable entities are sounds  $\lambda^{\text{first}}$  and  $\lambda^{\text{second}}$ , each varying on several levels, and responses “same” or “different” for each pair of these sounds’ levels. It is typically postulated, however, that the response is a function (in the mathematical sense of the word) of two unobservable random variables,  $A^{\text{first}}$  and  $A^{\text{second}}$ , interpreted as internal representations of the two sounds, their images. For instance, a model may postulate that the response “same” is given if and only if the distance between  $A^{\text{first}}$  and  $A^{\text{second}}$  in some metric is less than some epsilon. It is reasonable to hypothesize then that

$$\begin{array}{ccc} \lambda^{\text{first}} & & \lambda^{\text{second}} \\ \downarrow & & \downarrow \\ A^{\text{first}} & & A^{\text{second}} \end{array}$$

Otherwise, why would one interpret  $A^{\text{first}}$  and  $A^{\text{second}}$  as “separate” respective images of  $\lambda^{\text{first}}$  and  $\lambda^{\text{second}}$ , rather than speaking of  $A = (A^{\text{first}}, A^{\text{second}})$  as one image of the compound stimulus  $(\lambda^{\text{first}}, \lambda^{\text{second}})$ ?

Stochastic independence of random outputs is, of course, a special case of stochastic relationship. It is clear from our opening examples that this is one case when the issue of defining and testing for selective influences is conceptually transparent. Deterministic outputs are a special case of random outputs; moreover, they can be formally considered stochastically independent. To see that a deterministic output  $a$  is influenced by an input  $\lambda$  but not input  $\lambda'$ , see if its value changes in response to changes in  $\lambda$  but remains constant if  $\lambda'$  changes with  $\lambda$  fixed. The only reason for mentioning here this obvious consideration is this: there is a wide class of theoretical models which deal with deterministic inputs and random outputs, but in which selectiveness of influences is formulated as a relationship between deterministic entities, namely, between the inputs and some parameters of the distributions of the random outputs. Parameters of distributions are, by definition, deterministic quantities. Such models require no special theory of selective influences.

**Example 2.5** In multinomial processing tree models we see simple examples of random variables related to inputs through parameters describing these variables' distributions. A prototypical example is provided by R. Duncan Luce's (1963) two-state low threshold model of detection,



The processing flow is shown by the double-line arrows: from the root of the tree to the root's children nodes, labeled "detected" and "not detected," and from each of those to their children nodes, labeled "Yes" and "No." The labels  $p$ ,  $q$ , and  $r$  are probabilities. The information shown in the processing tree is sufficient for computations, except for one additional constraint: the model stipulates that  $qr = 0$  (i.e., when one of the  $q$  and  $r$  is nonzero the other one must be zero). The inputs  $\lambda^{\text{stimulus}}$  and  $\lambda^{\text{payoff}}$  are shown on the margins. A single-line arrow pointing at a node of the tree indicates influence on the random variable whose possible values are the children of this node. Stimulus influences the distribution of the (unobservable) binary random variable called "detection state." It has two values occurring with probabilities  $p$  and  $1 - p$ . Payoff is any procedure involving feedback and designed to bias to various degrees the participants towards or against saying "Yes." This input influences the (observable) random variable "response." The point to note here is this: there is no reason to consider the joint distributions of detection state and response for different combinations of stimuli and payoffs; all we need is to declare which of the three parameters of the model,  $p$ ,  $q$ ,  $r$  depends on which input,

$$p = p(\lambda^{\text{stimulus}}), q = q(\lambda^{\text{payoff}}), r = r(\lambda^{\text{payoff}}).$$

This is simple and clear, even though the outputs "detection state" and "response" are not stochastically independent.  $\square$

As it turns out, it is impossible to answer the questions posed in this introductory section without getting "back to basics," to the foundational concepts of probability, random variable, joint distribution, and dependence of joint distributions on deterministic variables. It is simply impossible not to make mistakes and not to get hopelessly confused in dealing with the issues of selective influences if one is only guided by intuitive and informal understanding of these notions. This applies even if the random variables involved are as simple as binary responses. The first part of this chapter (Sections 2.2–2.9) is dedicated to these foundational issues. The reader should be especially attentive when we discuss the fact that not all random variables are jointly distributed, that a set of random variables can always be assigned a joint distribution in the absence of any constraints, but that this may not be possible

if the joint distribution should agree with the known distributions of some subsets of this set of random variables. Essentially, the issue of selective influences boils down to establishing whether this is or is not possible in specific cases. We deal with this issue beginning with Section 2.10, as well as the issue of the methods by which one can determine whether a particular pattern of selective influences holds. In Section 2.17 we show how the theory of selective influences applies to a classical problem of cognitive psychology, the problem of determining, based on the overall response time, whether certain hypothetical processes involved in the formation of the response are concurrent or serial. The chapter concludes with a brief guide to the relevant literature.

## 2.2 What is a random variable?

Let us begin with the notion of a *distribution* of a random variable. The formal definition of this notion is as follows: the distribution of a random variable  $A$  is a triple

$$\bar{A} = (S, \Sigma, p),$$

where

1.  $S$  is some nonempty set, called the *set of possible values* of  $A$ ;
2.  $\Sigma$  is a *sigma-algebra over  $S$* , which means a collection of subsets of  $S$ , each called an *event* or a *measurable set*, such that
  - (a)  $S \in \Sigma$ ,
  - (b) if  $S' \in \Sigma$ , then  $S - S' \in \Sigma$ ,
  - (c) if  $S_1, S_2, \dots \in \Sigma$  (a finite or countably infinite sequence), then

$$\bigcup_{i=1,2,\dots} S_i \in \Sigma;$$

3.  $p$  is some function (called *probability measure*) from  $\Sigma$  to  $[0, 1]$ , such that  $p(S')$  for  $S' \in \Sigma$  is interpreted as the probability with which a value of  $A$  falls in (belongs to) event  $S'$ ; it is assumed that
  - (a)  $p(S) = 1$ ,
  - (b) (*sigma-additivity*) if  $S_1, S_2, \dots \in \Sigma$  (a finite or countably infinite sequence), and if in this sequence  $S_i \cap S_j = \emptyset$  whenever  $i \neq j$  (i.e., the subsets in the sequence are pairwise disjoint), then

$$p\left(\bigcup_{i=1,2,\dots} S_i\right) = \sum_{i=1,2,\dots} p(S_i).$$

The following consequences of this definition are easily derived:

1.  $\emptyset \in \Sigma$  and  $p(\emptyset) = 0$ ;
2. if  $S_1, S_2, \dots \in \Sigma$ , then  $\bigcap_{i=1}^{\infty} S_i \in \Sigma$ ;

3. if  $S_1, S_2, \dots \in \Sigma$  and  $S_1 \subset S_2 \subset \dots$ , then

$$\lim_{i \rightarrow \infty} p(S_i) = p\left(\bigcup_{i=1}^{\infty} S_i\right);$$

4. if  $S_1, S_2, \dots \in \Sigma$  and  $S_1 \supset S_2 \supset \dots$ , then

$$\lim_{i \rightarrow \infty} p(S_i) = p\left(\bigcap_{i=1}^{\infty} S_i\right);$$

5. if  $S_1, S_2 \in \Sigma$  and  $S_1 \subset S_2$ , then  $S_2 - S_1 \in \Sigma$  and

$$p(S_1) + p(S_2 - S_1) = p(S_2);$$

6. if  $S_1, S_2 \in \Sigma$ , then

$$p(S_1 \cap S_2) \leq \min(p(S_1), p(S_2)) \leq \max(p(S_1), p(S_2)) \leq p(S_1 \cup S_2).$$

Most of these consequences are known as *elementary properties of probability*. It is customary to write  $p(S')$  for  $S' \in \Sigma$  as  $\Pr(A \in S')$ , if the distribution of  $A$  is known from the context.

We see that in order to know the distribution of a random variable  $A$  we have to know its set of possible values  $S$  and a set of specially chosen subsets of  $S$ , called events. In addition, we should have a procedure “measuring” each event, that is, assigning to it a probability with which a value of  $A$  (an element of  $S$ ) falls within this event (which is also described by saying that the event in question “occurs”).

**Example 2.6** For a finite  $S$ , the sigma-algebra is usually defined as the power set, i.e., the set of all subsets of  $S$ . For example, the distribution of the outcome  $A$  of a roll of a fair die can be represented by the distribution

$$\bar{A} = (S = \{1, 2, 3, 4, 5, 6\}, \Sigma = \mathcal{P}(S), p),$$

where  $\mathcal{P}(S)$  denotes the power set of  $S$  and  $p(\{s_1, \dots, s_k\}) = k/6$  for any set  $\{s_1, \dots, s_k\} \in \Sigma$  of  $k$  elements in  $S$ . Similarly, the sum of two dice can be represented by the distribution  $\bar{A} = (S = \{2, \dots, 12\}, \Sigma = \mathcal{P}(S), p)$ , where

$$p(\{s_1, \dots, s_k\}) = \sum_{i=1}^k p(\{s_i\})$$

and  $p(\{s\}) = \frac{1}{36}(6 - |7 - s|)$  gives the probability of each singleton (one-element subset)  $\{s\}$ .  $\square$

**Example 2.7** Let  $S$  be an interval of real numbers, finite or infinite, perhaps the entire set  $\mathbb{R}$  of real numbers. For continuous distributions defined on  $S$ , at the very least we want to be able to measure the probability of all intervals  $(a, b) \subset S$ . This requirement implies that our sigma-algebra  $\Sigma$  of events must contain all so-called *Borel subsets* of  $S$ . The Borel sets form the smallest sigma-algebra  $\Sigma$  over  $S$  that contains all open (or, equivalently, all closed) intervals. One can construct this sigma algebra by the following *recursive procedure*: (1) include in  $\Sigma$  all intervals

in  $S$ ; (2) add to this set of intervals all countable unions of these intervals and of their complements; (3) add to the previously obtained sets all countable unions of these sets of their complements; (4) and so on. Clearly, these steps are recursive applications of the operations (b) and (c) in the definition of a sigma-algebra. Every Borel set will be obtained at some step of this procedure.

The Borel sigma-algebra is sufficient for most purposes, but often the sigma-algebra is further enlarged by adding to all Borel sets all *null sets*. The latter are sets that can be covered by a countable sequence of intervals with arbitrarily small total length (see Section 2.4). The motivation for this extension is that anything that can be covered by an arbitrarily small length should have its measure equal to zero (and for this it should be measurable). The smallest sigma-algebra containing intervals and null sets is called the *Lebesgue sigma-algebra*.

A continuous distribution on the real line can be defined using a density function  $f(a)$ . The distribution is given by  $\bar{A} = (S, \Sigma, p)$ , where  $\Sigma$  is the Lebesgue sigma-algebra, and the probability measure of a set  $S_A \in \Sigma$  is given by the integral of the density function  $f$  over the set  $S_A$ ,

$$p(S_A) = \int_{S_A} f(a)da.$$

(To be well defined for all Lebesgue-measurable sets  $S_A$ , the integral here should be understood in the Lebesgue sense, but we need not go into this.)  $\square$

We see that the measurability of a subset of  $S$  is not a property of the subset itself, but of this subset taken in conjunction with a sigma-algebra  $\Sigma$ . Examples of *non-measurable subsets* of  $S$  therefore are easily constructed: choose  $\Sigma$  which is not the entire power set of  $S$ , and choose a subset of  $S$  which is not in  $\Sigma$ . For instance, if  $\Sigma = \{\emptyset, \{1\}, \{2, 3\}, \{1, 2, 3\}\}$  over the set  $S = \{1, 2, 3\}$ , then the single-element subset  $\{3\}$  is non-measurable. This means that if  $A$  is distributed as  $(S, \Sigma, p)$ , the probability  $p(\{3\})$  with which  $A$  falls in  $\{3\}$  (or, simply, equals 3) is undefined. This example may seem artificial, as nothing prevents one from complementing  $\Sigma$  with all other subsets of  $S = \{1, 2, 3\}$  (i.e., to assume that  $p$  is defined for all of them even if it is only known for some). If  $S$  is an interval of reals, however, then there are deeper reasons for not including in  $\Sigma$  all subsets of  $S$ .

It is obvious that different random variables can have one and the same distribution. For instance, Peter and Paul can flip a fair coin each, and describe the outcomes by one and the same distribution

$$\bar{A} = \left( S = \{0, 1\}, \Sigma = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}, p(\Sigma) = \left\{0, \frac{1}{2}, \frac{1}{2}, 1\right\} \right).$$

To distinguish one random variable from another, therefore, it is not sufficient to know its distribution. We should, in addition, have a *label* or *name* for the random variable: for instance, we can identify one random variable as  $\text{coin}_1$ , distributed as  $\bar{A}$ , and another as  $\text{coin}_2$ , also distributed as  $\bar{A}$ .

Generally speaking, a random variable  $A$  can be viewed as a quadruple  $(\iota_A, S, \Sigma, p)$ , where  $\iota_A$  is its unique name and  $\bar{A} = (S, \Sigma, p)$  is its distribution. We

do not need to be that formal, however, as the notation for a random variable,  $A$ , also serves as its name. (The reader familiar with the conventional definition of a random variable as a measurable function on a sample space should wait patiently until Sections 2.6 and 2.7. A function may serve as an identifying label too.)

**Remark 2.1** Alternatively, one can assume that the name of a random variable is always (implicitly) part of the elements of its domain  $S$ . For instance, the domain for one of the two coins mentioned above may be defined as  $S^1 = \{(0, \text{coin}_1), (1, \text{coin}_1)\}$  and for another as  $S^2 = \{(0, \text{coin}_2), (1, \text{coin}_2)\}$ . The sigma-algebras  $\Sigma^1$  and  $\Sigma^2$  then have to be (re)defined accordingly. If this approach is followed consistently, every random variable is uniquely determined by its distribution. We do not follow this route in this chapter.

## 2.3 Jointly distributed random variables

Let  $A$ ,  $B$ , and  $C$  be random variables with distributions  $\bar{A} = (S^1, \Sigma^1, p_1)$ ,  $\bar{B} = (S^2, \Sigma^2, p_2)$ , and  $\bar{C} = (S^3, \Sigma^3, p_3)$ .

**Remark 2.2** We will consistently use numerical superscripts to refer to the domain sets for random variables, to sigma-algebras over these sets, and later to random variables and inputs. Notation  $S^3$ , for example, always refers to a domain set of some random variable, *not* to the Cartesian product  $S \times S \times S$ . This should not cause any difficulties, as we use numerical exponents in this chapter only twice, and both times this is explicitly indicated.

Let  $S_A \in \Sigma^1$ ,  $S_B \in \Sigma^2$ , and  $S_C \in \Sigma^3$  be three events. We know that  $p_1(S_A)$  is interpreted as the probability with which a value of  $A$  falls in  $S_A$  (or, the probability that the event  $S_A$  “occurs”); and analogously for  $p_2(S_B)$  and  $p_3(S_C)$ . We also speak of events *occurring jointly*, or *co-occurring*, a concept whose substantive meaning we will discuss in Section 2.7. For now we will take it formally. In order to speak of  $S_A$ ,  $S_B$ ,  $S_C$  co-occurring and to ask of the probabilities with which they co-occur, we have to introduce a new random variable, denoted  $D_{ABC}$ . As any random variable, it is defined by some unique name (e.g., “ $D_{ABC}$ ”) and a distribution

$$\overline{D_{ABC}} = (S^{123}, \Sigma^{123}, p_{123}).$$

The set  $S^{123}$  of possible values of  $D_{ABC}$  is the Cartesian product  $S^1 \times S^2 \times S^3$  (the set of all ordered triples with the first components chosen from  $S^1$ , the second from  $S^2$ , the third from  $S^3$ ). The sigma-algebra  $\Sigma^{123}$  is denoted  $\Sigma^1 \otimes \Sigma^2 \otimes \Sigma^3$  and defined as the *smallest sigma-algebra* containing the Cartesian products  $S_A \times S_B \times S_C$  for all  $S_A \in \Sigma^1$ ,  $S_B \in \Sigma^2$  and  $S_C \in \Sigma^3$ . This means that  $\Sigma^{123} = \Sigma^1 \otimes \Sigma^2 \otimes \Sigma^3$  is a set of subsets of  $S^1 \times S^2 \times S^3$ , such that

1. it contains all the Cartesian products  $S_A \times S_B \times S_C$  just mentioned;
2. with every subset  $S'$  it contains, it also contains the complement  $S^{123} - S'$ ;

3. with every sequence of subsets  $S_1, S_2 \dots$  it contains, it also contains their union,  $\bigcup_{i=1,2,\dots} S_i$ ;
4. it is included in any other set of subsets of  $S^1 \times S^2 \times S^3$  satisfying 1–3 above.

The probability measure  $p_{123}$  is called a *joint probability measure*. It should satisfy the general requirements of a probability measure, namely:

$$p_{123}(S^1 \times S^2 \times S^3) = 1,$$

and

$$p_{123}\left(\bigcup_{i=1,2,\dots} S_i\right) = \sum_{i=1,2,\dots} p(S_i)$$

for any sequence of pairwise disjoint elements  $S_1, S_2, \dots$  of  $\Sigma^{123}$ . In addition,  $p_{123}$  should satisfy the following *1-marginal probability equations*: for any  $S_A \in \Sigma^1$ ,  $S_B \in \Sigma^2$  and  $S_C \in \Sigma^3$ ,

$$\begin{aligned} p_{123}(S_A \times S^2 \times S^3) &= p_1(S_A), \\ p_{123}(S^1 \times S_B \times S^3) &= p_2(S_B), \\ p_{123}(S^1 \times S^2 \times S_C) &= p_3(S_C). \end{aligned}$$

**Example 2.8** Let

$$S = \{0, 1\}, \quad \Sigma = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\},$$

and let the random variables  $A$ ,  $B$ , and  $C$  be distributed as

$$\bar{A} = (S, \Sigma, p_1), \quad \bar{B} = (S, \Sigma, p_2), \quad \bar{C} = (S, \Sigma, p_3),$$

where

$$p_1(\Sigma) = \{0, 1/2, 1/2, 1\}, \quad p_2(\Sigma) = \{0, 1/4, 3/4, 1\}, \quad p_3(\Sigma) = \{0, 1, 0, 1\}.$$

A joint distribution of  $A, B, C$  is defined on the product sigma-algebra  $\Sigma^{123} = \Sigma \otimes \Sigma \otimes \Sigma$ , which is the smallest sigma-algebra containing all Cartesian products  $S_A \times S_B \times S_C$  such that  $S_A, S_B, S_C \in \Sigma$ . As the Cartesian products include those of all singletons (one-element subsets)  $\{(a, b, c)\} = \{a\} \times \{b\} \times \{c\}$ , and all subsets of  $S \times S \times S$  can be formed by finite unions of these, the product sigma algebra  $\Sigma \otimes \Sigma \otimes \Sigma$  is the full power set of  $S \times S \times S$ . One possible joint distribution for  $A, B, C$  is given by

$$\overline{D_{ABC}} = (S^{123} = S \times S \times S, \Sigma^{123} = \Sigma \otimes \Sigma \otimes \Sigma, p_{123}),$$

where

$$p_{123}(S_{ABC}) = \sum_{(a,b,c) \in S_{ABC}} p_{123}(\{(a, b, c)\})$$

and  $p_{123}(\{(a, b, c)\})$  is given by the table

$a$	$b$	$c$	$p_{123}(\{(a, b, c)\})$	$a$	$b$	$c$	$p_{123}(\{(a, b, c)\})$
0	0	0	1/16	1	0	0	3/16
0	0	1	0	1	0	1	0
0	1	0	7/16	1	1	0	5/16
0	1	1	0	1	1	1	0

Let us verify that this distribution satisfies the 1-marginal probability equations and is thus a proper joint distribution of  $A, B, C$ :

$$\begin{aligned} p_{123}(\{0\} \times S \times S) &= 1/16 + 0 + 7/16 + 0 = 1/2 = p_1(\{0\}), \\ p_{123}(\{1\} \times S \times S) &= 3/16 + 0 + 5/16 + 0 = 1/2 = p_1(\{1\}), \\ p_{123}(S \times \{0\} \times S) &= 1/16 + 0 + 3/16 + 0 = 1/4 = p_2(\{0\}), \\ p_{123}(S \times \{1\} \times S) &= 7/16 + 0 + 5/16 + 0 = 3/4 = p_2(\{1\}), \\ p_{123}(S \times S \times \{0\}) &= 1/16 + 7/16 + 3/16 + 5/16 = 1 = p_3(\{0\}), \\ p_{123}(S \times S \times \{1\}) &= 0 + 0 + 0 + 0 = 0 = p_3(\{1\}). \end{aligned}$$

For each 1-marginal, it suffices to verify the probabilities of the points 0 and 1 as the probability values for singletons fully determine the discrete distributions.  $\square$

The random variable  $D_{ABC}$  is commonly called a *vector of the (jointly distributed) random variables A, B, and C*, and it is denoted  $(A, B, C)$ . We will use this vectorial notation in the sequel. One should keep in mind, however, that any such vector is a random variable in its own right. Furthermore, one should keep in mind that the distribution  $\overline{(A, B, C)}$ , called the *joint distribution* with respect to the individual random variables  $A, B, C$ , is not uniquely determined by these  $A, B, C$ . Specifically, although the set  $S^{123} = S^1 \times S^2 \times S^3$  and the sigma-algebra  $\Sigma^{123} = \Sigma^1 \otimes \Sigma^2 \otimes \Sigma^3$  are uniquely determined by the sets and sigma-algebras in the distributions  $\bar{A}, \bar{B}$ , and  $\bar{C}$ , there can generally be more than one joint probability measure  $p_{123}$ . The individual  $p_1, p_2$ , and  $p_3$  only serve as constraints, in the form of the 1-marginal probability equations above.

$A, B$ , and  $C$  in  $(A, B, C)$  are called *stochastically independent* if, for any  $S_A \in \Sigma^1$ ,  $S_B \in \Sigma^2$  and  $S_C \in \Sigma^3$ ,

$$p_{123}(S_A \times S_B \times S_C) = p_1(S_A)p_2(S_B)p_3(S_C).$$

This joint probability measure always satisfies the 1-marginal probability equations.

**Example 2.9** Let  $A$  and  $B$  be standard normally distributed random variables. A bivariate normal joint distribution  $(A, B)(\rho)$  can be defined with the density function

$$f_{12}(a, b; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{a^2 + b^2 - 2\rho ab}{2(1-\rho^2)}\right),$$

where  $-1 < \rho < 1$  denotes the correlation coefficient. The sigma algebra  $\Sigma_{12} = \Sigma_1 \otimes \Sigma_2$  of the joint distribution is the product of two Lebesgue sigma-algebras

(called a Lebesgue sigma-algebra itself). The 1-marginal probability equations can be verified by checking that integrating out either  $a$  or  $b$  yields the standard normal density function with respect to the remaining variable. The probability measure for  $C = (A, B)(\rho)$  is obtained as

$$p_{12}(S_C) = \int_{(a,b) \in S_C} f_{12}(a, b; \rho) d(a, b).$$

Do  $C = (A, B)(\rho_1)$  and  $D = (A, B)(\rho_2)$  with  $\rho_1 \neq \rho_2$  exclude each other? Not in the sense that defining one of them makes the other meaningless. They both can be defined as variables of interest. However,  $C$  and  $D$  cannot be jointly distributed.  $\square$

The reverse relationship between joint and marginal distributions is more straightforward: the distribution  $\overline{(A, B, C)}$  uniquely determines the distributions and identities of  $A, B, C$ , called the *1-marginal random variables* with respect to  $(A, B, C)$ , as well as the distributions and identities of  $(A, B)$ ,  $(B, C)$ , and  $(A, C)$ , called the *2-marginal random variables* with respect to  $(A, B, C)$ . Thus, in the distribution  $\overline{A}$  the set  $S^1$  is the projection  $\text{Proj}_1$  of the set  $S^{123} = S^1 \times S^2 \times S^3$ , defined by

$$\text{Proj}_1(a, b, c) = a.$$

The sigma-algebra  $\Sigma^1$  consists of the projections  $\text{Proj}_1$  of the elements of the sigma-algebra  $\Sigma^{123} = \Sigma^1 \otimes \Sigma^2 \otimes \Sigma^3$  having the form  $S_A \times S^2 \times S^3$ . The probability measure  $p_1$  is determined by the 1-marginal probability equations. The 2-marginal distributions  $\overline{(A, B)}$ ,  $\overline{(B, C)}$ , and  $\overline{(A, C)}$  are found analogously. For example, if one defines function  $\text{Proj}_{23}$  by

$$\text{Proj}_{23}(a, b, c) = (b, c),$$

we have

$$\overline{(B, C)} = (S^{23}, \Sigma^{23}, p_{23}),$$

where

$$S^{23} = \text{Proj}_{23}(S^1 \times S^2 \times S^3),$$

$\Sigma^{23}$  consists of the sets of the form

$$\text{Proj}_{23}(S^1 \times S_{BC}), \quad S_{BC} \in \Sigma^2 \otimes \Sigma^3,$$

and

$$p_{23}(S_{BC}) = p_{123}(S^1 \times S_{BC}).$$

The last equality is one of the three *2-marginal probability equations* (the remaining two being for  $p_{12}$  and  $p_{13}$ ).

One can check that

$$S^{23} = S^2 \times S^3,$$

and

$$\Sigma^{23} = \Sigma^2 \otimes \Sigma^3,$$

which is the smallest sigma-algebra containing the Cartesian products  $S_B \times S_C$  for all  $S_B \in \Sigma^2$  and  $S_C \in \Sigma^3$ . In other words, the set  $S^{23}$  and the sigma-algebra  $\Sigma^{23}$  over it in the 2-marginal distribution are precisely the same as if they were formed for a joint distribution  $(\bar{B}, \bar{C})$  with respect to the 1-marginal distributions  $\bar{B}$  and  $\bar{C}$ . Moreover, the 2-marginal probability  $p_{23}$  is a joint probability satisfying the 1-marginal probability equations

$$p_{23}(S_B \times S^3) = p_2(S_B),$$

$$p_{23}(S^2 \times S_C) = p_3(S_C).$$

**Example 2.10** Continuing from Example 2.8, we can derive the following 2-marginals (and 1-marginals shown at the sides of the 2-marginals):

$p_{12}(\{(a, b)\})$		$b = 0$	$b = 1$	$p_{12}(\{(b, c)\})$		$c = 0$	$c = 1$
$a = 0$	1/16	7/16	1/2	$b = 0$	1/4	0	1/4
	3/16	5/16	1/2		3/4	0	3/4
	1/4	3/4			1	0	
$p_{12}(\{(a, c)\})$		$c = 0$	$c = 1$				
$a = 0$	1/2	0	1/2				
	1/2	0	1/2				
	1	0					

□

It should be clear now how one should generalize the notion of a joint distribution to an arbitrary number  $n$  of random variables,  $A^1, \dots, A^n$ , and how to define  $k$ -marginal distributions for  $k = 1, \dots, n$  ( $n$ -marginal distributions being permutations of the joint one, including itself).

**Remark 2.3** For an infinite set of random variables (countable or not) the definition of a joint distribution is less obvious. We will not deal with this notion in this chapter except for mentioning it occasionally, for completeness sake. With little elaboration, let  $(A^k : k \in K)$  be an indexed family of random variables (with an arbitrary indexing set  $K$ ), each distributed as  $(S^k, \Sigma^k, p_k)$ . We say that the random variables in  $(A^k : k \in K)$  are jointly distributed if  $A = (A^k : k \in K)$  is a random variable with the distribution

$$\bar{A} = \left( \prod_{k \in K} S^k, \bigotimes_{k \in K} \Sigma^k, p \right),$$

where

1.  $\prod_{k \in K} S^k$  is the Cartesian product of the sets  $S^k$  (its elements are functions choosing for each element of  $K$  an element of  $S^k$ );
2.  $\bigotimes_{k \in K} \Sigma^k$  is the smallest sigma-algebra containing sets of the form  $S' \times \prod_{k \in K - \{k_0\}} S^k$ , for all  $k_0 \in K$  and  $S' \in \Sigma^{k_0}$ ;

3.  $p$  is a probability measure on  $\bigotimes_{k \in K} \Sigma^k$  such that  $p(S' \times \prod_{k \in K - \{k_0\}} S^k) = p_{k_0}(S')$ , for all  $k_0 \in K$  and  $S' \in \Sigma^{k_0}$ .

The random variables  $A^k$  in  $A = (A^k : k \in K)$  are said to be stochastically independent if any finite subset of them consists of stochastically independent elements.

**Remark 2.4** Marginal random variables sometimes have to be defined hierarchically. Consider, for example,  $A' = (A, B)$  and  $B' = (C, D)$ . Then  $C' = (A', B')$  has the 1-marginal distributions  $\bar{A}' = \overline{(A, B)}$  and  $\bar{B}' = \overline{(C, D)}$ , and  $A'$ , in turn, has 1-marginal distributions  $\bar{A}$  and  $\bar{B}$ . It may sometimes be convenient to speak of all of  $(A, B)$ ,  $(C, D)$ ,  $A$ ,  $B$ ,  $C$ ,  $D$  as marginal random variables with respect to a random variable  $C' = ((A, B), (C, D))$ . Note that  $((A, B), (C, D))$ ,  $((A, B, C), D)$ ,  $(A, (B, (C, D)))$ , etc., are all distributed as  $(A, B, C, D)$ , because the Cartesian product  $S^1 \times S^2 \times S^3 \times S^4$  and the product sigma algebra  $\Sigma^1 \otimes \Sigma^2 \otimes \Sigma^3 \otimes \Sigma^4$  are associative. The random variables  $((A, B), (C, D))$ ,  $((A, B, C), D)$ ,  $(A, (B, (C, D)))$ , etc., differ in their labeling only. (In the infinite case (Remark 2.3) the formal definition is rather straightforward, but it involves potentially more than a finite number of hierarchical steps. We will assume that the notion is clear and a formal definition may be skipped.)

## 2.4 Random variables in the narrow sense

The concept of a random variable used in this chapter is very general, with no restrictions imposed on the sets and sigma-algebras in their distributions. Sometimes such random variables are referred to as *random entities*, *random elements*, or *random variables in the broad sense*, to distinguish them from *random variables in the narrow sense*. The latter are most important in applications. In particular, all our examples involve random variables in the narrow sense. They can be defined as follows. Let  $A$  be distributed as  $\bar{A} = (S, \Sigma, p)$ .

- (i) If  $S$  is countable,  $\Sigma$  is the power set of  $S$  (the set of all its subsets), then  $A$  is a random variable in the narrow sense;
- (ii) if  $S$  is an interval of real numbers,  $\Sigma$  is the *Lebesgue sigma-algebra* over  $S$  (as defined in Example 2.7), then  $A$  is a random variable in the narrow sense;
- (iii) if  $A_1, \dots, A_n$  are random variables in the narrow sense, then any jointly distributed vector  $(A_1, \dots, A_n)$  is a random variable (also referred to as a *random vector*) in the narrow sense.

Random variables satisfying (i) are called *discrete*. The distribution of such a random variable is uniquely determined by the probabilities assigned to its singleton (one-element) subsets. These probabilities can also be viewed as assigned to the elements themselves, in which case they form a *probability mass function*. An example of a discrete random variable is given in Example 2.6. However,  $S$  may also be countably infinite.

**Example 2.11** Let  $S$  be the set of positive integers  $\{1, 2, \dots, n, \dots\}$ , and let  $p(\{n\}) = \alpha^{n-1}(1 - \alpha)$ , where  $\alpha$  is a constant in  $[0, 1]$ . This defines a discrete random variable interpreted as the number of independent trials  $n$  with binary outcomes (success/failure) until the first failure. It is customary to replace (or even confuse)  $p(\{n\})$  with the probability mass function function  $p^*(n) = p(\{n\})$ .  $\square$

Random variables satisfying (ii) are called *continuous* (see Example 2.7). Any such variable can be viewed as having  $S$  extended to the entire set of reals, and its distribution is uniquely determined by the *distribution function*

$$F(x) = p((-\infty, x]),$$

for every real  $x$ . The function  $F(x)$  has the following properties:

1. it is non-decreasing;
2. as  $x \rightarrow -\infty$ ,  $F(x) \rightarrow 0$ ;
3. as  $x \rightarrow \infty$ ,  $F(x) \rightarrow 1$ ;
4. for any real  $x_0$ , as  $x \rightarrow x_0+$ ,  $F(x) \rightarrow F(x_0)$  (right-continuity);
5. for any real  $x_0$ , as  $x \rightarrow x_0-$ ,  $F(x)$  tends to a limit.

$F(x)$  generally is not left-continuous: as  $x \rightarrow x_0-$ , the limit of  $F(x)$  need not coincide with  $F(x_0)$ , the function may instead “jump” from the value of  $\lim_{x \rightarrow x_0-} F(x)$  to  $F(x_0)$ . The difference  $F(x_0) - \lim_{x \rightarrow x_0-} F(x)$  equals  $p(\{x_0\})$ , so the jumps occur if and only if  $p(\{x_0\}) > 0$ . A distribution function cannot have more than a countable set of jump points. For any two reals  $x_1 \leq x_2$ ,

$$F(x_2) - F(x_1) = p((x_1, x_2]).$$

**Example 2.12** A discrete random variable can always be redefined as a continuous one. Thus, the variable in the previous example can be redefined into a random variable  $X$  whose distribution is given by

$$F(x) = \begin{cases} 0 & \text{for } x < 1 \\ 1 - \alpha^n & \text{for } \lfloor x \rfloor = n \geq 1, \end{cases}$$

where  $\lfloor x \rfloor$  is the floor function (the largest integer not exceeding  $x$ ).  $\square$

The Lebesgue sigma-algebra over the reals, as defined in Example 2.7, is the smallest sigma-algebra including all intervals and all *null sets*. A subset  $S'$  of reals is a null set if, for any  $\varepsilon > 0$ , however small,  $S'$  is contained within a union of open intervals  $S_1, S_2, \dots$  whose overall length is less than  $\varepsilon$ . An empty set is, obviously a null set, and so is a single point, and a countable set of points.

**Remark 2.5** Let us prove that a countable set of points is a null set, to better understand the definition. Enumerate this set as  $x_1, x_2, \dots$ , choose an  $\varepsilon > 0$ , and enclose each  $x_i$  into interval  $\left]x - \frac{\varepsilon}{2^{i+1}}, x + \frac{\varepsilon}{2^{i+1}}\right[$ . The length of this interval is  $\frac{\varepsilon}{2^i}$ , whence the overall length of the system of such intervals cannot exceed

$$\sum_{i=1,2,\dots} \frac{\varepsilon}{2^i} \leq \varepsilon.$$

We conclude that a countable subset of  $S$  is a null set. There are uncountable null sets.

As should be clear from our discussion of jumps and Example 2.12, a null set may have a nonzero probability. If this does not happen, i.e., if  $F(x)$  has no jumps, the distribution of the random variable is called *absolutely continuous*.

Finally, the combination rule (iii) allows one to form vectors of discrete, continuous, and mixed jointly distributed random variables using the construction discussed in Section 2.3.

## 2.5 Functions of random variables

Let  $A$  be a random variable with distribution  $\bar{A} = (S^1, \Sigma^1, p_1)$ , let  $S^2$  be some set, and let  $f : S^1 \rightarrow S^2$  be some function. Consider some sigma-algebra  $\Sigma^2$  of events over  $S^2$ . For every  $S_B \in \Sigma^2$  one can determine the subset of all elements of  $S^1$  that are mapped by  $f$  into  $S_B$ ,

$$f^{-1}(S_B) = \{a \in S^1 : f(a) \in S_B\}.$$

This subset,  $f^{-1}(S_B)$ , does not have to be an event in  $\Sigma^1$ . If it is, for every  $S_B \in \Sigma^2$ , then  $f$  is said to be a *measurable function* (or  $\Sigma^1 \rightarrow \Sigma^2$ -measurable function, to be specific). Measurability of a function, therefore, is not a property of the function itself, but of the function taken in conjunction with two sigma-algebras. In particular, given  $S^1$  and  $\Sigma^1$ , any *one-to-one* function  $f : S^1 \rightarrow S^2 = f(S^1)$  will be measurable if we agree to define  $\Sigma^2 = f(\Sigma^1)$ , the set of all  $f$ -images of the elements of  $\Sigma^1$ ; it is easy to prove that  $f(\Sigma^1)$  is a sigma-algebra over  $f(S^1)$ , for any one-to-one  $f$ .

**Example 2.13** Let  $S^1 = S^2 = \{1, 2, 3\}$ ,

$$\Sigma^1 = \{\emptyset, \{1\}, \{2, 3\}, \{1, 2, 3\}\},$$

and

$$\Sigma^2 = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, \{1, 2, 3\}\}.$$

Then the function  $f : \Sigma^1 \rightarrow \Sigma^2$  defined by  $f(a) = a$  is not measurable, because  $\{2\} \in \Sigma^2$  but  $f^{-1}(\{2\}) = \{2\} \notin \Sigma^1$ . However, one can easily verify that  $f(a) = \min(a, 2)$  is a  $\Sigma^1 \rightarrow \Sigma^2$ -measurable function.  $\square$

Of course, with finite  $S^1, S^2$ , one can always define the sigma-algebras as full power sets and then all functions between these sets will be measurable.

Why is the notion of a measurable function important? Because measurable functions can be used to obtain new random variables from existing ones. Given a random variable  $A$  and a  $\Sigma^1 \rightarrow \Sigma^2$ -measurable function  $f : S^1 \rightarrow S^2$ , one can

define a random variable  $B = f(A)$  distributed as  $\bar{B} = (S^2, \Sigma^2, p_2)$  by putting, for any  $S' \in \Sigma^2$ ,

$$p_2(S') = p_1(f^{-1}(S')).$$

In other words, the probability with which the new variable  $B$  falls in an event belonging to  $\Sigma^2$  is defined as the probability with which  $A$  falls in the  $f$ -preimage of this event in  $\Sigma^1$  (which probability is well-defined because  $f$  is measurable). Of course, the notation  $B = f(A)$  serves as a unique identification of  $B$  once we agree that  $A$  is uniquely identified.

**Example 2.14** Let  $S^1$  and  $S^2$  be two intervals of reals, and let  $\Sigma^1$  and  $\Sigma^2$  be the Borel sigma-algebras over them (see Example 2.7). A function  $f : S^1 \rightarrow S^2$  which is  $\Sigma^1 \rightarrow \Sigma^2$ -measurable is called a *Borel-measurable function*. If in this definition  $\Sigma^1$  is the Lebesgue sigma-algebra over  $S^1$  while  $\Sigma^2$  continues to be the Borel sigma-algebra over  $S^2$  (note the asymmetry), then  $f$  is a *Lebesgue-measurable function*. It is sufficient to require in these two definitions that for any interval  $(a, b) \subset S^2$ , its preimage  $f^{-1}((a, b))$  be a Borel-measurable (respectively, Lebesgue-measurable) subset of  $S^1$ . It is easy to prove that if  $f$  is monotone or continuous, then it is Borel-measurable (hence also Lebesgue-measurable).

Now let  $A$  be a random variable with distribution  $\bar{A} = (\mathbb{R}, \Sigma^1, p)$ , where  $\Sigma^1$  is the Lebesgue sigma-algebra over  $\mathbb{R}$ . The function  $F(x) = p((-\infty, x])$  is called the *distribution function* for  $A$ . It is monotonically non-decreasing and maps into  $S^2 = [0, 1]$ . If we define  $\Sigma^2$  to be the Borel sigma-algebra over  $[0, 1]$ , then  $F$  (being non-decreasing) is Lebesgue-measurable. If we apply  $F$  to  $A$ , the resulting random variable  $B = F(A)$  is distributed on  $[0, 1]$ . If, furthermore,  $F$  is a continuous function, then the distribution of  $B = F(A)$  on  $[0, 1]$  is uniform. That is, its distribution is  $\bar{B} = ([0, 1], \Sigma^2, q)$ , where  $q((a, b)) = b - a$  for any  $(a, b) \subset [0, 1]$ .  $\square$

Let  $A$  be distributed as  $\bar{A} = (S^1, \Sigma^1, p_1)$ , and let  $B = f(A)$  and  $C = g(A)$  be two random variables with distributions  $\bar{B} = (S^2, \Sigma^2, p_2)$  and  $\bar{C} = (S^3, \Sigma^3, p_3)$ . This implies that both  $f$  and  $g$  are measurable functions in the sense of, respectively,  $\Sigma^1 \rightarrow \Sigma^2$  and  $\Sigma^1 \rightarrow \Sigma^3$ . For every  $S_B \in \Sigma^2$  and every  $S_C \in \Sigma^3$  we have

$$p_2(S_B) = p_1(f^{-1}(S_B)), \text{ and } p_3(S_C) = p_1(g^{-1}(S_C)).$$

A value  $b$  of  $B$  falls in  $S_B$  if and only if  $b = f(a)$  for some  $a \in f^{-1}(S_B)$ . A value  $c$  of  $C$  falls in  $S_C$  if and only if  $c = g(a)$  for some  $a \in g^{-1}(S_C)$ . This suggests a way of defining the notion of a *joint occurrence* of these events,  $S_B$  and  $S_C$ : they occur jointly if and only if  $a$  in the previous two sentences is one and the same. In other words, a value  $b$  of  $B$  falls in  $S_B$  and, jointly, a value  $c$  of  $C$  falls in  $S_C$  if and only if, for some  $a \in f^{-1}(S_B) \cap g^{-1}(S_C)$ ,  $b = f(a)$  and  $c = g(a)$ . Because  $f^{-1}(S_B) \cap g^{-1}(S_C)$  is  $\Sigma^1$ -measurable (belongs to  $\Sigma^1$ ), the probability

$$p_{23}(S_B \times S_C) = p_1(f^{-1}(S_B) \cap g^{-1}(S_C))$$

is well defined, and we can take it as the joint probability of  $S_B$  and  $S_C$ .

We now can construct the joint distribution of  $(B, C)$ ,

$$\overline{(B, C)} = (S^2 \times S^3, \Sigma^2 \otimes \Sigma^3, p_{23}),$$

where the set and the sigma-algebra are defined as required by the general notion of a joint distribution (Section 2.3). The joint probability measure  $p_{23}$  defined above for  $S_B \times S_C$ -type sets is extended to all other members of  $\Sigma^2 \otimes \Sigma^3$  by using the basic properties of a probability measure (Section 2.2). Equivalently, the joint probability measure  $p_{23}$  can be defined by

$$p_{23}(S') = p((f, g)^{-1}(S')),$$

for any  $S' \in \Sigma^2 \otimes \Sigma^3$ . The notation  $(f, g)^{-1}(S')$  designates the set  $S_A$  of all  $a \in S$ , such that  $(f(a), g(a)) \in S'$ . It can be shown that  $S_A \in \Sigma^1$ , that is,  $(f, g)$  is a measurable function.

It can easily be checked that  $p_{23}$  satisfies the 1-marginal probability equations,

$$p_{23}(S_B \times S^3) = p_1(f^{-1}(S_B) \cap g^{-1}(S^3)) = p_1(f^{-1}(S_B)) = p_2(S_B),$$

$$p_{23}(S^2 \times S_C) = p_1(f^{-1}(S^2) \cap g^{-1}(S_C)) = p_1(g^{-1}(S_C)) = p_3(S_C),$$

where we used the fact that

$$g^{-1}(S^3) = f^{-1}(S^2) = S^1.$$

We see that if two random variables are formed as functions of another random variable, their joint distribution is uniquely determined.

**Example 2.15** A simple but instructive example is the joint distribution of a random variable  $A$  and itself. Let  $A$  be distributed as  $(S, \Sigma, p)$ .  $(A, A)$  is a random variable, both components of which are functions of one and the same random variable,  $A = \text{id}(A)$ , where  $\text{id}$  is the identity function defined by  $\text{id}(a) = a$ . Let the distribution of  $(A, A)$  be  $(S \times S, \Sigma \otimes \Sigma, p_2)$ . By the general theory, for any  $S' \in \Sigma$  we have  $S' \times S' \in \Sigma \otimes \Sigma$  and

$$p_2(S' \times S') = p(\text{id}^{-1}(S') \cap \text{id}^{-1}(S')) = p(S'),$$

as it should be. It is not always true, however, that the probability measure  $p_2$  of the set of pairs

$$\text{diag}_{S \times S} = \{(a, a) : a \in S\}$$

equals 1, because this set is not necessarily an event in  $\Sigma \otimes \Sigma$ . As an example,  $\{(1, 1), (2, 2), (3, 3), (4, 4)\}$  is not such an event if  $\Sigma = \{\emptyset, \{1, 2\}, \{3, 4\}, \{1, 2, 3, 4\}\}$ . If, however,  $\text{diag}_{S \times S} \in \Sigma \otimes \Sigma$ , then

$$p_2(\text{diag}_{S \times S}) = p((\text{id}, \text{id})^{-1}(\text{diag}_{S \times S})) = p(S) = 1. \quad \square$$

The generalization to several functions of a random variable  $A$  is trivial. Thus, we can form a joint distribution not just of  $B, C$  but of  $A, B, C$  (for symmetry, we

can consider  $A$  the identity function of  $A$ ). In particular, the joint probability of  $S_B \in \Sigma^2$ ,  $S_C \in \Sigma^3$ , and  $S_A \in \Sigma^1$  is defined here as

$$p_{23}(S_A \times S_B \times S_C) = p_1(S_A \cap f^{-1}(S_B) \cap g^{-1}(S_C)).$$

One of the important classes of measurable functions of random variables are projections. We have already dealt with them in Section 2.3, when discussing marginal distributions. More generally, a vector of jointly distributed random variables  $A^1, A^2, \dots, A^n$  is a random variable with a distribution

$$(S^1 \times \dots \times S^n, \Sigma^1 \otimes \dots \otimes \Sigma^n, p_{1\dots n}),$$

where the notation should be clear from the foregoing. A projection function  $\text{Proj}_{i_1\dots i_k}$ , where  $k \leq n$  and  $i_1, \dots, i_k$  is a set of  $k$  distinct numbers chosen from  $(1, \dots, n)$ , is defined by

$$\text{Proj}_{i_1\dots i_k}(a_1, \dots, a_n) = (a_{i_1}, \dots, a_{i_k}).$$

Without loss of generality, let  $(i_1, \dots, i_k) = (1, \dots, k)$ ; if this is not the case, one can always make it so by renumbering the original set of  $n$  random variables. The function  $\text{Proj}_{1\dots k}$  creates a  $k$ -marginal random variable

$$\text{Proj}_{1\dots k}(A^1, \dots, A^n) = (A^1, \dots, A^k),$$

with the  $k$ -marginal distribution

$$(S^1 \times \dots \times S^k, \Sigma^1 \otimes \dots \otimes \Sigma^k, p_{1\dots k}).$$

where, for any measurable event  $S'$  in  $\Sigma^1 \otimes \dots \otimes \Sigma^k$ ,

$$p_{1\dots k}(S') = p_{1\dots n}(S' \times S^{k+1} \times \dots \times S^n).$$

## 2.6 Random variables as measurable functions

We have seen that if  $A^1, \dots, A^n$  are all functions of one and the same random variable  $R$ , then they posses a joint distribution. To recapitulate, if

$$A^1 = f_1(R), \dots, A^n = f_n(R),$$

$$\bar{R} = (S^*, \Sigma^*, p_*),$$

and

$$\overline{A^i} = (S^i, \Sigma^i, p_i), \quad i = 1, \dots, n,$$

then

$$\overline{(A^1, \dots, A^n)} = (S^1 \times \dots \times S^n, \Sigma^1 \otimes \dots \otimes \Sigma^n, p_{1\dots n}),$$

where

$$p_{1\dots n}(S') = p_*((f_1, \dots, f_n)^{-1}(S')),$$

for any  $S' \in \Sigma^1 \otimes \cdots \otimes \Sigma^n$ . In particular,

$$p_{1\dots n}(S_1 \times \cdots \times S_n) = p_* \left( \bigcap f_i^{-1}(S_i) \right),$$

for all

$$S_1 \in \Sigma^1, \dots, S_n \in \Sigma^n.$$

It is easy to see that the reverse of this statement is also true: if  $A^1, \dots, A^n$  have a joint distribution, they can be presented as functions of one and the same random variable. Indeed, denoting the random variable  $(A^1, \dots, A^n)$  by  $R$ , we have

$$A^1 = f_1(R), \dots, A^n = f_n(R),$$

where

$$f_i \equiv \text{Proj}_i.$$

These two simple observations constitute a proof of an important theorem.

**Theorem 2.1** *A vector  $(A^1, \dots, A^n)$  of random variables possesses a joint distribution if and only if there is a random variable  $R$  and a vector of functions  $\{f_1, \dots, f_n\}$ , such that  $A^1 = f_1(R), \dots, A^n = f_n(R)$ .*

Note that we need not specify here that the functions are measurable, because both  $A^i$  and  $R$  in  $A^i = f_i(R)$  are random variables (implying that  $f_i$  is measurable).

Although we do not deal in this chapter with infinite sets of jointly distributed random variables, it must be mentioned that Theorem 2.1 has the following *generalized formulation* (see Remark 2.3).

**Theorem 2.2** *A family  $(A^k : k \in K)$  of random variables possesses a joint distribution if and only if there is a random variable  $R$  and a family of functions  $\{f_k : k \in K\}$  such that  $A^k = f_k(R)$  for all  $k \in K$ .*

In probability textbooks, consideration is almost always confined to random variables that are jointly distributed. This enables what we may call the *traditional conceptualization of random variables*. It consists of choosing some distribution

$$\bar{R} = (S^*, \Sigma^*, p_*),$$

calling it a *sample (probability) space*, and identifying any random variable  $A$  as a  $(\Sigma^* \rightarrow \Sigma^1)$ -measurable function  $f : S^* \rightarrow S^1$ . The set and sigma-algebra pair  $(S^1, \Sigma^1)$  being chosen, the probability measure  $p_1$  satisfying, for every  $S' \in \Sigma^1$ ,

$$p_1(S') = p_*(f^{-1}(S')),$$

is referred to as an *induced probability measure*, and the distribution  $\bar{A} = (S^1, \Sigma^1, p_1)$  as an *induced (probability) space*.

The sample space  $\bar{R}$  is the distribution of some random variable  $R$ ; in the language just presented,  $R$  should be defined as the identity function  $\text{id} : S^* \rightarrow S^*$  (one that maps each element into itself) on the sample space  $\bar{R}$ ; its induced

probability space is, obviously, also  $\bar{R}$ . In our conceptual framework we simply define  $R$  by its distribution  $\bar{R}$  and some unique identifying label (such as “ $R$ ”). Note that the traditional language, too, requires an identifying label and a distribution (using our terminology) in order to define the sample space itself.

**Remark 2.6** The traditional language does not constitute a different approach. It is a terminological variant of the conceptual set-up adopted in this chapter and applied to a special object of study: a class  $\mathcal{A}$  of random variables that can be defined as functions of some “primary” random variable  $R$ . In accordance with Theorem 2.2,  $\mathcal{A}$  can also be described without mentioning  $R$ , as a class of random variables such that, for any indexed family of random variables  $(A^k : k \in K)$  with  $A^k \in \mathcal{A}(R)$  for all  $k \in K$ , there is a random variable  $A = (A^k : k \in K)$  that also belongs to  $\mathcal{A}$ .

## 2.7 Unrelated random variables and coupling schemes

There are two considerations to keep in mind when using the traditional language of random variables as measurable functions on sample spaces.

One of them is that sample spaces  $\bar{R}$  (or “primary” random variables  $R$ ) are more often than not nebulous: they need not be and usually are not explicitly introduced when dealing with collections of jointly distributed random variables, and they often have no substantive interpretation if introduced. Consider an experiment in which a participant is shown one of two stimuli, randomly chosen, and is asked to identify them by pressing one of two keys as soon as possible. In each trial we record two random variables: stimulus presented, and response time observed, RT. The joint distribution of stimuli and response times is well defined by virtue of pairing them trial-wise. But what would the “primary” random variable  $R$  be of which stimulus and RT would be functions? No one would normally attempt determining one, and it is difficult if one tries, except for the trivial choice  $R = (\text{stimulus}, \text{RT})$  or some one-to-one function thereof. The stimulus and RT then would be projections (i.e., functions) of  $R$ , but this hardly adds insights to our understanding of the situation. Moreover, as soon as one introduces a new random variable in the experimental design, say, “response key,” indicating which of the two keys was pressed, the “primary” random variable  $R$  has to be redefined. It may now be the jointly distributed triple  $R = (\text{stimulus}, \text{response key}, \text{RT})$ .

The second consideration is that there can be no such thing as a single “primary” random variable  $R$  allowing one to define all conceivable random variables as its functions. This is obvious from the cardinality considerations alone: the set  $S^*$  in  $\bar{R}$  would have to be “larger” than the set of possible values for any conceivable random variable (which can, of course, be chosen arbitrarily large). It is a mathematical impossibility. The universe of all conceivable random variables should necessarily include random variables that are not functions of a common “primary” one. In view of Theorem 2.2, this means that there must be random variables that do not

possess a joint distribution. The situation should look like the diagram below, with  $A^1, A^2, \dots$  being functions of some  $R^1, B^1, B^2, \dots$  being functions of some  $R^2$ , but  $R^1$  and  $R^2$  being *stochastically unrelated*, with no joint distribution.



It is true that, as explained below, once  $R^1$  and  $R^2$  are introduced (by their distributions and identifying labels), there is always a way to introduce a new random variable  $(H^1, H^2)$  (whose components are functions of some random variables) such that  $H^1$  has the same distribution as  $R^1$  and  $H^2$  has the same distribution as  $R^2$ . However, there is no way of conceiving all random variables in the form of functions of a single “primary” one.

Examples of random variables that normally are not introduced as jointly distributed are easy to find. If RTs in an experiment with two stimuli (say, “green” and “red”) are considered separately for stimulus “green” and stimulus “red”, we have two random variables:  $RT^{green}$  and  $RT^{red}$ . What “natural” stochastic relationship might they have? The answer is, none: the two random variables occur in mutually exclusive conditions, so there is no privileged way of coupling realizations of  $RT^{green}$  and  $RT^{red}$  and declaring them co-occurring. Once these random variables are introduced, one can impose a joint distribution on them. For example, one may consider them stochastically independent, essentially forcing on them the coupling scheme in which each realization of  $RT^{green}$  is considered as if it co-occurred with every realization  $RT^{red}$ . However, it is also possible to couple them differently; for instance, by the common quantile ranks, so that the  $q$ th quantile of  $RT^{red}$  is paired with and only with the  $q$ th quantile of  $RT^{green}$ . The two random variables then are functions of the quantile rank, which is a random variable uniformly distributed between 0 and 1. The point is, neither of these nor any of the infinity of other coupling schemes for the realizations of  $RT^{green}$  and  $RT^{red}$  is privileged, and none is necessary: one need not impose any joint distribution on  $RT^{green}$  and  $RT^{red}$ .

It can be shown that stochastic independence can be imposed on any set of pairwise *stochastically unrelated* random variables.

**Theorem 2.3** *For any vector  $(R^1, \dots, R^n)$  (more generally, any family  $(R^k : k \in K)$ ) of random variables that are pairwise stochastically unrelated there is a random variable  $H = (H^1, \dots, H^n)$  (generally,  $H = (H^k : k \in K)$ ) with stochastically independent  $H^k$ , such that  $\overline{H^k} = \overline{R^k}$  for all  $k \in K$ .*

$H$  is called the *independent coupling* for  $(R^k : k \in K)$ . In general, a *coupling* for a family of random variables  $(R^k : k \in K)$ , is any random variable  $H = (H^k : k \in K)$  whose every 1-marginal random variable  $H^k$  is distributed as  $R^k$ .

Theorem 2.3 must not be interpreted to mean that one can take all pairwise stochastically unrelated random variables and consider them stochastically independent. The reason for this is that this class is not a well-defined set, and

cannot be therefore indexed by any set. Indeed, if it were possible to present it as  $(R^k : k \in K)$ , then one could form a new random variable  $R = (R^k : k \in K)$  whose distribution is the same as  $\overline{(H^k : k \in K)}$  in Theorem 2.3, and it would follow that the set contains itself as an element (which is impossible for a set).

Summarizing, in practice random variables are often well-defined without their joint distribution being well-defined. There is nothing wrong in dealing with stochastically unrelated random variables without trying to embed them in jointly distributed system. When such an embedding is desirable, the joint distribution is “in the eyes of the beholder,” in the sense of depending on how one wishes to couple the realizations of the variables being interrelated.

## 2.8 On sameness, equality, and equal distributions

We have to distinguish two different meanings in which one can understand the equality of random variables,  $A = B$ .

One meaning is that  $A$  and  $B$  are different notations for one and the same variable; that is, that  $A$  and  $B$  have the same identifying label and the same distribution. This meaning of equality is implicit when we say “let  $D$  be  $(A, B, C)$ , jointly distributed” or “there is a random variable  $A = (A^k : k \in K)$ .”

The other meaning of  $A = B$  is that

1. these random variables have (or may have) different identifying labels (i.e., they are not or may not be the same);
2. they are identically distributed,  $\overline{A} = \overline{B} = (S, \Sigma, p_1)$ ;
3. they are jointly distributed, and their joint distribution has the form  $(S \times S, \Sigma \otimes \Sigma, p_2)$ ;
4. for any  $S' \in \Sigma$ ,

$$p_2(S' \times S') = p_1(S').$$

In some cases, if  $\text{diag}_S = \{(a, a) : a \in S\}$  is a measurable set (i.e., it belongs to  $\Sigma \otimes \Sigma$ ), one can replace the last property with

$$p_2(\text{diag}_S) = 1,$$

which can also be presented as

$$\Pr(A = B) = 1.$$

If  $A$  and  $B$  about which we know that  $A = B$  are represented as functions of some random variable  $R$ , then it is usually assumed that  $\text{diag}_S \in \Sigma \otimes \Sigma$ , and the two functions representing  $A$  and  $B$  are called *equal with probability 1* (or *almost surely*). Of course, if  $A$  and  $B$  are merely different notations for one and the same random variable, they are always jointly distributed and equal in the second sense of the term (see Example 2.15).

The equality of random variables, in either sense, should not be confused with the equality of distributions,  $\bar{A} = \bar{B}$ . The random variables  $A$  and  $B$  here may but do not have to be jointly distributed. They may very well be stochastically unrelated. We will use the symbol  $\sim$  in the meaning of “has the distribution” or “has the same distribution as.” Thus,  $A \sim \bar{A}$  always,  $A \sim B$  if and only if  $\bar{A} = \bar{B}$ , and  $A = B$  always implies  $A \sim B$ .

An important notational consideration applies to random variables with imposed on them or redefined joint distributions. One may write  $(A, B)$  either as indicating a pair of stochastically unrelated random variables, or some random variable  $C = (A, B)$ . The two meanings are distinguished by context. Nothing prevents one, in principle, from considering the same  $A$  and  $B$  as components of two differently distributed pairs,  $C = (A, B)$  and  $C' = (A, B)$ , or as components of a  $C = (A, B)$  possessing a joint distribution and a pair  $(A, B)$  of stochastically unrelated random variables. Doing this within the same context, however, will create conceptual difficulties. For one thing, we would lose the ability of presenting  $A$  and  $B$  as functions of some  $R$  (based on their joint distribution in  $C$ ).

There is a simple and principled way of avoiding this inconvenience: use different symbols for random variables comprising different pairs (more generally, vectors or indexed families), considering them across the pairs (vectors, families) as equally distributed stochastically unrelated random variables. In our example, we can write  $C = (A, B)$  and  $C' = (A', B')$ , where  $A \sim A'$  and  $B \sim B'$ , with  $C$  and  $C'$  being stochastically unrelated. The same principle was applied in the formulation of Theorem 2.3 and more generally, in the definition of a coupling: rather than saying that given a family of stochastically unrelated  $(R^k : k \in K)$ , its coupling is any random variable  $H = (R^k : k \in K)$  whose components are jointly distributed (e.g., independent), the definition says that a coupling is a random variable  $H = (H^k : k \in K)$  such that  $H^k \sim R^k$  for all  $k \in K$ . This means, in particular, that every vector of random variables is stochastically unrelated to any of its couplings.

## 2.9 Random outputs depending on inputs

Let a random variable be distributed as  $(S, \Sigma, p_\phi)$ , where  $\phi$  stands for some deterministic variable taking values in a set  $\Phi$ . This means that the probability measure on  $\Sigma$  (the entire function) is generally different for different values of  $\Phi$ . One could also write  $p(\phi)$  instead of  $p_\phi$ , but one should keep in mind that this is not a function from  $\Phi$  to a set of values of  $p$  (real numbers between 0 and 1), but rather a function from  $\Phi$  to the set of all possible probability measures on  $\Sigma$ . The dependence of  $p_\phi$  on  $\phi$  means that the distribution  $(S, \Sigma, p_\phi)$  of the random variable in question depends on  $\phi$ . We can present it as  $\bar{A}_\phi$ , and the random variable itself as  $A_\phi$ . One can say that the random variable  $A$  *depends on*  $\phi$ , which is equivalent to saying that there is an indexed family of random variables  $(A_\phi : \phi \in \Phi)$ .

Let  $\phi_1$  and  $\phi_2$  be two different elements of  $\Phi$ . We will assume throughout the rest of the chapter that the corresponding random variables  $A_{\phi_1}$  and  $A_{\phi_2}$  always have different identifying labels (such as “ $A$  at  $\phi = \phi_1$ ” and “ $A$  at  $\phi = \phi_2$ ”); that is, they are never one and the same variable. However, they may have one and the same distribution function, if  $p_{\phi_1} \equiv p_{\phi_2}$ . If  $A$  is a vector of jointly distributed random variables  $(A^1, \dots, A^n)$ , then its dependence on  $\phi$  can be shown as  $A_\phi = (A^1, \dots, A^n)_\phi$  or  $A_\phi = (A_\phi^1, \dots, A_\phi^n)$ .

In the following,  $\phi$  always represents mutually exclusive conditions under which  $A$  is observed, and the indexed family  $(A_\phi : \phi \in \Phi)$  abbreviated by  $A$  consists of pairwise stochastically unrelated random variables. The elements of  $\Phi$  are referred to as *treatments*, the term being used in the same way as in the analysis of variance: a combination of values of different *factors*, or *inputs*. We will use the latter term. An input is simply a variable  $\lambda$  with a set of possible values  $\Lambda$ . If the number of inputs considered is  $m$ , a treatment is a vector

$$\phi = (\lambda^1, \dots, \lambda^m),$$

with  $\lambda^1 \in \Lambda^1, \dots, \lambda^m \in \Lambda^m$ . The set of treatments is therefore

$$\Phi \subset \Lambda^1 \times \dots \times \Lambda^m.$$

**Remark 2.7** As it is commonly done in mathematics, we will use the same symbol to denote a variable and its specific values. For example, in  $\lambda^1 \in \Lambda^1$  the symbol  $\lambda^1$  refers to a value of  $\lambda^1$ , whereas in the sentence “ $A^1$  depends on  $\lambda^1$ ” the same symbol refers to the variable as a whole. This ambiguity is possible to avoid by using  $\Lambda^1$  in place of  $\lambda^1$  when referring to the entire variable, and using a pair  $(\lambda^1, \Lambda^1)$  when referring to an input value as that of a given input. We do not use this rigorous notation here, assuming context will be sufficient for disambiguation.

**Example 2.16** Let  $\phi$  describe a stimulus presented to a participant. Let it attain eight possible values formed by combinations of three binary attributes, such as

$$\begin{aligned}\lambda^1 &\in \Lambda^1 = \{\text{large, small}\}, \lambda^2 \in \Lambda^2 = \{\text{bright, dim}\}, \\ \lambda^3 &\in \Lambda^3 = \{\text{round, square}\}.\end{aligned}$$

Let the participant respond by identifying (correctly or incorrectly) these attributes, by saying  $A^1 = \text{“large”}$  or  $\text{“small”}$ ,  $A^2 = \text{“bright”}$  or  $\text{“dim”}$ , and  $A^3 = \text{“round”}$  or  $\text{“square”}$ . The response therefore is a vector of three binary random variables  $(A^1, A^2, A^3)_\phi$  that depends on stimuli  $\phi = (\lambda^1, \lambda^2, \lambda^3)$ . Equivalently, we can say that there are eight triples of random variables, one for each treatment,  $(A^1, A^2, A^3)_{\phi_1}, \dots, (A^1, A^2, A^3)_{\phi_8}$ .  $\square$

The set of all treatments  $\Phi$  may be equal to  $\Lambda^1 \times \dots \times \Lambda^m$ , but it need not be. Some of the logically possible combinations of input values may not be physically realizable or simply may not be of interest. The elements of  $\Phi$  therefore are referred to as *allowable treatments*. We will see later that this notion is important in pairing inputs with random outputs.

**Example 2.17** Suppose  $\Lambda^1$  and  $\Lambda^2$  denote the sets of possible lengths of two line segments presented side by side in the visual field of an observer. Let  $A^1$  and  $A^2$  denote the observer's numerical estimates of the two lengths. If the goal of the experiment is to study perceptual discrimination, it may be reasonable (and time-saving) to exclude the pairs with large values of  $|\lambda^1 - \lambda^2|$ . For example, if  $\Lambda^1 = \Lambda^2 = \{5, 6, 7, 8, 9\}$ , the set of allowable treatments may be defined as

$$\Phi = \{(\lambda^1, \lambda^2) \in \Lambda^1 \times \Lambda^2 : |\lambda^1 - \lambda^2| \leq 2\}.$$

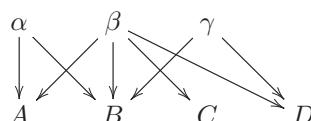
This set contains only 19 treatments of the 25 logically possible combinations.  $\square$

As explained in the introductory section, inputs may very well be random variables themselves, but only their possible values rather than their distributions are relevant in our analysis: the distributions of random outputs are always *conditioned* upon particular treatments. Therefore, all inputs are always treated as deterministic quantities.

## 2.10 Selectiveness in the dependence of outputs on inputs

We are interested in the relationship between (deterministic) inputs and random outputs. Specifically, we are interested in the selectiveness in this relationship: which input may and which may not influence a given output. Such selectiveness can be presented in the form of a *diagram of influences*, where an arrow from an input  $\lambda$  to a random output  $A$  means that  $\lambda$  *influences*  $A$  (note that the meaning of “influence” has not been as yet defined). The absence of an arrow from an input  $\lambda$  to a random output  $A$  excludes  $\lambda$  from the set of inputs that influence  $A$ .

Consider, for example the following arrow diagram

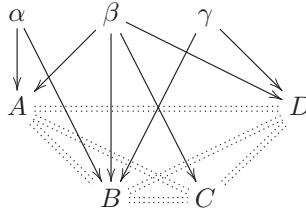


This diagram can be interpreted by saying that:

1. the random outputs  $(A, B, C, D)$  are jointly distributed, and their joint distribution (specifically, joint probability measure) depends on the inputs  $(\alpha, \beta, \gamma)$ ; in other words,  $(A, B, C, D)$  is in fact  $(A, B, C, D)_{\alpha\beta\gamma}$ , or  $(A_{\alpha\beta\gamma}, B_{\alpha\beta\gamma}, C_{\alpha\beta\gamma}, D_{\alpha\beta\gamma})$ .
2. output  $A$  is influenced by inputs  $\alpha, \beta$ , but not by input  $\gamma$ ;
3. output  $B$  is influenced by all inputs,  $\alpha, \beta, \gamma$ ;
4. output  $C$  is influenced by input  $\beta$ , but not by inputs  $\alpha, \gamma$ ;
5. output  $D$  is influenced by inputs  $\beta$  and  $\gamma$ , but not by  $\alpha$ .

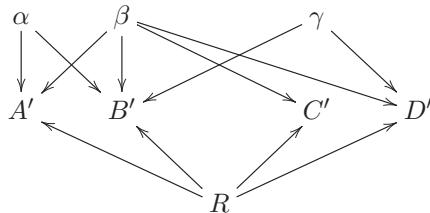
The first thing to do here is to ask the question we asked in the introductory section: does this even make sense? It certainly does if  $(A, B, C, D)_{\alpha\beta\gamma}$ , for every treatment  $(\alpha, \beta, \gamma)$ , is a vector of *independent* random variables. Then the points 2, 3, and 4

above simply translate into the statements: the marginal distribution of  $A$  depends on  $\alpha, \beta$  but not on  $\gamma$ ; the marginal distribution of  $B$  depends on  $\alpha, \beta, \gamma$ ; etc. However, does the selectiveness make sense if the random outputs are not stochastically independent? Look at the diagram below, the same as above, but with added point lines indicating stochastic interdependences.



We see, for instance, that output  $A$  is influenced by  $\alpha$ , and output  $C$  is stochastically dependent on  $A$ . In what sense, then, can one say that  $\alpha$  does not influence  $C$ ? The output  $B$  is influenced by all inputs, and every other output is stochastically dependent on  $B$ . Does not this mean that every output is influenced by every input?

This seemingly compelling line of reasoning is a conceptual confusion. It confuses two types of relations, both of which can be described using the word “dependence.” Stochastic dependence and dependence of outputs on inputs are different in nature. This is easy to understand if we consider the following diagram:



In this diagram, every random variable is a function of all the arguments from which the arrows leading to this random variable initiate:

$$\begin{aligned} A'_{\alpha\beta\gamma} &= f_1(\alpha, \beta, R), \\ B'_{\alpha\beta\gamma} &= f_2(\alpha, \beta, \gamma, R), \\ C'_{\alpha\beta\gamma} &= f_3(\beta, R), \\ D'_{\alpha\beta\gamma} &= f_4(\alpha, \beta, R). \end{aligned}$$

For every value of  $R$  and for every treatment  $(\alpha, \beta, \gamma)$ , the values of  $(A', B', C', D')_{\alpha\beta\gamma}$  are determined uniquely. Suppose now that we have, for every treatment,

$$(A', B', C', D')_{\alpha\beta\gamma} \sim (A, B, C, D)_{\alpha\beta\gamma}.$$

This assumption explains the coexistence of the stochastic relationship between the random outputs and the selectiveness in their dependence on the inputs. *For any given treatment, the components of  $(A, B, C, D)_{\alpha\beta\gamma}$  are generally stochastically interdependent because they are distributed as functions of one and the*

same random variable  $R$  (of course, as a special case, they may also be stochastically independent). At the same time, for any fixed value  $r$  of  $R$ , the value  $a = f_1(\alpha, \beta, r)$  of the output  $A'_{\alpha\beta\gamma}$  cannot depend on  $\gamma$ , the value  $c = f_3(\beta, r)$  of the output  $C'_{\alpha\beta\gamma}$  cannot depend on anything but  $\beta$ , etc. And because the distributions of  $(A', B', C', D')_{\alpha\beta\gamma}$  and  $(A, B, C, D)_{\alpha\beta\gamma}$  are the same, they share the same selectiveness pattern.

This consideration leads us to a rigorous definition of what it means for a vector of random outputs  $(A, B, C, D)_{\alpha\beta\gamma}$  to satisfy the pattern of selective influences represented in the opening diagram of this section: this pattern is satisfied if and only if the equations above are satisfied for some choice of a random variable  $R$  and functions  $f_1, f_2, f_3, f_4$ . This definition can be generalized to an arbitrary family of random outputs and an arbitrary family of inputs. However, we will confine our attention to the case when these families are finite vectors. And we will use a special (re-)arrangement of the inputs to make the definition especially simple.

**Remark 2.8** It should be kept in mind that the meaning of “ $\lambda$  influences  $A$ ” includes, as a special case the possibility of  $\lambda$  not influencing  $A$ . There is an asymmetry in saying that, in the example used in this section,  $C$  depends on  $\beta$ , and saying that  $C$  does not depend on  $\alpha$ . The latter is a definitive statement:  $\alpha$  is not within the list of arguments in the function  $c = f_3(\beta, r)$ . The dependence on  $\beta$  means that  $\beta$  is within this list. However, a constant function is a special case of a function. So  $c = f_3(\beta, r)$  may, as a special case, be constant at all values of  $R$ , or at all values of  $R$  except on a subset of measure zero. For instance, if  $R$  is uniformly distributed between 0 and 1 (we will see below that this choice is possible in a wide class of cases) and  $c = f_3(\beta, r)$  is a non-constant function of  $\beta$  only at rational  $r$ , then  $C$  does not depend on  $\beta$  with probability 1 (because the set of all rational points is countable, hence its Lebesgue measure is zero). This shows that the terms “depends on” and “influences” should generally be understood as “may depend on” and “may influence.”

## 2.11 Selective influences in a canonical form

Continuing with the same example, let us consider the random outputs one by one, and for each of them group together all inputs that influence it. We get

$$\begin{array}{cccc} \lambda^1 = (\alpha, \beta) & \lambda^2 = (\alpha, \beta, \gamma) & \lambda^3 = (\beta) & \lambda^4 = (\beta, \gamma) \\ \downarrow & \downarrow & \downarrow & \downarrow \\ A & B & C & D \end{array}$$

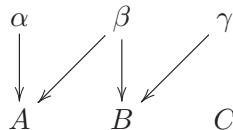
Let us assume that each of the inputs  $\alpha, \beta, \gamma$  has three possible values, crossed in all possible ways to form 27 treatments. Each of the newly formed groups of inputs can be viewed as a new input in its own right. Thus,  $\lambda^1$  and  $\lambda^4$  are inputs whose

sets of possible values  $\Lambda^1$  and  $\Lambda^4$  have nine possible values each,  $\lambda^2$  is an input with 27 possible values in  $\Lambda^2$ , and  $\lambda^3$  is an input with three values in  $\Lambda^3$ .

Such a rearrangement is always possible, whatever the original pattern of influences, and it achieves a one-to-one correspondence between random outputs and inputs. We call a diagram with such one-to-one correspondence a *canonical diagram of influences*. (The term “canonical” is used in mathematics to refer to a standard representation into which a variety of other representations can be transformed.) The problem of selectiveness with a canonical diagram acquires a simple form: is every random output selectively influenced by its corresponding input?

When dealing with canonical diagrams it is especially important to keep in mind that *allowable treatments* are generally just a subset of the Cartesian product of the sets of input values. In our example, this Cartesian product is  $\Lambda^1 \times \Lambda^2 \times \Lambda^3 \times \Lambda^4$  and it consists of  $9 \times 27 \times 3 \times 9$  elements. But, obviously, only 27 combinations of the new inputs’ values are allowable, corresponding to the 27 treatments formed by the completely crossed original inputs. Thus, if  $\lambda^2 = (\alpha, \beta, \gamma)$ , then the only allowable treatment containing this value of  $\lambda^2$  also contains  $\lambda^1 = (\alpha, \beta)$ ,  $\lambda^3 = (\beta)$ , and  $\lambda^4 = (\beta, \gamma)$ .

Another consideration related to the canonical diagrams of influences is that in order to ensure one-to-one correspondence between inputs and random outputs, we may need to allow for “dummy” inputs, with a single possible value. Consider the following example:



Not being influenced by any inputs (as it is the case with the output  $C$ ) is a special case of selectiveness, so this situation falls within the scope of our analysis. Presented in the canonical form, this diagram becomes

$$\begin{array}{ccc} \lambda^1 = (\alpha, \beta) & \lambda^2 = (\beta, \gamma) & \lambda^3 = () \\ \downarrow & \downarrow & \downarrow \\ A & B & C \end{array}$$

The new input  $\lambda^3$  represents an empty subset of original inputs. Therefore,  $\lambda^3$  does not change, and should formally viewed as an input whose set of possible values  $\Lambda^3$  contains a single element, that we may denote arbitrarily.

We are ready now to give a formal definition of selective influences. Let  $(\lambda^1, \dots, \lambda^n)$  be a vector of inputs, with values belonging to nonempty sets  $(\Lambda^1, \dots, \Lambda^n)$ , respectively. Let  $\Phi \subset \Lambda^1 \times \dots \times \Lambda^n$  be a nonempty set of allowable treatments. Let  $(A_\phi^1, \dots, A_\phi^n)$  be a vector of random variables jointly distributed for every  $\phi \in \Phi$ . (Recall that for  $\phi \neq \phi'$ , the random variables

$(A_\phi^1, \dots, A_\phi^n)$  and  $(A_{\phi'}^1, \dots, A_{\phi'}^n)$  are stochastically unrelated.) We say that the dependence of  $(A_\phi^1, \dots, A_\phi^n)$  on  $\phi$  satisfies the (canonical) diagram of influences

$$\begin{array}{ccc} \lambda^1 & \dots & \lambda^n \\ \downarrow & & \downarrow \\ A^1 & \dots & A^n \end{array}$$

if and only if one can find a random variable  $R$  and functions  $f_1, \dots, f_n$  such that

$$(A_\phi^1, \dots, A_\phi^n) \sim (f_1(\lambda^1, R), \dots, f_n(\lambda^n, R))$$

for every  $(\lambda^1, \dots, \lambda^n) = \phi \in \Phi$ .

**Remark 2.9** There is no implication of uniqueness in this definition: below, in the discussion of the linear feasibility test, we will reconstruct  $R$  explicitly, and we will see that it can, as a rule, be chosen in infinitely many ways. Theorem 2.6 below shows the non-uniqueness of  $R$  by another argument.

Instead of drawing diagrams, in the sequel we will present the same pattern of selective influences as

$$(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n),$$

and say that  $A^1, \dots, A^n$  are *selectively influenced* by  $\lambda^1, \dots, \lambda^n$  (respectively). If it is known that for a given vector of input–output pairs the definition above is not satisfied whatever  $R$  and  $f_1, \dots, f_n$  one chooses, then we write

$$(A^1, \dots, A^n) \not\leftrightarrow (\lambda^1, \dots, \lambda^n).$$

Note that for this schematic notation to make sense, the context in which it is used should specify the sets of input values, the distributions of  $(A_\phi^1, \dots, A_\phi^n)$ , and the set of allowable treatments.

**Example 2.18** Let  $R = (R_1, R_2, R_3)$  denote a vector of independent standard normal random variables, and suppose the input factors  $\Lambda^1$  and  $\Lambda^2$  are some subsets of  $\mathbb{R}$ . Then, the binary random variables

$$A_{(\lambda^1, \lambda^2)}^1 = \begin{cases} 1 & \text{if } R_1 < \lambda^1 + R_3, \\ 0 & \text{otherwise,} \end{cases}$$

$$A_{(\lambda^1, \lambda^2)}^2 = \begin{cases} 1 & \text{if } R_2 < \lambda^2 + R_3, \\ 0 & \text{otherwise,} \end{cases}$$

are selectively influenced by, respectively,  $\lambda^1 \in \Lambda^1$  and  $\lambda^2 \in \Lambda^2$ , because  $A^1$  depends only on  $(\lambda^1, R)$  and  $A^2$  depends only on  $(\lambda^2, R)$ . For any given  $(\lambda^1, \lambda^2)$ , the random variables  $A_{(\lambda^1, \lambda^2)}^1$  and  $A_{(\lambda^1, \lambda^2)}^2$  are not stochastically independent because  $R_1 - R_3$  and  $R_2 - R_3$  have a nonzero correlation.  $\square$

## 2.12 Joint distribution criterion

Let us begin by making sure that the simplest special case, when  $(A_\phi^1, \dots, A_\phi^n)$  are mutually independent random variables at every allowable treatment  $\phi$ , falls within the scope of the general definition. We expect, if our general definition is well constructed, that in this case selectiveness of influences,  $(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$ , follows from the fact that the distribution of  $A_\phi^k$  (for  $k = 1, \dots, n$ ) depends only on  $\lambda^k$ . In order not to deal with infinite indexed families, let us assume that  $\lambda^k$  has a finite number of values, enumerated as  $1, \dots, m_k$ .

Consider the random variable

$$H = (H_1^1, \dots, H_{m_1}^1, \dots, H_1^k, \dots, H_{m_k}^k, \dots, H_1^n, \dots, H_{m_n}^n)$$

with stochastically independent components, such that, for all  $i = 1, \dots, m_k$  and  $k = 1, \dots, n$ ,

$$H_i^k \sim A_\phi^k$$

whenever  $\lambda^k = i$  is in  $\phi$ . In other words, if the treatment  $\phi$  contains the  $i$ th value of the input  $\lambda^k$ , then we pick  $A_\phi^k$ , and change its identifying label with its distribution intact to create  $H_i^k$ . Clearly, the  $H_i^k$  will be the same (provided we always use the same label) irrespective of which  $\phi$  contains  $\lambda^k = i$ . The variable  $H$  above always exists by Theorem 2.3. Let us define function  $f_k$  for  $k = 1, \dots, n$  by

$$f_k(i, h_1^1, \dots, h_{m_1}^1, \dots, h_1^k, \dots, h_{m_k}^k, \dots, h_1^n, \dots, h_{m_n}^n) = h_i^k.$$

This can be understood as the “first-level”  $k$ th projection that selects from the range of the arguments the subrange  $h_1^k, \dots, h_{m_k}^k$ , followed by the “second-level”  $i$ th projection that selects from this subrange the argument  $h_i^k$ . It is obvious then that, for every  $\phi \in \Phi$ ,

$$A_\phi^k \sim f_k(i, H)$$

whenever  $\phi$  contains  $\lambda^k = i$ . But then

$$(A_\phi^1, \dots, A_\phi^n) \sim (f_1(\lambda^1, H), \dots, f_n(\lambda^n, H))$$

whenever  $(\lambda^1, \dots, \lambda^n) = \phi \in \Phi$ , as it is required by the general definition.

The vector  $H$  constructed in this analysis is a special case of the *reduced coupling vector* introduced next. As it turns out, the existence of such a vector, with one random variable per each value of each input is the general criterion for selective influences. A criterion for a statement is another statement which is equivalent to it. Put differently, a criterion is a condition which is both necessary and sufficient for a given statement.

Consider the statement that  $A^1, \dots, A^n$  are selectively influenced by  $\lambda^1, \dots, \lambda^n$ , respectively. By definition, for this to be true, there should exist functions  $f_1, \dots, f_n$  and a random variable  $R$  such that

$$(A_\phi^1, \dots, A_\phi^n) \sim (f_1(\lambda^1, R), \dots, f_n(\lambda^n, R))$$

for every  $(\lambda^1, \dots, \lambda^n) = \phi \in \Phi$ . We continue to assume that every input  $\lambda^k$  has a finite number of values, enumerated  $1, \dots, m_k$ . (Recall, from the discussion of dummy inputs, that  $m_k = 1$  is allowed.)

For each  $k$  and every value of  $\lambda^k$ , denote

$$H_{\lambda^k}^k = f_k(\lambda^k, R).$$

As  $\lambda^k$  runs from 1 to  $m_k$  and  $k$  runs from 1 to  $n$ , this creates  $m_1 + \dots + m_n$  random variables, one random variable per each value of each input, jointly distributed due to being functions of one and the same  $R$ . We have therefore a random variable

$$H = (H_1^1, \dots, H_{m_1}^1, \dots, H_1^k, \dots, H_{m_k}^k, \dots, H_1^n, \dots, H_{m_n}^n).$$

If follows from the definition of selective influences that if  $(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$ , then, for every allowable treatment  $\phi = (\lambda^1, \dots, \lambda^n)$ ,

$$(A_\phi^1, \dots, A_\phi^n) \sim (H_{\lambda^1}^1, \dots, H_{\lambda^n}^n).$$

In other words, the existence of a jointly distributed vector of random variables  $H$  with this property is a necessary condition for  $(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$ .

Let us now assume that a vector  $H$  with the above property exists. Let us define functions as we did it in the case with stochastic independence,

$$f_k(i, h_1^1, \dots, h_{m_1}^1, \dots, h_1^k, \dots, h_{m_k}^k, \dots, h_1^n, \dots, h_{m_n}^n) = h_i^k.$$

Then

$$(A_\phi^1, \dots, A_\phi^n) \sim (H_{\lambda^1}^1, \dots, H_{\lambda^n}^n) = (f_1(\lambda^1, H), \dots, f_n(\lambda^n, H))$$

for every  $(\lambda^1, \dots, \lambda^n) = \phi \in \Phi$ . This means that the existence of  $H$  is a sufficient condition for  $(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$ .

Summarizing, we have proved the following theorem.

**Theorem 2.4** (Joint distribution criterion) *Let  $(\lambda^1, \dots, \lambda^n)$  be a vector of inputs, with  $\lambda^k \in \Lambda^k = \{1, \dots, m_k\}$  ( $m_k \geq 1$ ,  $k = 1, \dots, n$ ). Let  $\Phi \subset \Lambda^1 \times \dots \times \Lambda^n$  be a nonempty set of allowable treatments. Let  $(A_\phi^1, \dots, A_\phi^n)$  be a set of random variables jointly distributed for every  $\phi \in \Phi$ . Then*

$$(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$$

*if and only if there exists a vector of jointly distributed random variables*

$$H = \left( \overbrace{H_1^1, \dots, H_{m_1}^1}^{\text{one variable per each value of each input}}, \dots, \overbrace{H_1^k, \dots, H_{m_k}^k}^{\text{one variable per each value of each input}}, \dots, \overbrace{H_1^n, \dots, H_{m_n}^n}^{\text{one variable per each value of each input}} \right),$$

*(one variable per each value of each input) such that*

$$(A_\phi^1, \dots, A_\phi^n) \sim (H_{\lambda^1}^1, \dots, H_{\lambda^n}^n)$$

*for every  $(\lambda^1, \dots, \lambda^n) = \phi \in \Phi$ .*

The vector  $H$  in this theorem is called a *reduced coupling vector* for the family  $((A_\phi^1, \dots, A_\phi^n) : \phi \in \Phi)$  (or for a given pattern of selective influences).

**Remark 2.10** According to the general definition of a coupling (Section 2.7), a coupling for the family  $((A_\phi^1, \dots, A_\phi^n) : \phi \in \Phi)$  is any random variable

$$H^* = ((H_\phi^1, \dots, H_\phi^n) : \phi \in \Phi)$$

such that, for all  $\phi \in \Phi$ ,

$$(A_\phi^1, \dots, A_\phi^n) \sim (H_\phi^1, \dots, H_\phi^n).$$

The vector  $H$  of Theorem 2.4 is obtained from such a coupling by imposing on it additional constraints: for any  $k = 1, \dots, n$  and any  $\phi, \phi' \in \Phi$  sharing the same value of input  $\lambda^k$ ,

$$H_\phi^k = H_{\phi'}^k.$$

These constraints allow one to reduce all different occurrences of  $H^k$  in  $H$  to one occurrence per each value of factor  $\lambda^k$ . Hence, the adjective “reduced” in the name for this special coupling. (In the literature on selective influences the reduced coupling was also called a *joint distribution vector*, and a *joint distribution criterion vector*. We will not use these terms here.)

Theorem 2.4 is much more important than may be suggested by its simple proof (essentially, by means of renaming functions of a random variable into random variables, and vice versa). The reasons for its importance are:

1. it is often easier to determine whether a coupling vector exists than whether one can find certain functions of a single random variable (unless the latter is taken to be the reduced coupling vector and the functions to be its projections); and
2. even when a reduced coupling vector is not explicitly constructed, its existence provides insights into the nature of the random variable  $R$  in the definition of selective influences.

The first of these reasons is yet another illustration of the fact that jointly distributed random variables are not, as a rule, introduced as functions of a single random variable (see Section 2.7). Take a simple example, when there are two binary inputs  $\lambda^1, \lambda^2$  (with values 1, 2 each) paired with two binary outputs (with values 1, 2 each). Let the set of allowable treatments consist of all four combinations,

$$(\lambda^1 = 1, \lambda^2 = 1), (\lambda^1 = 1, \lambda^2 = 2), (\lambda^1 = 2, \lambda^2 = 1), (\lambda^1 = 2, \lambda^2 = 2).$$

Note that 1 and 2 as values for the inputs are chosen merely for convenience. We could replace them by any numbers or distinct symbols (say,  $\boxtimes, \boxplus$  for  $\lambda^1$ , and  $\times, \times$  for  $\lambda^2$ ). The existence of the jointly distributed vectors  $(A_\phi^1, A_\phi^2)$  means that for each of the four treatments  $\phi$  we are given four probabilities of the form

$$\begin{aligned} \Pr(A_\phi^1 = 1, A_\phi^2 = 1), & \quad \Pr(A_\phi^1 = 1, A_\phi^2 = 2), \\ \Pr(A_\phi^1 = 2, A_\phi^2 = 1), & \quad \Pr(A_\phi^1 = 2, A_\phi^2 = 2). \end{aligned}$$

Of course, the four probabilities sum to 1. Again, the use of 1 and 2 for values here is arbitrary, other symbols, generally different for  $A_\phi^1$  and  $A_\phi^2$ , would do as well. According to the joint distribution criterion,  $(A^1, A^2) \leftrightarrow (A_\phi^1, A_\phi^2)$  means the existence of four jointly distributed random variables

$$H = (H_1^1, H_2^1, H_1^2, H_2^2),$$

with  $H_1^1$  corresponding to the first value of input  $\lambda^1$ ,  $H_2^1$  to the second value of input  $\lambda^1$ , etc., such that

$$(A^1, A^2)_{\lambda^1=1, \lambda^2=1} \sim (H_1^1, H_1^2), \quad (A^1, A^2)_{\lambda^1=1, \lambda^2=2} \sim (H_1^1, H_2^2), \\ (A^1, A^2)_{\lambda^1=2, \lambda^2=1} \sim (H_2^1, H_1^2), \quad (A^1, A^2)_{\lambda^1=2, \lambda^2=2} \sim (H_2^1, H_2^2).$$

This implies, of course, that  $H_1^1, H_2^1, H_1^2, H_2^2$  are all binary random variables, with values 1 and 2 each.

What is the meaning of saying that they are jointly distributed? The meaning is that for any of the  $2 \times 2 \times 2 \times 2$  possible combinations of values for  $H_1^1, H_2^1, H_1^2, H_2^2$  we can find a probability,

$$\Pr(H_1^1 = i, H_2^1 = i', H_1^2 = j, H_2^2 = j') = p_{ii'jj'},$$

where  $i, j, i', j' \in \{1, 2\}$ . It does not matter what these probabilities  $p_{ii'jj'}$  are, insofar as they

- (i) are legitimate probabilities, that is, they are nonnegative and sum to 1 across the 16 values of  $H$ ;
- (ii) satisfy the 2-marginal constraints

$$(A^1, A^2)_{\lambda^1=i, \lambda^2=j} \sim (H_i^1, H_j^2),$$

for all  $i, j \in \{1, 2\}$ .

The latter translates into

$$p_{i1j1} + p_{i1j2} + p_{i2j1} + p_{i2j2} = \Pr(H_1^1 = i, H_1^2 = j) = \Pr(A^1 = i, A^2 = j)_{\lambda^1=1, \lambda^2=1}, \\ p_{i1j1} + p_{i1j2} + p_{i2j1} + p_{i2j2} = \Pr(H_1^1 = i, H_2^2 = j) = \Pr(A^1 = i, A^2 = j)_{\lambda^1=1, \lambda^2=2}, \\ p_{1ij1} + p_{1ij2} + p_{2ij1} + p_{2ij2} = \Pr(H_2^1 = i, H_1^2 = j) = \Pr(A^1 = i, A^2 = j)_{\lambda^1=2, \lambda^2=1}, \\ p_{1ij1} + p_{1ij2} + p_{2ij1} + p_{2ij2} = \Pr(H_1^1 = i, H_2^2 = j) = \Pr(A^1 = i, A^2 = j)_{\lambda^1=2, \lambda^2=2}.$$

This is a simple system of four linear equations with 16 unknowns, subject to being legitimate probabilities (i.e., being nonnegative and summing to 1). We will discuss this algebraic structure in the next section, but it should be clear that this is a much more transparent task than the one of finding a random variable  $R$  and some functions, or proving that they cannot be found.

**Example 2.19** Let  $A^1, A^2$  have values in  $\{1, 2\}$  and depend on the factors  $\lambda^1 \in \Lambda^1 = \{1, 2\}$  and  $\lambda^2 \in \Lambda^2 = \{1, 2\}$ . Let all four possible treatments be allowable. Suppose we observe the following joint distributions of  $A^1, A^2$  for these treatments:

$\lambda^1$	$\lambda^2$	$A^1$	$A^2$	Pr	$\lambda^1$	$\lambda^2$	$A^1$	$A^2$	Pr
1	1	1	1	.140	1	2	1	1	.198
		1	2	.360			1	2	.302
		2	1	.360			2	1	.302
		2	2	.140			2	2	.198
$\lambda^1$	$\lambda^2$	$A^1$	$A^2$	Pr	$\lambda^1$	$\lambda^2$	$A^1$	$A^2$	Pr
2	1	1	1	.189	2	2	1	1	.460
		1	2	.311			1	2	.040
		2	1	.311			2	1	.040
		2	2	.189			2	2	.460

The question of whether  $(A^1, A^2) \leftrightarrow (\lambda^1, \lambda^2)$  now reduces to finding a solution for the system of linear equations mentioned above. Let us substitute the above observed probabilities into the system:

$$\begin{aligned}
p_{1111} + p_{1112} + p_{1211} + p_{1212} &= 0.140, & p_{1111} + p_{1121} + p_{1211} + p_{1221} &= 0.198, \\
p_{1121} + p_{1122} + p_{1221} + p_{1222} &= 0.360, & p_{1112} + p_{1122} + p_{1212} + p_{1222} &= 0.302, \\
p_{2111} + p_{2112} + p_{2211} + p_{2212} &= 0.360, & p_{2111} + p_{2121} + p_{2211} + p_{2221} &= 0.302, \\
p_{2121} + p_{2122} + p_{2221} + p_{2222} &= 0.140, & p_{2112} + p_{2122} + p_{2212} + p_{2222} &= 0.198, \\
\\
p_{1111} + p_{1112} + p_{2111} + p_{2112} &= 0.189, & p_{1111} + p_{1121} + p_{2111} + p_{2121} &= 0.460, \\
p_{1121} + p_{1122} + p_{2121} + p_{2122} &= 0.311, & p_{1112} + p_{1122} + p_{2112} + p_{2122} &= 0.040, \\
p_{1211} + p_{1212} + p_{2211} + p_{2212} &= 0.311, & p_{1211} + p_{1221} + p_{2211} + p_{2221} &= 0.040, \\
p_{1221} + p_{1222} + p_{2221} + p_{2222} &= 0.189, & p_{1212} + p_{1222} + p_{2212} + p_{2222} &= 0.460.
\end{aligned}$$

The values (found using the simplex linear programming algorithm)

$$\begin{aligned}
p_{1111} &= 0.067, & p_{1211} &= 0, & p_{2111} &= 0.122, & p_{2211} &= 0.04, \\
p_{1112} &= 0, & p_{1212} &= 0.073, & p_{2112} &= 0, & p_{2212} &= 0.198, \\
p_{1121} &= 0.131, & p_{1221} &= 0, & p_{2121} &= 0.14, & p_{2221} &= 0, \\
p_{1122} &= 0.04, & p_{1222} &= 0.189, & p_{2122} &= 0, & p_{2222} &= 0
\end{aligned}$$

satisfy these equations, and as they are nonnegative and sum to one, they represent a probability distribution. Thus, according to the joint distribution criterion, the observed joint distributions satisfy selective influences.  $\square$

To illustrate the second reason for the importance of Theorem 2.4, we consider the following question. By the definition of selective influences, the proposition  $(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$  means the existence of a random variable  $R$  and functions  $f_1, \dots, f_n$  such that

$$(A_\phi^1, \dots, A_\phi^n) \sim (f_1(\lambda^1, R), \dots, f_n(\lambda^n, R))$$

for every  $(\lambda^1, \dots, \lambda^n) = \phi \in \Phi$ . This definition says nothing about the nature and complexity of  $R$  and the functions involved, even for the simplest observable random variables  $(A^1, \dots, A^n)_\phi$ . In most applications  $(A^1, \dots, A^n)_\phi$  are random variables in the narrow sense (Section 2.4). It seems intuitive to expect that in such cases  $R$ , if it exists, is also a random variable in the narrow sense. However, this does not follow from the definition of selective influences. Even if one manages to

prove that for a given family of random variables  $(A^1, \dots, A^n)_\phi$  in the narrow sense this definition is satisfied by no random variable  $R$  in the narrow sense, we still do not know whether this means that the selectiveness  $(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$  is ruled out. What if there is a random variable  $R$  of a much greater complexity (say, a random function or a random set) for which one can find functions  $f_1, \dots, f_n$  as required by the definition?

The joint distribution criterion, however, allows one to rule out such a possibility. Because the reduced coupling vector

$$H = (H_1^1, \dots, H_{m_1}^1, \dots, H_1^n, \dots, H_{m_n}^n),$$

if it exists, should satisfy

$$(A_\phi^1, \dots, A_\phi^n) \sim (H_{\lambda^1}^1, \dots, H_{\lambda^n}^n),$$

it follows that, for any  $k$  and  $\lambda^k$ ,

$$H_{\lambda^k}^k \sim A_\phi^k,$$

whenever the treatment  $\phi$  contains  $\lambda^k$ . But this means that each  $H_{\lambda^k}^k$  is a random variable in a narrow sense, and from Section 2.4 we know then that  $H$  is a random variable in the narrow sense. This constitutes a proof of the following theorem, a simple corollary to the joint distribution criterion.

**Theorem 2.5** *Let  $(\lambda^1, \dots, \lambda^n)$ ,  $\Phi$ , and  $(A_\phi^1, \dots, A_\phi^n)$  be the same as in Theorem 2.4. Let, in addition,  $(A_\phi^1, \dots, A_\phi^n)$  be random variables in the narrow sense. Then*

$$(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$$

*if and only if there is a random variable  $R$  in the narrow sense and functions  $f_1, \dots, f_n$  such that*

$$(A_\phi^1, \dots, A_\phi^n) \sim (f_1(\lambda^1, R), \dots, f_n(\lambda^n, R))$$

*for every  $(\lambda^1, \dots, \lambda^n) = \phi \in \Phi$ .*

If one feels dissatisfied with considering vectors of random variables on a par with “single” random variables, this dissatisfaction is not well-grounded. The fact is, the dimensionality of vectors of random variables in the narrow sense is not essential. Consider, for example, the reduced coupling vector

$$H = (H_1^1, H_2^1, H_1^2, H_2^2),$$

constructed earlier for two binary random variables selectively influenced by two binary inputs. Clearly, in all considerations this four-component vector of binary random variables can be replaced with a single 16-valued random variable,  $H'$ . Let these 16 values be  $0, \dots, 15$ . The two variables are equivalent if one puts

$$\begin{aligned} \Pr(H_1^1 = i, H_2^1 = i', H_1^2 = j, H_2^2 = j') \\ = \Pr(H' = (i - 1)2^3 + (i' - 1)2^2 + (j - 1)2 + (j' - 1)). \end{aligned}$$

In particular, any functions of  $H$  can be presented as functions of  $H'$ .

In the case of continuous random variables the situation is, in a sense, even simpler, although we will have to omit the underlying justification. It follows from the theory of Borel-equivalent spaces (which is part of descriptive set theory), that any vector of continuous random variables

$$R = (R^1, \dots, R^k),$$

can be presented as a function of any continuous variable  $R'$  with an *atomless* distribution on an interval of real numbers. The “atomlessness” means that the sigma-algebra of  $R'$  contains no null-set whose probability measure is not zero. Simple examples are uniformly and normally distributed random variables. If the vector is discrete, the previous statement applies with no modifications (although we know that in this case one can also choose a discrete  $R'$ ). It follows that the statement also applies to mixed vectors, containing both discrete and continuous random variables (or vectors thereof, or vectors of vectors thereof, etc.).

We can complement, therefore, Theorem 2.5 with the following statement.

**Theorem 2.6** *Under the conditions of Theorem 2.5, the random variable  $R$  can always be chosen to be any continuous random variable with an atomless distribution on an interval of real numbers. If all the random variables  $A_\phi^1, \dots, A_\phi^n$  are discrete (in particular, have finite numbers of values), then  $R$  can be chosen to be discrete (respectively, have finite number of values).*

We have quite a bit more specificity now than based on the initial definition of selective influences. And it is achieved due to the joint distribution criterion almost “automatically.”

Theorem 2.4 is not restricted to finite-valued inputs, nor is it restricted to a finite number of inputs, or to outputs of a specific kind. It is completely general. For the reader’s convenience, we formulate here the general version of this theorem, avoiding all elaborations.

**Theorem 2.7** (Joint Distribution Criterion (general version)) *Let  $(\lambda^k : k \in K)$  be an indexed family of inputs, with  $\lambda^k \in \Lambda^k \neq \emptyset$ , for all  $k \in K$ . Let  $\Phi \subset \prod_{k \in K} \Lambda^k$  be a nonempty set of allowable treatments. Let  $(A_\phi^k : k \in K)$  be a family of random variables jointly distributed for every  $\phi \in \Phi$ . Then*

$$(A^k : k \in K) \leftrightarrow (\lambda^k : k \in K)$$

*if and only if there exists an indexed family of jointly distributed random variables*

$$H = (H_{\lambda^k}^k : \lambda^k \in \Lambda^k, k \in K),$$

*(one variable per each value of each input) such that*

$$(A_\phi^k : k \in K) \sim (H_{\lambda^k}^k : k \in K)$$

*for every  $(\lambda^k : k \in K) = \phi \in \Phi$ .*

## 2.13 Properties of selective influences and tests

Certain properties of selective influences (in the canonical form) are immediately obvious.

The first one is *nestedness with respect to input values*: if random outputs  $A^1, \dots, A^n$  are selectively influenced by inputs  $\lambda^1, \dots, \lambda^n$ , with sets of possible values  $\Lambda^1, \dots, \Lambda^n$ , then the same random outputs are selectively influenced by inputs  $\lambda'^1, \dots, \lambda'^n$  whose sets of possible values are  $\Lambda'^1 \subset \Lambda^1, \dots, \Lambda'^n \subset \Lambda^n$ . Every variable is essentially the set of its possible values. Inputs are no exception. In fact, in a more rigorous development  $\lambda$  would be reserved for input values, whereas input themselves, considered as variables, would be identified by  $\Lambda$  (see Remark 2.7). When a set of an input's values changes, the input is being replaced by a new one. The nestedness property in question tells us that if the change consists in removing some of the possible values of some of the inputs, the selectiveness pattern established for the original inputs cannot be violated. This does not, of course, work in the other direction: if we augment  $\Lambda^1, \dots, \Lambda^n$  by adding to them new elements, then the initial pattern of selectiveness may very well disappear.

The second property is *nestedness with respect to inputs and outputs* (in a canonical diagram they are in a one-to-one correspondence): if a vector of random outputs is selectively influenced by a vector of inputs, then any subvector of the random outputs is selectively influenced by the corresponding subvector of the inputs. In symbols, if

$$(A^1, \dots, A^n) \leftrightarrow \rho (\lambda^1, \dots, \lambda^n)$$

and  $i_1, \dots, i_k \in \{1, \dots, n\}$ , then

$$(A^{i_1}, \dots, A^{i_k}) \leftrightarrow \rho (\lambda^{i_1}, \dots, \lambda^{i_k}).$$

Note that the set of allowable treatments has to be redefined whether we eliminate certain input-output pairs or certain input values. In the latter case, the new set of allowable treatments is the largest  $\Phi' \subset \Lambda'^1 \times \dots \times \Lambda'^n$ , such that  $\Phi' \subset \Phi$ . In the case we drop input-output pairs, the new set of allowable treatments is the largest  $\Phi'' \subset \Lambda^{i_1} \times \dots \times \Lambda^{i_k}$ , such that every  $\phi'' \in \Phi''$  is a part of some  $\phi \in \Phi$ .

Both these nestedness properties follow from the fact that any subset of random variables that are components of a reduced coupling vector

$$H = (H_1^1, \dots, H_{m_1}^1, \dots, H_1^n, \dots, H_{m_n}^n),$$

are also jointly distributed. When we eliminate an  $i$ th value of input  $k$ , we drop from this vector  $H_i^k$ . When we eliminate an input  $k$ , we drop the subvector  $H_1^k, \dots, H_{m_k}^k$ . In both cases the resulting  $H'$  is easily checked to be a reduced coupling vector for the redefined sets of treatments and outputs.

By similar arguments one can establish that a pattern of selective influences is well-behaved in response to all possible groupings of the inputs, with or without a

corresponding grouping of outputs: thus, if

$$(A^1, \dots, A^k, \dots, A^l, \dots, A^n) \leftrightarrow \varphi (\lambda^1, \dots, \lambda^k, \dots, \lambda^l, \dots, \lambda^n),$$

then

$$(A^1, \dots, A^k, \dots, A^l, \dots, A^n) \leftrightarrow \varphi (\lambda^1, \dots, (\lambda^k, \lambda^l), \dots, (\lambda^k, \lambda^l), \dots, \lambda^n)$$

and

$$\begin{aligned} & (A^1, \dots, (A^k, A^l), \dots, (A^k, A^l), \dots, A^n) \\ & \leftrightarrow \varphi (\lambda^1, \dots, (\lambda^k, \lambda^l), \dots, (\lambda^k, \lambda^l), \dots, \lambda^n). \end{aligned}$$

We omit the details related to redefinitions of allowable treatments.

A simple consequence of the nestedness with respect to input–output pairs turns out to be of a great importance for determining if a selectiveness pattern is present. This consequence is called *complete marginal selectivity*: if  $(A^1, \dots, A^n) \leftrightarrow \varphi (\lambda^1, \dots, \lambda^n)$  and  $i_1, \dots, i_k \in \{1, \dots, n\}$ , then the distribution of  $(A_\phi^{i_1}, \dots, A_\phi^{i_k})$  depends only on  $(\lambda^{i_1}, \dots, \lambda^{i_k})$ . In other words, if  $\phi$  and  $\phi'$  include the same subset  $(\lambda^{i_1}, \dots, \lambda^{i_k})$ , then

$$(A_\phi^{i_1}, \dots, A_\phi^{i_k}) \sim (A_{\phi'}^{i_1}, \dots, A_{\phi'}^{i_k}).$$

In particular (*simple marginal selectivity*),

$$A_\phi^i \sim A_{\phi'}^i$$

for any  $\phi$  and  $\phi'$  that share a value of  $\lambda^i$  ( $i = 1, \dots, n$ ). The importance of marginal selectivity is that it is easy to check, ruling out selective influences whenever it is found violated.

**Example 2.20** Let  $A^1, A^2$  have values in  $\{1, 2\}$  and depend on the external factors  $\lambda^1 \in \Lambda^1 = \{1, 2\}$  and  $\lambda^2 \in \Lambda^2 = \{1, 2\}$ . Let the joint distribution of  $A^1, A^2$  for each treatment (all four being allowable) be as follows:

$\lambda^1 = 1, \lambda^2 = 1$	$A^2 = 1$	$A^2 = 2$		$\lambda^1 = 1, \lambda^2 = 2$	$A^2 = 1$	$A^2 = 2$	
$A^1 = 1$	.2	.2	.4		.3	.1	.4
$A^1 = 2$	.3	.3	.6		.2	.4	.6
	.5	.5			.5	.5	

$\lambda^1 = 2, \lambda^2 = 1$	$A^2 = 1$	$A^2 = 2$		$\lambda^1 = 2, \lambda^2 = 2$	$A^2 = 1$	$A^2 = 2$	
$A^1 = 1$	.4	.3	.7		.3	.4	.7
$A^1 = 2$	.1	.2	.3		.1	.2	.3
	.5	.5			.4	.6	

Marginal selectivity here is violated because the marginal distribution of  $A^2$  changes when  $\lambda^2 = 2$  and  $\lambda^1$  changes from 1 to 2.  $\square$

Marginal selectivity is strictly weaker than selective influences. The latter do imply marginal selectivity, but marginal selectivity can very well hold in the absence of selective influences.

**Example 2.21** Consider the following joint distributions:

$\lambda^1 = 1, \lambda^2 = 1$		$A^2 = 1 \ A^2 = 2$		$\lambda^1 = 1, \lambda^2 = 2$		$A^2 = 1 \ A^2 = 2$	
$A^1 = 1$	.5	0	.5	$A^1 = 1$	.5	0	.5
$A^1 = 2$	0	.5	.5	$A^1 = 2$	0	.5	.5
	.5	.5			.5	.5	
$\lambda^1 = 2, \lambda^2 = 1$		$A^2 = 1 \ A^2 = 2$		$\lambda^1 = 2, \lambda^2 = 2$		$A^2 = 1 \ A^2 = 2$	
$A^1 = 1$	.5	0	.5	$A^1 = 1$	0	.5	.5
$A^1 = 2$	0	.5	.5	$A^1 = 2$	.5	0	.5
	.5	.5			.5	.5	

Marginal selectivity is trivially satisfied as all marginals are uniform. However,  $(A^1, A^2) \not\sim (\lambda^1, \lambda^2)$  in this case. The joint distribution criterion would require the existence of a jointly distributed vector  $H$  whose components satisfy  $(A_{ij}^1, A_{ij}^2) \sim (H_i^1, H_j^2)$  for  $i, j \in \{1, 2\}$ . However, combining this with the above joint distributions, we obtain

$$H_1^1 = H_1^2, \quad H_1^1 = H_2^2, \quad H_2^1 = H_1^2, \quad H_2^1 = 3 - H_2^2,$$

which yields the contradiction

$$3 - H_2^2 = H_2^2.$$

□

Another property of selective influences is that if  $(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$ , and if, for all  $\phi = (\lambda^1, \dots, \lambda^n) \in \Phi$ ,

$$B_\phi^1 = g_1(\lambda^1, A_\phi^1), \dots, B_\phi^n = g_n(\lambda^n, A_\phi^n),$$

then  $(B^1, \dots, B^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$ . The functions  $g_1, \dots, g_n$  are referred to as *input-value-specific transformations of random outputs*. The property in question, therefore, is the invariance of selective influences, if established, with respect to such transformations.

Let us make sure that this property is true. According to the general definition, we have a random variable  $R$  and functions  $f_1, \dots, f_n$  such that

$$(A_\phi^1, \dots, A_\phi^n) \sim (f_1(\lambda^1, R), \dots, f_n(\lambda^n, R)),$$

for every  $\phi = (\lambda^1, \dots, \lambda^n) \in \Phi$ . But then

$$(B_\phi^1, \dots, B_\phi^n) \sim (g_1(\lambda^1, f_1(\lambda^1, R)), \dots, g_n(\lambda^n, f_n(\lambda^n, R))),$$

and every  $g_k(\lambda^k, f_k(\lambda^k, R))$  is some function  $f_k^*(\lambda^k, R)$ . The vectors  $(B_\phi^1, \dots, B_\phi^n)$  therefore satisfy the definition too.

As a special case, the transformation may not depend on input values,

$$B_\phi^1 = g_1(A_\phi^1), \dots, B_\phi^n = g_n(A_\phi^n).$$

This would include all possible *renamings* and *groupings* of the values of the random outputs: a pattern of selective influences is preserved under all such transformations. For instance, one can rename values 1, 2 of a binary output into  $\sqcup, \sqcap$ , or one can group values 1, 2, 3, 4 into “cruder” values, by means of a transformation like

$$1 \mapsto \sqcup, 2 \mapsto \sqcup, 3 \mapsto \sqcap, 4 \mapsto \sqcap.$$

The meaning of the input-value-specificity is this. We choose a  $k \in \{1, \dots, n\}$  and assume, for simplicity, that  $\lambda^k$  has discrete values, 1, 2, ... Let  $A_\phi^k$  be transformed into random variables  $B_{1,\phi}^k, B_{2,\phi}^k$ , etc., all sharing the same set of possible values and the same sigma-algebra. We know that one can replace  $A^k$  in

$$(A^1, \dots, A^k, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^k, \dots, \lambda^n)$$

with any of these new random variables,

$$\begin{aligned} (A^1, \dots, B_{1,\phi}^k, \dots, A^n) &\leftrightarrow (\lambda^1, \dots, \lambda^k, \dots, \lambda^n), \\ (A^1, \dots, B_{2,\phi}^k, \dots, A^n) &\leftrightarrow (\lambda^1, \dots, \lambda^k, \dots, \lambda^n), \\ &\text{etc.} \end{aligned}$$

The input-value-specificity is involved if one forms a random variable

$$B_\phi^k = \begin{cases} B_{1,\phi}^k & \text{if } \lambda^k = 1 \\ B_{2,\phi}^k & \text{if } \lambda^k = 2 \\ \text{etc.} & \end{cases}$$

The invariance property says that this random variable, too, can replace  $A^k$  in a pattern of selective influences,

$$(A^1, \dots, B^k, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^k, \dots, \lambda^n).$$

Note that the property in question works in one direction only: if  $(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$  then  $(B^1, \dots, B^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$ . It is perfectly possible (if we use grouping of values) that  $(A^1, \dots, A^n) \not\leftrightarrow (\lambda^1, \dots, \lambda^n)$  but following an input-value-specific transformation,  $(B^1, \dots, B^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$ . However, if the transformation  $B_\phi^1 = g_1(\lambda^1, A_\phi^1), \dots, B_\phi^n = g_n(\lambda^n, A_\phi^n)$ , is reversible, that is, there exists another transformation  $A_\phi^1 = h_1(\lambda^1, B_\phi^1), \dots, A_\phi^n = h_n(\lambda^n, B_\phi^n)$  back to the original variables, then  $(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$  if and only if  $(B^1, \dots, B^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$ .

**Example 2.22** Consider the random variables  $A^1, A^2$  with values in {1, 2}, depending on the input factors  $\lambda^1 \in \{1, 2\}$ ,  $\lambda^2 \in \{1, 2\}$ , and having the following joint distributions at the four possible treatments:

$\lambda^1 = 1, \lambda^2 = 1$		$A^2 = 1$	$A^2 = 2$	$\lambda^1 = 1, \lambda^2 = 2$		$A^2 = 1$	$A^2 = 2$
$A^1 = 1$	0.3	0.4	0.7	$A^1 = 1$	0.35	0.35	0.7
$A^1 = 2$	0.1	0.2	0.3	$A^1 = 2$	0.15	0.15	0.3
	0.4	0.6			0.5	0.5	
$\lambda^1 = 2, \lambda^2 = 1$		$A^2 = 1$	$A^2 = 2$	$\lambda^1 = 2, \lambda^2 = 2$		$A^2 = 1$	$A^2 = 2$
$A^1 = 1$	0.32	0.48	0.8	$A^1 = 1$	0.45	0.35	0.8
$A^1 = 2$	0.08	0.12	0.2	$A^1 = 2$	0.05	0.15	0.2
	0.4	0.6			0.5	0.5	

We will see in the next section that  $(A^1, A^2) \leftrightarrow \rho(\lambda^1, \lambda^2)$  is satisfied in this case. Let us define the input-value-specific transformations  $B^1 = g_1(\lambda^1, A^1)$  and  $B^2 = g_2(\lambda^2, A^2)$ , where

$$\begin{aligned} g_1(1, \{1, 2\}) &= \{+1, -1\}, & g_2(1, \{1, 2\}) &= \{7, 3\}, \\ g_1(2, \{1, 2\}) &= \{-1, +1\}, & g_2(2, \{1, 2\}) &= \{3, 7\}. \end{aligned}$$

As we see,  $A^1 = 1$  is mapped into  $B^1 = +1$  or  $B^1 = -1$  according as  $\lambda^1$  is 1 or 2,  $A^2 = 1$  is mapped into  $B^2 = 7$  or  $B^2 = 3$  according as  $\lambda^2$  is 1 or 2, etc. We obtain the following joint distributions

$\lambda^1 = 1, \lambda^2 = 1$		$B^2 = 7$	$B^2 = 3$	$\lambda^1 = 1, \lambda^2 = 2$		$B^2 = 7$	$B^2 = 3$
$B^1 = +1$	0.3	0.4	0.7	$B^1 = +1$	0.35	0.35	0.7
$B^1 = -1$	0.1	0.2	0.3	$B^1 = -1$	0.15	0.15	0.3
	0.4	0.6			0.5	0.5	
$\lambda^1 = 2, \lambda^2 = 1$		$B^2 = 7$	$B^2 = 3$	$\lambda^1 = 2, \lambda^2 = 2$		$B^2 = 7$	$B^2 = 3$
$B^1 = +1$	0.08	0.12	0.2	$B^1 = +1$	0.15	0.05	0.2
$B^1 = -1$	0.32	0.48	0.8	$B^1 = -1$	0.35	0.45	0.8
	0.4	0.6			0.5	0.5	

We know that the transformed variables satisfy  $(B^1, B^2) \leftrightarrow \rho(\lambda^1, \lambda^2)$  because  $(A^1, A^2) \leftrightarrow \rho(\lambda^1, \lambda^2)$ .  $\square$

In the subsequent sections we will consider several *tests of selective influences*. Such a test is always a statement whose truth value (whether it is true or false) determines whether a given pattern of selective influences holds or does not hold. The truth value of the test statement must be determinable from the distributions of  $(A_\phi^1, \dots, A_\phi^n)$  for all allowable  $\phi$ . If its truth implies  $(A^1, \dots, A^n) \leftrightarrow \rho(\lambda^1, \dots, \lambda^n)$ , then the test provides a sufficient condition for selective influences; if its falsity implies  $(A^1, \dots, A^n) \not\leftrightarrow \rho(\lambda^1, \dots, \lambda^n)$ , then the test provides a necessary condition for selective influences. If the test provides both necessary and sufficient condition, it is a criterion.

The distribution of  $(A_\phi^1, \dots, A_\phi^n)$ , if the random variables are known from their observed realizations, cannot be known precisely, because probabilities are never observable. All our tests require that the distributions of  $(A_\phi^1, \dots, A_\phi^n)$ , or at least some parameters thereof, be known precisely. Therefore, they can only be applied to empirical observations if the latter are replaced by theoretical distributions. This

can be done based on statistical considerations, outside the scope of the tests themselves. In particular, if all sample sizes are sufficiently large, theoretical distributions can be assumed to be so close to the empirical ones that their difference cannot affect the outcome of a test.

As follows from the discussion above, the most basic and obvious test of selective influences is the (complete) *marginal selectivity test*. This is a necessary condition for selective influences: if, at least for one pair of distinct treatments  $\phi$  and  $\phi'$  that include one and the same subvector  $(\lambda^{i_1}, \dots, \lambda^{i_k})$ , the distributions of the  $k$ -marginal random variables  $(A_\phi^{i_1}, \dots, A_\phi^{i_k})$  and  $(A_{\phi'}^{i_1}, \dots, A_{\phi'}^{i_k})$  are not the same, then  $(A^1, \dots, A^n) \not\sim (\lambda^1, \dots, \lambda^n)$ .

## 2.14 Linear feasibility test

In this section we will discuss a test which is both a necessary and sufficient condition for the selective influences in the case when the number of input-output pairs, the set of values of each input, and the set of possible values of each random output are all finite. Let us enumerate, for  $k = 1, \dots, n$ , the values of each input  $\lambda^k$  as  $1, \dots, m_k$ , and the values of each random output  $A^k$  as  $1, \dots, v_k$ . In Section 2.12 we discussed the case  $n = 2$ ,  $m_1 = m_2 = 2$ , and  $v_1 = v_2 = 2$ . We determined there that the question of whether  $(A^1, A^2) \leftrightarrow (\lambda^1, \lambda^2)$  translates into a question of whether certain linear equations have a solution subject to certain constraints. We will see that this is the case generally.

The observable distributions of  $(A_\phi^1, \dots, A_\phi^n)$  are represented by the probabilities of the events that can be described as

$$\left( \overbrace{A^1 = a_1, \dots, A^k = a_k, \dots, A^n = a_n}^{\text{output values}}, \overbrace{\lambda^1 = l_1, \dots, \lambda^k = l_k, \dots, \lambda^n = l_n}^{\text{input values}} \right),$$

where  $a_k \in \{1, \dots, v_k\}$  (output values) and  $l_k \in \{1, \dots, m_k\}$  (input values). Let us form a matrix  $M$  whose rows are enumerated (labeled) by all such vectors. We only consider the vectors with allowable treatments,

$$\phi = (\lambda^1 = l_1, \dots, \lambda^k = l_k, \dots, \lambda^n = l_n) \in \Phi.$$

If the number of the allowable treatments is  $t$  (between 1 and  $m_1 \times \dots \times m_n$ ), then the number of the rows in  $M$  is  $t \times v_1 \times \dots \times v_n$ .

The columns of the matrix  $M$  are enumerated (labeled) by the vectors of the form

$$\left( \overbrace{H_1^1 = h_1^1, \dots, H_{m_1}^1 = h_{m_1}^1}^{\text{coupling vector } H}, \dots, \overbrace{H_1^n = h_1^n, \dots, H_{m_n}^n = h_{m_n}^n}^{\text{coupling vector } H} \right),$$

where  $h_i^k \in \{1, \dots, v_k\}$ . Such vectors represent events whose probabilities define the distribution of a reduced coupling vector  $H$  (if one exists). The number of such

events, hence the number of the columns in  $M$  is  $(v_1)^{m_1} \times \cdots \times (v_n)^{m_n}$  (where the superscripts represent conventional exponents).

We also form a column vector  $P$  whose elements are labeled in the same way and in the same order as the rows of the matrix  $M$ , and a column vector  $Q$  whose elements are labeled in the same way and in the same order as the columns of the matrix  $M$ .

Let us now fill in the entries of the vectors  $P$ ,  $Q$ , and the matrix  $M$ . The matrix  $M$  is Boolean: it is filled with 1's and 0's. Consider a cell  $(I, J)$  belonging to the column labeled

$$J = \left( \overbrace{H_1^1 = h_1^1, \dots, H_{m_1}^1 = h_{m_1}^1}^{}, \dots, \overbrace{H_1^n = h_1^n, \dots, H_{m_n}^n = h_{m_n}^n}^{} \right)$$

and to the row labeled

$$I = \left( \overbrace{A^1 = a_1, \dots, A^k = a_k, \dots, A^n = a_n}^{}; \overbrace{\lambda^1 = l_1, \dots, \lambda^k = l_k, \dots, \lambda^n = l_n}^{} \right).$$

In the vector-label  $J$  pick the entries

$$H_{l_1}^1 = h_{l_1}^1, \dots, H_{l_k}^k = h_{l_k}^k, \dots, H_{l_n}^n = h_{l_n}^n$$

corresponding to the values of  $(\lambda^1, \dots, \lambda^n)$  indicated in the vector-label  $I$ . If

$$(h_{l_1}^1, \dots, h_{l_k}^k, \dots, h_{l_n}^n) = (a_1, \dots, a_k, \dots, a_n)$$

then the cell  $(I, J)$  should be filled with 1; otherwise its value is 0.

The vector  $P$  is filled with the probabilities

$$\Pr(A^1 = a_1, \dots, A^n = a_n)_{\phi=(\lambda^1=l_1, \dots, \lambda^n=l_n)}.$$

For any allowable  $\phi$ , the probabilities across all possible combinations of  $(a_1, \dots, a_n)$  sum to 1. These probabilities are assumed to be known.

The vector  $Q$  is filled with the probabilities

$$\Pr(H_1^1 = h_1^1, \dots, H_{m_1}^1 = h_{m_1}^1, \dots, H_1^n = h_1^n, \dots, H_{m_n}^n = h_{m_n}^n),$$

which sum to 1 across all possible values of  $(h_1^1, \dots, h_{m_1}^1, \dots, h_1^n, \dots, h_{m_n}^n)$ . These probabilities are not known, they have to be found or determined not to exist.

**Example 2.23** Let us now apply these general definitions to the simplest nontrivial case  $n = 2$ ,  $m_1 = m_2 = 2$ ,  $v_1 = v_2 = 2$  considered in Section 2.12. The matrix  $M$  filled with binary values is (replacing 0 with “.” for better legibility)

	$H_1^1$	1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2
	$H_2^1$	1 1 1 1 2 2 2 2 1 1 1 1 2 2 2 2
	$H_1^2$	1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2
	$H_2^2$	1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
	$A^1 = 1, A^2 = 1$	1 1 . . 1 1 . . . . . . . . .
$\lambda^1 = 1, \lambda^2 = 1$	$A^1 = 1, A^2 = 2$	. . 1 1 . . 1 1 . . . . . . . .
	$A^1 = 2, A^2 = 1$	. . . . . . . . 1 1 . . 1 1 . .
	$A^1 = 2, A^2 = 2$	. . . . . . . . . . 1 1 . . 1 1
	$A^1 = 1, A^2 = 1$	1 . 1 . 1 . 1 . . . . . . . . .
$\lambda^1 = 1, \lambda^2 = 2$	$A^1 = 1, A^2 = 2$	. 1 . 1 . 1 . 1 . . . . . . . .
	$A^1 = 2, A^2 = 1$	. . . . . . . . . 1 . 1 . 1 . 1 .
	$A^1 = 2, A^2 = 2$	. . . . . . . . . . 1 . 1 . 1 . 1
	$A^1 = 1, A^2 = 1$	1 1 . . . . . . 1 1 . . . . . . .
$\lambda^1 = 2, \lambda^2 = 1$	$A^1 = 1, A^2 = 2$	. . 1 1 . . . . . . 1 1 . . . .
	$A^1 = 2, A^2 = 1$	. . . . . 1 1 . . . . . . 1 1 . .
	$A^1 = 2, A^2 = 2$	. . . . . . 1 1 . . . . . . 1 1
	$A^1 = 1, A^2 = 1$	1 . 1 . . . . . 1 . 1 . . . . . .
$\lambda^1 = 2, \lambda^2 = 2$	$A^1 = 1, A^2 = 2$	. 1 . 1 . . . . . 1 . 1 . . . .
	$A^1 = 2, A^2 = 1$	. . . . . 1 . 1 . . . . . . 1 . 1 .
	$A^1 = 2, A^2 = 2$	. . . . . . 1 . 1 . . . . . . 1 . 1

The vector  $P$  consists of the observed probabilities corresponding to the row labels of the matrix, and the vector  $Q$  consists of the joint probabilities of the coupling vector  $H = (H_1^1, H_2^1, H_1^2, H_2^2)$  as indicated in the column labels of the matrix. Using the observed probabilities of Example 2.22 we obtain

$$P = [.3, .4, .1, .2, .35, .35, .15, .15, .08, .12, .32, .48, .15, .05, .35, .45]^T.$$

□

**Theorem 2.8** *If the sets of values for all  $n$  inputs and all  $n$  random outputs are finite, then, using the notation of this section,*

$$(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$$

*holds if and only if the system of linear equations*

$$MQ = P$$

*has a solution  $Q \geq 0$  (the inequality meaning that the elements of  $Q$  are nonnegative).*

Without the nonnegativity constraint, the system  $MQ = P$  always has solutions, because the number of the unknowns (elements of  $Q$ ) equals or exceeds the rank of the matrix  $M$ , which can be shown to never exceed

$$(m_1(v_1 - 1) + 1) \times \dots \times (m_n(v_n - 1) + 1).$$

Moreover, the structure of the matrix  $M$  is such that any solution for  $Q$  should automatically have its elements summing to 1. The latter, therefore, is not a constraint. However, it is not guaranteed that  $Q \geq 0$ : it is possible that all solutions

for  $Q$  have some of the elements negative, in which case our test establishes that  $(A^1, \dots, A^n) \not\leftrightarrow (\lambda^1, \dots, \lambda^n)$ .

Let us introduce a function

$$\text{Sol}(M, P)$$

that attains two values: “True,” if  $MQ = P$  has a nonnegative solution, and “False,” if such a solution does not exist. Note that  $M$  is an argument that is determined uniquely by the format of the problem: the number of input–output pairs and number of possible values for inputs and outputs. The task of computing  $\text{Sol}(M, P)$  is a standard *feasibility problem* of the area of linear algebra called linear programming. Due to this term, the test in question is called the *linear feasibility test*,

$$(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n) \text{ if and only if } \text{Sol}(M, P) = \text{True}.$$

It is known from linear programming that  $\text{Sol}(M, P)$  can always be computed.

**Example 2.24** Let us apply the linear feasibility test to the matrix  $M$  and vector  $P$  of Example 2.23. Using the simplex linear programming algorithm, we obtain the solution

$$Q = [.03, 0, 0, 0, 0, .27, .32, .08, 0, .05, .12, 0, 0, .05, .03, .05]^T \geq 0$$

satisfying  $MQ = P$ . This means that  $\text{Sol}(M, P) = \text{“True”}$ , hence  $(A^1, A^2) \leftrightarrow (\lambda^1, \lambda^2)$ .

The (complete) marginal selectivity test mentioned in the previous section is part of the linear feasibility test. If the former is violated, so will also the latter. It follows from the structure of the matrix  $M$ , as explained in the following example.

**Example 2.25** Consider the matrix of Example 2.23. If  $(A^1, A^2) \leftrightarrow (\lambda^1, \lambda^2)$  is satisfied for a given vector  $P$  of observed probabilities, then we know that there exists a vector  $Q \geq 0$  such that  $MQ = P$ . The marginal probabilities of  $A^1$  and  $A^2$  within each treatment are obtained by summing certain elements of  $P$ . However, as  $MQ = P$ , we can obtain these marginal probabilities also by summing certain rows of  $M$  and then multiplying these summed rows by  $Q$ . Thus, if we sum the rows of  $M$  corresponding to the same value of  $A^1$  within each treatment, we obtain

$\lambda^1 = 1, \lambda^2 = 1$	$A^1 = 1$	1 1 1 1 1 1 1 1 . . . . . . . .
	$A^1 = 2$	. . . . . . . . 1 1 1 1 1 1 1 1 1
$\lambda^1 = 1, \lambda^2 = 2$	$A^1 = 1$	1 1 1 1 1 1 1 1 . . . . . . . .
	$A^1 = 2$	. . . . . . . . 1 1 1 1 1 1 1 1 1
$\lambda^1 = 2, \lambda^2 = 1$	$A^1 = 1$	1 1 1 1 . . . 1 1 1 1 . . .
	$A^1 = 2$	. . . . 1 1 1 1 . . . 1 1 1 1
$\lambda^1 = 2, \lambda^2 = 2$	$A^1 = 1$	1 1 1 1 . . . 1 1 1 1 . . .
	$A^1 = 2$	. . . . 1 1 1 1 . . . 1 1 1 1

As the rows corresponding to the marginal probabilities of  $A^1$  are identical between the treatments with  $\lambda^1 = 1$  and between the treatments with  $\lambda^1 = 2$ , we see that the

marginal distribution of  $A^1$  does not depend on  $\lambda_2$ . If we then sum the rows of  $M$  corresponding to the same value of  $A^2$  within each treatment, we obtain

$\lambda^1 = 1, \lambda^2 = 1$	$A^2 = 1$	1 1 . . 1 1 . . 1 1 . . 1 1 . .
	$A^2 = 2$	. . 1 1 . . 1 1 . . 1 1 . . 1 1
$\lambda^1 = 1, \lambda^2 = 2$	$A^2 = 1$	1 . 1 . 1 . 1 . 1 . 1 . 1 . 1 .
	$A^2 = 2$	. 1 . 1 . 1 . 1 . 1 . 1 . 1 . 1 .
$\lambda^1 = 2, \lambda^2 = 1$	$A^2 = 1$	1 1 . . 1 1 . . 1 1 . . 1 1 . .
	$A^2 = 2$	. . 1 1 . . 1 1 . . 1 1 . . 1 1
$\lambda^1 = 2, \lambda^2 = 2$	$A^2 = 1$	1 . 1 . 1 . 1 . 1 . 1 . 1 . 1 .
	$A^2 = 2$	. 1 . 1 . 1 . 1 . 1 . 1 . 1 . 1 .

and we can see that the marginal distribution of  $A^2$  does not depend on  $\lambda^1$ . Thus, linear feasibility test includes the test for marginal selectivity, so if the latter is violated, the former fails.  $\square$

One may feel that  $\text{Sol}(M, P)$  is not a “true” function, as it requires a computer algorithm to be computed, and it is not presented in an analytic form. Such a misgiving is not well-founded. An analytic (or closed-form) solution is merely one that can be presented in terms of familiar functions and operations. For example, if a solution of a problem involves the standard normal integral

$$N(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t \exp(-z^2/2) dz,$$

the solution may or may not be called analytic depending on how familiar and easily computable this function is. In the past,  $N(t)$  could be viewed as “less analytic” than  $\exp(x)$ , and in Napier’s time  $\exp(x)$  would be viewed as “less analytic” than  $x^2$ . Familiarity is not a mathematical category, and the existence of a rigorous definition of a function combined with an algorithm allowing one to compute it to a desired level of precision is all one needs to use it in a solution to a problem. The computational complexity, of course, may be a concern. In our case, however, it is known that as the size of the matrix  $M$  increases, the computational time required to compute  $\text{Sol}(M, P)$  increases only as a polynomial function of this size (rather than exponentially or even faster). This makes the linear feasibility test practical.

It still may be of interest to see whether the linear feasibility test could be formulated in terms of a system of equalities and inequalities involving the entries of the vector  $P$  alone. This can always be achieved, with every linear feasibility problem. These equalities and inequalities, in fact, can be generated by a computer algorithm (called a *facet enumeration algorithm*).

**Example 2.26** Geometrically, the linear feasibility test checks if  $P$  is within the *convex polytope* determined by points  $MQ$  such that  $Q \geq 0$ ,  $\sum Q = 1$ . The columns of  $M$  correspond to the vertices of this polytope. A facet enumeration

algorithm transforms this *vertex representation* of the polytope to the so-called *half-plane representation*, that is, to a representation of the form

$$M_1 P \geq Q_1, \quad M_2 P = Q_2,$$

where  $M_1, M_2$  are matrices and  $Q_1, Q_2$  are vectors. For our  $16 \times 16$  example matrix, this yields

$$M_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \quad Q_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

and

$$M_2 = \begin{bmatrix} 1 & 1 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & -1 & 0 & -1 & 0 & -1 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \quad Q_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

The equations  $M_2 P = Q_2$  of this representation always correspond to the marginal selectivity constraints. Thus, a vector  $P$  of observed probabilities satisfying marginal selectivity satisfies selective influences if and only if  $M_1 P \geq Q_1$ . Assuming marginal selectivity,  $M_1 P \geq Q_1$  can in this case also be simplified into the four double-inequalities

$$0 \leq p_{i\cdot} + p_{\cdot j} + p_{i'j'} - p_{ij'} - p_{i'j} - p_{i'j'} \leq 1, \quad i \neq i', j \neq j',$$

where we denote

$$\begin{aligned} p_{i\cdot} &= \Pr(A^1 = 1)_{\phi=(\lambda^1=i, \lambda^2=\cdot)}, \\ p_{\cdot j} &= \Pr(A^2 = 1)_{\phi=(\lambda^1=\cdot, \lambda^2=j)}, \\ p_{ij} &= \Pr(A^1 = 1, A^2 = 1)_{\phi=(\lambda^1=i, \lambda^2=j)} \end{aligned}$$

(the definition of  $p_{i\cdot}$  and  $p_{\cdot j}$  presupposes marginal selectivity). These are known as the *Bell/CHSH/Fine inequalities* in quantum mechanics.  $\square$

In the same way, the representation as inequalities can be obtained for any linear feasibility test matrix  $M$ . It should be noted, however, that the number of the inequalities increases explosively as the size of the matrix  $M$  increases. Thus, for three pairs of completely crossed binary inputs and three binary random outputs, the number of independent equalities representing marginal selectivity is 42, and the number of inequalities is 53792. From a practical point of view, therefore, computing  $\text{Sol}(M, P)$  directly is a better approach in all but the simplest cases.

## 2.15 Distance tests

Let us establish some general terminology. A *pseudo-quasi-metric* (or *p.q.-metric*, for short) on a nonempty set  $X$  is defined as a function  $d : X \times X \rightarrow \mathbb{R}^+$  (set of nonnegative real numbers), such that, for any  $x, y, z \in X$ ,

- (1) (zero property)  $d(x, x) = 0$ ,
- (2) (triangle inequality)  $d(x, y) + d(y, z) \geq d(x, z)$ .

A p.q.-metric that satisfies, in addition,

- (3) (symmetry)  $d(x, y) = d(y, x)$ ,

is called a *pseudo-metric*. A p.q.-metric that satisfies

- (4) (positivity) if  $x \neq y$ , then  $d(x, y) > 0$ ,

is called a *quasi-metric*. Finally, a p.q.-metric that satisfies both (3) and (4) is called a *metric*. The terminology is not well-established and varies from one area or application to another.

**Remark 2.11** To refer to the value  $d(x, y)$  of a metric, pseudo-metrics, quasi-metrics, or a p.q.-metric at a specific pair of points  $(x, y)$ , one usually uses the generic term “distance,” adding the corresponding prefixes (pseudo, quasi, or p.q.) only if it is required for disambiguation. Thus, the value of a p.q.-metric for a specific pair  $(x, y)$  can be called the distance from  $x$  to  $y$ , or the p.q.-distance from  $x$  to  $y$ . (For pseudo-metrics, “from  $x$  to  $y$ ” can be replaced with “between  $x$  and  $y$ .”) The term “distance” can also be used (with or without the prefixes) to refer to the functions themselves. Therefore “p.q.-metric tests” below can also be referred to as “distance tests” or “p.q.-distance tests.”

The nature of the set  $X$  in the definition is entirely arbitrary. We are interested in a set of jointly distributed random variables, that is, those representable as functions of one and the same random variable. A p.q.-metric on such a set is a function  $d$  mapping pairs of random variables into nonnegative real numbers, such that  $d(R, R) = 0$  and  $d(R^1, R^2) + d(R^2, R^3) \geq d(R^1, R^3)$ , for any random variables  $R^1, R^2, R^3$  in the set. We assume that  $d(R^1, R^2)$  is entirely determined by the joint distribution of  $(R^1, R^2)$ . In other words, it does not depend on the identifying label of the pair (or on how  $R^1$  and  $R^2$  are presented as functions of a common random variable).

An immediate consequence (and generalization) of the triangle inequality is the following *chain inequality*: if  $R^1, \dots, R^l$  are elements of  $X$  ( $l \geq 3$ ), not necessarily distinct, then

$$d(R^1, R^l) \leq \sum_{i=2}^l d(R^{i-1}, R^i).$$

This inequality, as it turns out, can be utilized to construct tests of selective influences.

Suppose that the random outputs  $(A_\phi^1, \dots, A_\phi^n)$  across all  $\phi \in \Phi$  belong to a certain type, or class of random variables (e.g., those in the narrow sense, or with a finite number of values, etc.). We continue to consider, for simplicity, inputs with finite number of values each. We know that  $(A^1, \dots, A^n) \leftrightarrow^\rho (\lambda^1, \dots, \lambda^n)$  if and only if there exists a reduced coupling vector  $H$ . Assuming that it does exist, its elements are of the same type, or class, as  $(A_\phi^1, \dots, A_\phi^n)$ , and any  $l \geq 3$  of these elements,

$$H_{j_1}^{k_1}, H_{j_2}^{k_2}, \dots, H_{j_l}^{k_l}$$

can be used to form a chain inequality,

$$d(H_{j_1}^{k_1}, H_{j_l}^{k_l}) \leq \sum_{i=2}^l d(H_{j_{i-1}}^{k_{i-1}}, H_{j_i}^{k_i}).$$

Let us choose these elements of  $H$  so that  $\lambda^{k_1} = j_1$  and  $\lambda^{k_l} = j_l$  belong to some allowable treatment  $\phi_{1k}$ , and each pair  $\lambda^{k_{i-1}} = j_{i-1}, \lambda^{k_i} = j_i$  belongs to some allowable treatment  $\phi_{i-1,i}$  ( $i = 2, \dots, l$ ). The allowable treatments  $\phi_{1k}, \phi_{12}, \dots, \phi_{l-1,l}$  need not be pairwise distinct. Such a sequence of input values,

$$\lambda^{k_1} = j_1, \lambda^{k_2} = j_2, \dots, \lambda^{k_l} = j_l$$

is called *treatment-realizable*. This choice ensures that

$$(H_{j_1}^{k_1}, H_{j_l}^{k_l}) \sim (A_{\phi_{1k}}^{k_1}, A_{\phi_{1l}}^{k_l})$$

and

$$(H_{j_{i-1}}^{k_{i-1}}, H_{j_i}^{k_i}) \sim (A_{\phi_{i-1,i}}^{k_{i-1}}, A_{\phi_{i-1,i}}^{k_i}), \text{ for } i = 2, \dots, l.$$

But then

$$d(H_{j_1}^{k_1}, H_{j_l}^{k_l}) = d(A_{\phi_{1k}}^{k_1}, A_{\phi_{1l}}^{k_l})$$

and

$$d(H_{j_{i-1}}^{k_{i-1}}, H_{j_i}^{k_i}) = d(A_{\phi_{i-1,i}}^{k_{i-1}}, A_{\phi_{i-1,i}}^{k_i}), \text{ for } i = 2, \dots, l,$$

whence the chain inequality can be rewritten using only observable pairwise distributions,

$$d(A_{\phi_{1k}}^{k_1}, A_{\phi_{1l}}^{k_l}) \leq \sum_{i=2}^l d(A_{\phi_{i-1,i}}^{k_{i-1}}, A_{\phi_{i-1,i}}^{k_i}).$$

This inequality is a necessary condition for the existence of  $H$ . If it is found violated for at least one treatment-realizable sequence of input values, then the existence of  $H$  is ruled out, and one should conclude that  $(A^1, \dots, A^n) \not\leftrightarrow^\rho (\lambda^1, \dots, \lambda^n)$ .

There are numerous ways of constructing p.q.-metrics for jointly distributed random variables. We will confine our consideration to only two examples.

If all random outputs have one and the same set of possible values  $S$ , then one way of creating a p.q.-metric on a set  $X$  of such random variables is to use any p.q.-metric  $D$  on  $S$  and put, for any random variables  $Q, R \in X$ ,

$$d(Q, R) = E[D(Q, R)].$$

The right-hand expression is the expected value of the random variable  $D(Q, R)$ . The underlying assumption is, of course, that this random variable is well-defined (that is,  $D$  is a measurable function from  $S \times S$  to nonnegative real numbers), and that its expectation is finite. It can easily be proved then that  $d$  is a p.q.-metric on  $X$ .

As a simple example, consider the p.q.-metric

$$D(x, y) = \begin{cases} |x - y|^p & \text{if } x < y \\ 0 & \text{otherwise} \end{cases}$$

on the set of real numbers, with  $0 \leq p \leq 1$  (a power exponent). It is a p.q.-metric because  $D(x, x) = 0$ , and

$$D(x, y) + D(y, z) \geq D(x, z),$$

as one can prove by considering various arrangements of numbers  $x, y, z$ . Using  $D$ , one can construct a p.q.-metric for any set  $X$  of random variables whose (common) set of possible values is a subset of reals. Let this set be a subset of integers. Then the p.q.-metric on  $X$  derived from  $D$  is

$$d_p(Q, R) = \sum_{q < r} |q - r|^p p(q, r),$$

where

$$p(q, r) = \Pr(Q = q, R = r).$$

**Example 2.27** Let the outputs  $A^1, A^2$  have the following distributions for treatments in  $\Lambda^1 \times \Lambda^2 = \{1, 2\} \times \{1, 2\}$ :

$\lambda^1 = 1, \lambda^2 = 1$			
	$A^2 = 0$	$A^2 = 1$	$A^2 = 2$
$A^1 = 0$	.24	.07	0
$A^1 = 2$	.07	.24	.07
$A^1 = 4$	0	.07	.24

$\lambda^1 = 1, \lambda^2 = 2$			
	$A^2 = 0$	$A^2 = 1$	$A^2 = 2$
$A^1 = 0$	.24	.07	0
$A^1 = 2$	.07	.24	.07
$A^1 = 4$	0	.07	.24

$\lambda^1 = 2, \lambda^2 = 1$			
	$A^2 = 0$	$A^2 = 1$	$A^2 = 2$
$A^1 = 0$	.24	.07	0
$A^1 = 2$	.07	.24	.07
$A^1 = 4$	0	.07	.24

$\lambda^1 = 2, \lambda^2 = 2$			
	$A^2 = 0$	$A^2 = 1$	$A^2 = 2$
$A^1 = 0$	0	.07	.24
$A^1 = 2$	.07	.24	.07
$A^1 = 4$	.24	.07	0

Let us put  $p = 1$  and compute the values of the  $d_1$ -p.q.-metric. For any  $\lambda^1, \lambda^2$  here,

$$\begin{aligned} d_1(A^1, A^2) &= \sum_{a_1 < a_2} |a_1 - a_2|^1 p(a_1, a_2) \\ &= |1 - 0|^1 p(0, 1) + |2 - 0|^1 p(0, 2) \end{aligned}$$

and

$$\begin{aligned} d_1(A^2, A^1) &= \sum_{a_2 < a_1} |a_1 - a_2|^1 p(a_1, a_2) \\ &= |2 - 0|^1 p(2, 0) + |4 - 0|^1 p(4, 0) + \cdots + |4 - 2|^1 p(4, 2). \end{aligned}$$

The calculations yield the following distances:

	$\lambda^1 = 1, \lambda^2 = 1$	$\lambda^1 = 1, \lambda^2 = 2$	$\lambda^1 = 2, \lambda^2 = 1$	$\lambda^1 = 2, \lambda^2 = 2$
$d_{p=1}(A^1, A^2)$	.07	.07	.07	.55
$d_{p=1}(A^2, A^1)$	1.07	1.07	1.07	1.55

Using this table, all possible distance test inequalities are of the form  $a \leq b + c + d$ , where  $a, b$ , and  $d$  belong to one row and  $c$  to another, provided all four values are in distinct columns. It is easy to see that all the inequalities are passed.  $\square$

Pq.-metrics can be introduced directly in probabilistic terms rather than derived from “deterministic” metrics on sets of possible values. Consider, as an example, the following construction. Let  $(S^1, \Sigma^1), \dots, (S^m, \Sigma^m)$  be the sets of possible values with sigma-algebras for random variables  $R^1, \dots, R^m$ , respectively, and let us partition each  $S^k$  into  $l_k > 1$  measurable subsets  $S^{1k}, \dots, S^{l_k k} \in \Sigma^k$ . It follows that the joint probabilities of any pair  $S^{ik}, S^{i'k'}$ ,

$$\Pr(R^k \in S^{ik}, R^{k'} \in S^{i'k'}),$$

are well defined. It can easily be proved that the function

$$d_{\text{class}}(R^k, R^{k'}) = \sum_{i < i'} \Pr(R^k \in S^{ik}, R^{k'} \in S^{i'k'})$$

is a p.q.-metric. It is called a *classification p.q.-metric*, and it can be applied to all types of random variables without restrictions.

**Example 2.28** Consider the case with two real-valued random variables  $R^1, R^2$  and define the partition of  $S^1 = \mathbb{R}$  and  $S^2 = \mathbb{R}$  as, respectively,

$$S^{11} = (-\infty, x), \quad S^{21} = [x, \infty)$$

and

$$S^{12} = (-\infty, y), \quad S^{22} = [y, \infty).$$

Then, the classification distance is simply

$$d_{\text{class}}(R^1, R^2) = \Pr(R^1 \in S^{11}, R^2 \in S^{22}) = \Pr(R^1 < x, R^2 \geq y).$$

Different choices of  $x, y$  give us different classification distances.  $\square$

**Remark 2.12** A classification p.q.-metric can also be viewed as a limit case of the metric  $d_p$  introduced above, provided we first map by a measurable function  $f_k$  each  $S^k$  into a set  $\{1, \dots, l_k\}$ , and then define all the transformed random variables  $f_k(R^k)$  as distributed on  $\{1, \dots, l\}$ , with  $l = \max(l_1, \dots, l_m)$ . The latter is always possible by assigning to the “redundant” integers probability zero. Following this

transformation and equalization of domains,  $d_{class}$  is obtained as  $d_{p=0}$ . Another way of introducing the classification metric is as a special case of an *order-distance*. Without elaborating, the latter involves a relation of strict order  $\prec$  between values of one random variable and values of another. The order-distance is defined as

$$d_{ord}(Q, R) = \Pr(Q \prec R).$$

Recall that a sequence  $\lambda^{k_1} = j_1, \lambda^{k_2} = j_2, \dots, \lambda^{k_l} = j_l$  of input values is treatment-realizable if  $\{\lambda^{k_1} = j_1, \lambda^{k_k} = j_k\}$  and  $\{\lambda^{k_{i-1}} = j_{i-1}, \lambda^{k_i} = j_i\}$  for  $i = 2, \dots, l$  belong to allowable treatments. If the elements of all these pairs are distinct, and if these pairs are the only subsequences of more than one element that have the property of being a subset of an allowable treatment, then the sequence is called *irreducible*. It turns out that one only has to check the chain inequalities for irreducible sequences: these inequalities are satisfied for all treatment-realizable sequences if and only if they are satisfied for all irreducible ones.

The set of irreducible sequences may be significantly smaller than the set of all treatment-realizable sequences. Thus, it can be shown that if the set  $\Phi$  consists of all possible combinations of input values, then the only irreducible sequences are quadruples of the form

$$\lambda^k = j_1, \lambda^{k'} = j_2, \lambda^k = j_3, \lambda^{k'} = j_4,$$

with  $k \neq k'$ ,  $j_1 \neq j_3$  and  $j_2 \neq j_4$ . The only inequalities to check then are of the form,

$$d\left(A_{\phi_{14}}^k, A_{\phi_{14}}^{k'}\right) \leq d\left(A_{\phi_{12}}^k, A_{\phi_{12}}^{k'}\right) + d\left(A_{\phi_{23}}^{k'}, A_{\phi_{23}}^k\right) + d\left(A_{\phi_{34}}^k, A_{\phi_{34}}^{k'}\right),$$

where  $\phi_{14}, \phi_{12}, \phi_{23}, \phi_{34}$  are any allowable treatments that contain, respectively,

$$\begin{aligned} & \left\{ \lambda^k = j_1, \lambda^{k'} = j_4 \right\}, \left\{ \lambda^k = j_1, \lambda^{k'} = j_2 \right\}, \\ & \left\{ \lambda^{k'} = j_2, \lambda^k = j_3 \right\}, \left\{ \lambda^k = j_3, \lambda^{k'} = j_4 \right\}. \end{aligned}$$

## 2.16 (Non)Invariance of tests with respect to transformations

In this section we introduce another class of tests of selective influences, called *cosphericity tests*. Prior to introducing them, however, we should discuss an important issue.

We know from Section 2.13 that if  $(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$ , then  $(B^1, \dots, B^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$ , where the  $B$ 's are input-value-specific transformations of the  $A$ 's, that is,

$$B_\phi^1 = g_1(\lambda^1, A_\phi^1), \dots, B_\phi^n = g_n(\lambda^n, A_\phi^n),$$

for all  $\phi = (\lambda^1, \dots, \lambda^n) \in \Phi$ . It follows that if a test provides a necessary condition for selective influences, then its failure for any of the input-value-specific

transformations of  $(A^1, \dots, A^n)_\phi$  establishes  $(A^1, \dots, A^n) \not\leftrightarrow (\lambda^1, \dots, \lambda^n)$ . If the outcome of a test is not invariant with respect to some of such transformations, this consideration automatically expands this test into a multitude of tests, one for each of these transformations. This may enormously increase the ability of a test to detect violations of selective influences. This might sound paradoxical, or at least unexpected, but this is generally true for any test that provides a necessary but not sufficient condition for a tested proposition: the lack of invariance in the test's outcome with respect to transformations that preserve the tested proposition is an advantage rather than a drawback.

**Remark 2.13** If a test provides a sufficient condition for  $(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$ , and it is not invariant with respect to input-value-specific transformations, then one should apply it to a variety of  $(B^1, \dots, B^n)_\phi$  from which  $(A^1, \dots, A^n)_\phi$  can be obtained by such a transformation. At the time this was written (end of 2012), we did not have nontrivial tests that provided sufficient but not necessary conditions. If a test is a criterion when applied to input–output pairs of a particular type, then its (non)invariance with respect to transformations is immaterial for establishing or rejecting selective influences for original random variables (although transformed ones may be of interest for their own sake).

Of the two distance tests considered in the previous section,  $d_p$ -test is not invariant (for any fixed  $p$ ) with respect to numerical transformations of the random outputs.

**Example 2.29** Continuing Example 2.27, let us transform the outputs  $A^1, A^2$  as  $B^1 = g_1(A^1), B^2 = g_2(A^2)$ , where  $g_1$  is given by  $0 \mapsto 2, 2 \mapsto 1, 4 \mapsto 1$  and  $g_2$  is given by  $0 \mapsto 2, 1 \mapsto 1, 2 \mapsto 1$ . We get the joint distributions

$\begin{array}{c cc} \lambda^1 = 1, \lambda^2 = 1 & B^2 = 1 & B^2 = 2 \\ \hline B^1 = 1 & .62 & .07 \\ B^1 = 2 & .07 & .24 \end{array}$	$\begin{array}{c cc} \lambda^1 = 1, \lambda^2 = 2 & B^2 = 1 & B^2 = 2 \\ \hline B^1 = 1 & .62 & .07 \\ B^1 = 2 & .07 & .24 \end{array}$
$\begin{array}{c cc} \lambda^1 = 2, \lambda^2 = 1 & B^2 = 1 & B^2 = 2 \\ \hline B^1 = 1 & .62 & .07 \\ B^1 = 2 & .07 & .24 \end{array}$	$\begin{array}{c cc} \lambda^1 = 2, \lambda^2 = 2 & B^2 = 1 & B^2 = 2 \\ \hline B^1 = 1 & .38 & .31 \\ B^1 = 2 & .31 & 0 \end{array}$

and the corresponding  $d_{p=1}$  distances are

	$\lambda^1 = 1, \lambda^2 = 1$	$\lambda^1 = 1, \lambda^2 = 2$	$\lambda^1 = 2, \lambda^2 = 1$	$\lambda^1 = 2, \lambda^2 = 2$
$d_{p=1}(B^1, B^2)$	.07	.07	.07	.31
$d_{p=1}(B^2, B^1)$	.07	.07	.07	.31

Now the distance test inequality  $.31 \leq .07 + .07 + .07 = .21$  fails, implying  $(B^1, B^2) \not\leftrightarrow (\lambda^1, \lambda^2)$ , which in turn implies  $(A^1, A^2) \not\leftrightarrow (\lambda^1, \lambda^2)$ . Thus, the  $d_p$ -test is not invariant with respect to transformations of the variables.  $\square$

The second distance test considered in the previous section,  $d_{class}$ -test, is invariant (for any given partition scheme) with respect to any transformations of the possible values of random outputs. The obvious proviso for this statement is that a

transformed value is always classified into a partition with the same number as the original value. If this proviso is violated, it would amount to changing the partition scheme for the original outputs. The power of the  $d_{class}$ -test to detect violations of selective influences does not come from different transformations. Rather, it comes from complete flexibility in the partitioning scheme. Another way of looking at this test (see Remark 2.12) is that a transformation of the random outputs (different mappings into natural numbers) is built into the identity of the test. If the transformation changes, we apply a different test.

**Example 2.30** Consider the system  $(A^1, A^2)$  of Example 2.27. Let us partition  $S^1$  into  $S^{11} = \{0\}$ ,  $S^{21} = \{2, 4\}$ , and  $S^2$  into  $S^{12} = \{0, 1\}$ ,  $S^{22} = \{2\}$ . We obtain the following joint probabilities for the partition memberships  $A^k \in S^{ik}$ :

$\lambda^1 = 1, \lambda^2 = 1$	$A^2 \in S^{21}$	$A^2 \in S^{22}$	$\lambda^1 = 1, \lambda^2 = 2$	$A^2 \in S^{21}$	$A^2 \in S^{22}$
$A^1 \in S^{11}$	.31	0	$A^1 \in S^{11}$	.31	0
$A^1 \in S^{12}$	.38	.31	$A^1 \in S^{12}$	.38	.31

$\lambda^1 = 2, \lambda^2 = 1$	$A^2 \in S^{21}$	$A^2 \in S^{22}$	$\lambda^1 = 2, \lambda^2 = 2$	$A^2 \in S^{21}$	$A^2 \in S^{22}$
$A^1 \in S^{11}$	.31	0	$A^1 \in S^{11}$	.07	.24
$A^1 \in S^{12}$	.38	.31	$A^1 \in S^{12}$	.62	.07

This yields the classification distances

	$\lambda^1 = 1, \lambda^2 = 1$	$\lambda^1 = 1, \lambda^2 = 2$	$\lambda^1 = 2, \lambda^2 = 1$	$\lambda^1 = 2, \lambda^2 = 2$
$d_{class}(A^1, A^2)$	0	0	0	.24
$d_{class}(A^2, A^1)$	.38	.38	.38	.62

which can be seen to satisfy all distance test inequalities, as in Example 2.27.

Consider now the partitioning of  $S^1$  into  $S^{11} = \{0, 2\}$ ,  $S^{21} = \{4\}$ , and of  $S^2$  into  $S^{12} = \{0, 1\}$ ,  $S^{22} = \{2\}$ . The partition membership indicator  $B^k$  (given by  $B^k = i$  when  $A^k \in S^{ik}$ ) corresponds to the transformed variables  $B^k$  of Example 2.29. As a result, we get the same joint distribution tables as there. We know that  $d_{class}$  corresponds to  $d_{p=0}$  (see Remark 2.12), and it is easy to see that  $d_{p=0}$  is identical to  $d_{p=1}$  when the sets are partitioned into only two classes each. Therefore, the  $d_{class}$  distance table we obtain is identical to the  $d_{p=1}$  table shown in Example 2.29, and we conclude that the  $d_{class}$ -distance test fails, implying  $(A^1, A^2) \not\sim (\lambda^1, \lambda^2)$ .  $\square$

We conclude this section by presenting a test based on pairwise correlations between random outputs. It is called the *cospHERicity test*, and confined to random variables for which conventional correlations can be computed. These are all variables that are defined (or can be redefined) on the set of real numbers with the Lebesgue sigma-algebra. Discrete random variables can always be redefined to fall within this category.

The primary application of the cospHERicity test is to two input–output pairs, with two values per input, and all four treatments allowable. That is, we test the assumption  $(A^1, A^2) \sim (\lambda^1, \lambda^2)$ , with  $\Lambda^1 = \{1, 2\}$ ,  $\Lambda^2 = \{1, 2\}$ , and allowable

treatments  $\phi_{11} = (\lambda_1^1, \lambda_1^2)$ ,  $\phi_{12} = (\lambda_1^1, \lambda_2^2)$ , etc. The use of the test for larger designs will be discussed later.

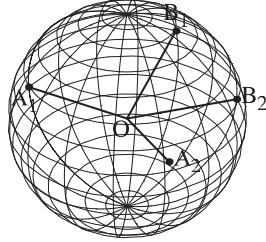
Denote the correlation between  $A_{\phi_{ij}}^1$  and  $A_{\phi_{ij}}^2$  (as the two are jointly distributed) by  $\rho_{ij}$ ,  $i, j \in \{1, 2\}$ . The cosphericity test is the proposition: if  $(A^1, A^2) \leftrightarrow_P (\lambda^1, \lambda^2)$ , then

$$|\rho_{11}\rho_{12} - \rho_{21}\rho_{22}| \leq \sqrt{1 - (\rho_{11})^2} \sqrt{1 - (\rho_{12})^2} + \sqrt{1 - (\rho_{21})^2} \sqrt{1 - (\rho_{22})^2}.$$

Superscript 2 here indicates squaring. If this inequality is violated, then the initial assumption  $(A^1, A^2) \leftrightarrow_P (\lambda^1, \lambda^2)$  should be rejected.

The explanation for the name “cosphericity” is this: the inequality above holds if and only if one can place four points,  $\mathbf{A}_1, \mathbf{A}_2, \mathbf{B}_1, \mathbf{B}_2$ , on the surface of a unit sphere (in the Euclidean three-dimensional space) centered at point  $\mathbf{O}$ , so that

$$\begin{aligned} \cos \angle \mathbf{A}_1 \mathbf{OB}_1 &= \rho_{11}, & \cos \angle \mathbf{A}_1 \mathbf{OB}_2 &= \rho_{12}, \\ \cos \angle \mathbf{A}_2 \mathbf{OB}_1 &= \rho_{21}, & \cos \angle \mathbf{A}_2 \mathbf{OB}_2 &= \rho_{22}. \end{aligned}$$



**Example 2.31** Consider the following output distributions of  $A^1, A^2$  for the treatments in  $\Lambda^1 \times \Lambda^2 = \{1, 2\} \times \{1, 2\}$ :

$\lambda^1 = 1, \lambda^2 = 1$				
		$A^2 = 0$	$A^2 = 1$	$A^2 = 5$
$A^1 = 0$		.24	.07	0
$A^1 = 1$		.07	.24	.07
$A^1 = 5$		0	.07	.24

$\lambda^1 = 1, \lambda^2 = 2$				
		$A^2 = 0$	$A^2 = 1$	$A^2 = 5$
$A^1 = 0$		.24	.07	0
$A^1 = 1$		.07	.24	.07
$A^1 = 5$		0	.07	.24

$\lambda^1 = 2, \lambda^2 = 1$				
		$A^2 = 0$	$A^2 = 1$	$A^2 = 5$
$A^1 = 0$		.24	.07	0
$A^1 = 1$		.07	.24	.07
$A^1 = 5$		0	.07	.24

$\lambda^1 = 2, \lambda^2 = 2$				
		$A^2 = 0$	$A^2 = 1$	$A^2 = 5$
$A^1 = 0$		0	.07	.24
$A^1 = 1$		.07	.24	.07
$A^1 = 5$		.24	.07	0

The correlation coefficients of the four distributions are  $\rho_{11} = \rho_{12} = \rho_{21} \approx .7299$  and  $\rho_{22} \approx -.6322$ . Substituting these in the cosphericity test, we obtain

$$\begin{aligned} .9942 &\approx |.7299 \cdot .7299 - .7299(-.6322)| \\ &\leq \sqrt{1 - .7299^2} \sqrt{1 - .7299^2} + \sqrt{1 - .7299^2} \sqrt{1 - .6322^2} \approx .9969, \end{aligned}$$

so the test is passed.  $\square$

Correlation between two random variables is not invariant with respect to any but affine transformations of the random variables. This allows one to expand the

single cosphericity test into a potential infinity of tests, corresponding to different nonlinear input-value-specific transformations  $g_1(\lambda^1, A_\phi^1)$  and  $g_2(\lambda^2, A_\phi^2)$ . An interesting fact is that if, by means of some *reversible* transformations  $g_1, g_2$ , the random variables  $(A^1, A^2)_\phi$  can be made bivariate-normally distributed at all four treatments, then the cosphericity test performed on thus transformed random outputs provides both a necessary and sufficient condition for  $(A^1, A^2) \leftrightarrow (\lambda^1, \lambda^2)$ .

**Example 2.32** The system of Example 2.31 passed the cosphericity test. However, if we apply the nonlinear transformation  $B^1 = g(A^1), B^2 = g(A^2)$ , where  $g$  is given by  $0 \mapsto 0, 1 \mapsto 1, 5 \mapsto 2$ , we get

$\lambda^1 = 1, \lambda^2 = 1$					$\lambda^1 = 1, \lambda^2 = 2$				
		$B^2 = 0$	$B^2 = 1$	$B^2 = 2$			$B^2 = 0$	$B^2 = 1$	$B^2 = 2$
$B^1 = 0$	.24	.07	0	$B^1 = 0$	.24	.07	0		
$B^1 = 1$	.07	.24	.07	$B^1 = 1$	.07	.24	.07		
$B^2 = 2$	0	.07	.24	$B^2 = 2$	0	.07	.24		

$\lambda^1 = 2, \lambda^2 = 1$					$\lambda^1 = 2, \lambda^2 = 2$				
		$B^2 = 0$	$B^2 = 1$	$B^2 = 2$			$B^2 = 0$	$B^2 = 1$	$B^2 = 2$
$B^1 = 0$	.24	.07	0	$B^1 = 0$	0	.07	.24		
$B^1 = 1$	.07	.24	.07	$B^1 = 1$	.07	.24	.07		
$B^2 = 2$	0	.07	.24	$B^2 = 2$	.24	.07	0		

and the correlations for these joint distributions are  $\rho_{11} = \rho_{12} = \rho_{21} \approx .7742$  and  $\rho_{22} \approx -.7742$ . Substituting these in the cosphericity test, we obtain

$$\begin{aligned} 1.1988 &\approx |.7742 \cdot .7742 - .7742(-.7742)| \\ &\leq \sqrt{1 - .7742^2} \sqrt{1 - .7742^2} + \sqrt{1 - .7742^2} \sqrt{1 - .7742^2} \approx .8012. \end{aligned}$$

We see that the cosphericity test is not passed for the transformed variables. As a result, selective influences are ruled out for the original variables as well.  $\square$

The cosphericity test can also be applied to more than two input-output pairs. If we assume that  $(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$ , then, by the nestedness property for input-output pairs, for any two of them,  $(A^k, \lambda^k)$  and  $(A^{k'}, \lambda^{k'})$ , we should have  $(A^k, A^{k'}) \leftrightarrow (\lambda^k, \lambda^{k'})$ . The test only applies if there are two values  $i$  and  $i'$  of  $\lambda^k$  and two values  $j$  and  $j'$  of  $\lambda^{k'}$  such that, for some allowable treatments  $\phi_{ij}, \phi_{ij'}, \phi_{i'j}, \phi_{i'j'}$ ,

$$\lambda^k = i, \lambda^{k'} = j \in \phi_{ij}, \lambda^k = i, \lambda^{k'} = j' \in \phi_{ij'}, \text{ etc.}$$

In other words, the inputs and their values should be chosen so that  $\{\lambda^k = i, \lambda^k = i'\}$  and  $\{\lambda^{k'} = j, \lambda^{k'} = j'\}$  form a completely crossed subdesign within the set of allowable treatments. By the nestedness property for input values, we have  $(A^k, A^{k'}) \leftrightarrow (\lambda^k, \lambda^{k'})$  with the input values restricted to  $\{i, i'\}$  and  $\{j, j'\}$  and the new set of allowable treatments consisting of all four possible combinations. If this

cosphericity inequality is violated for all least one combination of  $k, k', i, i', j, j'$ , then the initial assumption  $(A^1, \dots, A^n) \leftrightarrow (\lambda^1, \dots, \lambda^n)$  should be rejected.

## 2.17 Conditional determinism and conditional independence of outcomes

The definition of selective influences (in the canonical form) requires the existence of a random variable  $R$  and functions  $f_1, \dots, f_n$  such that, for all allowable treatments  $\phi$ ,

$$(A_\phi^1, \dots, A_\phi^n) \sim (f_1(\lambda^1, R), \dots, f_n(\lambda^n, R)).$$

One obvious consequence of this definition is that, conditioned on any value  $r$  of  $R$ , the outputs become (equal to) deterministic functions of the corresponding factors,

$$f_1(\lambda^1, r), \dots, f_n(\lambda^n, r).$$

It is sometimes easy to deal with these deterministic quantities, derive certain inequalities that hold for every value of  $r$ , and then show that they are preserved as  $R$  randomly varies. It is an especially useful approach if the distributions of  $(A_\phi^1, \dots, A_\phi^n)$  at allowable treatments  $\phi$  are not known, and instead we know distributions of certain functions of these random variables, such as their sums or maxima.

Let us discuss this on an example from studies of mental architectures. This is a traditional area of psychology dealing with decomposing performance of a task into a network of subprocesses when we only know the distributions of the overall performance time (referred to as response time) at different treatments. Let us assume that we observe response times  $T$  in an experiment with two factors,  $\lambda^1, \lambda^2$ , manipulated at two levels each, denoted in both cases by 1 and 2. All four treatments are allowable. Let us postulate that there are two processes involved, with their durations  $A^1$  and  $A^2$  being random variables, and that  $(A^1, A^2) \leftrightarrow (\lambda^1, \lambda^2)$ . We want to determine which of the three “architectures,” or composition schemes, is being employed:

1. serial,  $T_\phi = A_\phi^1 + A_\phi^2$
2. parallel-OR,  $T_\phi = \min(A_\phi^1, A_\phi^2)$ , or
3. parallel-AND,  $T_\phi = \max(A_\phi^1, A_\phi^2)$ .

One tool traditionally used for this purpose is the *interaction contrast*,

$$c(t) = \Pr(T_{11} \leq t) + \Pr(T_{22} \leq t) - \Pr(T_{12} \leq t) - \Pr(T_{21} \leq t),$$

where  $t$  is any nonnegative number, and  $T_{ij}$  abbreviates  $T_{\phi=(i,j)}$ .

We do not know the joint distribution of  $A_\phi^1, A_\phi^2$  at any of the four treatments, but we can write

$$(A_{ij}^1, A_{ij}^2) \sim (f_1(\lambda^1 = i, R), f_2(\lambda^2 = j, R)) = (g_i^1(R), g_j^2(R)), \quad i, j \in \{1, 2\}.$$

We need one additional assumption: that  $R$  can be chosen in such a way that, for any of its possible values  $r$ ,

$$g_1^1(r) \leq g_2^1(r), \quad g_1^2(r) \leq g_2^2(r).$$

In other words, switching either factor from level 1 to level 2 prolongs the corresponding processing time. We call this assumption *prolongation constraints*. Various analogs of this assumption are common in studies of mental architectures.

Deterministic real-valued quantities can be viewed as random variables with Heaviside distribution functions:

$$\Pr(g_i^k(r) \leq t) = \begin{cases} 0 & \text{if } t < g_i^k(r), \\ 1 & \text{if } t \geq g_i^k(r). \end{cases}$$

Analogously,

$$\Pr(\text{comp}(g_i^1(r), g_j^2(r)) \leq t) = \begin{cases} 0 & \text{if } t < \text{comp}(g_i^1(r), g_j^2(r)), \\ 1 & \text{if } t \geq \text{comp}(g_i^1(r), g_j^2(r)), \end{cases}$$

where  $\text{comp}$  stands for one of the three composition rules of interest, plus, maximum, or minimum. This allows us to form the *conditional interaction contrast*,

$$c^*(t, r) = \Pr(t_{11}(r) \leq t) + \Pr(t_{22}(r) \leq t) - \Pr(t_{12}(r) \leq t) - \Pr(t_{21}(r) \leq t),$$

where

$$t_{ij} = \text{comp}(g_i^1(r), g_j^2(r)).$$

It is easy to see that

$$\Pr(T_{ij} \leq t) = \int_{S_R} \Pr(t_{ij}(r) \leq t) dp_R(r)$$

and

$$c(t) = \int_{S_R} c^*(t, r) dp_R(r),$$

where the Lebesgue integral is taken over the entire domain  $S_R$  of  $R$ , and  $p_R$  is the probability measure in the distribution of  $R$ . (The reader not familiar with Lebesgue integrals can think of  $dp_R(r)$  above as a generalized version of  $f_R(r)dr$ , where  $f_R$  is the density function of  $R$  over the set of real numbers.)

Using this observation we can easily establish that if the composition rule is min (parallel-OR architecture), then  $c(t) \leq 0$ , for all  $t$ , because  $c^*(t, r) \leq 0$  at any  $t$  and any fixed  $r$ . Indeed, consider all possible arrangements of  $g_1^1(r), g_2^1(r), g_1^2(r), g_2^2(r)$ , keeping in mind the prolongation constraints and assuming, with no loss of generality, that  $g_1^1(r) \leq g_2^2(r)$ . These possible arrangements are

- (i)  $g_1^1(r) \leq g_2^1(r) \leq g_1^2(r) \leq g_2^2(r)$ ,
- (ii)  $g_1^1(r) \leq g_2^2(r) \leq g_1^1(r) \leq g_2^2(r)$ ,
- (iii)  $g_1^1(r) \leq g_2^2(r) \leq g_2^2(r) \leq g_1^1(r)$ .

Thus, for (ii), we have

$$\begin{aligned} t_{11}(r) &= \min(g_1^1(r), g_1^2(r)) = g_1^1(r), \\ t_{12}(r) &= \min(g_1^1(r), g_2^2(r)) = g_1^1(r), \\ t_{21}(r) &= \min(g_2^1(r), g_1^2(r)) = g_1^2(r), \\ t_{22}(r) &= \min(g_2^1(r), g_2^2(r)) = g_2^1(r). \end{aligned}$$

Then, substituting for the numerical values

$$\begin{aligned} c^*(t, r) &= \Pr(t_{11}(r) \leq t) + \Pr(t_{22}(r) \leq t) - \Pr(t_{12}(r) \leq t) - \Pr(t_{21}(r) \leq t) \\ &= \begin{cases} 0 + 0 - 0 - 0 = 0 & \text{if } t < g_1^1(r), \\ 1 + 0 - 1 - 0 = 0 & \text{if } g_1^1(r) \leq t < g_1^2(r), \\ 1 + 0 - 1 - 1 < 0 & \text{if } g_1^2(r) \leq t < g_2^1(r), \\ 1 + 1 - 1 - 1 = 0 & \text{if } g_2^1(r) \leq t < g_2^2(r), \\ 1 + 1 - 1 - 1 = 0 & \text{if } t \geq g_2^2(r). \end{cases} \end{aligned}$$

In the same way, one proves that  $c^*(t, r)$  is never positive in cases (i) and (iii).

By analogous reasoning we can show that if the composition rule is max (parallel-AND architecture), then  $c(t) \geq 0$ , for all  $t$ , because  $c^*(t, r) \geq 0$  at any  $t$  and any fixed  $r$ .

For the serial architecture (the composition rule +)  $c^*(t, r)$  does not preserve its sign, but the analysis of the arrangements shows that, for any  $t$  and  $r$ ,

$$\int_0^t c^*(t, r) dt \geq 0$$

and

$$\int_0^\infty c^*(t, r) dt = 0.$$

Then the same properties should hold for  $c(t)$ , because

$$\int_0^t c(t) dt = \int_0^t \left( \int_{S_R} c^*(t, r) dp_R(r) \right) dt = \int_{S_R} \left( \int_0^t c^*(t, r) dt \right) dp_R(r).$$

However, dealing with deterministic quantities is not always convenient. If a deterministic quantity changes as a function of  $r$ , the probability with which it falls within a given measurable subset may jump from 0 to 1, or vice versa. In some cases it may be desirable to deal with “well-behaved” distributions only, with associated probabilities that change continuously or even sufficiently smoothly. (The term “smooth” refers to the highest order of continuous derivative a function possesses.) To make this desideratum achievable in the context of selective influences, we begin by stating the following equivalence.

**Theorem 2.9**  $(A^1, \dots, A^n) \leftrightarrow_P (\lambda^1, \dots, \lambda^n)$  if and only if one can find stochastically independent random variables  $R, R^1, \dots, R^n$  and functions  $w_1, \dots, w_n$ , such that

$$(A_\phi^1, \dots, A_\phi^n) \sim (w_1(\lambda^1, R, R^1), \dots, w_n(\lambda^n, R, R^n))$$

for all allowable treatments  $\phi = (\lambda^1, \dots, \lambda^n)$ .

By analogy with factor analysis, we can call  $R^1, \dots, R^n$  *specific sources of variability*, and call  $R$  a *common source of variability*. The proof of the theorem is very simple. If a representation

$$(A_\phi^1, \dots, A_\phi^n) \sim (f_1(\lambda^1, R), \dots, f_n(\lambda^n, R))$$

exists, one can choose arbitrary  $R^1, \dots, R^n$  (combined together and with  $R$  by an independent coupling) and put  $w_k(\lambda, r, r') = f_k(\lambda, r)$ ,  $k = 1, \dots, n$ . If a representation stated in the theorem exists, then define  $R^* = (R, R^1, \dots, R^n)$  and put  $f_k(\lambda, (r, r^1, \dots, r^n)) = w_k(\lambda, r, \text{Proj}_k(r^1, \dots, r^n))$ .

The consequences of this simple theorem are significant. Once the possibility of splitting a single source of randomness into common and specific components has been established, it becomes possible that in certain situations this split can be more than a formal redefinition of a single source. It follows from the theorem that conditioned upon any value  $r$  of  $R$ , the random variables  $w_1(\lambda^1, r, R^1), \dots, w_n(\lambda^n, r, R^n)$  are stochastically independent. One can hypothesize now that these independent random variables have distributions with desired properties. For example, if all random variables  $(A_\phi^1, \dots, A_\phi^n)$  are real-valued and continuous,  $w_1(\lambda^1, r, R^1), \dots, w_n(\lambda^n, r, R^n)$  may be assumed to possess densities, or have the property that the probability with which  $w_k(\lambda^k, r, R^k)$  falls within any interval of reals is a continuously differentiable function of  $r$ . Such assumptions may be important in studying mental architectures or random variables underlying comparisons of stimuli.

## 2.18 Related literature

There are many textbooks treating measure theory and probability (e.g., Chung, 1974). However, the reader should be aware that (a) older textbooks usually deal with random variables in the narrow sense only; and (b) in most textbooks, random variables are defined as measurable functions on a single sample space, thereby restricting the consideration to jointly distributed random variables. For random variables that need not be jointly distributed and the associated theory of coupling them into jointly distributed entities, see Thorisson (2000).

The earliest explicit discussions of selective influences in psychology can be found in Sternberg (1969) and Townsend (1984). Marginal selectivity for two random variables was first mentioned in Townsend and Schweickert (1989). Other

historical details and relations can be found in Dzhafarov (2003a), where the theory of selective influences presented in this chapter was first proposed. In this earlier work (and its elaboration in Dzhafarov and Gluhovsky, 2006), the “is distributed as” relation in the defining representation for selective influences,

$$(A_\phi^1, \dots, A_\phi^n) \sim (f_1(\lambda^1, R), \dots, f_n(\lambda^n, R))$$

was somewhat carelessly replaced with equality. For a mathematically rigorous and maximally general version of the definition and joint distribution criterion, see Dzhafarov and Kujala (2010).

The tests of selective influences were first introduced in Kujala and Dzhafarov (2008). They included the cosphericity tests and a special form of distance tests. A general version of distance tests (p.q.-metric tests) was introduced in Dzhafarov and Kujala (2013). The linear feasibility test is described in Dzhafarov and Kujala (2012b). For applications of the theory of selective influences to discrimination judgments and to mental processing architectures, see Dzhafarov (2003b, 2003c) and Dzhafarov *et al.* (2004).

The parallels between the theory of selective influences and the analysis of the so-called Bohmian version of the Einstein–Podolsky–Rosen entanglement paradigm of quantum physics are described in Dzhafarov and Kujala (2012a, 2012b). The history there dates back to Bell’s (1964) epoch-making inequalities, and then to their elaborations in Clauser *et al.* (1969) and Fine (1982).

Mathematically, the theory of selective influences is subsumed by the general theory of contextuality, developed primarily in quantum physics. The relevant literature includes Dzhafarov & Kujala (2014a, 2014b, 2014c, 2015) and Dzhafarov *et al.* (2015).

## 2.19 Acknowledgments

This work was supported by NSF grant SES-1155956 and AFOSR grant FA9550-14-1-0318. We are grateful to Jing Chen, Shree Frazier, Nicole Murchison, Alison Schroeder, and Ru Zhang for pointing out numerous typos and imprecisions in the original draft of the chapter.

## References

- Bell, J. S. (1964). On the Einstein Podolsky Rosen paradox. *Physics*, **1**(3), 195–200.
- Chung, K. L. (1974). *A Course in Probability Theory*. New York, NY: Academic Press.
- Clauser, J. F., Horne, M. A., Shimony, A. and Holt, R. A. (1969). Proposed experiment to test local hidden-variable theories. *Physical Review Letters*, **23**, 880–884.
- Dzhafarov, E. N. (2003a). Selective influence through conditional independence. *Psychometrika*, **68**(1), 7–25.

- Dzhafarov, E. N. (2003b). Thurstonian-type representations for “same–different” discriminations: deterministic decisions and independent images. *Journal of Mathematical Psychology*, **47**, 184–204.
- Dzhafarov, E. N. (2003c). Thurstonian-type representations for “same–different” discriminations: probabilistic decisions and interdependent images. *Journal of Mathematical Psychology*, **47**, 205–219.
- Dzhafarov, E. N. and Gluhovsky, I. (2006). Notes on selective influence, probabilistic causality, and probabilistic dimensionality. *Journal of Mathematical Psychology*, **50**, 390–401.
- Dzhafarov, E. N. and Kujala, J. V. (2010). The joint distribution criterion and the distance tests for selective probabilistic causality. *Frontiers in Psychology*, **1**, 151. doi: 10.3389/fpsyg.2010.00151
- Dzhafarov, E. N. and Kujala, J. V. (2012a). Quantum entanglement and the issue of selective influences in psychology: an overview. In Busemeyer, J. R., Dubois, F., Lambert-Mobilansky, A. and Melucci, M. (eds), *Quantum Interaction*, Vol. 7620, pp. 184–195. New York, NY: Springer.
- Dzhafarov, E. N. and Kujala, J. V. (2012b). Selectivity in probabilistic causality: where psychology runs into quantum physics. *Journal of Mathematical Psychology*, **56**, 54–63.
- Dzhafarov, E. N. and Kujala, J. V. (2013). Order-distance and other metric-like functions on jointly distributed random variables. *Proceedings of the American Mathematical Society*, **141**, 3291–3301.
- Dzhafarov, E. N. and Kujala, J. V. (2014a). Contextuality is about identity of random variables. *Physica Scripta*, **T163**, 014009.
- Dzhafarov, E. N. and Kujala, J. V. (2014b). Embedding quantum into classical: contextualization vs conditionalization. *PLoS ONE*, **9**(3), e92818.
- Dzhafarov, E. N. and Kujala, J. V. (2014c). A qualified Kolmogorovian account of probabilistic contextuality. *Lecture Notes in Computer Science*, **8369**, 201–212.
- Dzhafarov, E. N. and Kujala, J. V. (2015). Random variables recorded under mutually exclusive conditions: contextuality-by-default. In Liljenström, H. (ed.), *Advances in Cognitive Neurodynamics IV*, pp. 405–410. Dordrecht: Springer.
- Dzhafarov, E. N., Kujala, J. V. and Larsson, J.-Å. (2015). Contextuality in three types of quantum-mechanical systems. *Foundations of Physics*, **7**, 762–782.
- Dzhafarov, E. N., Schweickert, R. and Sung, K. (2004). Mental architectures with selectively influenced but stochastically interdependent components. *Journal of Mathematical Psychology*, **48**, 51–64.
- Fine, A. (1982). Joint distributions, quantum correlations, and commuting observables. *Journal of Mathematical Physics*, **23**, 1306–1310.
- Kujala, J. V. and Dzhafarov, E. N. (2008). Testing for selectivity in the dependence of random variables on external factors. *Journal of Mathematical Psychology*, **52**, 128–144.
- Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, **70**, 61–79.
- Sternberg, S. (1969). The discovery of processing stages: extensions of Donders’ method. *Acta Psychologica*, **30**, 276–315.
- Thorisson, H. (2000). *Coupling, Stationarity, and Regeneration*. New York, NY: Springer.

- Townsend, J. T. (1984). Uncovering mental processes with factorial experiments. *Journal of Mathematical Psychology*, **28**(4), 363–400.
- Townsend, J. T. and Schweickert, R. (1989). Toward the trichotomy method of reaction times: laying the foundation of stochastic mental networks. *Journal of Mathematical Psychology*, **33**(3), 309–327.

# 3 Functional equations

Che Tat Ng

3.1	Introduction	151
3.2	Cauchy's functional equations	153
3.2.1	The additive Cauchy equation on the real line	153
3.2.2	The additive Cauchy equation on a restricted domain	156
3.2.3	The fundamental Cauchy equations	158
3.3	The Pexider equation	158
3.4	A generalized Pexider equation	160
3.5	Some uniqueness and regularity theorems	161
3.6	Differentiable solutions	168
3.7	Regularity conditions	172
3.8	Some functional equations in binocular space perception	173
3.9	A functional equation involving three means	181
3.10	Utility of gains and losses	182
3.11	Miscellaneous comments	186
	Appendix	187
	References	190

## 3.1 Introduction

A functional equation is an equation in which the unknown objects are functions. Unlike differential equations, it is not a standard stand-alone topic in an undergraduate syllabus. Nonetheless, most students in mathematics would have encountered and solved some functional equations. For example, in an algebra course, we may have learnt how to obtain all the group homomorphisms from one group to another. A group homomorphism  $f$  is a function satisfying the algebraic identity  $f(xy) = f(x)f(y)$ . When the mission is to find all such  $f$ , the identity is considered a functional equation and the unknown is  $f$ .

Each function comes with its domain and codomain. Variables are used to point to elements in the domain and codomain. In psychological sciences we are interested in people. Their psychological properties or qualities are the true variables. How the variables are measured, or how numbers are assigned to them, belongs to the theory of measurement. The numerically measured values are often, but not

always, the variables of our functional equations. For this reason, in this chapter we treat functional equations with real variables more often. The function values may be a measure of some psychological attributes, and the codomain may then be the real numbers.

The real line,  $\mathbb{R}$ , will be endowed with the usual algebraic, topological and ordering structures. This chapter does not deal with how and in what yardstick psychological properties are measured. If  $x$  represents how strong a stimulus is, we will skip the question whether  $2x$  has the interpretation of “twice as strong.” Perhaps we should first address what is meant by “twice as strong” and ask if stimulus can be measured and scaled so that the interpretation holds. Tone intensity  $x$  and a respondent’s perception of loudness,  $f(x)$ , are measured along the psychological continuum. The articles by Hayes and Embretson (2012), Sowden (2012) and Van Zandt and Townsend (2012) offer excellent accounts on psychological measurement.

This chapter is intended to introduce and cover in depth some basic skills in handling functional equations. It is not a survey article. Several factors influenced our choices of the equations covered, the first being how basic and useful some methods are and whether they are appropriate at the senior undergraduate levels. The Cauchy and the Pexider equations are selected as an entry point. The following example serves to demonstrate how the Pexider functional equation arises in mathematical psychology and how its solutions may be reconciled with data collected.

**Example 3.1** (Luce (2000), §4.5. Dependence of Utility on Money) Let  $x$  and  $y$  be money and let  $x \oplus y$  stand for the joint receipt. Thaler’s (1985) classroom study involving judgments used survey questions like:

Individual A was given tickets to lotteries involving the World Series. He won \$50 in one lottery and \$25 in the other.

Individual B was given a ticket to a single, larger World Series lottery. He won 75.

Who was happier? A, B or indifferent?

The outcome of the survey is that a substantial majority of the respondents said A would be happier than B. This would suggest that  $x \oplus y > x + y$  when both  $x$  and  $y$  are positive. For losses, the data led Thaler to the opposite conclusion, that  $x + y > x \oplus y$  for  $x < 0, y < 0$ .

A side note: in Luce’s studies the preference order,  $\succsim$ , is a primitive.

On the joint receipt  $\oplus$  of money two assumptions are made. First, it admits an additive representation:

$$V(x \oplus y) = V(x) + V(y) \tag{3.1}$$

where  $V$  is continuous, strictly increasing and  $V(0) = 0$ . Second, it is invariant under changes of units of money, i.e., for all  $r > 0$ ,

$$(rx) \oplus (ry) = r(x \oplus y).$$

The two assumptions together yield

$$V(r(x \oplus y)) = V(rx) + V(ry).$$

Letting  $V_r(x) = V(rx)$ , it becomes

$$V_r(x \oplus y) = V_r(x) + V_r(y). \quad (3.2)$$

Comparing (3.1) with (3.2) and using the invertibility and continuity of  $V$  and  $V_r$  we get that  $V_r = \theta(r)V$  for some constant  $\theta$  which may depend on the fixed  $r$ . This gives us a Pexider equation

$$V(rx) = \theta(r)V(x).$$

The solutions of  $V$  are given by

$$V(x) = \begin{cases} \alpha x^\beta, & x \geq 0 \\ \alpha' |x|^\beta, & x < 0 \end{cases},$$

where  $\alpha > 0$ ,  $\beta > 0$  and  $\alpha' < 0$  are constants. Putting that back in (3.1), we get

$$x \oplus y = \begin{cases} (x^\beta + y^\beta)^{\frac{1}{\beta}}, & x \geq 0, y \geq 0 \\ \left(x^\beta + \frac{\alpha'}{\alpha} |y|^\beta\right)^{\frac{1}{\beta}}, & x \geq 0, y < 0, \alpha x^\beta + \alpha' |y|^\beta \geq 0 \\ -\left(\frac{\alpha'}{\alpha} x^\beta + |y|^\beta\right)^{\frac{1}{\beta}}, & x \geq 0, y < 0, \alpha x^\beta + \alpha' |y|^\beta < 0 \\ \left(\frac{\alpha'}{\alpha} |x|^\beta + y^\beta\right)^{\frac{1}{\beta}}, & x < 0, y \geq 0, \alpha' |x|^\beta + \alpha y^\beta \geq 0 \\ -\left(|x|^\beta + \frac{\alpha}{\alpha'} y^\beta\right)^{\frac{1}{\beta}}, & x < 0, y \geq 0, \alpha' |x|^\beta + \alpha y^\beta < 0 \\ -(|x|^\beta + |y|^\beta)^{\frac{1}{\beta}}, & x < 0, y < 0 \end{cases}.$$

The preference order observed in Thaler's survey,  $x \oplus y \succ x + y$  for gains  $x$  and  $y$ , and the reversal  $x + y \succ x \oplus y$  for losses  $x$  and  $y$ , is equivalent to  $\beta < 1$ .

The fundamental equation of information measures is chosen to reinforce the use of differential equations. The equation is used in a characterization of Shannon's entropy. The entropy plays an important role in models of cognition and signal detection theory. Beyond the entry point, several equations are chosen to demonstrate the use of a uniqueness theorem in solving functional equations.

## 3.2 Cauchy's functional equations

### 3.2.1 The additive Cauchy equation on the real line

A starting point for the theory of functional equations is the additive Cauchy equation

$$f(x + y) = f(x) + f(y) \quad (\forall x, y \in \mathbb{R}) \quad (3.3)$$

for “unknown” functions  $f : \mathbb{R} \rightarrow \mathbb{R}$ . A function satisfying this equation is called an *additive* function on the real line. We shall give its general solution via the following theorem.

**Theorem 3.1** *If  $f : \mathbb{R} \rightarrow \mathbb{R}$  satisfies*

$$f(x+y) = f(x) + f(y) \quad (\forall x, y \in \mathbb{R}) \quad (3.4)$$

*then*

$$f(rx) = rf(x) \quad (\forall x \in \mathbb{R}, r \in \mathbb{Q}). \quad (3.5)$$

*Hence  $f$  is a linear function on  $\mathbb{R}$  which is considered as a vector space over the rational field  $\mathbb{Q}$ .*

*Conversely, by definition, a linear function  $f$  on  $\mathbb{R}$  over  $\mathbb{Q}$  is one that satisfies both (3.4) and (3.5), and is a solution of the additive Cauchy equation (3.4).*

*Proof* Suppose that  $f$  satisfies (3.4). Putting in (3.4)  $x = 0$  gives

$$f(0) = 0. \quad (3.6)$$

Letting in (3.4)  $y = -x$  and using (3.6) we get that  $f$  is an odd function

$$f(-x) = -f(x) \quad (\forall x \in \mathbb{R}). \quad (3.7)$$

We shall show by simple induction that

$$f(nx) = nf(x) \quad (\forall x \in \mathbb{R}, n \in \mathbb{N}). \quad (3.8)$$

At  $n = 0$  it holds true by (3.6). Suppose that it holds true for some  $n = k \geq 0$ :

$$f(kx) = kf(x) \quad (\forall x \in \mathbb{R}). \quad (3.9)$$

Putting in (3.4)  $y = kx$  and using (3.9) we get

$$f((k+1)x) = f(kx) + f(x) = kf(x) + f(x) = (k+1)f(x) \quad (3.10)$$

and so the claim holds for  $n = k + 1$ . This completes the inductive proof for (3.8). Combining (3.7) and (3.8) we get

$$f(nx) = nf(x) \quad (\forall x \in \mathbb{R}, n \in \mathbb{Z}). \quad (3.11)$$

Let  $r$  be a rational number. Then  $r = m/n$  for some integers  $m, n, n \neq 0$ . Now

$$\begin{aligned} f(rx) = rf(x) &\quad \text{iff} \quad f(mx/n) = mf(x)/n \\ &\quad \text{iff} \quad nf(mx/n) = mf(x) \\ &\quad \text{iff} \quad f(n(mx/n)) = f(mx) \quad \text{by (3.11)} \\ &\quad \text{iff} \quad f(mx) = f(mx). \end{aligned}$$

The latter is true and this proves (3.5).  $\square$

One may ask in what sense the above theorem solved the additive Cauchy Equation (3.3). Under the Axiom of Choice every vector space over a field admits a

(Hamel) basis. In particular, there exists a subset  $B$  of  $\mathbb{R}$  which is a basis over  $\mathbb{Q}$ . Every  $x \in \mathbb{R}$  can be represented uniquely as a linear combination of elements of  $B$

$$x = \sum_{b \in B} x_b b \quad (3.12)$$

where  $x_b \in \mathbb{Q}$  and that only a finite number of  $x_b$  are nonzero. Every linear function  $f$  is uniquely determined by its values on  $B$  via

$$f(x) = \sum_{b \in B} x_b f(b). \quad (3.13)$$

The initial values of  $f$  on  $B$  are free. It is in this sense that we understand how an additive function  $f$  on the real line can be constructed.

By cardinality considerations,  $B$  is an uncountable set.

If the initial values of  $f$  on  $B$  satisfy the constant ratio condition

$$f(b)/b = c \quad (\forall b \in B) \quad (3.14)$$

for some constant  $c$ , then the resulting solution, as determined by (3.13), takes the form

$$f(x) = cx \quad (\forall x \in \mathbb{R}) \quad (3.15)$$

and is thus linear over the real field  $\mathbb{R}$ .

If the constant ratio condition is not satisfied, then there exist  $b_1 \neq b_2$  in  $B$  such that

$$f(b_1)/b_1 \neq f(b_2)/b_2. \quad (3.16)$$

In that case, the span  $V$  of the two elements  $(b_1, f(b_1))$  and  $(b_2, f(b_2))$  over  $\mathbb{Q}$  is dense in the plane  $\mathbb{R} \times \mathbb{R}$ . Because  $V$  is a subset of the graph  $\{(x, f(x)) \mid x \in \mathbb{R}\}$  of  $f$ , it implies that the graph of  $f$  is dense in the plane. The above theorem and the follow-up observations lead to the following.

**Corollary 3.1** *Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a solution of the additive Cauchy equation (3.3). Either it satisfies the constant ratio condition (3.14) on some basis  $B$ , and is thus linear, (3.15), or that it fails the constant ratio condition, and thus its graph is dense in the plane.*

**Corollary 3.2** *Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a solution of the Cauchy equation (3.3). If  $f$  is monotonic on some proper subinterval, locally bounded from above or from below at some point, or continuous at some point, then it is linear, (3.15).*

The conditions stated in the above corollary imply that the graph of  $f$  is not dense in the plane. We shall come across some other regularity conditions that force an additive function to be linear (over  $\mathbb{R}$ ).

**Corollary 3.3** *The only field automorphism  $f : \mathbb{R} \rightarrow \mathbb{R}$  is the identity map.*

*Proof* A field automorphism  $f : \mathbb{R} \rightarrow \mathbb{R}$ , by definition, is a bijection satisfying both identities  $f(x + y) = f(x) + f(y)$  and  $f(xy) = f(x)f(y)$  for all  $x, y$ . The second identity implies in particular that  $f(x^2) = f(x)^2 \geq 0$ . So  $f$  is bounded from below by zero on the half-line  $[0, \infty[$ . Being additive, by the first identity,  $f$  must be linear:  $f(x) = cx$  for some constant  $c$ . The second equation yields  $cxy = c^2xy$  for all  $x, y$ . So  $c = 0, 1$ .  $f$  being a bijection,  $c = 0$  is excluded.  $\square$

Let  $(G, \cdot)$  and  $(H, \cdot)$  be groups. Functions  $f : G \rightarrow H$  satisfying  $f(xy) = f(x)f(y)$  for all  $x, y \in G$  are called (group-) homomorphisms. The real line under addition is an example of a group. An additive function on the real line is, in this general context, a group homomorphism of the real line into itself. A property like (3.8), seen in the proof of the theorem, holds for general group homomorphisms.

One may ask the following questions. If  $f$  is an additive function on the real line satisfying also the identity  $f(f(x)) = x$  for all  $x$ , must  $f$  be one of the linear maps  $f(x) = x$  for all  $x$ , or  $f(x) = -x$  for all  $x$ ? Does there exist additive functions on the real line satisfying also the identity  $f(f(x)) = -x$  for all  $x$ ? Results reported in Ng and Zhang (2006) resolved such questions.

### 3.2.2 The additive Cauchy equation on a restricted domain

In applications we may encounter the Cauchy equation on a restricted domain. For example, let us consider the equation

$$f(x + y) = f(x) + f(y) \quad (\forall x, y \in ]0, \infty[) \quad (3.17)$$

where  $f : ]0, \infty[ \rightarrow \mathbb{R}$ . How can we handle the equation? One standard method is to ask if every function satisfying this equation admits an extension  $\bar{f} : \mathbb{R} \rightarrow \mathbb{R}$  which satisfies the equation (3.3) on the full line. If it does, then we regard equation (3.17) as being solved. The solutions of (3.17) will then be the solutions of (3.3) restricted to  $]0, \infty[$ . It turns out that the answer is indeed positive. Here we state the result formally and include a proof.

**Theorem 3.2** *Every function  $f$  satisfying the restricted Cauchy equation (3.17) can be extended to a function which is additive on the full real line.*

*Proof* Suppose that  $f : ]0, \infty[ \rightarrow \mathbb{R}$  satisfies (3.17). Let  $x \in \mathbb{R}$  be given. We may find two positive real numbers  $p_1, p_2$  such that

$$x = p_1 - p_2. \quad (3.18)$$

Define  $\bar{f}$  by

$$\bar{f}(x) = f(p_1) - f(p_2). \quad (3.19)$$

The first step is to check if  $\bar{f}$  is well-defined. For that purpose, suppose that there are two ways of writing  $x$  in the form (3.18), say

$$x = p_1 - p_2 \quad \text{and} \quad x = p'_1 - p'_2. \quad (3.20)$$

We must confirm that  $f(p_1) - f(p_2) = f(p'_1) - f(p'_2)$ . Because  $p_1 - p_2 = p'_1 - p'_2$ , rearranging terms we get  $p_1 + p'_2 = p'_1 + p_2$ . Applying  $f$  to both sides and using (3.17) we get  $f(p_1) + f(p'_2) = f(p'_1) + f(p_2)$ . Thus  $f(p_1) - f(p_2) = f(p'_1) - f(p'_2)$  is confirmed and  $\bar{f}$  is well-defined.

The second step is to show that  $\bar{f}$  indeed satisfies (3.3). Let  $x$  and  $y$  be arbitrarily given. Choose positive  $p_1, p_2, q_1$  and  $q_2$  such that

$$x = p_1 - p_2 \quad \text{and} \quad y = q_1 - q_2. \quad (3.21)$$

Then

$$x + y = (p_1 + q_1) - (p_2 + q_2) \quad (3.22)$$

and  $p_1 + q_1 > 0, p_2 + q_2 > 0$ . So, using the definition of  $\bar{f}$  and (3.17), we have

$$\begin{aligned} \bar{f}(x+y) &= \bar{f}(x) + \bar{f}(y) \\ \text{iff } f(p_1+q_1) - f(p_2+q_2) &= (f(p_1) - f(p_2)) + (f(q_1) - f(q_2)) \\ \text{iff } (f(p_1) + f(q_1)) - (f(p_2) + f(q_2)) &= (f(p_1) - f(p_2)) + (f(q_1) - f(q_2)). \end{aligned}$$

The last identity holds and thus  $\bar{f}(x+y) = \bar{f}(x) + \bar{f}(y)$ .

The third and last step is to show that  $\bar{f}$  is indeed an extension of  $f$ . Let  $x > 0$  be given. Evidently we can write

$$x = p_1 - p_2 \quad \text{where} \quad p_1 = x + 1 > 0, \quad p_2 = 1 > 0. \quad (3.23)$$

So  $\bar{f}(x) = f(p_1) - f(p_2) = f(x+1) - f(1)$ . By (3.17),  $f(x+1) = f(x) + f(1)$ . Hence  $\bar{f}(x) = (f(x) + f(1)) - f(1) = f(x)$ . This proves that  $\bar{f}$  is an extension of  $f$ .  $\square$

Not every Cauchy equation on a restricted domain admits an extension to the full line. Here is an example.

**Example 3.2** Consider the equation

$$f(x+y) = f(x) + f(y) \quad (\forall x \in ]1, 2[, y \in ]3, 4[) \quad (3.24)$$

where  $f$  is a real-valued function defined on the union  $]1, 2[ \cup ]3, 4[ \cup ]4, 6[$ . A particular solution  $f_0$  is the step function defined by

$$f_0(x) = \begin{cases} 1 & \text{if } x \in ]1, 2[ \\ 3 & \text{if } x \in ]3, 4[ \\ 4 & \text{if } x \in ]4, 6[ \end{cases}$$

The function  $f_0$  does not admit an additive extension to the full real line. Because if it had an additive extension  $f$ ,  $f$  will be constantly equal to 1 on the interval  $]1, 2[$ . Being locally bounded, by the previous corollary,  $f$  must be linear. But no linear function is constantly equal to 1 on the interval  $]1, 2[$  and we get a contradiction.

### 3.2.3 The fundamental Cauchy equations

The real line under addition is isomorphic to the interval  $]0, \infty[$  under multiplication. The natural exponential function serves as an isomorphism. The following are simple variations of the additive Cauchy equation.

$$f : \mathbb{R} \rightarrow ]0, \infty[, \quad f(x+y) = f(x)f(y) \quad (\forall x, y \in \mathbb{R}). \quad (3.25)$$

$$f : ]0, \infty[ \rightarrow \mathbb{R}, \quad f(xy) = f(x) + f(y) \quad (\forall x, y \in ]0, \infty[). \quad (3.26)$$

$$f : ]0, \infty[ \rightarrow ]0, \infty[, \quad f(xy) = f(x)f(y) \quad (\forall x, y \in ]0, \infty[). \quad (3.27)$$

We may conveniently label them as the exponential, the logarithmic, and the multiplicative Cauchy equation, respectively. Together with the additive Cauchy equation, the four are known as the fundamental Cauchy equations.

By using the stated isomorphism, the natural exponential function, each equation can be solved by converting to the additive Cauchy equation. Here we exhibit their general solutions.

**Corollary 3.4** *The general solution of the above three variations of the additive Cauchy equation are given by  $f(x) = e^{a(x)}$ ,  $f(x) = a(\ln(x))$ , and  $f(x) = e^{a(\ln(x))}$ , respectively. Here  $a$  stands for an additive function on the real line.*

## 3.3 The Pexider equation

The Pexider equation on the full real line is

$$f, g, h : \mathbb{R} \rightarrow \mathbb{R}, \quad f(x+y) = g(x) + h(y) \quad (\forall x, y \in \mathbb{R}). \quad (3.28)$$

**Theorem 3.3** *The general solution of the Pexider equation (3.28), is given by*

$$f(x) = a(x) + c_1 + c_2,$$

$$g(x) = a(x) + c_1,$$

$$h(x) = a(x) + c_2,$$

where  $a$  is an additive function and  $c_1, c_2$  are constants.

*Proof* Putting in (3.28)  $x = y = 0$  we get

$$f(0) = g(0) + h(0). \quad (3.29)$$

Subtracting this equation from (3.28) side by side and letting  $f_0 = f - f(0)$ ,  $g_0 = g - g(0)$ ,  $h_0 = h - h(0)$  we get

$$f_0(x+y) = g_0(x) + h_0(y) \quad (\forall x, y \in \mathbb{R}). \quad (3.30)$$

Putting  $x = 0$  in (3.30) and using the fact  $g_0(0) = 0$  we get  $f_0 = h_0$ . Similarly, putting  $y = 0$  in (3.30) yields  $f_0 = g_0$ . Let  $a := f_0 = g_0 = h_0$  and (3.30)

becomes the additive Cauchy equation  $a(x + y) = a(x) + a(y)$ . Thus  $a$  is additive. Now  $f(x) = a(x) + f(0)$ ,  $g(x) = a(x) + g(0)$  and  $h(x) = a(x) + h(0)$ . Letting  $c_1 = g(0)$ ,  $c_2 = h(0)$  and using (3.29), we arrive at the stated form of the solutions.

The converse statement is simple to check.  $\square$

When a functional equation containing one unknown function occurring several times is generalized by replacing each occurrence with a possibly different function, we frequently called the resulting equation a *Pexiderized* version. Similar to the above Pexiderization of the additive Cauchy equation, each of the remaining fundamental Cauchy equations has its Pexiderized version. Their solutions can be obtained via Theorem 3.3. The table below exhibits the equations and their general continuous solutions. The continuity assumption can be replaced by other conditions such as monotonicity.

$f(x + y) = g(x)h(y)$ $f, g, h : \mathbb{R} \rightarrow ]0, \infty[$	$f(x) = c_1c_2e^{cx}$ , $g(x) = c_1e^{cx}$ $h(x) = c_2e^{cx}$ , $c_1, c_2 > 0$
$f(xy) = g(x) + h(y)$ $f, g, h : ]0, \infty[ \rightarrow \mathbb{R}$	$f(x) = c \log x + c_1 + c_2$ , $g(x) = c \log x + c_1$ $h(x) = c \log x + c_2$
$f(xy) = g(x)h(y)$ $f, g, h : ]0, \infty[ \rightarrow ]0, \infty[$	$f(x) = c_1c_2x^c$ , $g(x) = c_1x^c$ $h(x) = c_2x^c$ , $c_1, c_2 > 0$

Attention should be paid to the domain and codomain of the functions. For example, the continuous solution of

$$f(x + y) = g(x)h(y) \quad (\forall x, y \in \mathbb{R}) \\ f : \mathbb{R} \rightarrow \mathbb{R}, \quad g : \mathbb{R} \rightarrow ]0, \infty[, \quad h : \mathbb{R} \rightarrow ]-\infty, 0[$$

is given by  $f(x) = c_1c_2e^{cx}$ ,  $g(x) = c_1e^{cx}$ ,  $h(x) = c_2e^{cx}$ ,  $c_1 > 0$ ,  $c_2 < 0$ .

For a much fuller account on Pexider's and related equations we refer readers to the book of Aczél (1966), section 3.1. More examples can be found in Aczél (1987).

**Example 3.3** Dzhafarov and Colonius (2011) reported on their understanding of Fechner's main idea. They see that as the summation of differential sensitivity values along an interval of stimulus values. Below is an outline of their observation.

Let  $a$  and  $b$  be stimuli above or at the threshold value 0. The hypothesis is that a respondent's subjective dissimilarity,  $D(a, b)$ , has the properties of a unidimensional distance:  $D(a, b) = 0$  if and only if  $a = b$ ;  $D(a, b) = D(b, a)$ ; and for all  $a \leq b \leq c$ , the additivity property

$$D(a, c) = D(a, b) + D(b, c). \tag{3.31}$$

This additivity property is central for Fechner's theory.

Weber's law is the assertion that the subjective dissimilarity between stimuli with physical magnitudes  $a$  and  $b$  (where  $0 < a \leq b$ ) is determined by the ratio of these magnitudes,  $b/a$ . In mathematical term it means that

$$D(a, b) = F(b/a)$$

for some function  $F$ .

Weber's law together with Fechner's additivity hypothesis then yield the functional equation

$$F(c/a) = F(b/a) + F(c/b). \quad (3.32)$$

With the change of variables  $x = b/a$ ,  $y = c/b$ , it becomes the logarithmic Cauchy equation

$$F(xy) = F(x) + F(y).$$

It holds for all  $x, y$  in some interval  $[1, k]$ ,  $k > 1$ . It comes natural with some regularity assumption, say  $D(a, b) \geq 0$ . So  $F(x) \geq 0$ . The solution of the equation is given by  $F(x) = K \log x$ . This yields Fechner's difference formula

$$D(a, b) = K \log(b/a) \quad (0 < a \leq b).$$

On the measuring scales of the variables we refer to Stevens (1973), Hayes and Embretson (2012), Sowden (2012) and Van Zandt and Townsend (2012). Weber's law and Fechner's logarithmic relation are also mentioned.

### 3.4 A generalized Pexider equation

The generalized Pexider equation

$$f(x + y) = g(x) + h(x)k(y) \quad (3.33)$$

was covered by Falmagne in (1985, theorem 3.6):

**Theorem 3.4** *Let  $I$  be an open interval containing 0 and let  $f, g, h, k$  be real-valued functions defined on  $I$ . Suppose that*

$$f(x + y) = g(x) + h(x)k(y) \quad (\forall x, y \in I \text{ with } x + y \in I) \quad (3.34)$$

*and  $f$  and  $k$  are strictly monotonic (or nonconstant and continuous). Then either  $h$  is a constant function and*

$$f(x) = \beta_0 x + \beta_1 + \beta_2,$$

$$g(x) = \beta_0 x + \beta_2,$$

$$h(x) = \beta_3,$$

$$k(x) = (\beta_0/\beta_3)x + (\beta_1/\beta_3)$$

for some constants  $\beta_0, \beta_1, \beta_2$  and  $\beta_3 \neq 0$ ; or

$$\begin{aligned} f(x) &= \beta_0(1 - e^{\lambda x}) + \beta_1 + \beta_2, \\ g(x) &= (\beta_0 + \beta_1)(1 - e^{\lambda x}) + \beta_2, \\ h(x) &= \beta_3 e^{\lambda x}, \\ k(x) &= (\beta_0/\beta_3)(1 - e^{\lambda x}) + (\beta_1/\beta_3) \end{aligned}$$

for some constants  $\beta_0, \beta_1, \beta_2, \beta_3 \neq 0$  and  $\lambda$ .

We choose not to include the proof of Theorem 3.4 here. Instead, we shall show the steps used in solving a very similar equation, (3.42), in the next section.

In Chudziak and Tabor (2008), (3.33) is studied in the following setting. Let  $X$  be a normed space. Let  $D$  be a nonempty, open and connected subset of  $X \times X$ . Let  $D_1 := \{x : (x, y) \in D \text{ for some } y\}$ ,  $D_2 := \{y : (x, y) \in D \text{ for some } x\}$  and  $D_+ := \{x + y : (x, y) \in D\}$ . The equation (3.33) is supposed to hold for all  $(x, y) \in D$ , and  $f, g, h, k$  are complex-valued functions with domains  $D_+, D_1, D_1$ , and  $D_2$ , respectively. A theorem of Radó and Baker (1987) is applied.

### 3.5 Some uniqueness and regularity theorems

Let  $X$  and  $Y$  be topological spaces,  $T : X \times Y \rightarrow \mathbb{R}$ . We consider the functional equation

$$f(x) + g(y) = h[T(x, y)] \quad (\forall x \in X, y \in Y) \quad (3.35)$$

where  $f, g$  and  $h$  are real-valued functions.

We state some uniqueness and regularity condition theorems.

**Theorem 3.5** (Pfanzagl, 1970) *Let  $X$  be a locally connected Hausdorff space with the property that any pair of points is contained in some open and connected set with compact closure. Let  $T : X^2 \rightarrow \mathbb{R}$  be continuous in each variable. If the functional equation*

$$f(x) + f(y) = h[T(x, y)] \quad (\forall x, y \in X)$$

*has two solutions  $(f_i, h_i)$  with continuous  $f_i$  ( $i = 0, 1$ ) and nonconstant  $f_0$  then  $f_1 = \alpha f_0 + \beta$  for some constants  $\alpha$  and  $\beta$ .*

**Theorem 3.6** (Ng, 1973b, theorem 2.0) *Let  $X$  be a locally connected Hausdorff space with the property that any pair of points is contained in some open and connected set with compact closure. Let  $Y$  be a connected topological space. Let  $T : X \times Y \rightarrow \mathbb{R}$  be continuous in each variable. If the functional equation (3.35)*

$$f(x) + g(y) = h[T(x, y)] \quad (\forall x \in X, y \in Y)$$

*has two solutions  $(f_i, g_i, h_i)$  with continuous  $f_i$  and nonconstant  $f_0$ , then  $g_1 = \alpha g_0 + \beta$  for some constants  $\alpha$  and  $\beta$ .*

The proof of the above theorem is led by the following lemma.

**Lemma 3.1** (*Ng, 1973b, lemma 2.1*) Let  $X$  be a locally connected Hausdorff space with the property that any pair of points is contained in some open and connected set with compact closure. Let  $Y$  be a connected space. Suppose that  $T : X \times Y \rightarrow \mathbb{R}$  is continuous in each variable. Let  $(f, g, h)$  be a solution of the functional equation (3.35)

$$f(x) + g(y) = h[T(x, y)] \quad (\forall x \in X, y \in Y)$$

with continuous  $f$ . If there exist  $x_1, x_2 \in X$  such that  $f(x_1) = f(x_2)$  and  $T(x_1, y) \neq T(x_2, y)$  for all  $y \in Y$  then  $g$  is a constant function on  $Y$ .

**Theorem 3.7** (*Ng, 1973b, theorem 2.1*) Let  $X$  be a locally connected Hausdorff space with the property that any pair of points is contained in some open and connected set with compact closure. Let  $Y$  be a connected space. Suppose that  $T : X \times Y \rightarrow \mathbb{R}$  is continuous in each variable. If  $(f_0, g_0, h_0)$  is a particular solution of (3.35)

$$f(x) + g(y) = h[T(x, y)] \quad (\forall x \in X, y \in Y)$$

with nonconstant continuous  $f_0$  and nonconstant  $g_0$  then the general solution  $(f, g, h)$  with continuous  $f$  is given by

$$f = c f_0 + c_1, \quad g = c g_0 + c_2, \quad h = c h_0 + c_1 + c_2 \quad (3.36)$$

for some constants  $c, c_1$  and  $c_2$ . Here the domain of  $h$  is the range of  $T$ .

**Remark 3.1** Theorem 3.6 and Theorem 3.7 are extensions of Theorem 3.5. The proofs, detailed in (Ng, 1973b), are based on the arguments seen in Pfanzagl (1970) and will not be reproduced here. Instead, the current article is focused on some applications of these theorems. The topological condition on  $X$  given in the above theorems is perhaps not always easy to observe. A more familiar class of spaces are the pathwise connected spaces. A space  $X$  is pathwise connected if for any two points  $x_1, x_2 \in X$ , there exists a continuous function (path)  $\phi : [0, 1] \rightarrow X$  such that  $\phi(0) = x_1$  and  $\phi(1) = x_2$ . We do not require that  $\phi$  is a topological embedding. It was mentioned in Ng (1973b), in passing, that the above theorems holds true for path-connected spaces  $X$ , and was an observation made by M. A. McKiernan. Here we shall include an argument to support that claim. We begin by showing that Lemma 3.1 holds true under the assumption that  $X$  is pathwise connected.

Let  $x_1, x_2 \in X$  be such that  $f(x_1) = f(x_2)$  and  $T(x_1, y) \neq T(x_2, y)$  for all  $y \in Y$ . Let  $\phi$  be a path that connects  $x_1$  to  $x_2$  in  $X$ . Replacing  $x$  by  $\phi(t)$  in the functional equation we get

$$f(\phi(t)) + g(y) = h[T(\phi(t), y)] \quad (\forall t \in [0, 1], y \in Y).$$

Letting  $\tilde{X} = [0, 1]$ ,  $\tilde{f}(t) = f(\phi(t))$  and  $\tilde{T}(t, y) = T(\phi(t), y)$  the equation becomes

$$\tilde{f}(t) + g(y) = h[\tilde{T}(t, y)] \quad (\forall t \in \tilde{X}, y \in Y). \quad (3.37)$$

At  $t_1 = 0$  and  $t_2 = 1$  we have  $\tilde{f}(t_1) = f(\phi(0)) = f(x_1)$  and similarly  $\tilde{f}(t_2) = f(x_2)$ . Thus the assumption  $f(x_1) = f(x_2)$  passes to  $\tilde{f}(t_1) = \tilde{f}(t_2)$ . Likewise, the assumption  $T(x_1, y) \neq T(x_2, y)$  for all  $y \in Y$  passes to  $\tilde{T}(t_1, y) \neq \tilde{T}(t_2, y)$  for all  $y \in Y$ . The continuity of  $f$  and  $T$  passes to  $\tilde{f}$  and  $\tilde{T}$ , respectively. The interval  $\tilde{X} = [0, 1]$  is a locally connected Hausdorff space where any pair of points is contained in some open and connected set with compact closure. Lemma 3.1 is applicable to (3.37), implying that  $g$  is constant on  $Y$ .

Having delivered Lemma 3.1 for pathwise connected spaces  $X$  we get the subsequent Theorem 3.6 and Theorem 3.7 for pathwise connected spaces by the same steps shown in Ng (1973b).

Although the above uniqueness theorems are stated for continuous solution, the continuity may follow from weaker assumptions. The following are regularity results reported in Ng (1973a) that shows boundedness may imply continuity.

**Theorem 3.8** *Let  $X$  be a space such that each pair of points is contained in the continuous image of some connected and locally connected space (e.g., when  $X$  is connected and locally connected, or when  $X$  is pathwise connected), and let  $Y$  be a topological space. Suppose that  $T : X \times Y \rightarrow \mathbb{R}$  is continuous in each variable. Let  $(f, g, h)$  be a solution of the equation (3.35)*

$$f(x) + g(y) = h[T(x, y)] \quad (\forall x \in X, y \in Y).$$

*If  $f$  is nonconstant and locally bounded from above (or from below) at each point of  $X$ , then  $g$  is continuous on  $Y$ .*

**Theorem 3.9** *For (3.35), if each pair of points of  $X$  is contained in some compact connected subset of  $X$ ,  $T$  is jointly continuous on the product space  $X \times Y$  and  $f$  is nonconstant and locally bounded from above on  $X$  (or locally bounded from below on  $X$ ), then  $g$  is continuous on  $Y$ .*

**Theorem 3.10** *For (3.35), if  $X$  is connected,  $T$  is continuous in each variable and  $f$  is nonconstant and bounded on  $X$  from both sides, then  $g$  is continuous on  $Y$ .*

**Example 3.4** Let  $X$  and  $Y$  be nondegenerated real intervals. Let  $T(x, y) := x + y$  for all  $x \in X, y \in Y$ . Equation (3.35) is the Pexider equation

$$f(x) + g(y) = h(x + y) \quad (\forall x \in X, y \in Y)$$

on a restricted domain. The functions  $f_0(x) = x$ ,  $g_0(y) = y$  and  $h_0(t) = t$  ( $x \in X$ ,  $y \in Y$  and  $t$  in the range of  $T$ ) constitute a particular solution with nonconstant and continuous  $f_0$  and nonconstant  $g_0$ . Thus we may conclude from the uniqueness theorem, Theorem 3.7, that the general solution of the equation with continuous  $f$

is given by

$$\begin{aligned} f(x) &= cf_0(x) + c_1 = cx + c_1 \quad (x \in X), \\ g(y) &= cg_0(y) + c_2 = cy + c_2 \quad (y \in Y), \\ h(t) &= ch_0(t) + c_1 + c_2 = ct + c_1 + c_2 \quad (t \in T(X, Y) = X + Y). \end{aligned}$$

**Example 3.5** Let  $X, Y$  be nondegenerated real intervals. Let  $T(x, y) := x$  for all  $x \in X, y \in Y$ . Equation (3.35) becomes

$$f(x) + g(y) = h(x) \quad (\forall x \in X, y \in Y).$$

Clearly, the equation implies that  $g = c$ , a constant, and that  $h = f + c$ . The converse is also clear. If we seek solutions with continuous  $f$ , then any continuous function  $f$  may serve. In particular, there are solutions  $(f_i, g_i, h_i)$ ,  $(i = 1, 2)$ , in which  $f_1$  and  $f_2$  are continuous and not affinely dependent. This shows that in Theorem 3.7, the assumption that  $g_0$  is nonconstant is not redundant for the conclusion.

**Example 3.6** *Plateau's experiment.* This example is extracted from Falmagne (1985). Plateau (1872) gave a pair of painted disks – one white and one black – to eight artists and instructed them to paint a gray disk midway between the two. The resulting gray disks, reported Plateau, were virtually identical for all eight artists, inspite of the variation in the illumination conditions. Let us suppose that such results would hold for any pair of gray disks. A formalization of these data leads to the functional equations

$$\lambda M(x, y) = M(\lambda x, \lambda y), \quad (3.38)$$

$$u[M(x, y)] = \frac{u(x) + u(y)}{2}. \quad (3.39)$$

Here,  $(x, y)$  stands for a pair of gray disks in some specified viewing condition, and  $M(x, y)$  is the resulting gray disk under the same viewing condition. The first equation corresponds to the statement that the result is independent of the illumination.  $u$  is a psychophysical scale, mapping the lux scale into the reals, such that midway is realized. It is shown in Falmagne (1985) that under mild conditions  $u$  must be either a power function  $u(x) = \alpha x^\beta + \gamma$  or logarithmic  $u(x) = \alpha \log x + \gamma$ .

Here we outline a deduction of the asserted form of  $u$  using the uniqueness theorem. Putting  $x = \lambda x$ , and  $y = \lambda y$  into the second equation and using the first equation we get

$$2u[\lambda M(x, y)] = u(\lambda x) + u(\lambda y). \quad (3.40)$$

Treating  $\lambda$  momentarily as a parameter and consider  $f_\lambda(x) := u(\lambda x)$ ,  $g_\lambda(y) := u(\lambda y)$ ,  $h_\lambda(t) := 2u(\lambda t)$  and  $T = M$  the equation takes the form

$$f_\lambda(x) + g_\lambda(y) = h_\lambda[T(x, y)] \quad (\forall x \in X, y \in Y), \quad (3.41)$$

where we see a family of solutions  $(f_\lambda, g_\lambda, h_\lambda)$  indexed by  $\lambda$ . Assume that  $X = Y$  and is a nondegenerated interval and that  $u$  is continuous and strictly increasing.

The uniqueness theorem is applicable with  $f_0 = u$ ,  $g_0 = u$  and  $h_0 = 2u$ . Hence for each  $\lambda$  there exist constants  $c(\lambda)$ ,  $c_1(\lambda)$ ,  $c_2(\lambda)$  such that

$$\begin{aligned}f_\lambda(x) &= c(\lambda)f_0(x) + c_1(\lambda) \quad (x \in X), \\g_\lambda(y) &= c(\lambda)g_0(y) + c_2(\lambda) \quad (y \in Y), \\h_\lambda(t) &= c(\lambda)h_0(t) + c_1(\lambda) + c_2(\lambda) \quad (t \in X + Y := T(X, Y)).\end{aligned}$$

Returning to the definitions of  $f_\lambda$ ,  $g_\lambda$  and  $h_\lambda$  it means

$$\begin{aligned}u(\lambda x) &= c(\lambda)u(x) + c_1(\lambda) \quad (x \in X), \\u(\lambda y) &= c(\lambda)u(y) + c_2(\lambda) \quad (y \in Y), \\2u(\lambda t) &= 2c(\lambda)u(t) + c_1(\lambda) + c_2(\lambda) \quad (t \in X + Y := T(X, Y)).\end{aligned}$$

Because  $X = Y$  is assumed, comparing the first two equations we get  $c_1 = c_2 =: \delta$ , say. The three equations become one:

$$u(\lambda x) = c(\lambda)u(x) + \delta(\lambda). \quad (3.42)$$

Equation (3.42) is very much like (3.33) and we proceed to solve it.

To be on a more concrete setting, let us assume that (3.42) holds for all  $\lambda \in ]0, 1]$  and for all  $x \in X = [0, 1]$ . The continuity of the nonconstant function  $u$  implies that  $c$  and  $\delta$  are continuous functions. Putting  $x = \mu x$  in (3.42) and using the equation again we get

$$\begin{aligned}u(\lambda(\mu x)) &= c(\lambda)u(\mu x) + \delta(\lambda) \\&= c(\lambda)[c(\mu)u(x) + \delta(\mu)] + \delta(\lambda) \\&= c(\lambda)c(\mu)u(x) + c(\lambda)\delta(\mu) + \delta(\lambda).\end{aligned}$$

On the other hand we also have

$$\begin{aligned}u(\lambda(\mu x)) &= u((\lambda\mu)x) \\&= c(\lambda\mu)u(x) + \delta(\lambda\mu).\end{aligned}$$

Comparison gives

$$c(\lambda)c(\mu)u(x) + c(\lambda)\delta(\mu) + \delta(\lambda) = c(\lambda\mu)u(x) + \delta(\lambda\mu)$$

and thus

$$c(\lambda)c(\mu) = c(\lambda\mu), \quad (3.43)$$

$$c(\lambda)\delta(\mu) + \delta(\lambda) = \delta(\lambda\mu). \quad (3.44)$$

The first equation, (3.43), with continuous  $c$ , yields

$$c(\lambda) = \lambda^\beta \quad (\forall \lambda \in ]0, 1]) \quad (3.45)$$

for some constant  $\beta$ , or  $c$  is the constant function 0. We discard the latter as it would lead to a constant function  $u$  when we put it back into (3.42). Putting (3.45) back into (3.42) we get

$$\lambda^\beta \delta(\mu) + \delta(\lambda) = \delta(\lambda\mu) \quad (\forall \lambda, \mu \in ]0, 1]). \quad (3.46)$$

The symmetry of the right-hand side in  $\lambda$  and  $\mu$  implies

$$\lambda^\beta \delta(\mu) + \delta(\lambda) = \mu^\beta \delta(\lambda) + \delta(\mu) \quad (\forall \lambda, \mu \in ]0, 1]).$$

Thus

$$[\lambda^\beta - 1]\delta(\mu) = [\mu^\beta - 1]\delta(\lambda) \quad (\forall \lambda, \mu \in ]0, 1]). \quad (3.47)$$

There are two cases to consider.

Case 1. Suppose that  $\beta = 0$ . In this case we have  $c(\lambda) = 1$  for all  $\lambda$  and (3.46) reduces to the logarithmic Cauchy equation

$$\delta(\mu) + \delta(\lambda) = \delta(\lambda\mu) \quad (\forall \lambda, \mu \in ]0, 1]). \quad (3.48)$$

With continuity of  $\delta$ , it implies

$$\delta(\lambda) = \alpha \log(\lambda) \quad (\forall \lambda \in ]0, 1])$$

for some constant  $\alpha$ . We discard the case  $\alpha = 0$ , as that would lead to a constant function  $u$ . Putting  $c(\lambda) = 1$  and  $\delta(\lambda) = \alpha \log(\lambda)$  into (3.42) we get

$$u(\lambda x) = \alpha \log(\lambda) + u(x). \quad (3.49)$$

Theorem 3.6 is applicable with  $T(\lambda, x) = \lambda x$ . A particular solution of (3.49) is  $u_0(x) = \alpha \log x$ . Thus  $u$  is given by

$$u(x) = \alpha \log(x) + \gamma \quad (3.50)$$

for some constant  $\gamma$ .

Case 2. Suppose that  $\beta \neq 0$ . Fixing in (3.46)  $\mu = \mu_0 \neq 1$  gives

$$\delta(\lambda) = \frac{\delta(\mu_0)}{\mu_0^\beta - 1} [\lambda^\beta - 1] \quad (\forall \lambda \in ]0, 1]).$$

Putting that back into (3.42) we get

$$u(\lambda x) = \lambda^\beta u(x) + \frac{\delta(\mu_0)}{\mu_0^\beta - 1} [\lambda^\beta - 1].$$

Letting  $\gamma := -\frac{\delta(\mu_0)}{\mu_0^\beta - 1}$  and rearranging terms it becomes

$$u(\lambda x) - \gamma = \lambda^\beta [u(x) - \gamma].$$

Letting  $v(x) := u(x) - \gamma$  it becomes

$$v(\lambda x) = \lambda^\beta v(x).$$

Its solution is given by

$$v(x) = \alpha x^\beta$$

for some constant  $\alpha$ . Recall that  $v(x) = u(x) - \gamma$  and so

$$u(x) = \alpha x^\beta + \gamma$$

follows.

The lemma below is a small observation used to support the next example.

**Lemma 3.2** *Let  $]a, b[$  ( $a < b$ ) be an open interval and let  $f : ]a, b[ \rightarrow \mathbb{R}$  be a nonconstant continuous function. Then there exists a point  $c \in ]a, b[$  such that  $f$  is nonconstant on both subintervals  $]a, c[$  and  $]c, b[$ .*

*Proof* Let  $c_0$  be a sample point taken from  $]a, b[$ . If  $f$  is nonconstant on both  $]a, c_0[$  and  $]c_0, b[$  then we are done. Else, without loss of generality, say that  $f$  is constant on  $]a, c_0[$ . There exists a highest point  $a'$  in  $]a, b[$  such that  $f$  is constant on  $]a, a'[$ . Then  $a' \geq c_0$  and  $f$  is nonconstant on  $]a', b[$ .

Let  $c_1$  be a sample point taken from  $]a', b[$ . By the choice of  $a'$ ,  $f$  is nonconstant on  $]a, c_1[$ . If  $f$  is nonconstant on  $]c_1, b[$  we are done. Else  $f$  is constant on  $]c_1, b[$ . Then there exists a lowest point  $b'$  in  $]a, b[$  such that  $f$  is constant on  $]b', b[$ . It is clear that  $b' \leq c_1$  and  $a' < b'$ . Now, for all  $c$  belong to  $]a', b[$  we have the nonconstancy of  $f$  on both  $]a, c[$  and  $]c, b[$ .  $\square$

**Example 3.7** This example is taken from Ng *et al.*, (2008, lemma 13).

Let  $\alpha_1, \alpha_2 : ]0, 1[ \rightarrow ]0, 1[$  be homeomorphisms and let  $f : ]0, 1[ \rightarrow \mathbb{R}$  be a continuous function satisfying the functional equation

$$f(\alpha_1(r)x) + f(\alpha_2(r)x) = f(x) \quad (r, x \in ]0, 1[). \quad (3.51)$$

We shall show that the solution is given by either  $f = 0$  or

$$f(x) = cx^\rho \quad \text{and} \quad \alpha_1(r)^\rho + \alpha_2(r)^\rho = 1 \quad (r, x \in ]0, 1[) \quad (3.52)$$

for some constants  $c \neq 0, \rho > 0$ .

Here are the steps. Letting  $y := \alpha_2(r)x$  and  $q := \alpha_1 \circ \alpha_2^{-1}$ , the equation takes the form

$$f(xq(y/x)) = f(x) - f(y) \quad (y, x \in ]0, 1[, y < x) \quad (3.53)$$

where  $q$  is again a homeomorphism on  $]0, 1[$ . The only constant function  $f$  satisfying the equation is  $f = 0$ . Assume now that  $f$  is nonconstant. By Lemma 3.2, applied twice, there exist  $a_1, a_2$  in  $]0, 1[$ , with  $a_1 < a_2$ , such that  $f$  is nonconstant on each of the subintervals  $]0, a_1[$ ,  $]a_1, a_2[$  and  $]a_2, 1[$ . By inspection, for each fixed  $\mu \in ]0, 1[$ , the map  $f_\mu$  defined by  $f_\mu(x) := f(\mu x)$  is again a continuous solution of (3.53). Applying Theorem 3.7 with  $T(x, y) = xq(y/x)$  over the two rectangular subdomains  $D_i := \{(x, y) | x, y \in ]0, 1[, y < a_i < x\}$ , ( $i = 1, 2$ ), we get that there exist constants  $\phi(\mu, a_i)$ ,  $\psi_1(\mu, a_i)$  and  $\psi_2(\mu, a_i)$  such that

$$\begin{aligned} f(\mu x) &= \phi(\mu, a_i)f(x) + \psi_1(\mu, a_i) \quad (x \in ]a_i, 1[), \\ f(\mu y) &= \phi(\mu, a_i)f(y) - \psi_2(\mu, a_i) \quad (y \in ]0, a_i[), \\ f(\mu u) &= \phi(\mu, a_i)f(u) + \psi_1(\mu, a_i) + \psi_2(\mu, a_i) \quad (u \in T(D_i)). \end{aligned} \quad (3.54)$$

The consistency of the first equation holding for  $a_1$  and for  $a_2$  and the nonconstancy of  $f$  on the common interval  $]a_2, 1[$  yields  $\phi(\mu, a_1) = \phi(\mu, a_2) =: \phi(\mu)$  and  $\psi_1(\mu, a_1) = \psi_1(\mu, a_2)$ . Similarly, from the second equation for  $a_1$  and for  $a_2$  we get also  $\psi_2(\mu, a_1) = \psi_2(\mu, a_2)$ . Comparing the first equation for  $a_1$  with the second for  $a_2$  over the common interval  $]a_1, a_2[$  on which  $f$  is nonconstant, we get that  $\psi_1(\mu, a_1) = -\psi_2(\mu, a_2)$ . The third equation thus becomes  $f(\mu u) = \phi(\mu)f(u)$ . When this third equation is compared with the first two, we get  $\psi_i = 0$ . The union of the equations of (3.54) is, therefore, simply

$$f(\mu t) = \phi(\mu)f(t) \quad (\mu, t \in ]0, 1[).$$

This is a Pexider multiplicative equation, and  $f$  has the form

$$f(t) = ct^\rho \quad (t \in ]0, 1[)$$

for some constants  $c \neq 0, \rho \neq 0$ . Putting it back into (3.51) we arrive at  $\alpha_1(r)^\rho + \alpha_2(r)^\rho = 1$  as the necessary and sufficient condition for the equation to hold. This condition itself rules out the use of negative  $\rho$  values and we have proved (3.52).

It would be interesting to know what the general solution of (3.51) is without the continuity assumption on  $f$ . For example, with  $\alpha_1(r) = r$  and  $\alpha_2(r) = 1 - r$ ,  $f$  is additive.

### 3.6 Differentiable solutions

When the functions in a functional equation are differentiable, we often obtain its (differentiable) solution by solving some resulting differential equations. We give some examples to illustrate the ideas.

**Example 3.8** We shall show that the general differentiable solution of the Pexider equation, (3.28),

$$f, g, h : \mathbb{R} \rightarrow \mathbb{R}, \quad f(x+y) = g(x) + h(y) \quad (\forall x, y \in \mathbb{R})$$

is given by

$$f(x) = cx + c_1 + c_2,$$

$$g(x) = cx + c_1,$$

$$h(x) = cx + c_2$$

where  $c, c_1$  and  $c_2$  are arbitrary constants.

Differentiating the two sides of the equation with respect to the variable  $x$  we get

$$f'(x+y) = g'(x) \quad (\forall x, y \in \mathbb{R}).$$

Putting in it  $x = 0$  and setting  $c := g'(0)$  we get the differential equation  $f'(y) = c$  for all  $y \in \mathbb{R}$ . Hence  $f(y) = cy + d$  for some constant  $d$ . Putting that form of  $f$  back into the Pexider equation we get

$$cx + cy + d = g(x) + h(y) \quad (\forall x, y \in \mathbb{R}).$$

Separating the variables we get  $cy + d - h(y) = g(x) - cx$ . The left does not depend on  $x$  and the right side does not depend on  $y$ . Thus  $cy + d - h(y) = g(x) - cx = c_1$ , a constant. Letting  $c_2 := d - c_1$  we arrive at the asserted affine-ness of the solutions.

We have thus obtained all the differentiable solutions without passing through Theorem 3.3, and in particular, without prior knowledge of the general solution.

**Proposition 3.1** *The general twice differentiable solution of d'Alembert's functional equation*

$$f : \mathbb{R} \rightarrow \mathbb{R}, \quad f(x+y) + f(x-y) = 2f(x)f(y) \quad (\forall x, y \in \mathbb{R}) \quad (3.55)$$

is given by

$$f(x) = 0, \quad f(x) = \cos(ax) \quad \text{and} \quad f(x) = \cosh(ax)$$

where  $a$  is a constant.

*Proof* Putting  $y = 0$  in the equation we get  $2f(x) = 2f(x)f(0)$  for all  $x$ . If  $f(0) = 0$  we get the trivial solution  $f(x) = 0$ .

Else we get  $f(0) = 1$ . Putting  $x = 0$  in the equation and using  $f(0) = 1$  we get that  $f$  is even:  $f(y) = f(-y)$  for all  $y$ . The evenness of  $f$  implies that  $f''$  is also even:  $f''(y) = f''(-y)$ . Differentiating the equation with respect to  $x$  twice we get

$$f''(x+y) + f''(x-y) = 2f''(x)f(y).$$

Putting in it  $x = 0$  gives  $f''(y) + f''(-y) = 2f''(0)f(y)$ . Because  $f''$  is even we get  $2f''(y) = 2f''(0)f(y)$  and thus the differential equation  $f''(x) = cf(x)$  where  $c = f''(0)$ . Solving the differential equation we reach the conclusion.  $\square$

**Proposition 3.2** *Let  $f : ]0, 1[ \rightarrow \mathbb{R}$ . The general twice differentiable solution of the functional equation*

$$f(x) + (1-x)f\left(\frac{y}{1-x}\right) = f(y) + (1-y)f\left(\frac{x}{1-y}\right) \quad (3.56)$$

$(\forall x, y \in ]0, 1[ \text{ with } x+y < 1)$

is given by

$$f(x) = cx \log x + c(1-x) \log(1-x) + c_1x + c_2 \quad (3.57)$$

where  $c, c_1$  and  $c_2$  are constants.

*Proof* Differentiating the functional equation with respect to the variable  $x$  we get

$$f'(x) - f\left(\frac{y}{1-x}\right) + \frac{y}{1-x}f'\left(\frac{y}{1-x}\right) = f'\left(\frac{x}{1-y}\right).$$

Differentiating the above equation with respect to  $y$  we get

$$\frac{y}{(1-x)^2}f''\left(\frac{y}{1-x}\right) = \frac{x}{(1-y)^2}f''\left(\frac{x}{1-y}\right).$$

Letting  $u = \frac{y}{1-x}$  and  $v = \frac{x}{1-y}$  the equation becomes

$$u(1-u)f''(u) = v(1-v)f''(v).$$

It holds for all  $u$  and  $v$  in  $]0, 1[$  as we can taking  $x = (v - vu)/(1 - vu)$  and  $y = (u - vu)/(1 - vu)$ . The successful separation of variables yields  $u(1-u)f''(u) = c$ , a constant. Solving this differential equation we obtain the asserted form  $f(x) = cx \log x + c(1-x) \log(1-x) + c_1x + c_2$ .  $\square$

The general solution of (3.56) is reported in Maksa and Ng (1986).

Here is a short introduction as how Equation (3.56), known as the fundamental equation of information measures, is motivated by the characterizations of the Shannon entropy. Let  $(p_1, p_2, \dots, p_n)$  be a complete probability distribution:  $p_i \geq 0$  and  $\sum p_i = 1$ . The Shannon entropy associated with an experiment with underlying distribution  $(p_1, p_2, \dots, p_n)$  is given by  $S_n(p_1, p_2, \dots, p_n) = \sum_{i=1}^n -p_i \log_2 p_i$ , where the convention  $0 \log_2 0 = 0$  is used to handle the zero probability. This entropy has many properties. For example,  $S_n(p_1, p_2, \dots, p_n)$  is symmetric in the variables  $p_i$  and satisfies the recursive relations

$$\begin{aligned} S_n(p_1, p_2, \dots, p_n) &= S_{n-1}(p_1 + p_2, p_3, \dots, p_n) \\ &\quad + (p_1 + p_2)S_2(p_1/(p_1 + p_2), p_2/(p_1 + p_2)). \end{aligned}$$

In particular

$$\begin{aligned} S_3(p_1, p_2, p_3) &= S_2(p_1 + p_2, p_3) \\ &\quad + (p_1 + p_2)S_2(p_1/(p_1 + p_2), p_2/(p_1 + p_2)). \end{aligned}$$

Letting  $f(p) := S_2(p, 1-p)$  we get  $S_3(p_1, p_2, p_3) = f(p_3) + (p_1 + p_2)f(p_1/(p_1 + p_2))$ . The symmetry  $S_3(p_1, p_2, p_3) = S_3(p_3, p_2, p_1)$  then yields

$$f(p_3) + (p_1 + p_2)f(p_1/(p_1 + p_2)) = f(p_1) + (p_3 + p_2)f(p_3/(p_3 + p_2)).$$

Letting  $x = p_1$ ,  $y = p_2$ , and staying with strictly positive probabilities, we arrive at (3.56).

Let  $\alpha \neq 1$  be a real number. The entropies of degree  $\alpha$  (cf. Havrda and Charvát, 1967; Tsallis, 1988) is the one-parameter family of measures

$$E_n^\alpha(p_1, \dots, p_n) = \frac{1}{2^{1-\alpha} - 1} \left( \sum_{i=1}^n p_i^\alpha - 1 \right). \quad (3.58)$$

Because

$$\lim_{\alpha \rightarrow 1} E_n^\alpha = S_n \quad (3.59)$$

it is natural to define  $E_n^1$  by  $E_n^1 = S_n$ , the Shannon entropy.

Entropies of degree  $\alpha$  satisfies the recursivity

$$\begin{aligned} E_n^\alpha(p_1, \dots, p_n) &= E_{n-1}^\alpha(p_1 + p_2, p_3, \dots, p_n) \\ &\quad + (p_1 + p_2)^\alpha E_2^\alpha(p_1/(p_1 + p_2), p_2/(p_1 + p_2)) \end{aligned} \quad (3.60)$$

and are symmetric functions. Parallel to the derivation of (3.56), the function  $f(p) := E_2^\alpha(p, 1-p)$  satisfies

$$\begin{aligned} f(x) + (1-x)^\alpha f\left(\frac{y}{1-x}\right) &= f(y) + (1-y)^\alpha f\left(\frac{x}{1-y}\right) \quad (3.61) \\ (\forall x, y \in ]0, 1[ \text{ with } x+y < 1). \end{aligned}$$

Solving this equation for  $\alpha \neq 1$  leads to an axiomatic characterization of the entropies of degree  $\alpha \neq 1$ . More characterizations can be found in Aczél and Daróczy (1975). We shall extend the equation to allow higher dimensions and then show how it is solved. The main purpose of this chapter is to demonstrate the diverse methods in treating functions equations. A reason to pick the topic of entropies for the demonstration is that in recent works, the use of entropies gains some attention, e.g. Ng *et al.* (2008, 2009), Luce *et al.* (2008a, 2008b).

Let  $m \geq 1$  be fixed and let  $I = ]0, 1[^m$  be the  $m$ -dimensional unit interval. Let  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m)$  be a fixed  $m$ -tuple of real numbers. For  $x \in I$ , let  $(1-x)^\alpha$  denote  $\prod_{i=1}^m (1-x_i)^{\alpha_i}$ . The functional equation

$$\begin{aligned} f(x) + (1-x)^\alpha f\left(\frac{y}{1-x}\right) &= f(y) + (1-y)^\alpha f\left(\frac{x}{1-y}\right) \quad (3.62) \\ (\forall x, y \in I \text{ with } x+y \in I) \end{aligned}$$

is called the  $m$ -dimensional fundamental equation of information of degree  $\alpha$ . The algebraic operations in the arguments of  $f$  as well as the exponents are performed coordinatewise. Equation (3.61) is the special case of (3.62) at  $m = 1$ . While the one-dimensional equation is motivated by the study of the entropies of degree  $\alpha$  which are functions of one-dimensional probability distributions, the two-dimensional equation is motivated by the study of functions, or measures, based on two probability distributions. An example of such a measure is the error function introduced by Kerridge (1961) (also known as the divergence, introduced by Kullback (1959)):

$$E_n(p_1, \dots, p_n; q_1, \dots, q_n) = \sum_{i=1}^n p_i \log_2 \frac{p_i}{q_i}. \quad (3.63)$$

**Proposition 3.3** (Ng, 1980a, 1980b) Suppose that  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m)$  is not a basic unit vector (a vector with exactly one  $\alpha_i = 1$  and the rest are zero). The general solution of (3.62) is given by

$$f(x) = \begin{cases} ax^\alpha + b(1-x)^\alpha - b & \text{if } \alpha \neq (0, \dots, 0), \\ a + \sum_{i=1}^m \ell_i(1-x_i) & \text{if } \alpha = (0, \dots, 0). \end{cases}$$

Here, each  $\ell_i$  is a solution of the logarithmic Cauchy equation

$$\ell(xy) = \ell(x) + \ell(y) \quad (\forall x, y \in I).$$

The proof is given in the appendix.

The error function, (3.63), falls under the case of  $\alpha = (1, 0)$ , a unit vector. For its treatment we refer to Aczél and Ng (1981), Ebanks *et al.* (1998) and chapter 10 of Kannappan (2009) provide a rich source of references for functional equations arising from information theory.

### 3.7 Regularity conditions

Sometimes from a functional equation itself we can infer from one regularity property to another regularity property. We give some illustrations.

**Proposition 3.4** *Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be an additive function. If it is Lebesgue measurable on some subset of positive (Lebesgue) measure, then it is bounded in some neighborhood of 0.*

*Proof* We shall use Steinhaus' theorem: if  $E \subset \mathbb{R}$  has positive Lebesgue measure, then  $E - E := \{x - y \mid x, y \in E\}$  is a neighbourhood of 0.

Another theorem about Lebesgue measurable functions is that they are nearly bounded: if  $g : D \rightarrow \mathbb{R}$  is Lebesgue measurable, and the (Lebesgue) measure  $m(D)$  is finite. Then, for all given  $\epsilon > 0$ , there exists a subset  $E$  of  $D$  so that  $m(D \setminus E) < \epsilon$  and  $g$  is bounded on  $E$ .

Now let  $f$  be an additive function on the real line, and suppose that it is Lebesgue measurable on a set  $S$  of positive measure. Without loss of generality, we may assume that  $m(S)$  is finite. Applying the above near-boundedness result with  $\epsilon = m(S)/2$ , we get that there exists a subset  $E$  of  $S$  such that  $m(S \setminus E) < m(S)/2$  and  $|f|$  is bounded by a finite constant  $M$  on  $E$ . Then  $m(E) \geq m(S)/2 > 0$  and so  $E - E$  is a neighborhood of 0. For arbitrary  $x, y \in E$ , using the additivity of  $f$  we get  $f(x - y) = f(x) - f(y)$ , thus  $|f(x - y)| = |f(x) - f(y)| \leq |f(x)| + |f(y)| \leq M + M = 2M$ . This shows that  $|f|$  is bounded on the neighborhood  $E - E$  of 0 by  $2M$ .  $\square$

Proposition 3.4 and Corollary 3.2 together imply that Lebesgue measurable additive functions are linear.

Lee (1964) gives the Lebesgue measurable solutions of Equation (3.56) by first showing that each Lebesgue measurable solution is in fact bounded on some proper interval. The proof again illustrates the use of Steinhaus-type theorems: the mapping  $(s, t) \mapsto m(A \cap (1 - sB) \cap tC)$  is continuous (cf. Aczél and Daróczy, 1975, theorem 3.4.16).

**Proposition 3.5** (Andrade, 1900) *Continuous solutions of d'Alembert's functional equation, (3.55), have derivatives of all orders.*

*Proof* Integrate the equation with respect to  $y$  on the interval 0 to  $z$  yields

$$\int_x^{x+z} f(t) dt - \int_x^{x-z} f(t) dt = 2f(x) \int_0^z f(t) dt.$$

By the fundamental theorems of calculus, the left-hand side is a differentiable function of  $x$ . Thus, the right-hand side is differentiable in  $x$ . If  $\int_0^z f(t)dt$  vanishes for all  $z$  then  $f$  is the constant function 0 which has derivatives of all orders. Else, fix one  $z = z_0$  for which  $\int_0^{z_0} f(t)dt \neq 0$  and we get the differentiability of  $f$ . Knowing that  $f$  is differentiable once, the left-hand side is then twice-differentiable in  $x$ . This in turn implies that  $f(x)$  on the right-hand side is twice-differentiable. Iterating this process, we conclude that  $f$  is differentiable as many times as we wish.  $\square$

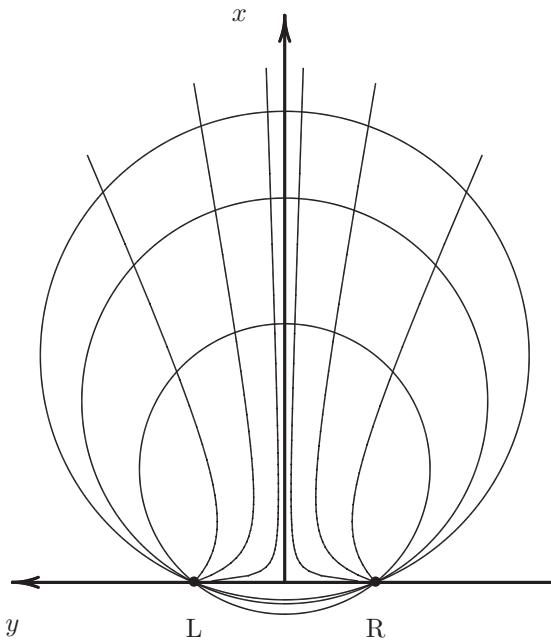
It is apparent that, based on the same proof lines, Lebesgue integrable solutions of d'Alembert's functional equation are continuous. For more examples see, e.g., Aczél (1966, section 4.2).

### 3.8 Some functional equations in binocular space perception

Most introductory materials in this section are extracted from Aczél *et al.* (1999). We only give a brief outline of the derivation of the equations. Readers need to refer to the paper itself and to Luneburg (1947) for full detail. Here, our attention is devoted to the solving of the equations.

The model is about visual observation of stimuli in the horizontal plane at eye level. Core to the Luneburg theory of binocular vision (Luneburg, 1947) are the postulates: visual space, coordinatized with respect to perceived egocentric distance and perceived direction, is a Riemannian space of constant Gaussian curvature. Its metric determines the perceived interpoint distances and its geodesics are the perceptual straight lines. The points lying on a Vieth–Müller circle are perceived as being equidistant from the observer, and the hyperbolae of Hillebrand are perceived as radial lines of constant direction.

The Vieth–Müller circles go through  $L$  and  $R$  (the rotation centers of the left and right eye, respectively; see Figure 3.1). Each Hillebrand hyperbola branch goes through either  $L$  or  $R$ , except for the  $x$ -axis, which is a degenerated hyperbola of Hillerbrand. The point of intersection of a hyperbola branch and of a Vieth–Müller circle, connected with that circle's point at the negative vertical axis, yields a straight line meeting the vertical axis under constant angle for a given hyperbola. Under idealized assumptions on the optics of the eye, the Vieth–Müller circle through the fixation point is the locus of stimuli that are projected on the corresponding retinal points (cf. Howard and Rogers, 1995). For this to be true, the center of rotation and the optical node have to be assumed to coincide with the center of curvature of a spherical eye. Then, corresponding retinal points are congruent when the two retinas are superimposed so that the foveæ, onto which the fixation point is projected, coincide. Accordingly, the hyperbola of Hillebrand is the locus of symmetric retinal points, which deviate to the same extent but in opposite directions from the foveæ.



**Figure 3.1** Vieth–Müller circles and hyperbolæ of Hillebrand. The left eye is centered at L, the right eye at R (Aczél et al., 1999).

The most often used modified bipolar coordinates are then given by the linear combinations

$$\begin{aligned}\gamma &= \alpha - \beta, \\ \varphi &= \frac{\alpha + \beta}{2}.\end{aligned}\tag{3.64}$$

The bipolar parallax  $\gamma$  is the difference of the monocular directions and the bipolar latitude  $\varphi$  is their average. Note that the vertex of the angle  $\varphi$  in Figure 3.2 is not at the origin of the coordinate system, but at the intersection of the Vieth–Müller circle with the negative part of the  $x$ -axis.

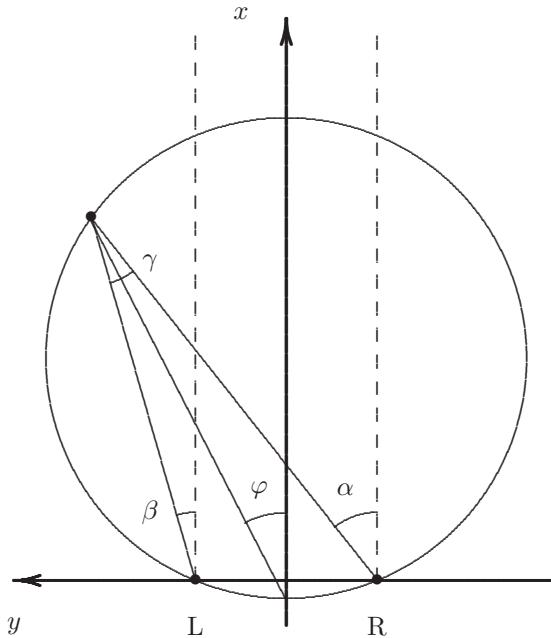
The trajectories  $\gamma = \text{constant}$  are the Vieth–Müller circles, while the trajectories  $\varphi = \text{constant}$  are the hyperbolæ of Hillebrand. Luneburg's postulate (L2) is then equivalent to the statement that the perceived egocentric distance only depends on  $\gamma$ , and the perceived direction only on  $\varphi$ .

The loci of perceived equidistance and constant perceived direction, respectively, as equivalence classes that are induced by ordering relations.

The set of stimuli is given by

$$S = \{(\alpha, \beta) \mid \alpha, \beta \in ]-\pi/2, \pi/2[, \alpha > \beta\}.$$

We define  $\preceq_\rho$ ,  $\preceq_\theta$  on  $S$  which describe an observer's qualitative order judgments with respect to the stimuli. By  $(\alpha, \beta) \preceq_\rho (\alpha', \beta')$  we express that the perceived egocentric distance of the stimulus  $(\alpha, \beta)$  is judged not to exceed that of  $(\alpha', \beta')$ .



**Figure 3.2** The bipolar coordinates  $\alpha, \beta$  denote the monocular directions with respect to the right and the left eye. The bipolar parallax  $\gamma$  associated to a stimulus is the angle subtended by the visual axis when the eyes converge on the stimulus. The bipolar latitude  $\varphi$  describes the lateral deviation of a stimulus from the  $x$ -axis (Aczél et al., 1999).

Similarly,  $(\alpha, \beta) \preceq_\theta (\alpha', \beta')$  means that  $(\alpha, \beta)$  is not perceived as lying to the right of  $(\alpha', \beta')$ . The relation  $\preceq_\rho$  thus orders the stimuli from near to far, and  $\preceq_\theta$  from left to right. Let  $\sim_\rho$  and  $\sim_\theta$  denote the symmetric parts of  $\preceq_\rho$  and  $\preceq_\theta$ , respectively. Using definitions (3.64) we may now reformulate Luneburg's psychophysical assumptions concerning the relations  $\preceq_\rho$  and  $\preceq_\theta$  in the language of representational measurement theory. For all  $(\alpha, \beta), (\alpha', \beta') \in S$

$$\begin{aligned} (\alpha, \beta) \preceq_\rho (\alpha', \beta') &\iff \alpha - \beta \geq \alpha' - \beta', \\ (\alpha, \beta) \preceq_\theta (\alpha', \beta') &\iff \alpha + \beta \geq \alpha' + \beta'. \end{aligned} \quad (3.65)$$

As an immediate consequence of (3.65) we observe that the interesting empirical relationship

$$(\alpha, \beta) \preceq_\rho (\alpha', \beta') \iff (\alpha, \beta') \preceq_\theta (\alpha', \beta) \quad (3.66)$$

holds, whenever  $(\alpha, \beta), (\alpha', \beta'), (\alpha, \beta'), (\alpha', \beta) \in S$ . Following the terminology of Krantz et al. (1971), we call binary relations  $\preceq_\rho, \preceq_\theta$  dual.

In a careful and well-documented study, Foley (1966) found that judgments of perceived equidistance exhibit remarkable and consistent deviations from the Vieth–Müller circle and large interindividual differences. The locus of perceived

equidistance shows a tendency toward physical equidistance, as already reported by Hardy *et al.* (1953), and may be considerably skewed with respect to the  $x$ -axis. Several approaches have been devised to account for these empirical facts.

Our generalization is obtained by assuming that an observer's judgments are not based on a direct comparison of the respective bipolar coordinates as in (3.65), but on comparing some transformation or evaluation of these coordinates. In other words, we assume that there exist strictly increasing functions  $f, g$ , mapping  $] -\pi/2, \pi/2[$  onto real intervals such that for all  $(\alpha, \beta), (\alpha', \beta') \in S$

$$\begin{aligned} (\alpha, \beta) \preceq_{\rho} (\alpha', \beta') &\iff f(\alpha) - g(\beta) \geq f(\alpha') - g(\beta'), \\ (\alpha, \beta) \preceq_{\theta} (\alpha', \beta') &\iff f(\alpha) + g(\beta) \geq f(\alpha') + g(\beta'). \end{aligned} \quad (3.67)$$

This approach amounts to a psychologically significant recoordination of physical space. Generalizing Equation (3.64) we may define

$$\begin{aligned} \Gamma &= f(\alpha) - g(\beta), \\ \Phi &= \frac{f(\alpha) + g(\beta)}{2}. \end{aligned} \quad (3.68)$$

The coordinates  $\Gamma, \Phi$  of the horizontal plane at eye level are supposed to share some of the fundamental properties that Luneburg originally intended when he introduced the modified bipolar coordinates  $\gamma, \varphi$ . To be specific, we have the following modification of (L2):

- (L2') The points lying on a trajectory  $\Gamma = \text{constant}$  are perceived as being equidistant from the observer, while the trajectories  $\Phi = \text{constant}$  are perceived as radial lines of constant direction.

A psychophysical invariance is a physical transformation of the stimuli that leaves perception invariant. The psychophysical invariance can be translated into a functional equation that has to be satisfied by the psychophysical function and thereby restricts its possible form (cf. Falmagne, 1985). In the present case we may derive restrictions for the functions  $f$  and  $g$  appearing in the representation (3.67).

The Luneburg theory as expressed in (3.65) suggests the following definition of psychophysical invariances.

**Definition 3.1** A binary relation  $\preceq$  on  $S$  is called

1.  $\gamma$ -shift invariant if

$$(\alpha, \beta) \preceq (\alpha', \beta') \iff (\alpha + \tau, \beta - \tau) \preceq (\alpha' + \tau, \beta' - \tau),$$

2.  $\varphi$ -shift invariant if

$$(\alpha, \beta) \preceq (\alpha', \beta') \iff (\alpha + \tau, \beta + \tau) \preceq (\alpha' + \tau, \beta' + \tau),$$

3.  $\alpha$ -shift invariant if

$$(\alpha, \beta) \preceq (\alpha', \beta') \iff (\alpha + \tau, \beta) \preceq (\alpha' + \tau, \beta'),$$

4.  $\beta$ -shift invariant if

$$(\alpha, \beta) \preceq (\alpha', \beta') \iff (\alpha, \beta + \tau) \preceq (\alpha', \beta' + \tau)$$

whenever the involved pairs are in  $S$ .

All of these invariances are self-dual properties, i.e.,  $\preceq_\rho$  is preserved if and only if  $\preceq_\theta$  is preserved.

The invariances give rise to functional equations. Let us demonstrate that with the  $\varphi$ -shift invariance of  $\preceq_\rho$ .

By the representation formulated in (3.67),  $\varphi$ -shift invariance of  $\preceq_\rho$  translates into

$$\begin{aligned} f(\alpha) - g(\beta) &\geq f(\alpha') - g(\beta') \\ \iff f(\alpha + \tau) - g(\beta + \tau) &\geq f(\alpha' + \tau) - g(\beta' + \tau). \end{aligned}$$

The real valued function  $H$  given by

$$H[f(\alpha) - g(\beta), \tau] = f(\alpha + \tau) - g(\beta + \tau) \quad (3.69)$$

is thus well-defined and strictly increasing in both variables (i.e.  $H(x, \tau)$  strictly increases both in  $x$  and in  $\tau$ ). It provides a complete characterization of  $\varphi$ -shift invariance in the context of the proposed conjoint representation. The domains in which the variables in (3.69) may move are confined by  $(\alpha, \beta) \in S$  and also by  $(\alpha + \tau, \beta + \tau) \in S$ .

We now proceed to solve this equation.

**Theorem 3.11** *The general strictly increasing solutions  $f, g$  of (3.69):*

$$H[f(\alpha) - g(\beta), \tau] = f(\alpha + \tau) - g(\beta + \tau) \quad (3.70)$$

for all  $(\alpha, \beta) \in S$ ,  $(\alpha + \tau, \beta + \tau) \in S$ , which map  $]-\pi/2, \pi/2[$  onto real intervals, are given by

$$f(\alpha) = a\beta + b, \quad g(\beta) = a'\beta + b' \quad (a > 0, a' > 0); \quad (3.71)$$

and

$$\begin{aligned} f(\alpha) &= ae^{C\alpha} + b, \quad g(\beta) = a'e^{C\beta} + b', \\ (a > 0, a' > 0, C > 0 \text{ or } a < 0, a' < 0, C < 0). \end{aligned} \quad (3.72)$$

Note:  $H$  emerges only as an auxiliary function when deriving (3.70) from the  $\varphi$ -shift invariance and has no direct psychophysical relevance, we skip listing it along with the forms of  $f$  and  $g$ .

*Proof* Define  $T$  by

$$f(\alpha) - g(\beta) = T(\alpha, \beta). \quad (3.73)$$

Keep  $\tau \in ]0, \pi/2[$  constant for the moment. Then equation (3.70),

$$f(\alpha + \tau) - g(\beta + \tau) = H[T(\alpha, \beta), \tau],$$

has the form (3.35). It is supposed to hold whenever

$$\alpha, \beta, \alpha + \tau, \beta + \tau \in ] -\frac{\pi}{2}, \frac{\pi}{2} [ , \quad \alpha > \beta. \quad (3.74)$$

In order to get  $X$  and  $Y$  in Theorem 3.7, fix an arbitrary  $\delta \in ] -\frac{\pi}{2}, \frac{\pi}{2} - \tau [$  and define

$$X = ]\delta, \frac{\pi}{2} - \tau [ , \quad Y = ] -\frac{\pi}{2}, \delta [ . \quad (3.75)$$

Then for all  $\alpha \in X, \beta \in Y$  the restrictions (3.74) is satisfied.

So we can apply Theorem 3.7 to the two solutions

$$\begin{aligned} f_0(\alpha) &= f(\alpha), \quad g_0(\beta) = -g(\beta), \quad h_0(z) = z, \\ f_1(\alpha) &= f(\alpha + \tau), \quad g_1(\beta) = -g(\beta + \tau), \quad h_1(z) = H(z, \tau) \end{aligned}$$

of (3.35) and get, from (3.36), the relations

$$f(\alpha + \tau) = c(\tau)f(\alpha) + c_1(\tau), \quad (3.76)$$

$$g(\beta + \tau) = c(\tau)g(\beta) - c_2(\tau), \quad (3.77)$$

$$H(z, \tau) = c(\tau)z + c_1(\tau) + c_2(\tau). \quad (3.78)$$

By (3.75), Equation (3.76) holds for  $\alpha \in X = ]\delta, \frac{\pi}{2} - \tau [$ . Letting, however,  $\delta$  tend to  $-\frac{\pi}{2}$ , we have (3.76) for  $\alpha \in ] -\frac{\pi}{2}, \frac{\pi}{2} - \tau [$ . Similarly, letting  $\delta$  tend to  $\frac{\pi}{2} - \tau$ , we have also (3.77) for all  $\beta \in ] -\frac{\pi}{2}, \frac{\pi}{2} - \tau [$ .

Notice that we have implicitly let  $\tau$  vary again, so the ‘constants’  $c, c_1, c_2$  are now functions of  $\tau$  and (3.76), (3.77), (3.78) hold for all  $\tau \in ]0, \frac{\pi}{2} [$ .

Thus our task reduces to solving Equations (3.76), (3.77), (3.78) on the indicated domains. We need to worry only about (3.76) and (3.77), as the  $c, c_1, c_2$  obtained from them determine  $H$  in (3.78).

The domain for (3.76) is

$$\tau \in ]0, \frac{\pi}{2} [ , \quad \alpha \in ] -\frac{\pi}{2}, \frac{\pi}{2} - \tau [ . \quad (3.79)$$

We combine two frequently used steps to get to the solution. The first is to bring down the number of functions occurring in the equation. Putting  $\alpha = 0$  into (3.76) and subtract the resulting equation from it we get

$$k(\alpha + \tau) = c(\tau)k(\alpha) + k(\tau) \quad \left( \tau \in ]0, \frac{\pi}{2} [ , \quad \alpha \in ] -\frac{\pi}{2}, \frac{\pi}{2} - \tau [ \right) \quad (3.80)$$

where

$$k(\alpha) := f(\alpha) - f(0). \quad (3.81)$$

The second step is to bring out the algebraic property of addition in the domain of the functions. By (3.80), using  $\alpha + (\sigma + \tau) = (\alpha + \sigma) + \tau$  and thus

$k(\alpha + (\sigma + \tau)) = k((\alpha + \sigma) + \tau)$ , we get

$$\begin{aligned} c(\sigma + \tau)k(\alpha) + k(\sigma + \tau) &= c(\tau)k(\alpha + \sigma) + k(\tau) \\ &= c(\tau)c(\sigma)k(\alpha) + c(\tau)k(\sigma) + k(\tau). \end{aligned}$$

Comparing the coefficients of  $k(\alpha)$ , which is not constant on any proper interval, we get

$$c(\sigma + \tau) = c(\sigma)c(\tau) \quad \text{whenever } \sigma, \tau, \sigma + \tau \in ]0, \frac{\pi}{2}[. \quad (3.82)$$

This is an exponential Cauchy equation in the function  $c$ . We shall establish the continuity of  $c$ , which we do not know yet, from that of  $k$ , which we do know by (3.81). Equation (3.81) shows also that  $k$  is strictly increasing on  $]-\frac{\pi}{2}, \frac{\pi}{2}[$  and that  $k(0) = 0$ , so  $k(\alpha) \neq 0$  for  $\alpha \neq 0$ .

Take an arbitrary  $\tau_0 \in ]0, \frac{\pi}{2}[$  and a non-zero  $\alpha_0$  in  $]-\frac{\pi}{2}, \frac{\pi}{2} - \tau_0[$ . By (3.80), because  $k$  is continuous at  $\tau_0$  and at  $\alpha_0 + \tau_0$ , and  $k(\alpha_0) \neq 0$ , we get that  $c$  is continuous at  $\tau_0$ . Because  $\tau_0$  is arbitrary,  $c$  is continuous on all of  $]0, \frac{\pi}{2}[$ .

The general continuous solutions of (3.82) are (see e.g., Aczél, 1987, sections 4, 5)  $c(\tau) \equiv 0$  which,  $k$  being strictly increasing, is excluded by (3.80), and

$$c(\tau) = e^{C\tau} \quad \left( \tau \in ]0, \frac{\pi}{2}[ \right) \quad (3.83)$$

for some constant  $C$ .

We can now write (3.80) as

$$k(\alpha + \tau) = e^{C\tau}k(\alpha) + k(\tau) \quad \left( \tau \in ]0, \frac{\pi}{2}[, \quad \alpha \in ]-\frac{\pi}{2}, \frac{\pi}{2} - \tau[ \right) \quad (3.84)$$

Knowing that  $k$  is continuous, we prove that it is twice-differentiable. Readers may recall the method shown in Proposition 3.5.

Take an arbitrary  $\alpha_0 \in ]-\frac{\pi}{2}, \frac{\pi}{2}[$ . Select a  $\tau_0 > 0$  and interval neighborhoods  $U, V$  of  $\alpha_0$  and  $\tau_0$ , respectively, such that (3.79) is satisfied by all  $\alpha \in U, \tau \in V$ , that is, this rectangular neighborhood of  $(\alpha_0, \tau_0)$  is in the domain of (3.84). Now integrate (3.84) with respect to  $\tau$  from  $\tau_0$  to  $\tau_1$  ( $\neq \tau_0$ ) in  $V$  to get

$$k(\alpha) \int_{\tau_0}^{\tau_1} e^{C\tau} d\tau + \int_{\tau_0}^{\tau_1} k(\tau) d\tau = \int_{\tau_0}^{\tau_1} k(\alpha + \tau) d\tau = \int_{\alpha + \tau_0}^{\alpha + \tau_1} k(\sigma) d\sigma \quad (3.85)$$

with  $\sigma = \alpha + \tau$ . Because the right-hand side is differentiable in  $\alpha$  at  $\alpha_0$  and the coefficient of  $k(\alpha)$  on the left is a nonzero constant, while the second term on the left is also constant, we get that  $k$  is differentiable at  $\alpha_0$  and thus,  $\alpha_0$  being arbitrary, everywhere on  $]-\frac{\pi}{2}, \frac{\pi}{2}[$ . Because now  $k$  is differentiable, the same argument involving (3.85) shows that  $k$  is twice-differentiable on  $]-\frac{\pi}{2}, \frac{\pi}{2}[$  as asserted.

This fact, that  $k$  is twice-differentiable, furnishes us easily with the *general solution*  $k$ . Indeed, differentiate equation (3.84) with respect to  $\alpha$  and get

$$k'(\alpha + \tau) = e^{C\tau} k'(\alpha), \text{ that is, } l(\alpha + \tau) = l(\alpha) \quad (3.86)$$

on the domain (3.79), where

$$l(\alpha) = e^{-C\alpha} k'(\alpha). \quad (3.87)$$

Differentiating (3.86) with respect to  $\tau$  (which we can do because  $k$  is twice-differentiable), we get  $l'(\alpha + \tau) = 0$  on (3.79). Because for all  $\sigma \in ] -\frac{\pi}{2}, \frac{\pi}{2}[$  there exist  $\alpha$  and  $\tau$  satisfying (3.79) such that  $\alpha + \tau = \sigma$ , we get  $l'(\sigma) = 0$  thus  $l(\sigma) = A$  (constant) on  $] -\frac{\pi}{2}, \frac{\pi}{2}[$  and, by (3.87) and (3.81)

$$f'(\alpha) = Ae^{C\alpha}.$$

If  $C = 0$ , then  $f'(\alpha) = A$  and so

$$f(\alpha) = A\alpha + b. \quad (3.88)$$

If  $C \neq 0$  then

$$f(\alpha) = \frac{A}{C}(e^{C\alpha} - 1) + B = a(e^{C\alpha} - 1) + b \quad (3.89)$$

where, for convenience, we wrote  $a$  for  $A/C$ ,  $b$  for  $B - A/C$ .

Putting (3.83) and (3.88) or (3.89) back into (3.76), we get

$$f(\alpha) = a\alpha + b, \quad c(\tau) = 1, \quad c_1(\tau) = a\tau \quad (3.90)$$

(we wrote here  $a$  in place of  $A$ , the notations in (3.88) and (3.89) are independent) and

$$f(\alpha) = ae^{C\alpha} + b, \quad c(\tau) = e^{C\tau}, \quad c_1(\tau) = b(1 - e^{C\tau}) \quad (3.91)$$

as general solution of equation (3.76) on (3.79), noting that  $f$  is strictly increasing if  $a > 0$  in (3.90) and either  $a > 0, C > 0$  or  $a < 0, C < 0$  in (3.91).

Equation (3.77) is the same as (3.76) where  $c_2$  plays the role of  $-c_1$  and  $\alpha$  plays the role of  $\beta$ . The domains of the equations are the same, (3.79).

So, applying (3.90) and (3.91) we see that (3.77) has the general strictly increasing solutions

$$g(\beta) = a'\beta + b', \quad c(\tau) = 1, \quad c_2(\tau) = -a'\tau \quad (3.92)$$

and

$$g(\beta) = a'e^{C\beta} + b', \quad c(\tau) = e^{C\tau}, \quad c_2(\tau) = -b'(1 - e^{C\tau}). \quad (3.93)$$

Here  $a' > 0$  in (3.92) and either  $a' > 0, C > 0$  or  $a' < 0, C < 0$  in (3.93).  $\square$

### 3.9 A functional equation involving three means

Let  $I, J \subset \mathbb{R}$  be proper intervals and let  $f$  be a one-to-one continuous function mapping  $I$  onto  $J$ .

$$a(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i$$

and     $b(x_1, \dots, x_n) = f^{-1} \left( \frac{1}{n} \sum_{i=1}^n f(x_i) \right) \quad (x_i \in I),$

respectively, are the arithmetic mean and the quasiarithmetic mean generated by  $f$ . The functional equation

$$\frac{1}{n} \sum_{i=1}^n t(x_i) = h \left( \frac{1}{n} \sum_{i=1}^n x_i, f^{-1} \left( \frac{1}{n} \sum_{i=1}^n f(x_i) \right) \right), \quad (3.94)$$

where  $f$  is given and the functions  $t$  and  $h$  are unknown, was initiated by a question of Udo Ebert (Universität Oldenburg, Germany) to J. Aczél. His consideration about a problem of inequality measurement in economics resulted in a problem which was formulated by Aczél in the form of the functional equation (3.94). We refer readers to Ebert (1988, 1997) for other related works of Ebert where quasiarithmetic means are often used.

In Járai *et al.* (2000) the equation was solved for all fixed  $n$  under certain regularity conditions.

It was shown that for  $n \geq 4$ , under the continuity assumption of  $t$ , its solution is given by  $t(x) = c_1 f(x) + c_2 x + c_3$  for some constants  $c_1, c_2, c_3$ . The proof of this result was again based on an application of Theorem 3.7 and will not be included here.

For  $n = 3$ , it was solved under the stronger regularity conditions:  $f$  is continuously differentiable with nonvanishing derivative and not affine on any proper subinterval and  $h$  is continuously differentiable. Its solution is given by  $t(x) = c_1 f(x) + c_2 x + c_3$ . We shall include its proof to illustrate the method. Without the stronger regularity conditions, the question remains open.

Using  $(x, y, z)$  instead of  $(x_1, x_2, x_3)$  and differentiating both sides of the equation with respect to the variable  $x$ , we get

$$t'(x) = \partial_1 h(a(x, y, z), b(x, y, z)) + \partial_2 h(a(x, y, z), b(x, y, z)) \frac{f'(x)}{f'(b(x, y, z))} \quad (3.95)$$

for all  $x, y, z \in I$ . First we shall prove that each  $x_0 \in I$  has a neighborhood, such that for each  $x$  from this neighborhood there exist continuously differentiable functions  $\varphi$  and  $\psi$  such that  $x \mapsto a(x, \varphi(x), \psi(x))$  and  $x \mapsto b(x, \varphi(x), \psi(x))$  are constant. Such neighborhood can be found by the implicit function theorem. Let us choose  $y_0$

and  $z_0$  such that  $f'(y_0) \neq f'(z_0)$ . Their existence is assured because of the supposition that  $f$  is not an affine function on any proper interval. Let  $u_0 = a(x_0, y_0, z_0)$ ,  $v_0 = b(x_0, y_0, z_0)$ . The implicit function theorem (Krantz and Parks, 1951) is then applied for the equation system

$$\begin{aligned} 3u_0 &= x + y + z, \\ 3f(v_0) &= f(x) + f(y) + f(z), \end{aligned}$$

allowing us to see  $y$  and  $z$  as functions  $\varphi$  and  $\psi$  of  $x$ . Now, let us substitute in (3.95) the values  $y = \varphi(x)$  and  $z = \psi(x)$ . Then we obtain the equation

$$t'(x) = \alpha f'(x) + \beta,$$

where  $\beta = \partial_1 h(u_0, v_0)$  and  $\alpha = \partial_2 h(u_0, v_0)/f'(v_0)$ . This equation is valid on a neighborhood of  $x_0$ . Now, this equation implies, that there exists a constant  $\gamma$  such that  $t(x) = \alpha f(x) + \beta x + \gamma$  on a neighborhood of  $x_0$ . The constants  $\alpha$ ,  $\beta$  and  $\gamma$  may depend on  $x_0$ , but we shall prove that they do not.

Suppose that  $t(x) = \alpha_1 f(x) + \beta_1 x + \gamma_1$  on some subinterval  $I_1$  of  $I$  and  $t(x) = \alpha_2 f(x) + \beta_2 x + \gamma_2$  on another subinterval  $I_2$  of  $I$ . Suppose that  $I_1$  and  $I_2$  are not disjoint. If  $\alpha_1 \neq \alpha_2$ , then this means that  $f$  is locally affine, a contradiction. Hence  $\alpha_1 = \alpha_2$  and so  $\beta_1 = \beta_2$  and  $\gamma_1 = \gamma_2$ .

Now if  $I_1$  and  $I_2$  are disjoint, then choosing a chain of subintervals connecting them we obtain that  $\alpha_1 = \alpha_2$ ,  $\beta_1 = \beta_2$  and  $\gamma_1 = \gamma_2$ . So we have proved that  $\alpha$ ,  $\beta$  and  $\gamma$  do not depend on  $x_0$ .

### 3.10 Utility of gains and losses

Luce, in his book (2000) on the utility of gains and losses, generated a rich collection of functional equations. Some are very challenging. Here we give a snapshot on the topic of rank dependent utility representations.

The mathematical model invokes many basic items and primitives. Some are descriptive and may come without a fully satisfactory definition. We list some of them along with some general comments.

- Chance experiments and events. An example of a chance experiment is a toss of a die. The events are the six possible faces that come up. A finite experiment may be represented as  $\{E_1, E_2, \dots, E_n\}$  where the  $E_i$  are the disjoint chance events associated with the experiment. The terminology is the familiar one dealing with probability theory. The study here deals with experiments that can be repeated to establish frequencies as well as those that cannot be replicated. The former are labeled as experiments under risk, the latter as experiments under uncertainty. The corresponding events can be labeled as risky or uncertain accordingly. The notion of joint experiments seen in probability theory is not invoked.

- Outcomes and consequences. Outcome refers to the event that occurs or is observed at the end of the experiment. Consequence refers to the consequence resulting from the outcome. They come together in a formulation of gambles.
- Gambles. Gambles  $g$  may be represented in the form

$$g = (g_1, E_1, g_2, E_2, \dots, g_n, E_n)$$

where  $g_i$  is the consequence assigned to the event  $E_i$  by the gamble  $g$ . The set of consequences, and the structures on it, is often left unspecified. A consequence may be money, as in a lottery ticket. It may also be itself a gamble, as in the case that the payout of a lottery could be another chance or a free lottery. There is no apparent clear separation between consequences and gambles.

- Preference order. The most important primitive structure is the preference order. It is a weak order  $\succsim$  on a set of gambles and consequences. For gambles  $g$  and  $h$  the relation  $g \succsim h$  stands for the description  $g$  is preferred to  $h$ . For consequences  $x$  and  $y$  that are money,  $x \succsim y$  means that  $x$  is preferred to  $y$ . The preference order is a rather personal matter, so  $\succsim$  differs from individual to individual. It is, however, assumed that when it comes to money, the preference order is universal, that  $x \succsim y$  iff  $x \geq y$ . The preference relation induces an equivalence relation.  $g \sim h$  means that  $g \succsim h$  and  $h \succsim g$ . In mathematical psychology, we are interested in the behavior of people. Whether a respondent prefers  $g$  to  $h$  is observable, but the rationale behind that usually is not. In practice, the set of gambles and consequences used in a study is a small set. The theory founded is mostly based on a larger set.
- Status quo, gains and losses. There is a particular consequence,  $e$ , called the status quo. It could stand for what an individual has before participating in the gamble. Consequences  $x$  with  $x \succsim e$  are called gains, those with  $e \succsim x$  are called losses. Consequence is a vague notion which may include money, goods, and lottery tickets; a gamble may be classified as a consequence and thus may be a gain or a loss.
- Joint receipt. For consequences  $x$  and  $y$ , their joint receipt  $x \oplus y$  is given its literal meaning – receiving both  $x$  and  $y$ . If  $x$  is a cup of tea,  $y$  is a pack of sugar,  $x \oplus y$  simply means being in receipt of the two items. No implied meaning is behind what one may do with the two items, and adding sugar to the tea is not necessarily the resulting new consequence. We also speak of the joint receipt,  $g \oplus h$ , of two gambles in the literal sense.
- A joint-receipt preference structure (Luce, 2000, p. 135). This refers to a set  $\mathcal{D}$  which is closed under  $\oplus$ , consisting of consequences, gambles, and their joint receipts, along with a preference order satisfying the following axioms. (1) The preference order  $\succsim$  is transitive and connected, (2) joint receipt is (weakly) commutative:  $g \oplus h \sim h \oplus g$ , (3) (weak) associativity:  $f \oplus (g \oplus h) \sim (f \oplus g) \oplus h$ , (4) (weak) monotonicity:  $f \succsim g$  iff  $f \oplus h \succsim g \oplus h$ , and (5) (weak) identity  $g \oplus e \sim g$ .

$\mathcal{D}^+ := \{f \mid f \in \mathcal{D}, f \succsim e\}$  denotes the substructure consisting of gains.

- The Archimedean axiom: For all  $f, g, h, k \in \mathcal{D}^+$  with  $f > g$ , there exists large enough positive integer  $n$  such that

$$h \oplus (f \oplus f \oplus f \cdots \oplus f \text{ } n\text{-fold}) \succsim k \oplus (g \oplus g \oplus g \cdots \oplus g \text{ } n\text{-fold}).$$

The following theorem is due to Roberts and Luce.

**Theorem 3.12** (Roberts and Luce, 1968) *Suppose that  $(\mathcal{D}^+, e, \succsim, \oplus)$  is an Archimedean joint-receipt preference structure, then there is a representation  $V : \mathcal{D}^+ \rightarrow [0, \infty[$  such that for all  $f, g \in \mathcal{D}^+$ ,*

$$\begin{aligned} f \succsim g &\text{ iff } V(f) \geq V(g), \\ V(f \oplus g) &= V(f) + V(g), \\ V(e) &= 0. \end{aligned}$$

The essence of the theorem is to give conditions upon which the indifference classes of the structure can be embedded, via a mapping  $V$ , into a sub-semigroup of  $[0, \infty[$  under the standard linear order and addition operation.

By a rank-dependent utility representation on a set  $S$  of binary gambles  $(g, E_1, h, E_2) \succsim e$  (of gains) we refer to a function  $U : S \rightarrow [0, \infty[$  and mappings  $W_E$  ( $E \neq \emptyset$ ) from the underlying space of events into  $[0, 1]$  with the following properties

$$W_E(\emptyset) = 0, \quad W_E(E) = 1,$$

$$U(f) \geq U(g) \text{ iff } f \succsim g,$$

$$U(e) = 0$$

$$U(g, E_1, h, E_2) = \begin{cases} U(g)W_{E_1 \cup E_2}(E_1) + U(h)[1 - W_{E_1 \cup E_2}(E_1)], & g \succsim h \\ U(g), & g \sim h \\ U(g)[1 - W_{E_1 \cup E_2}(E_2)] + U(h)W_{E_1 \cup E_2}(E_2), & h \succsim g \end{cases}$$

The representation is said to be dense in intervals if the range of  $U$  and the range of each  $W_E$  are dense in intervals. The representation is said to be onto intervals if the range of  $U$  is a (proper) interval, and that there is at least one experiment such that  $W_E$  has range  $[0, 1]$ . For a more detailed definition and results on the topic we refer to Luce (2000), chapter 3 in particular.

The weighting functions  $W_E$  need not be complementary:  $W_{E_1 \cup E_2}(E_1) + W_{E_1 \cup E_2}(E_2) = 1$ . When they are, rank dependence is lost.

Assuming the existence of  $V$  as stated in Theorem 3.12, and because utility functions  $U$  and  $V$  are functions preserving the preference order  $\succsim$ ,  $U$  and  $V$  are indirectly functions of each other. That is to say, there exists a strictly increasing function  $\phi$  so that

$$U(g) = \phi(V(g)). \tag{3.96}$$

Luce sets as a task to find some behavioral properties relating joint receipt and binary gambles that tells what  $\phi$  must be, or at the very least, narrows the choice of  $\phi$ . Indirectly instead of stating what  $\phi$  is, an alternative is to say what formula

for  $U(f \oplus g)$  could be in terms of  $U(f)$  and  $U(g)$ :

$$\begin{aligned} U(f \oplus g) &= \phi(V(f \oplus g)) \\ &= \phi(V(f) + V(g)) \\ &= \phi(\phi^{-1}(U(f)) + \phi^{-1}(U(g))). \end{aligned}$$

An example of an axiom that links binary gambles to the joint-receipt structure is the property called segregation by Kahneman and Tversky (1979)

$$(f \oplus g, E_1, g, E_2) \sim (f, E_1, e, E_2) \oplus g. \quad (3.97)$$

This axiom is a behavioral condition testable in a lab. Other such axioms are:

Duplex decomposition:

$$(x, C, y, D) \oplus (e, C', e, D') \sim (x, C, e, D) \oplus (e, C', y, D') \quad (3.98)$$

where primes on an event mean independent realizations of that event.

Qualitative event additivity I (Ng *et al.*, 2009):

$$x \oplus (e, C, e, D) \oplus (e, C', e, D') \sim (x, C, e, D) \oplus (e, C', x, D'). \quad (3.99)$$

An example of a formula for  $U(f \oplus g)$  is the p-additive representation

$$U(f \oplus g) = U(f) + U(g) + \delta U(f)U(g) \quad (3.100)$$

where  $\delta > 0$  is a constant.

A first result on the rank-dependent utility representation is due to Marley and Luce (2002).  $U$  and  $W$  mapping onto intervals are part of the axioms. Their next result on a characterization involved the use of joint-receipts formulated in Luce and Marley (2000). A functional equation arising from their studies is

$$f(v) = f(vw) + f(g^{-1}(g(v)q(vw))) \quad (v \in [0, k[, w \in [0, 1]) \quad (3.101)$$

where  $f : [0, k[ \rightarrow [0, \infty[, q : [0, 1] \rightarrow [0, 1]$  and  $g : [0, k[ \rightarrow [0, k'[$  is a strictly monotonic surjection with inverse function  $g^{-1}$  and  $k$  and  $k'$  are fixed in  $]0, \infty]$ . The equation is solved in Aczél *et al.* (2000).

Its solution is given by

- (i) either  $f = 0$  and  $g, q$  arbitrary, or
- (ii)  $g$  is arbitrary and there exists a constant  $c > 0$  such that  $f(0) = 0$  and  $f(v) = c$  for  $v > 0$ ,  $0 < q(0) \leq 1$  and  $q(w) = 0$  for  $w > 0$ , or
- (iii) there exist constants  $\alpha > 0$ ,  $c > 0$ ,  $d > 0$  and  $\mu \geq -k^{-c}$  such that

$$\begin{aligned} q(w) &= (1 - w^c)^d, \\ g(0) &= f(0) = 0, \quad \text{and for } v > 0, \\ g(v) &= \delta(\mu + v^{-c})^{-d}, \\ f(v) &= \begin{cases} \frac{\alpha}{\mu} \ln(1 + \mu v^c) & \text{if } \mu \neq 0 \\ \alpha v^c & \text{if } \mu = 0. \end{cases} \end{aligned}$$

The convention  $k^{-c} = 0$  is adapted when  $k = \infty$ . If  $k' = \infty$  then  $\mu = -k^{-c}$  and  $\delta > 0$  is arbitrary. If  $k'$  is finite then  $\mu > -k^{-c}$  and  $\delta = k'(\mu + k^{-c})^d$ .

A mathematical theorem used in solving the equation is the fact that a strictly convex (or concave) function has a right-sided derivative and a left-sided derivative at all points, and except for at most a countable number of exceptions, is differentiable. The basic Cauchy equations and the generalized Pexider equation (3.33) are encountered fairly often. Of the general solutions, only those given in the third family (iii) are useful in the final representation theorem because of the strict monotonicity requirement  $U(x) \geq U(y)$  iff  $x \succsim y$  demanded on the utility functions. In writing this chapter, we refrain from going over the proof, considering that convex functions and one-sided derivatives are not commonly accessible to non-mathematics major readers.

Generally speaking, a functional equation with nested layers of unknown functions, e.g., (3.101), is hard to solve.

In models where gambles with no payoff have utility, we do not assume  $(e, E_1, e, E_2, \dots, E_n) \sim e$  and we will encounter functional equations dealing with events alone. For example, in Ng *et al.* (2008), a set of axioms lead to functional equations whose solution give rise to the following representation of  $(e, E_1, e, E_2, \dots, E_n)$ :

$$1 + \delta U(e, E_1, e, E_2, \dots, e, E_n) = \frac{1}{\lambda(\Omega)} \sum_{i=1}^n \lambda(E_i) \quad (3.102)$$

for some function  $\lambda$ . Here  $\Omega = \cup_{i=1}^n E_i$ . For  $(e, E_1, e, E_2, \dots, E_n) \succ e$  we may opt for positive  $\delta$  and positive and subadditive measure  $\lambda$ . It remains open as what reasonable properties may further narrow the choice of the function  $\lambda$ , particularly for uncertain events.

The earlier section on binocular space perception and the current section on the utility of gambles have something in common. They are within the broader theory of conjoint measurement. The variables in the cited models run in connected intervals and is an invitation to the use of analytic tools. The theory of conjoint measurement is by no means restricted to such models. The articles by Slinko (2009), Bouyssou *et al.* (2009), and by Luce *et al.* (2009), in a book in honor of Peter Fishburn, provide a good source of reading materials and references.

Another model studied by Luce is based on tone inputs to the two ears of a respondent and the perceived loudness. Let  $(x, u)$  stands for tone magnitude  $x$  applied to the left ear and magnitude  $y$  to the right ear. Let the preference order  $(x, u) \succsim (y, v)$  stands for that  $(x, u)$  is perceived as louder than or equal to  $(y, v)$  by a respondent. Functional equations again play a part in the theory. We refer readers to Narens (1996), Ng (2012), and Steingrimsson and Luce (2007) for relevant literatures.

### 3.11 Miscellaneous comments

In addition to Falmagne (1985), Aczél and Luce (2001) offer another good source of references. In Luce's studies, as seen in example 1 and in section 10, the preference order,  $\succsim$ , is a primitive structure. In the theory of ethical

inequality indices (cf. Ebert, 1988) the preference order is also a primitive. Many concepts have common grounds. For example the concept of the equally distributed equivalent income (Ebert, 1988) is analogous to symmetric matching (Luce, 2012). Entropies are used in Ebert (1988) and in Luce *et al.* (2009). *The Journal of Mathematical Psychology* is a key source of references for papers where functional equations play a role. Luce has been a prominent contributor. He has generated many interesting and challenging functional equations through his axiomatic approaches. *The British Journal of Mathematical and Statistical Psychology* is also a good source for papers in the area. For functional equations used in economics, we refer to Eichhorn (1978).

## Appendix

*Proof of Proposition 3.3.* Suppose that  $f$  satisfies the equation (3.62). Let

$$D^0 := \{(x, y) : x, y \in I \text{ with } x + y \in I\}.$$

Consider the function  $\Delta : D^0 \rightarrow \mathbb{R}$  defined by

$$\Delta(x, y) = f(x) + (1 - x)^\alpha f\left(\frac{y}{1 - x}\right) - f(x + y). \quad (3.103)$$

We shall show that it satisfies the following identities

$$\Delta(x, y) = \Delta(y, x), \quad (3.104)$$

$$\Delta(x, y) + \Delta(x + y, z) = \Delta(x, y + z) + \Delta(y, z), \quad (3.105)$$

$$\Delta(tx, ty) = t^\alpha \Delta(x, y) \quad (3.106)$$

whenever all the arguments are within the domain of  $\Delta$ .

The first identity (3.104) comes immediately from (3.62). Using (3.62) we make the following computations

$$\begin{aligned} & \Delta(x, y) + \Delta(x + y, z) \\ &= \left[ f(x) + (1 - x)^\alpha f\left(\frac{y}{1 - x}\right) - f(x + y) \right] \\ &+ \left[ f(x + y) + (1 - x - y)^\alpha f\left(\frac{z}{1 - x - y}\right) - f(x + y + z) \right] \\ &= f(x) + (1 - x)^\alpha \left[ f\left(\frac{y}{1 - x}\right) + \left(1 - \frac{y}{1 - x}\right)^\alpha f\left(\frac{z/(1 - x)}{1 - \frac{y}{1 - x}}\right) \right] \\ &\quad - f(x + y + z) \\ &= f(x) + (1 - x)^\alpha \left[ \Delta\left(\frac{y}{1 - x}, \frac{z}{1 - x}\right) + f\left(\frac{y+z}{1-x}\right) \right] \\ &\quad - f(x + y + z) \\ &= \Delta(x, y + z) + (1 - x)^\alpha \Delta\left(\frac{y}{1 - x}, \frac{z}{1 - x}\right) \end{aligned}$$

and get

$$\Delta(x, y) + \Delta(x + y, z) = \Delta(x, y + z) + (1 - x)^\alpha \Delta\left(\frac{y}{1 - x}, \frac{z}{1 - x}\right). \quad (3.107)$$

The left-hand side is symmetric in  $x$  and  $y$  and the right-hand side is symmetric in  $y$  and  $z$ . The equation implies that both sides are symmetric in the three variables  $x$ ,  $y$  and  $z$ . The full symmetry of the left side and (3.104) then gives (3.105). Moreover, comparing (3.107) with (3.105) we get

$$\Delta(y, z) = (1 - x)^\alpha \Delta\left(\frac{y}{1 - x}, \frac{z}{1 - x}\right)$$

which is equivalent to (3.106).

We now extend  $\Delta$  to  $\tilde{\Delta} : ]0, \infty[^m \times ]0, \infty[^m \rightarrow \mathbb{R}$  by the definition

$$\tilde{\Delta}(x, y) = s^\alpha \Delta\left(\frac{x}{s}, \frac{y}{s}\right)$$

for all  $s$  provided that  $(\frac{x}{s}, \frac{y}{s}) \in D^0$ .

Because of the homogeneity, (3.106), the function  $\tilde{\Delta}$  is well defined and is an extension of  $\Delta$ . Moreover, it is straightforward to check that the extension satisfies the same identities

$$\begin{aligned} \tilde{\Delta}(x, y) &= \tilde{\Delta}(y, x), \\ \tilde{\Delta}(x, y) + \tilde{\Delta}(x + y, z) &= \tilde{\Delta}(x, y + z) + \tilde{\Delta}(y, z), \\ \tilde{\Delta}(tx, ty) &= t^\alpha \tilde{\Delta}(x, y) \end{aligned}$$

for all  $x, y, z, t$  in  $]0, \infty[^m$ . Using the first and second identity repeatedly we get

$$\begin{aligned} \tilde{\Delta}(sx, sy) + \tilde{\Delta}(sx + sy, tx + ty) \\ &= \tilde{\Delta}(sx, sy + tx + ty) + \tilde{\Delta}(sy, tx + ty) \\ &= [\tilde{\Delta}(sx, sy + tx + ty) + \tilde{\Delta}(tx, (s+t)y)] - \tilde{\Delta}(tx, (s+t)y) \\ &\quad + [\tilde{\Delta}(tx, ty) + \tilde{\Delta}(tx + ty, sy)] - \tilde{\Delta}(tx, ty) \\ &= [\tilde{\Delta}(sx, tx) + \tilde{\Delta}(sx + tx, (s+t)y)] - \tilde{\Delta}(tx, (s+t)y) \\ &\quad + [\tilde{\Delta}(tx, ty + sy) + \tilde{\Delta}(ty, sy)] - \tilde{\Delta}(tx, ty) \\ &= \tilde{\Delta}(sx, tx) + \tilde{\Delta}(sx + tx, (s+t)y) + \tilde{\Delta}(ty, sy) - \tilde{\Delta}(tx, ty). \end{aligned}$$

Hence

$$\begin{aligned} \tilde{\Delta}(sx, sy) + \tilde{\Delta}(tx, ty) - \tilde{\Delta}((s+t)x, (s+t)y) \\ &= \tilde{\Delta}(sx, tx) + \tilde{\Delta}(sy, ty) - \tilde{\Delta}(s(x+y), t(x+y)). \end{aligned}$$

Applying the third identity we arrive at

$$[s^\alpha + t^\alpha - (s+t)^\alpha] \tilde{\Delta}(x, y) = [x^\alpha + y^\alpha - (x+y)^\alpha] \tilde{\Delta}(s, t). \quad (3.108)$$

There are two cases to consider.

Case 1. Suppose that there exist  $s_0, t_0$  such that

$$s_0^\alpha + t_0^\alpha - (s_0 + t_0)^\alpha \neq 0.$$

Then from (3.108) we get

$$\tilde{\Delta}(x, y) = a[x^\alpha + y^\alpha - (x + y)^\alpha]$$

where  $a := \tilde{\Delta}(s_0, t_0)/[s_0^\alpha + t_0^\alpha - (s_0 + t_0)^\alpha]$  is a constant. In particular,  $\Delta(x, y) = a[x^\alpha + y^\alpha - (x + y)^\alpha]$ . Putting this back in (3.103) we get

$$\begin{aligned} a[x^\alpha + y^\alpha - (x + y)^\alpha] &= f(x) + (1-x)^\alpha f\left(\frac{y}{1-x}\right) - f(x+y) \\ (\forall x, y \in I \text{ with } x+y \in I). \end{aligned}$$

Letting

$$\phi(x) := f(x) - ax^\alpha$$

the above equation is reduced to

$$\phi(x) + (1-x)^\alpha \phi\left(\frac{y}{1-x}\right) - \phi(x+y) = 0 \quad (\forall x, y \in I \text{ with } x+y \in I). \quad (3.109)$$

Replacing  $y$  by  $(1-x)z$  we rewrite it as

$$\phi(x) + (1-x)^\alpha \phi(z) = \phi(x+z-xz) \quad (\forall x, z \in I). \quad (3.110)$$

We now take up two subcases.

Subcase 1.1. Suppose that  $\alpha$  is not the zero vector. The symmetry of the right-hand side of (3.110) yields

$$\phi(x) + (1-x)^\alpha \phi(z) = \phi(z) + (1-z)^\alpha \phi(x) \quad (\forall x, z \in I).$$

Fixing in it  $z = z_0$  with  $(1-z)^\alpha \neq 1$  we get  $\phi(x) = b(1-x)^\alpha - b$  where  $b = \phi(z_0)/((1-z_0)^\alpha - 1)$  is a constant. Putting this back to the definition of  $\phi$  we obtain  $f(x) = ax^\alpha + b(1-x)^\alpha - b$  as asserted.

Subcase 1.2. Suppose that  $\alpha$  is the zero vector. Then (3.110) becomes

$$\phi(x) + \phi(z) = \phi(x+z-xz) \quad (\forall x, z \in I).$$

Letting  $L(x) := \phi(1-x)$  the equation becomes

$$L(1-x) + L(1-z) = L((1-x)(1-z)) \quad (\forall x, z \in I).$$

Introducing the new variables  $u = 1-x$ ,  $v = 1-z$ , it becomes the logarithmic Cauchy equation  $L(uv) = L(u) + L(v)$  on the  $m$ -dimensional interval. Its solution is given by  $L(x) = \sum_{i=1}^m \ell_i(x_i)$  where each  $\ell_i$  satisfies the one-dimensional logarithmic Cauchy equation  $\ell(uv) = \ell(u) + \ell(v)$  (Kuczma, 1972). Tracing back to

the definition of  $\phi$  and of  $L$  we get  $f(x) = ax^\alpha + \phi(x) = a + \sum_{i=1}^m \ell_i(1 - x_i)$  as asserted.

Case 2. Suppose that

$$s^\alpha + t^\alpha - (s + t)^\alpha = 0 \quad (\forall s, t \in ]0, \infty[^m).$$

This means that  $g(s) := s^\alpha = \prod_{i=1}^m s_i^{\alpha_i}$  satisfies the additive Cauchy equation  $g(s + t) = g(s) + g(t)$  on the  $m$ -dimensional interval. Being continuous, it must be of the form  $g(s) = \sum_{i=1}^m c_i s_i$  for some constants  $c_i$ . The product form and the sum form coincide only when  $\alpha$  is a basic unit vector. The case has been excluded by the assumption made in the proposition.

We have deduced that all solutions must have the stated forms. The converse statement is easy to check.

## References

- Aczél, J. (1966). *Lectures on Functional Equations and Their Applications*. New York, NY: Academic Press.
- Aczél, J. (1987). *A Short Course on Functional Equations Based upon Recent Applications to the Social and Behavioral Sciences*. Dordrecht-Boston: Reidel-Kluwer.
- Aczél, J. and Daróczy, Z. (1975). *On Measures of Information and Their Characterizations*. Mathematics in Science and Engineering Volume 115. New York, NY: Academic Press.
- Aczél, J., Boros, Z., Heller, J. and Ng, C. T. (1999). Functional equations in binocular space perception. *Journal of Mathematical Psychology*, **43**, 71–101.
- Aczél, J. and Luce, R. D. (2001). Functional equations in behavioral and social sciences. In Smelser, N. J. and Baltes, P. B. (eds), *International Encyclopedia of the Social & Behavioral Sciences*. Oxford: Elsevier Ltd.
- Aczél, J., Maksa, G., Ng, C. T. and Páles, Z. (2000). A functional equation arising from ranked additive and separable utility. *Proceedings of the American Mathematical Society*, **129**, 989–998.
- Aczél, J. and Ng, C. T. (1981). On general information functions. *Utilitas Math*, **19**, 157–170.
- Andrade, J. (1900). Sur l'équation fonctionnelle de Poisson. *Bulletin de la Société Mathématique de France*, **28**, 58–63.
- Bouyssou, D., Marchant, T. and Pirlot, M. (2009). A conjoint measurement approach to the discrete Sugeno integral. In Brams, S., Gehrlein, W. V. and Roberts, F. S. (eds), *The Mathematics of Preference, Choice and Order*. Berlin/Heidelberg: Springer-Verlag.
- Chudziak, J. and Tabor, J. (2008). Generalized Pexider equation on a restricted domain. *Journal of Mathematical Psychology*, **52**, 389–392.
- Dzhafarov, E. N. and Colonius, H. (2011). The Fechnerian Idea. *American Journal of Psychology*, **124**, 127–140.
- Ebanks, B., Sahoo, P. and Sander, W. (1998). *Characterizations of Information Measures*. Singapore: World Scientific.

- Ebert, U. (1997). *Linear Inequality Concepts and Social Welfare*. London School of Economics, Suntory and Toyota International Centres for Economics and Related Disciplines, Discussion Paper Series (ISSN 1352-2469) No. DARP 33.
- Ebert, U. (1988). Measurement of inequality: an attempt at unification and generalization. *Social Choice and Welfare*, **5**, 147–169.
- Eichhorn, W. (1978). *Functional Equations in Economics*. Reading: Addison-Wesley.
- Falmagne, J. (1985). *Elements of Psychophysical Theory*. New York, NY: Oxford University Press.
- Foley, J. M. (1966). Locus of perceived equidistance as a function of viewing distance. *Journal of the Optical Society of America*, **56**, 822–827.
- Hardy, L. H., Rand, G., Ritter, M. C., Blank, A. A. and Boeder, P. (1953). *The Geometry of Binocular Space Perception*. New York, NY: Knapp Memorial Laboratories.
- Havrda, J. and Charvát, F. (1967). Quantification method of classification processes. Concept of Structural  $\alpha$ -entropy. *Kybernetika (Prague)*, **3**, 30–35.
- Hayes, H. and Embretson, S. E. (2012). Psychological measurement: scaling and analysis. In Cooper, H., Camic, P. M., Long, D. L., Panter, A. T., Rindskopf, D. and Sher, K. J. (eds), *APA handbook of research methods in psychology*, vol. 3, Ch. 10. Washington, DC: American Psychological Association.
- Howard, I. P. and Rogers, B. J. (1995). *Binocular Vision and Stereopsis*. New York, NY: Oxford University Press.
- Járai, A., Ng, C. T. and Zhang, W. (2000). A functional equation involving three means. *Rocznik Naukowo-Dydaktyczny Akademii Pedagogicznej w Krakowie, Prace Matematyczne*, **17**, 117–123.
- Kahneman, D. and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, **47**, 263–291.
- Kannappan, Pl. (2009). *Functional Equations and Inequalities with Applications*. Springer Monographs in Mathematics. Dordrecht: Springer.
- Kerridge, D. F. (1961). Inaccuracy and inference. *Journal of the Royal Statistical Society, Series B*, **23**, 184–194.
- Krantz, D., Luce, R. D., Suppes, P. and Tversky, A. (1971). *Foundations of Measurement*, vol. I. San Diego, CA: Academic Press.
- Krantz, S. G. and Parks, H. R. (1951). *The implicit function theorem, history, theory, and applications*. Boston, MA: Birkhäuser.
- Kuczma, M. (1972). Note on additive functions of several variables. *Uniw. Śląski, w Katowicach-Prace Mat.*, **2**, 49–51.
- Kullback, S. (1959). *Information theory and Statistics*. New York: Wiley.
- Lee, P. M. (1964). On the axioms of information theory. *Annals of Mathematics and Statistics*, **35**, 415–418.
- Luce, R. D. (2000). *Utility of Gains and Losses, Measurement-theoretical and Experimental approaches*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Luce, R. D. (2012). Predictions about bisymmetry and cross-modal matches from global theories of subjective intensities. *Psychological Review*, **119**, 373–387.
- Luce, R. D. and Marley, A. A. J. (2000). Separable and additive representations of binary gambles of gains. *Math Social Sciences*, **40**, 277–295.
- Luce, R. D., Marley, A. A. J. and Ng, C. T. (2009). Entropy-related measures of the utility of gambling. In Brams, S., Gehrlein, W. V. and Roberts, F. S. (eds), *The Mathematics of Preference, Choice and Order*. Berlin: Springer.

- Luce, R. D., Ng, C. T., Marley, A. A. J. and Aczél, J. (2008a). Utility of gambling I: entropy modified linear weighted utility. *Economic Theory*, **36**, 1–33.
- Luce, R. D., Ng, C. T., Marley, A. A. J. and Aczél, J. (2008b). Utility of gambling II: risk, paradoxes, and data. *Economic Theory*, **36**, 165–187.
- Luneburg, R. K. (1947). *Mathematical Analysis of Binocular Vision*. Princeton, NJ: Princeton University Press.
- Maksa, Gy. and Ng, C. T. (1986). The fundamental equation of information on open domain. *Publicationes Mathematicae Debrecen*, **33**, 9–11.
- Marley, A. A. J. and Luce, R. D. (2002). A simple axiomatization of binary rank-dependent utility of gains (losses). *Journal of Mathematical Psychology*, **46**, 40–55.
- Narens, L. (1996). A theory of magnitude estimation. *Journal of Mathematical Psychology*, **40**, 109–129.
- Ng, C. T., Luce, R. and Marley, A. A. J. (2008). On the utility of gambling: extending the approach of Meghnissi (1976). *Aequationes Math.*, **76**, 281–304.
- Ng, C. T., Luce, R. and Marley, A. A. J. (2009). Utility of gambling when events are valued: an application of inset entropy. *Theory and Decision*, **67**, 23–63.
- Ng, C. T. and Zhang, W. (2006). An algebraic equation for linear operators. *Linear Algebra and its Applications*, **412**, 303–325.
- Ng, C. T. (1973a). Local boundedness and continuity for a functional equation on topological spaces. *Proceedings of the American Mathematical Society*, **39**, 525–529.
- Ng, C. T. (1973b). On the functional equation  $f(x) + \sum_{i=1}^n g_i(y_i) = h(T(x, y_1, y_2, \dots, y_n))$ . *Ann. Polon. Math.*, **27**, 329–336.
- Ng, C. T. (1980a). Information functions on open domain. *Comptes rendus mathématiques de l'Academie des Sciences*, **2**, 119–123.
- Ng, C. T. (1980b). Information functions on open domains II. *Comptes rendus mathématiques de l'Academie des Sciences*, **2**, 155–158.
- Ng, C. T. (2012). On an intensity attribute – loudness. In Rudolph, L. (ed.), *Qualitative Mathematics for the Social Sciences, Mathematical Models for Research on Cultural Dynamics*. London: Routledge, Taylor & Francis group.
- Pfanzagl, J. (1970). On a functional equation related to families of exponential probability measures. *Aequationes Math.*, **4**, 139–142.
- Radó, F. and Baker, J. A. (1987). Pexider's equation and aggregation of allocations. *Aequationes Math.*, **32**, 227–239.
- Roberts, F. S. and Luce, R. D. (1968). Axiomatic thermodynamics and extensive measurement. *Synthese*, **18**, 311–326.
- Steingrimsson, R. and Luce, R. D. (2007). Empirical evaluation of a model of global psychophysical judgments: IV. Forms for the weighting function. *Journal of Mathematical Psychology*, **51**, 29–44.
- Slinko, A. (2009). Additive representability of finite measurement structures. In Brams, S., Gehrlein, W. V. and Roberts, F. S. (eds), *The Mathematics of Preference, Choice and Order*. Berlin: Springer.
- Sowden, P. T. (2012). Psychophysics. In Cooper, H., Camic, P. M., Long, D. L., Panter, A. T., Rindskopf, D. and Sher, K. J. (eds), *APA Handbook of Research Methods in Psychology*, vol. 1, Ch. 23. Washington, DC: American Psychological Association.
- Stevens, S. S. (1973). *Psychophysics: Introduction to its perceptual, Neural, and Social Prospects*. New York, NY: Wiley.

- Tsallis, C. (1988). Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics*, **52**, 479–487.
- Van Zandt, T. and Townsend, J. T. (2012). Mathematical psychology. In Cooper, H. (ed.), *APA Handbook of Research Methods in Psychology*, vol. 2, Ch. 20. Washington, DC: American Psychological Association.

# 4 Network analysis

John P. Boyd and William H. Batchelder

4.1	Introduction	195
4.2	Discrete network models	196
4.2.1	Relations, digraphs, and graphs	198
4.2.1.1	Relations	198
4.2.1.2	Digraphs	202
4.2.1.3	Graphs	202
4.2.2	Examples of networks	211
4.2.2.1	Bottlenose dolphins in Doubtful Sound	212
4.2.2.2	Karate students	214
4.2.2.3	Human disease network	214
4.2.2.4	Internet	215
4.2.3	Probabilistic properties of degree sequences	217
4.2.4	Matrix methods	223
4.2.4.1	Matrix operations	223
4.2.4.2	Eigenvalues and eigenvectors	226
4.2.4.3	Laplacians	234
4.3	Probabilistic graph models	240
4.3.1	The Hammersley–Clifford theorem	244
4.3.1.1	Step 1: Special representation of probability distributions on $\mathbf{X}$	245
4.3.1.2	Step 2: The dependency structure of $\mathbf{X}$	247
4.3.1.3	Step 3: The H-C theorem for 0–1-valued random variables	247
4.3.2	The H-C theorem and joint probability distributions for models	248
4.3.3	The H-C theorem for PGMs and PDMs	252
4.3.4	Conditionally uniform models	259
4.4	Further topics	265
	Notation	270
	References	270

## 4.1 Introduction

The history of network analysis can be traced back to the Book of Genesis and the list of who “begat” whom (Freeman, 2004). More recently, in 1851 Lewis Henry Morgan, one of the founders of American anthropology, found that tribes of native Americans, such as the Iroquois and Ojibwa, had a consistent system of naming relatives that was entirely different from European nomenclature. The Iroquois and Ojibwa system has a feature called *bifurcate merging*, which means that the same word is used for “father” as for “father’s brother” (i.e., they are *merged*), while distinct words are used meaning “father’s brother” and “mother’s brother” (they are *bifurcated*; Morgan, 1851/1997).

Modern network analysis started in the 1930s as an attempt to formally capture with a diagram of nodes and edges between nodes the more nebulous interpersonal relationships among members of a social group. Moreno (1934) invented the sociogram as a data structure to represent relationships between pairs of members of a well-defined set. For example, the set might consist of members of an elementary school, and the relationships might be necessarily reciprocal such as playing together or being enrolled in the same class, or not necessarily reciprocal such as expressing friendship toward or receiving advice from. From a mathematical perspective, a sociogram is a type of graph (described formally later), and soon after Moreno’s and other’s work, the area of sociometry was established. Later, the interdisciplinary area of social networks flowered, which Freeman (2004) characterized by the following properties:

- intuition based on ties linking social actors,
- systematic empirical data,
- graphic imagery,
- mathematical and/or computational models.

Since the late 1990s, there has been an explosion of interest in networks of all kinds, e.g., neural networks, health-related networks, economic networks, semantic networks, as well as networks in computer science, physics, sociology, and psychology. In fact, the term *network science* is well-established today, and it refers to a rapidly growing interdisciplinary area that reaches well beyond the areas of sociometry and social networks. There has been a confluence of forces that have led to the explosion of interest in network science. First, there have been substantial recent increases in mathematical, computational, and statistical tools that can be used to analyze networks. Second, on the empirical side, through the Internet and related information technologies, it has become possible to collect very large sets of network data. Finally, due to advances in computer science and statistics, new computational and simulation methods have been developed for bringing the new mathematical and statistical tools to bear on these large data sets.

It would be an impossible task for a single chapter, let alone a single book, to cover all the details of the current state of network science. In this chapter, there will be no effort to review the field of networks; however, in the concluding

section, references to several important books and articles will be offered for the reader interested in further knowledge. Instead, between this short introduction and the concluding section, the chapter will have two major sections. In the first, Discrete Network Models, a formal definition of networks will be given in terms of notions in graph theory. Then graph-theoretic properties of networks will be defined and some study of computational complexity in discerning these properties in a network will be given. Certain idealized small networks will be defined, and it will be discussed how complex networks can be regarded as composed from these smaller networks. Then examples of real empirical networks will be given and discussed. Central to this presentation will be the representation of networks as matrices, where the presence or absence of ties between nodes in the network are encoded in a node-by-node square matrix. Finally, the section will describe computational approaches from linear algebra that analyze networks represented as adjacency matrices. To facilitate this, the elements of matrix algebra will be reviewed up to the study of the decomposition of a square matrix in terms of its eigenvalues and eigenvectors. It will be shown how manipulations of the matrix representation of a network can reveal a great deal of its structural properties.

The second major section will discuss statistical models for networks. In particular it will be explained that the usual way that statistical models are developed in an area is not possible for networks because they contain large amounts of dependency between the edge and arc random variables. One solution to this difficulty is the use of the Hammersley–Clifford (H-C) theorem that was originally developed in the area of spatial statistics. We will provide the formal details of the H-C theorem and then show how it can be used to develop statistical models for networks. This development will lead to the presentation of the so-called exponential random graph model (ERGM), which currently is one of the most studied types of probabilistic graph models. In addition, it will be shown how one can set up non-parametric, null hypotheses for assessing the over- or underabundance of certain idealized network structures in an empirical network.

## 4.2 Discrete network models

Intuitively, a network is any interconnected system of nodes. Networks can be classified into social networks (friendship networks, collaboration networks, social media ties, sexual relations, kinship ties), group networks (corporate board interlock, alliances between countries), biological networks (biochemical networks, neural networks, food webs), technological networks (the Internet, power grids, telephone networks, transportation networks, local area networks or LANs), and many more.

There are many ways to describe the ties that link the nodes in a network. The most elementary kind of tie is the presence or absence of a tie. This Boolean off/on property is often coded as a 0 or a 1. There are two broad types of these 0–1 networks, symmetric and nonsymmetric. The symmetric ties are illustrated by

collaboration networks which are by definition symmetric. Kinship networks, however, can be asymmetric as with the “mother” relation: if Mary is John’s mother, you can be sure that John is not Mary’s mother, for at least two reasons. Other networks allow, but do not require, individual ties to be symmetric: “love” can be mutual or unrequited. Another property of networks is the number of relations considered. Often only one type of relation is described at a time, but with kinship networks, for example, one may have to include several basic kinship ties such as “mother,” “father,” “sister,” “brother,” “daughter,” “son,” plus words for “in-laws,” “step,” and “half” relations. Finally, some ties require more than two states to define them: for example, a “signed” network has “positive,” “negative,” and “neutral” or “absence” ties, indicated by +1, -1, and 0, respectively. Continuing in this direction, ties could be rated on a seven-point strength scale, or even on a continuous scale. A variation on this theme is the idea of multiple ties, as in counting the number of times one person emails another. Positive and negative can sometimes be better represented by splitting them into two or more separate relations: the presence of a positive tie and a negative tie may not be equivalent to a neutral tie.

A description of a network must include a definition of the nodes (also called vertices or points) as well as a specification of the connections or ties between them. This description can be extremely precise, such as the chemical bonds between elements of a molecule, or it can be expressed in everyday language. Every English sentence containing two or more proper nouns defines a relation (of a type determined by the rest of the sentence) between these two nouns. For example, the sentence “Adam is the father of Seth” gives the name, in order, of two people in the fatherhood relation. Note that if this sentence is true, then from what we know about fatherhood, the sentence “Seth is the father of Adam” is false, assuming the names refer to the same two people. Other common relationship names are “loves,” “hates,” and “works with.” The last relation is different from the other examples because it is generally a symmetric relation. Still, “ $x$  works with  $y$ ” where  $x$  and  $y$  are members of some well-defined set of individuals determines a definite set of unordered pairs, assuming the “works with” relation has been operationalized.

Although many networks are binary, or two place, relations, there are relations involving more than two variables. For example, an English sentence of the form “ $x$  gives  $y$  to  $z$ ” defines a ternary relation between a set of “givers,” “gifts,” and “receivers.” Even the concept of a “love triangle” is more than a series of three relations: say Adam and Bob are friends, Bob is married to Chloe, and that Adam and Chloe are having an affair. Some would say that it is not really a love triangle unless Bob knows about the affair. Another example of a ternary relation that is an important type of network data is based on not only asking person  $x$  if he likes (or works with, etc.)  $y$ , but also asking  $x$  if he thinks that  $y$  likes  $z$  (Krackhardt, 1987).

Having given the reader a wide view of networks, the rest of this chapter will focus on 0–1 binary relations described by graphs or digraphs. The reasons for this restriction are twofold. First, most of the examples and empirical results in network sciences deal with 0–1 binary relations; and second, many of the formal tools for

analyzing binary relations carry over with minimal change to cases where the ties are not binary.

### 4.2.1 Relations, digraphs, and graphs

#### 4.2.1.1 Relations

Graphs and digraphs are both special cases of (binary) relational structures, which provide a unifying framework. A *relational structure*  $R$  on a set  $V$  is an ordered pair,  $R = (V, A)$ , where  $V = V_R$  is a set called the *underlying set* or the *universe* of  $R$  and where  $A = A_R$ , is a (binary) *relation*, defined as a subset of the Cartesian product  $V \times V$ . Some authors require the universe to be nonempty, but applications later in this chapter force us to allow the possibility of an empty universe. In the context of graph theory, elements of the universe are called *nodes* or *vertices*. Elements of the relation are called *arcs* or *arrows*. The subscripts on  $V$  and  $A$  are only necessary when more than one relation is being discussed. Note that if  $S = (V_S, A_S)$  is another relational system, then  $R = S$  if and only if  $V_R = V_S$  and  $A_R = A_S$ . That is, it is not enough to have just the relations equal, but the universes also have to be equal. This is the weakness of the common, more informal, definition of a relational system as a “set of ordered pairs,” namely,  $A$  alone without reference to  $V$ . Our definition requires that both the ordered pairs and the underlying set of nodes be made explicit. In their full generality, relational structures can include more than one relation, and also functions, on the universe.

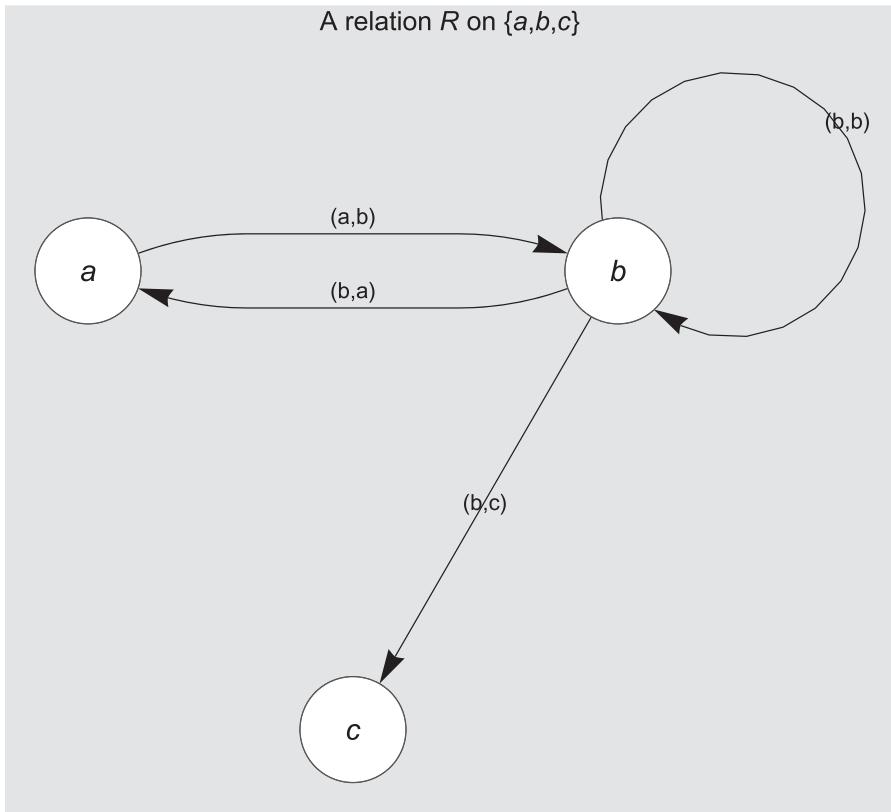
In a relational structure  $R = (V, A)$ , the statement  $(u, v) \in A$  is often denoted by  $uAv$  or, if  $A$  is fixed by context, by  $u \rightsquigarrow v$ . This arc is said to *begin* at  $u$  and *end* at  $v$ . For example, if  $(\mathbb{R}, <)$  is the “usual strict ordering relation on the reals,” then the fact that  $e$  is less than  $\pi$  is captured by the expression  $e < \pi$ . Figure 4.1 gives an example of a relation and a two-dimensional diagram that visually represents it. These diagrams are very useful for presenting data, but the geometry (such as the node coordinates or the shape, length, or thickness of the arrows) has no meaning in the theory of relations, digraphs, or graphs.

Let the nodes of a relational structure,  $R = (V, A)$ , be given in a definite order, say,  $V = \{v_1, \dots, v_N\}$ , then its *adjacency matrix*  $\mathbf{A} = (a_{ij})$  can be defined by  $a_{ij} = 1$  if  $(v_i, v_j) \in A$ , otherwise  $a_{ij} = 0$ . Conversely, a  $\{0, 1\}$ -valued  $N$  by  $N$  matrix determines a unique relational structure on the set  $[N] \stackrel{\text{def}}{=} \{1, \dots, N\}$ . Note that the adjacency matrix  $\mathbf{A}$  is just the indicator function of the relation  $A$  as a subset of the Cartesian product  $[N] \times [N]$ , namely,

$$\mathbf{A} = I_A : [N] \times [N] \rightarrow \{0, 1\} : (i, j) \mapsto 1 \text{ if } i \rightsquigarrow j, \text{ else } 0.$$

For example, the adjacency matrix for the relational structure in Figure 4.1, given the usual ordering of letters, is

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \quad (4.1)$$



**Figure 4.1** A diagram of a relational structure  $R$  on  $V_R = \{a, b, c\}$  with arcs  $A_R = \{a \rightsquigarrow b, b \rightsquigarrow a, b \rightsquigarrow b, b \rightsquigarrow c\}$ .

The *transpose* of an  $M$  by  $N$  matrix  $\mathbf{A}$  is the  $N$  by  $M$  matrix  $\mathbf{A}^T$  defined by  $(\mathbf{A}^T)_{ij} = a_{ji}$  for all  $i, j$ . A relational structure  $R = (V, A)$  is *symmetric* if its matrix satisfies  $\mathbf{A} = \mathbf{A}^T$ .

Some useful properties of a relational structure  $R = (V, A)$  are often imposed as axioms which, given an empirically derived relation, can be tested. In the list below, the properties on the left are true if the formulas on the right hold for all  $x$ ,  $y$ , and  $z$  in  $V$ ,

- |   |       |
|---|-------|
| <i>reflexive:</i> $xAx$ ,<br><i>irreflexive:</i> not $xAx$ ,<br><i>symmetric:</i> $xAy$ implies $yAx$ ,<br><i>asymmetric:</i> $xAy$ implies not $yAx$ ,<br><i>transitive:</i> $xAy$ and $yAz$ imply $xAz$ . | (4.2) |
|---|-------|

While empirical relations may not satisfy any of the properties in (4.2), because of a few “exceptions,” we can give examples that illustrate the concepts. An example of the reflexive property, defined on a set  $V$  of people, is given by the relation

$x$  “knows the name of”  $y$ , assuming that these people know their own names. An irreflexive relation might be “is the father of,” because no one can be the father of oneself. Note that a relation can be neither reflexive nor irreflexive. For example, the relation  $x$  “shoots”  $y$  is neither reflexive nor irreflexive, because some people do shoot themselves, while most, fortunately, do not.

Similarly, a relation can be neither symmetric nor asymmetric. The possibility of nonsymmetry for the “knows the name of” relation arises because some  $x$  can name  $y$ , but not vice versa, say because  $y$  is a celebrity and  $x$  is not, or because  $x$  has a better memory than does  $y$ . However, a relation such as “belongs to the same club” is inherently symmetric.

How hard is it to determine whether or not a relation on  $N$  nodes satisfies one of the properties in (4.2)? Obviously, the answer may depend on the number of arcs in the relation and on the property being investigated. This question is especially important in network theory because some of the networks being studied today, such as neural or Internet connections, may have billions of nodes, and furthermore, the properties that one might want to examine may be much more complex than the simple relational properties in (4.2). One approach is to specify an algorithm in ordinary mathematical prose. Another approach, which is somewhat closer to the ideal of a Turing machine program, is to use computer-like “pseudo-code” and count, as a function of the number of nodes  $N$ , how many steps it takes to answer each question. To determine reflexivity, for example, suppose the relation is given in terms of its  $N$  by  $N$  adjacency matrix  $\mathbf{A}$ . If we just run down the diagonal of this matrix checking that all the entries are 1s, then we are done. This takes exactly  $N$  steps if we use the following algorithm written in English: scan all the diagonal entries of  $\mathbf{A}$  and if a 0 is found, then  $\mathbf{A}$  is not reflexive; otherwise, it is reflexive. Here is the same algorithm expressed in the slightly more precise C style pseudo-code, where  $i++$  means to increment the index  $i$  by one, where  $x = y$  assigns the value of  $y$  to  $x$ , and where  $x == y$  is *true* if  $x$  and  $y$  are equal and *false* if not:

```

boolean reflexive_test(adjacency_matrix A){
1   N = dimension(A);
2   r = true;
3   for(i = 1; i ≤ N; i++)
4     if aii == 0 then r = false;
5   return r}

```

This algorithm takes  $2N + 4$  steps, where a *step* is defined as a single operation or test in the code, to decide whether the Boolean variable  $r$ , representing the property of reflexivity, is true or false. Of course, we are assuming that each “step” takes the same amount of time. If we did know the time for each operation, say,  $a_1$  nanoseconds for line 1,  $a_2$  for line 2,  $a_3$  for the initialization of  $i$  in the first part of line 3,  $b_3N$  for the tests and updates in the second part of line 3, and  $a_4$  for the if–then statement in line 4, and  $a_5$  for the return operation, then the total running time for this algorithm would be  $f(N) = a_1 + a_2 + a_3 + a_5 + (b_3 + a_4)N$  nanoseconds. An added complication is that this is not the best possible algorithm,

because we could just jump out of the for-loop the first time we encounter a 0. Depending on the location of the first of 0, this might save a lot of time on average. Still, the function  $f(N)$  gives the time of the worst-case scenario for completing the task, which is when the relation is in fact reflexive. This function  $f$  is called the (*time-*) *computational complexity* of the algorithm. The *memory complexity*, the amount of computer storage that is needed to do the computation, is also of interest, but will not be discussed here.

Another difficulty with measuring the time to complete an algorithm is that it is dependent on the primitive operations available. For example, if our computer program has the *trace* of a matrix  $\mathbf{A}$ , denoted by  $\text{tr}(\mathbf{A})$ , and defined as the sum of the diagonal elements, then our test is greatly simplified: an  $N$  by  $N$  adjacency matrix represents a reflexive relation if and only if  $\text{tr}(\mathbf{A}) = N$ . This reduces the algorithm for deciding reflexivity to a one-step procedure, independent of the number of nodes. However, this merely disguises the fact that  $\text{tr}$  is just a name for the procedure, in the form of a series of steps, of adding up the diagonal elements and which may have the same number of steps as does our algorithm and takes the same “order of magnitude” of time to execute.

Given that we have agreed upon a small set of operations for our algorithm, say from a specific programming language like C (without libraries that have subroutines like *trace* above), or from a Turing machine, then we may be able to give an exact measure of its run-time complexity in terms of a function of node count  $N$ . However, the expression we derived for the run-time complexity for the reflexive property for binary relations was rather involved and hence difficult to compare with other algorithms. This is where the *O*-notation, pronounced “big-oh,” comes in. It gives a convenient way for describing the order of magnitude of a function  $f : \mathbb{N} \rightarrow \mathbb{R}_{>0}$  in terms of another, presumably simpler, function  $g : \mathbb{N} \rightarrow \mathbb{R}_{>0}$ . We write  $f(n) = O(g(n))$ , which means that there is a constant  $c > 0$  and a natural number  $n_0$  such that if  $n \geq n_0$ , then  $f(n) \leq cg(n)$ . The function  $g$  is said to be an *asymptotic upper bound* for  $f$ , or, more simply,  $f$  is of *order*  $g$ . An equivalent way of defining  $f(n) = O(g(n))$  is to say that  $f/g$  is bounded as  $n \rightarrow \infty$ .

To see how the big-oh notation applies to our reflexivity algorithm above, we want to say that its running time is “linear.” Our  $f$  is the function  $N \mapsto a_1 + a_2 + a_3 + a_5 + (b_3 + a_4)N$  and we can let  $g(N) = N$ . We have to choose a constant  $c$  such that  $cg(N)$  is eventually larger than  $f(N)$ . As a first try, let  $c = b_3 + a_4 + 1$ . Equating  $f$  and  $cg = cN$  and solving for  $N$ , we get  $N = a_1 + a_2 + a_3 + a_5$ . So if we choose  $N_0 = a_1 + a_2 + a_3 + a_5$ , we can see that for  $N \geq N_0$ ,  $f(N) \leq cN$  (because the slope of  $cg$  is greater than the slope of  $f$  for  $N > N_0$ ). Note that the choice of  $c$  and  $N_0$  is not unique: if we choose  $c = a_1 + a_2 + a_3 + a_5 + (b_3 + a_4)$ , then we can choose  $N_0 = 1$ . The conclusion is still that reflexivity is of order  $O(N)$ .

Other common functions that are used inside the expression  $O(g)$  are  $O(\exp(n))$ ,  $O(\log(n))$ , or  $O(n \log(n))$ . Also, it is useful to know that if  $f(n) = a_k n^k + a_{k-1} n^{k-1} + \dots$  is a polynomial of degree  $k$  with constant coefficients, then  $f(n) = O(n^k)$ , ignoring the coefficient  $a_k$  and all the terms of lower order. See Knuth

(1973) for rules of using the  $O$ -notation and the warning that it involves “one-way equalities,” in that we never write  $O(n^2) = \frac{1}{2}n^2 + n$ , while the converse is perfectly correct. Note that while it is technically correct to write  $\frac{1}{2}n^2 + n = O(n^3)$ , this is never intentionally done, because the 3 on the right-hand side is not the lowest possible exponent, 2, which would give a tighter upper bound.

The  $O$ -notation can be used as part of approximations. For example, Stirling’s formula for  $n! \stackrel{\text{def}}{=} 1 \times 2 \times \cdots \times n$  is

$$n! \approx \sqrt{2\pi n} n^n e^{-n}.$$

Taking logs (in this chapter all logs are taken to the base  $e$ , Euler’s constant) and using the  $O$ -notation, this can be rewritten more cleanly as

$$\log n! = n \log n - n + O(\log n) = O(n \log n).$$

#### 4.2.1.2 Digraphs

A *digraph*  $D = (V, A)$  is an irreflexive relation on a finite set  $V = V_D$  of nodes that is disjoint from arcs  $A = A_D$ . Our notation and definitions for graphs and digraphs primarily follows Bollobás (1998), except that cardinality of sets and the dimensionality of matrices are indicated with capital letters, such as  $M$  and  $N$  instead of  $m$  or  $n$ . The *in-neighborhood*  $N^-(v)$  of a node  $v$  is the set of nodes that end at  $v$ , while its *out-neighborhood*  $N^+(v)$  is the set of nodes that begin at  $v$ . The cardinality of the in- and out-neighborhoods of  $v$  are called its *in-degree* and *out-degree*, denoted by  $d^-(v)$  and  $d^+(v)$ , respectively. In terms of the adjacency matrix  $A$  of  $D$ , the in-degree of a node equals its column-sum,  $d^-(v) = \sum_u a_{uv}$ , while the out-degree is its row-sum,  $d^+(v) = \sum_u a_{vu}$ .

A *walk* in a digraph is a sequence of nodes,  $\mathbf{v} = (v_0, v_1, \dots, v_k)$ , such that for all  $i \in [k]$ ,  $v_{i-1} \rightsquigarrow v_i$ . Walks can be denoted more simply by the string,  $v_0 v_1 \cdots v_k$ , and the *length* of this walk is defined to be  $k$ . Note that a walk can have repeated nodes or repeated arcs, but if all its nodes are distinct, then all its arcs are also distinct. If all the arcs of a walk are distinct, it is called a *trail*. Similarly, if all the nodes of a walk are distinct, it is said to be a *path*. It is easy to see that every path is a trail and that every trail is a walk. The reader should construct minimal counterexamples to the converse statements.

There are two interesting types of walks. A *circuit* is a *closed* trail, defined as a trail such that the first node equals the last node. The second, more restrictive, concept is a *cycle*, defined as a circuit of length at least 3 where all the nodes are distinct, except for the first and last.

#### 4.2.1.3 Graphs

A *graph*  $G$  is an ordered pair of finite disjoint sets  $(V, E)$  such that  $E$  is a subset of the set  $V^{(2)}$  of distinct unordered pairs of elements of  $V$ . Elements of  $V$  are called *nodes* (or *vertices*), just as for digraphs, but elements of the set  $E$  are called *edges*, instead of arcs. An equivalent definition of graphs as irreflexive and symmetric relations would emphasize the close relation between graphs and digraphs, but is

not as simple or easy to work with as the standard definition. An edge  $\{u, v\} \in E$  of a graph  $G$  is said to *join* or *connect* the nodes  $u$  and  $v$  and is denoted by  $u \sim v$  or, equivalently, by  $v \sim u$ . If  $u \sim v$ , then  $u$  and  $v$  are said to be *adjacent* or *neighboring*.

Many definitions for graphs are the same as for digraphs with minor changes. Certainly, walks, trails, and paths apply to graphs in the obvious way. Usually, there is little difficulty in carrying over definitions from digraphs to graphs. For example, the in-neighborhood and the out-neighborhood of a node  $v$  in a graph are identical to each other, so the in- and out- prefixes can be dropped, and its *neighborhood* is denoted by  $N(v)$ . The same is the true of in- and out-degree: for graphs the *degree* of a node  $v$  is denoted by  $d_v$ . If a node has degree zero, then it is said to be *isolated*.

An important global property of a graph is the vector  $\mathbf{d}$  of  $N$  nodal degrees. This vector is called the *degree sequence* of the graph. The distribution of the degree sequence of large graphs can exhibit such properties as “self-similarity” and being “scale-free” (Estrada, 2011). It is often convenient to reorder the degree sequence so that it is non-increasing. For example, the graph  $P_5$ , consisting of nothing but a path on five nodes, has degree sequence  $\mathbf{d} = (2, 2, 2, 1, 1)$ . Note, however, that not every sequence of positive integers is a degree sequence. Can you see why neither  $(3, 2, 1)$  nor  $(5, 4, 3, 2, 1, 1)$  is a degree sequence? If a sequence is the degree sequence of some graph, then it is said to be *graphical*. The necessary and sufficient conditions for a sequence to be graphical is given by the famous theorem (Erdős and Gallai, 1960):

**Theorem 4.1** (Erdős–Gallai) *Let  $\mathbf{d}$  be a sequence of  $N$  nonincreasing, nonnegative integers. Then  $\mathbf{d}$  is graphical if and only if it sums to an even number and for all natural numbers  $k < N$*

$$\sum_{i=1}^k d_i \leq k(k-1) + \sum_{i=k+1}^N \min(k, d_i). \quad (4.3)$$

*Proof* Suppose  $\mathbf{d}$  is graphical. Then  $\mathbf{d}$  is the degree sequence of a graph  $G = (V, E)$ . The sum of the degrees is even because  $\sum_{i=1}^N d_i = 2|E|$ , which holds because each edge adds one to the degree of two nodes. This known as Euler’s “handshaking lemma,” the first theorem of graph theory.

To show that (4.3) holds, choose any positive  $k < N$  and let  $U = \{v_1, \dots, v_k\}$  be the  $k$  nodes of largest degree and  $U^c = \{v_{k+1}, \dots, v_N\}$  be the remaining nodes. Let  $e$  be the number of edges within  $U$  and let  $f$  be the number of edges from  $U$  to  $U^c$ . Then  $\sum_{i=1}^k d_i = 2e + f$ , where the factor 2 again comes from the handshaking lemma. Now  $e \leq \binom{k}{2}$ , because this is the maximum possible number of edges for  $k$  nodes. Therefore,  $2e \leq k(k-1)$ . Now let  $i$  be a natural number such that  $k < i \leq N$ , noting that  $f_i$ , the number of edges from  $v_i$  to  $U$ , must be less than both  $k$ , the number of nodes in  $U$ , and its own degree,  $d_i$ ; i.e.,  $f_i \leq \min(k, d_i)$ . This implies that  $f = \sum_{i=k+1}^N f_i \leq \sum_{i=k+1}^N \min(k, d_i)$ , accounting for the second term in the right-hand side of (4.3).

The converse is more difficult, but see Choudum (1986) or Harary (1969) for proofs.  $\square$

If we consider the case  $k = 1$  in (4.3), we see that this implies the obvious condition that  $d_1 < N$ , which is why the first example sequence,  $(3, 2, 1)$ , is not graphical. However, in the second example sequence,  $(5, 4, 3, 2, 1, 1)$ , we see that it satisfies the condition for  $k = 1$ , but not for  $k = 2$ .

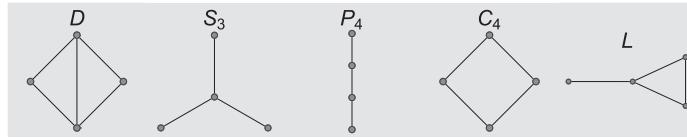
A distance function  $d$  can be defined on the nodes in any graph  $G = (V, E)$  as follows: for all  $u, v \in V$ , if  $u = v$ , let  $d(v, v)$  be 0, else if there is at least one path in  $G$  from  $u$  to  $v$ , then let  $d(v, v)$  be the minimal length of all paths in  $G$  from  $u$  to  $v$ , else (for the case  $u \neq v$  and there being no paths from  $u$  to  $v$ ) let  $d(u, v)$  be  $\infty$ . The *diameter* of a graph  $\text{diam } G \stackrel{\text{def}}{=} \max_{u,v} d(u, v)$  is the maximum distance in  $G$ , while the *radius* is  $\text{rad } G \stackrel{\text{def}}{=} \min_u \max_v d(u, v)$ . The definition of radius is a little tricky until you consider the closed unit disk, the set of points in the plane at a distance of 1 or less from the origin. Now every point on the unit disk has a maximum distance, ranging from 1 to 2, to other points on the disk, and the minimum among such maximum distances is exactly 1, achieved by the distance from center to any point on the boundary of the unit disk.

There are several families of graphs that are defined for each natural number. While naturally occurring networks rarely correspond to these idealized graphs, they can serve as building blocks that can be used to describe these more complicated structures. Certain networks are characterized by having more, or fewer, of these graphs or *motifs* than would be expected by chance (Milo *et al.*, 2002).

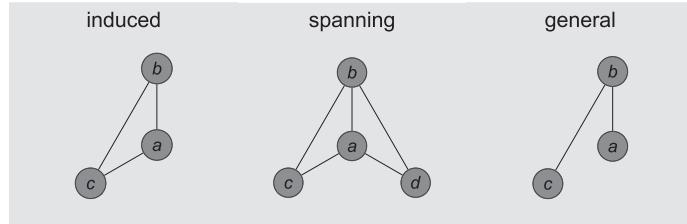
An example of such a family of graphs is the *complete* graph  $K_N$ , which has  $N \geq 0$  nodes together with every one of the  $\binom{N}{2}$  possible edges. At the other extreme is the *empty* graph  $E_N$ , which also consists of  $N \geq 0$  nodes, but has no edges at all. The graph  $K_1 = E_1$  is called the *trivial* graph, while  $K_0 = E_0 = (\emptyset, \emptyset)$  is called the *null* graph. Finally, the *path* and *cycle* graphs,  $P_N$  and  $C_N$ , are the paths and cycles (defined above) on  $N$  nodes.

Sometimes a family has just one member that is prominent as a motif. One of these is constructed by removing just one edge from the complete graph  $K_4$ . The resulting graph is called a *diamond* and is denoted by  $D$ . Yes, it is just one of a family of graphs called *fans*, but the “diamond” name has stuck for the four-node version. Another such motif is the *lollipop* graph, formed from  $C_3$  by adding another node and an edge to this node from any node in  $C_3$ .

While not defined uniquely for each natural number, another interesting type of graph is a *bipartite* graph  $(V, E)$ , where the node set  $V$  can be partitioned into two nonempty subsets,  $V_1$  and  $V_2$ , of size  $N_1$  and  $N_2$ , respectively, such that every edge in  $E$  connects a node in  $V_1$  with one in  $V_2$ . If, in addition, the edge set of a bipartite graph contains every possible pair  $\{v_1, v_2\}$ , where  $v_1 \in V_1$  and  $v_2 \in V_2$ , then we have a *complete bipartite graph*, denoted by  $K_{N_1, N_2}$ . Obviously,  $K_{N_1, N_2}$  has  $N_1 + N_2$  nodes and  $N_1 N_2$  edges. A special case of a complete bipartite graph is the *N-star* graph,  $S_N \stackrel{\text{def}}{=} K_{1, N}$ . The graph  $S_N$  has  $N + 1$  nodes and  $N$  edges. Warning: some authors define the *N-star* to have  $N$  nodes and  $N - 1$  edges.



**Figure 4.2** Small graphs on four nodes: the diamond, star, path, cycle, and lollipop graphs.



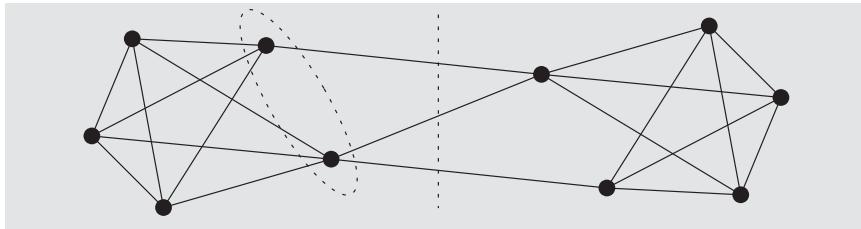
**Figure 4.3** Three types of subgraphs of  $K_4$ : an induced subgraph  $\langle a, b, c \rangle$ , a spanning subgraph that is not induced, and a general subgraph that is neither induced nor spanning.

Some of these small graphs that occur often as motifs are pictured in Figure 4.2. Note that  $C_4$  is isomorphic (formally defined below) to the complete bipartite graph  $K_{2,2}$  and that the diamond  $D$  is  $K_4$  with one edge removed. Finally, it is assumed that distinct dots in the figure represent distinct nodes: they are not labeled because we are only interested in each graph's isomorphism class, not the labeled graph itself.

A graph  $G' = (V', E')$  is a *subgraph* of the graph  $G = (V, E)$ , written  $G' \subseteq G$ , if  $V' \subseteq V$  and  $E' \subseteq E$ . Because  $G'$  is stipulated to be a graph, it is by definition symmetric and all edges must join members of  $V'$ . This latter statement means that if  $u \xrightarrow{G'} v$ , then both  $u$  and  $v$  must be in  $V'$ .

If  $G'$  contains all the edges in  $E$  that join two nodes in  $V'$ , then  $G'$  is said to be the subgraph *induced*, or *generated*, by  $V'$  and is denoted by  $G(V')$ , or, if the context is clear, by  $\langle V' \rangle$ . Thus a subgraph  $G'$  is an induced subgraph of  $G$  if  $G' = G(V')$ . If  $V' = V$  then  $G'$  is a *spanning* subgraph of  $G$ . These three types of subgraph are illustrated for the graph  $K_4$  in Figure 4.3. Warning: some authors define “subgraph” to be what we call an “induced subgraph,” perhaps because they do not use the more general concept.

Another interesting property of graphs is whether or not they are *connected*, which means that there is a path between any two distinct nodes. For any graph  $G = (V, E)$ , it can be shown that there is a partition  $\mathcal{P} = (V_1, V_2, \dots, V_n)$  of the nodes such that each generated subgraph  $\langle V_k \rangle$  is a maximal connected subgraph of  $G$ . These subgraphs are called the *connected components* of  $G$ . If  $n = 1$ , then the graph is connected. At the other extreme, if  $n = N$ , then  $G$  must be the empty graph on  $N$  nodes. The computational complexity of finding all the connected components of a graph with  $N$  nodes is shown, by a very sophisticated algorithm (Kocay and Kreher, 2005), to be linear, i.e., to have computational complexity  $O(N)$ .



**Figure 4.4** A graph whose connectivity measures are all distinct. Node connectivity  $\kappa = 2$  (the nodes in the ellipse), edge connectivity  $\kappa' = 3$  (the edges crossed by the dotted line), and the minimal degree  $\delta = 4$  (any one of eight nodes).

Because many graphs derived from empirical networks are connected, it is of interest to measure the degree of connectivity. A graph has *connectivity*  $\kappa(G) \in \mathbb{N}$  if  $\kappa$  is the smallest number of nodes needed to disconnect  $G$ . That is, there is a subset  $W \subseteq V$  of nodes such that  $\langle V \setminus W \rangle$  is not connected,  $|W| = \kappa$ , and if  $U$  is another set of nodes such that  $\langle V \setminus U \rangle$  is not connected, then  $|W| \leq |U|$ . A related concept is the *edge connectivity*, denoted by  $\kappa'$ , and defined as the smallest number of *edges* needed to disconnect the graph. Note that if the graph is already disconnected, then  $\kappa = \kappa' = 0$ . A final related measure is  $\delta(G)$ , the minimal degree of nodes in  $G$ . The measures are related by the following theorem.

**Theorem 4.2** *For any graph  $G$ ,*

$$\kappa(G) \leq \kappa'(G) \leq \delta(G). \quad (4.4)$$

For many graphs, the three inequalities in (4.4) are in fact equalities, but Harary (1969) gives the minimal-node example in Figure 4.4 where they are all strict inequalities.

If a subgraph  $H$  of a graph  $G$  is itself a complete graph, then  $H$  is called a *clique*. Notice that a clique  $H$  is always a generated subgraph,  $H = \langle V_H \rangle$ , because it has all possible edges. This allows us to speak informally of a clique as a *subset*, even though we know it is really a *subgraph*. The computation complexity of finding all the cliques in a graph with  $N$  nodes would seem to be  $O(2^N)$  by simply listing all possible subsets of nodes. This time complexity is said to be *exponential* and exceeds any given polynomial, as can be seen by considering the limit of the ratio  $\lim_{N \rightarrow \infty} 2^N/N^k = \infty$ , which can be shown by  $k$  applications of l'Hôpital's rule from calculus. Such a bound is called NP, for *nondeterministic polynomial*. One might think that, as in the case above of connected components, that a cleverer algorithm could give a lower bound for the complexity of the clique problem. However, it can be shown that if there were such a polynomial time algorithm, i.e., one of complexity  $O(N^k)$  for some  $k \in \mathbb{N}$ , for the clique problem, then all other NP-problems would also have a polynomial solution (Kocay and Kreher, 2005). In sum, the clique problem is NP-*complete*, which makes it very difficult to solve

except for small  $N$ . What “small” is depends on factors such as the size and speed of the computer, the algorithm employed, and properties of the graph.

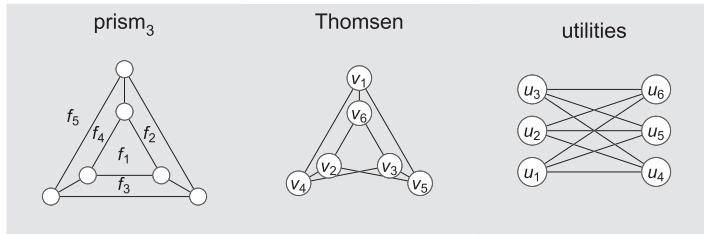
Of particular interest are *maximal* cliques, cliques that are not proper subgraphs of any other clique. For example, if the path  $P_4$  is defined on the nodes  $\{1, 2, 3, 4\}$  with edges  $E = \{1 \sim 2, 2 \sim 3, 3 \sim 4\}$ , then its maximal cliques are the three subgraphs generated by the edges themselves, namely,  $\langle \{1, 2\} \rangle$ ,  $\langle \{2, 3\} \rangle$ , and  $\langle \{3, 4\} \rangle$ . Its nonmaximal cliques are the four singleton subgraphs, plus the empty graph  $E_0$ , making a total of eight cliques in all. As a slightly more challenging example, the disjoint union of two complete graphs,  $K_{N_1} \cup K_{N_2}$ , has exactly two maximal cliques,  $K_{N_1}$  and  $K_{N_2}$ , while its total number of cliques is  $2^{N_1} + 2^{N_2}$ . Warning: some authors (Harary, 1969) define clique to mean what we call “maximal clique,” but our more inclusive definition predominates in applications.

Two graphs  $G = (V, E)$  and  $G' = (V', E')$  are *isomorphic* if there is a one-to-one correspondence, or *bijection*, between the nodes that preserves adjacency. That is, there is a bijection  $\phi: V \rightarrow V'$  such that for all  $u, v \in V$ ,  $u \overset{G}{\sim} v$  if and only if  $\phi(u) \overset{G'}{\sim} \phi(v)$ . Two isomorphic graphs have the same abstract properties, except for relabeling the nodes. That is, each node  $v \in V$  is “relabelled”  $\phi(v)$ . Examples of properties preserved by isomorphisms are connectedness, the number of 3-cycles, and the degree sequence. The kinds of property not preserved by isomorphism are those that refer to particular nodes, such as “node  $v$  has degree 3.” Another application of the isomorphism concept is the number of nonisomorphic cliques for the complete graph  $K_N$ : every subgraph of  $K_N$  is isomorphic to every other subgraph of the same size, so the total number of non-isomorphic cliques is  $N + 1$ .

A good way to visualize an isomorphism  $\phi: G \rightarrow G'$  is to observe the adjacency matrices  $\mathbf{A}$  and  $\mathbf{A}'$ . Assuming that the nodes of  $G$  are ordered,  $V = \{v_1, \dots, v_N\}$ , then the nodes of  $G'$  can be given the induced ordering,  $\phi(v_i) < \phi(v_j) \Leftrightarrow i < j$ . By the definition of the adjacency matrix, we have  $a_{ij} = 1$  if and only if  $(v_i, v_j) \in E$ , which, because  $\phi$  is an isomorphism, holds if and only if  $(\phi(v_i), \phi(v_j)) \in E'$ , which, by the definition of the adjacency matrix of  $G'$  and the induced ordering, holds if and only if  $a'_{ij} = 1$ . That is, the two adjacency matrices are identical! Of course, if the ordering of the  $V'$  is different from the one induced from the ordering of  $V$ , then the adjacency matrices may be different.

If a graph isomorphism  $\phi: G \rightarrow G'$  has the property that  $G = G'$ , i.e., that  $\phi$  is a permutation on  $V_G$  and that  $E_G = E_{G'}$ , then  $\phi$  is called an *automorphism*. The set of all automorphisms of a graph  $G$  is denoted by  $\text{Aut}(G)$ , a well-studied algebraic structure useful in the enumeration of graphs (Harary, 1969).

As one can see from the figures so far, it is useful to be able to depict a graph by *embedding* its nodes and edges into a two-dimensional planar configuration. That is, we have a function  $f: V \rightarrow \mathbb{R}^2$  such that no two nodes are sent to the same point. Then each edge is represented by a (possibly curved) line between its two nodes. However, as mentioned, it is meaningless to interpret most aspects of the geometric configuration such as the distance between nodes or the length or shape of the edges, because there are always many other different embeddings of the same graph that also leave invariant all its graph theoretic properties.



**Figure 4.5** The mapping given by the  $v_i \mapsto u_i$  is an isomorphism between the Thomsen graph and the utilities graph, but no isomorphisms can exist from the prism graph to either of the other two.

Figure 4.5 shows three different graph embeddings. The first is an embedding of a 3-prism graph, while the next two are different embeddings of the complete bipartite graph  $K_{3,3}$ . The labels for the last two embeddings, “Thomsen” and “utilities,” are two alternative names for  $K_{3,3}$ . The name “utilities” arises from the puzzle of how can one connect all the nodes on the left, thought of as three “utilities,” say, gas, electric, and water, to all of the three “cottages” on the right without any of the utility lines (wires or pipes) crossing. In other words, is there a *planar* embedding, defined as an embedding, with curved lines allowed, such that no lines cross? The answer is no, as proven in Kuratowski’s planar graph theorem presented below in Theorem 4.3. The interesting thing about the first embedding of  $K_{3,3}$  is that there is just one crossing, which is the least possible. The other minimal nonplanar graph in Kuratowski’s theorem is  $K_5$ , which has five crossings in the circular embedding. There is an embedding of  $K_5$  with just one crossing. Can you find it?

Kuratowski’s theorem requires the following definition: two graphs are *homeomorphic* if they both can be formed from a third graph by a sequence of subdivisions of their edges. That is, if  $G = (V, E)$  is a graph with an edge  $e = \{u, v\}$  and if  $w \notin V$ , then the graph

$$G' = (V \cup \{w\}, [E \cup \{\{u, w\}, \{w, v\}\}] \setminus \{e\})$$

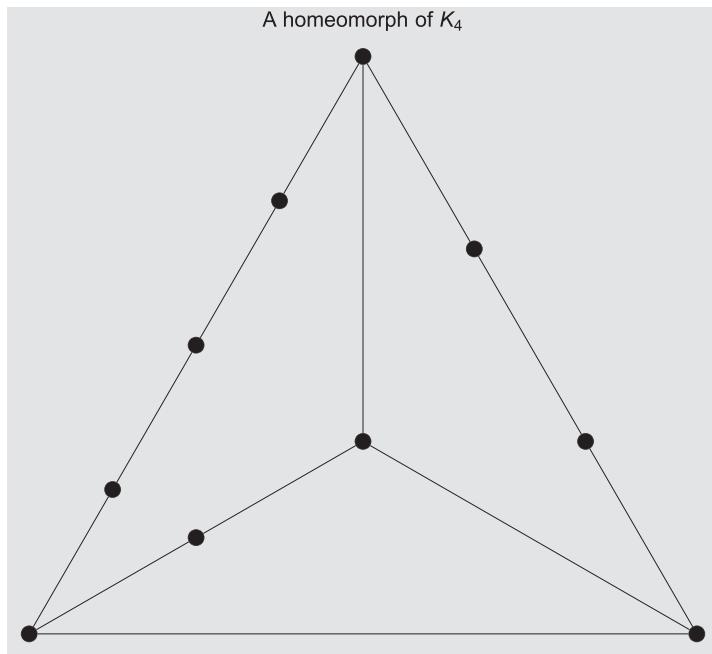
is a homeomorph of  $G$  formed by subdividing  $e$ . Figure 4.6 is an example of a homeomorph of  $K_4$  formed by a sequence of six subdivisions.

We are now able to state the theorem:

**Theorem 4.3** (Kuratowski) *A graph is planar if and only if it has no subgraph homeomorphic to  $K_5$  or  $K_{3,3}$ .*

*Proof* See Harary (1969). □

We know that the last two graphs of Figure 4.5 are isomorphic, because we can specify an isomorphism. The reader can verify that there are  $23!3! = 72$  distinct isomorphisms. Hint: the mapping  $v_i \mapsto w_i$  given in the caption can be followed by the symmetry on the vertical axis:  $w_i \leftrightarrow w_{i+4} : i = 1, 2, 3$ . This gives two

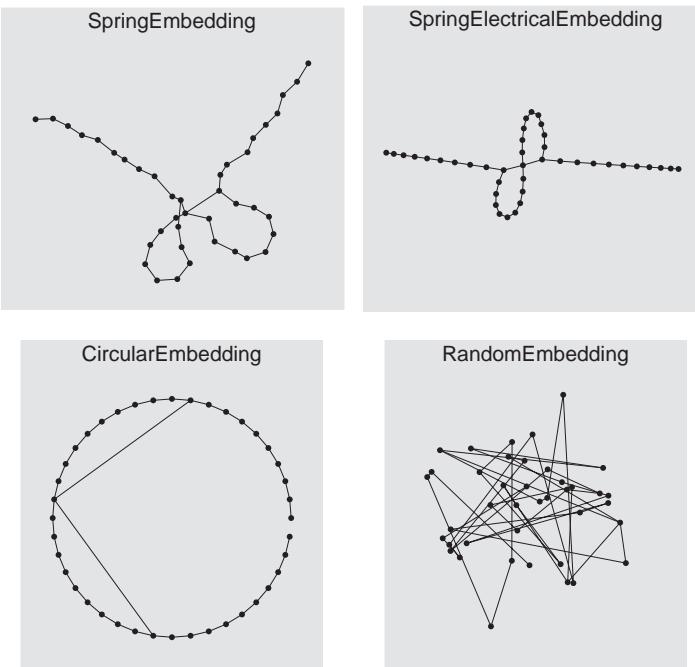


**Figure 4.6** A homeomorph of  $K_4$  formed by subdividing six edges.

isomorphisms. But wait, there's more! The left-hand nodes can be permuted in any of  $3!$  ways and, independently, the right-hand nodes can also be permuted in  $3!$  ways.

It is harder to prove that there is no possible isomorphism between the 3-prism graph and the second (and hence the third) graph. The first two graphs were drawn so as to emphasize their superficial similarities, giving a false hope that they are isomorphic. How can one show they are not? One method is to check all  $6! = 720$  possible bijections between the six nodes of each graph to determine if one is an isomorphism. Alternatively, a useful heuristic for proving two graphs to be nonisomorphic is to look for some abstract property (i.e., one not having to do with node labels) that one graph has and the other does not. For example, do the graphs have a different number of nodes or lines, or do the degree sequences differ? The graphs of Figure 4.5 agree on all these properties. However, note that the 3-prism graph has two 3-cycles, while the last two do not, proving that the first is not isomorphic to either of the others.

There are many different algorithms for graph embedding, but we shall discuss only four, illustrated in Figure 4.7 for the graph  $G_{39}$ , which is the path  $P_{39}$  with addition of the edges  $10 \sim 20$  and  $20 \sim 30$ . The four graph embedding methods are called the *spring*, *spring-electrical*, *circular*, and *random* embeddings. These embedding methods (and others) are all available in *Mathematica*® (Wolfram Research, 2014), and *Pajek*. The latter program is freely available at <http://pajek.imfm.si/>.



**Figure 4.7** Four methods of embedding a graph.

The spring embedding is based on Hooke's law for a spring, which states that the force needed to stretch or compress a spring a distance  $x$  from its resting position is proportional to  $x$ . The potential energy of such a deflection is proportional to  $x^2$ . Each edge in the graph is considered to be a spring whose resting length is 1, so that the potential energy of an embedding  $f$  is proportional to

$$U_f = \sum_{u \sim v} (1 - \|f(u) - f(v)\|)^2, \quad (4.5)$$

where the sum is over all edges in the graph. A spring embedding is a graph embedding  $f$  which minimizes the potential energy (4.5). One can start with an initial embedding and proceed by Newton's method of “steepest descent” in a series of steps that converges to a minimal energy configuration. Of course, the embedding is not unique, because any rigid motion preserves the distances between points. One weakness of the spring embedding is that the embedding of a path can have zigzags instead of a straight path and have the same energy as defined in (4.5). This can be seen in Figure 4.7.

The spring-electrical embedding is a modification of the spring embedding that overcomes some of its problems: in addition to viewing edges as springs, each node is given the same electrical charge. The force between two point electrical charges is inversely proportional to the square of the distance between them. Note that this force acts on all nodes whether they are adjacent or not. This has the effect of straightening out paths, as can be seen in the top right embedding of Figure 4.7.

---

The spring-electrical embedding is the default embedding used in all the graphs in this chapter, unless otherwise stated.

Next, the circular embedding assumes an ordering of the nodes, say,  $V = \{v_1, v_2, \dots, v_N\}$  ordered by the indices, so that the embedding is defined in polar coordinates by  $f(v_k) = (1, 2\pi k/N)$ . That is, the nodes are mapped into the unit circle with angle  $2\pi k/N$ . Because  $G_{39}$  is defined in terms of the natural order, the circular embedding has a nice appearance in Figure 4.7. However, isomorphic images of  $G_{39}$  might be a maze of edges crisscrossing the circle. The circular embedding cannot be expected to give a “good” embedding of graphs in general.

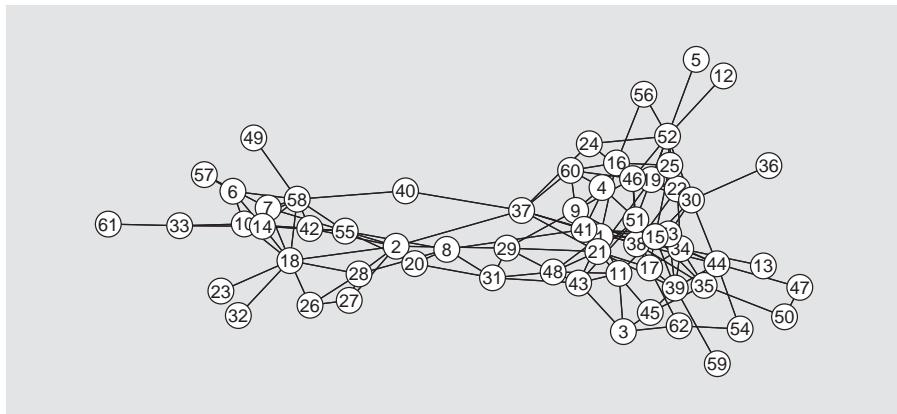
Finally, a random embedding is a function  $f$  from nodes into the unit square (or sometimes the unit disk) chosen from the uniform distribution. Figure 4.7 gives a sample random embedding of  $G_{39}$ .

The spring-electrical embedding is the default way of representing graphs of the real-world networks in this chapter. However, the reader should be aware that its virtues are primarily the visual appearance of graphs and does not consistently represent graphical properties in any way interpretable in terms of the distances between nodes in the embedding. For example, the spring, spring-electrical, and circular embeddings keep nodes separated even if they are identical in every way. Section 4.2.4 will present a method of embedding graphs based on the eigenvectors corresponding to the second and third smallest Laplacian eigenvalues. This method gives a fairly good visual display, but more importantly has an interesting graph-theoretic interpretation.

### 4.2.2 Examples of networks

We shall look at four interesting networks, which are presented in this section but analyzed in later sections. The four groups are bottlenose dolphins in a fjord, karate students in a class, genetic relations between human diseases, and the Internet at the level of autonomous systems. Many other examples could have been chosen, and here are a few other types of networks that are commonly studied: coauthors in a given field of study, formal hierarchies and informal contacts in organizations, words linked by free association norms, molecules in a cell linked by metabolic processes, and neural networks in brains.

In order to describe an empirical network with either a graph or a digraph, one has to decide on a set of objective criteria for determining the nodes and edges (or arcs). Often, the nodes have a natural definition, such as the set of girls in a classroom or the set of nerve cells in the brain. At other times, however, one has many choices in determining the working definition of a node, such as in the study of international trade: do you lump the countries in the EU together, or consider the individual nation states as the nodes? Finally, when one begins to consider the edges of the network, one might be forced to include more nodes because they are strongly connected to the initial set of nodes, e.g., adding boys in the classroom example, or adding spinal cord and peripheral nervous system cells to the study of brain cells.



**Figure 4.8** A network of friendships among dolphins.

Having provisionally defined the nodes of empirical network, the next step is to determine the presence or absence of edges or arcs between pairs of nodes. Sometimes these are defined formally; for example, in a coauthor network, an edge between two authors exists if and only if they have been coauthors on at least one paper from a specific set of journals in predefined years. In other networks, more arbitrary empirical criteria are used for ascribing edges or arcs in the network. For example, in some studies of social networks each person is asked to list their top three friends, so that the arc  $x \rightsquigarrow y$  exists if and only if person  $y$  is on the list of the three friends listed by person  $x$ . Of course, this does not imply that  $y$  is a friend of  $x$  in any deeper sense of the meaning of friendship. Certainly, it is difficult to believe that everyone in a group would have exactly three friends. We will see in the examples to follow that the choice of criteria for edges or arcs in a network is a central issue in the study of networks.

#### 4.2.2.1 Bottlenose dolphins in Doubtful Sound

The first example of a network is the set of “friendships” among 62 bottlenose dolphins Lusseau *et al.* (2003) found in Doubtful Sound, a New Zealand fjord. Their social network is shown in Figure 4.8. These dolphins were observed over a period of seven years, swimming in schools unsegregated by sex. Thus, the set of nodes is the set of dolphins (male and female) found in a well-defined geographical area. However, juvenile dolphins and those that died during the study period were not included in the final sample.

The edge set was determined from the schools they swam in. A “school” here means a set of dolphins seen together on a particular day, but the next day the schools would change. The median school size was 14. Thus, the raw data consisted of an  $M$  by  $N$  0–1 matrix  $\mathbf{B} = (b_{ij})_{M,N}$ , where  $N = 62$  is the number of dolphins,  $M$  is the number of schools, and  $b_{ij} = 1$  if the  $j$ th dolphin was seen in the  $i$ th school, else  $b_{ij} = 0$ . This is an example of two-mode data that can be analyzed

by using several methods of analysis (Greenacre, 2007). One method within the graph-theoretic framework would be to view the data as a bipartite graph. This bipartite graph might have too many nodes to be clearly displayed, although one could use two different shapes to signal which nodes are dolphins and which are schools.

However, if one wanted to define a one-mode dolphin by dolphin matrix, the most natural choice would be  $\mathbf{C} = \mathbf{B}^T \mathbf{B}$ . Note that  $c_{ij}$  is the number of schools that dolphins  $i$  and  $j$  were seen together in (dually, one could examine the school by school matrix  $\mathbf{B}\mathbf{B}^T$ ). One could consider the graph on the set of dolphins defined by the “ever swam together” relation: i.e., for any two dolphins  $i$  and  $j$ , we have  $i \sim j$  if and only if  $c_{ij} > 0$ . Unfortunately, in this case this would result in the complete graph, or nearly so, because almost every pair of dolphins eventually did swim together. Alternatively, one could choose an integer cutoff  $k > 0$  such that the adjacency matrix would be neither too dense or too sparse:  $i \sim_k j$  if and only if  $c_{ij} \geq k$ . However, this is a very crude procedure that does not take account of the variance in school size (row sums of  $\mathbf{B}$ ) or schools per dolphin (column sums). In this case, as well as in many examples of empirical graphs or digraphs, one needs a better measure of association, or similarity, than  $\mathbf{C}$ .

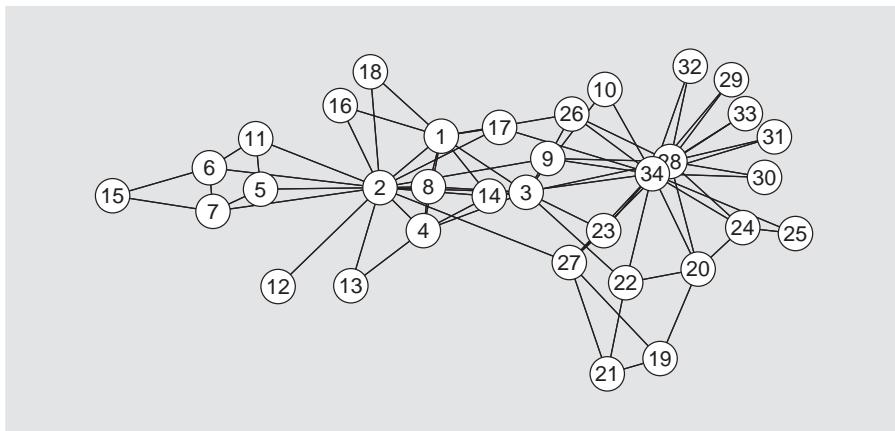
The measure of association chosen by Lusseau *et al.* (2003) is called the *half-weight index* (HWI), defined by

$$\text{HWI}(i, j) = 2|S_i \cap S_j|/(|S_i| + |S_j|), \quad (4.6)$$

where  $S_i$  is the set of schools where dolphin  $i$  was seen. In terms of the data matrix  $\mathbf{B}$ , if the first dolphin above is represented by column  $j$ , then  $S_j = \{i \mid b_{ij} = 1\}$ . The intersection  $S_i \cap S_j$  is the set of schools where dolphins  $i$  and  $j$  were seen together.

Note that there are many alternatives to HWI as an index of similarity, because there are many possible combinations of union, intersection, and relative complement that make good sense for such measures. For example, a criticism of HWI is that the total number of schools does not appear in the formula. In addition, one might also argue for a “dissimilarity” measure where the larger the number the less similar are the two sets. There are seven such dissimilarity measures between two sets in the program *Mathematica*<sup>®</sup>: “Dice,” “Jaccard,” “Matching,” “Rogers-Tanimoto,” “Russel-Rao,” “Sokal/Sneath,” and “Yule.” All have values between 0 and 1 except for the Hamming distance based on the symmetric difference of sets,  $S_i \Delta S_j \stackrel{\text{def}}{=} (S_i \setminus S_j) \cup (S_j \setminus S_i)$ .

The final form of the dolphin data is a 0–1 matrix, transformed from the observations into HWI numbers and then into a 0 or a 1 by a statistical algorithm. It may be that the many kinds of (dis)similarity measures would lead to very similar adjacency matrices at the end of this process. Still, it seems that much of the rich information in these exceptional data has been lost by massaging the data into a 0–1 matrix. Still, transforming a raw two-mode matrix into an adjacency matrix is a frequent step seen in network analysis.



**Figure 4.9** A network of friendships among karate students.

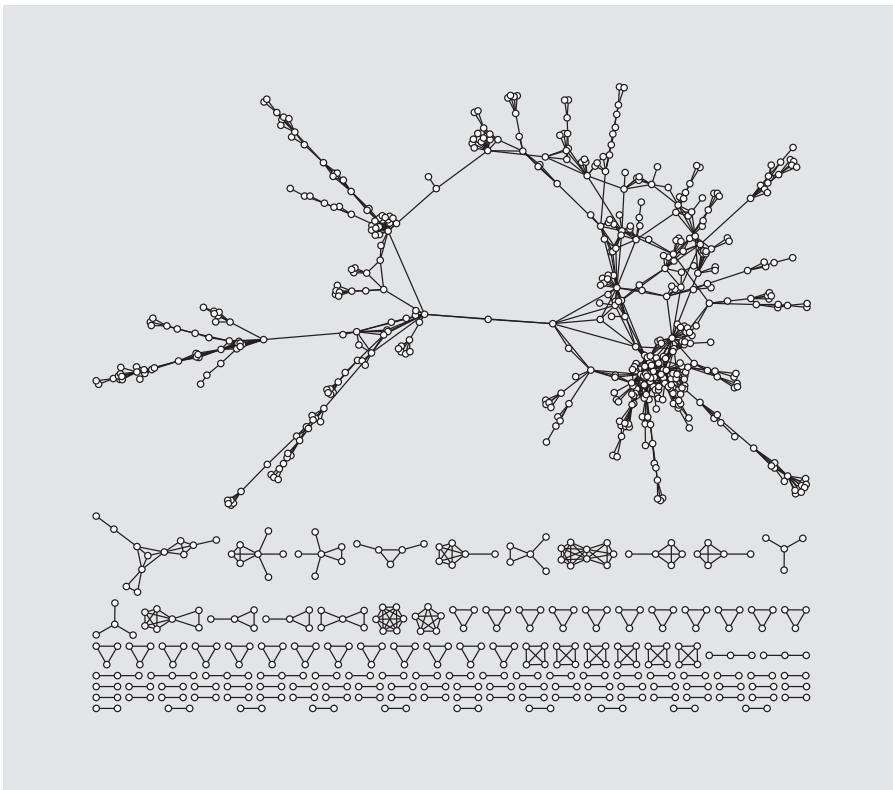
#### 4.2.2.2 Karate students

The second example of a network is friendships among members of a university-based karate club (Zachary, 1977). The observations were taken for a period of three years, from 1970 to 1972. The club had 60 members at the time of the study. The 26 members of the club that had no outside interaction with members of the club were excluded from the network, leaving 34 who did. The network for the 34 who did have contact with other members outside of club activities are depicted in Figure 4.9.

Zachary (1977) considered two members of the karate club to be “friends” if they were observed to consistently associate *outside* the club activities such as karate classes or club meetings. There were eight such outside activities ranging from “Association in and between academic classes at the university” to “Attendance at intercollegiate karate tournaments.” This definition emphasizes that friendship is characterized by voluntary acts rather than formal criteria like common membership in a club. The two most prominent members of the karate club, the club president and the informal karate instructor, are represented in Figure 4.9 by nodes 2 and 34, respectively. These two nodes have the highest degrees, 16 and 17. One of their main points of contention was the price of karate instruction, and the members of the club formed two informal factions around this issue.

#### 4.2.2.3 Human disease network

The third example of a network is derived from a study (Goh *et al.*, 2007) of human diseases and the genes whose mutations are associated with them. The website <http://www.omim.org/> keeps updated versions of this data. As of December 2005, a total of  $N = 1284$  humans diseases were linked with  $M = 1777$  disease genes. A gene and a disease are linked if mutations in the gene are associated with the disease. As with the dolphin data above, the raw data are an  $M$  by  $N$  0–1 matrix  $\mathbf{B} = (b_{ij})$ , where  $b_{ij} = 1$  if the  $j$ th disease was associated with the  $i$ th gene, else

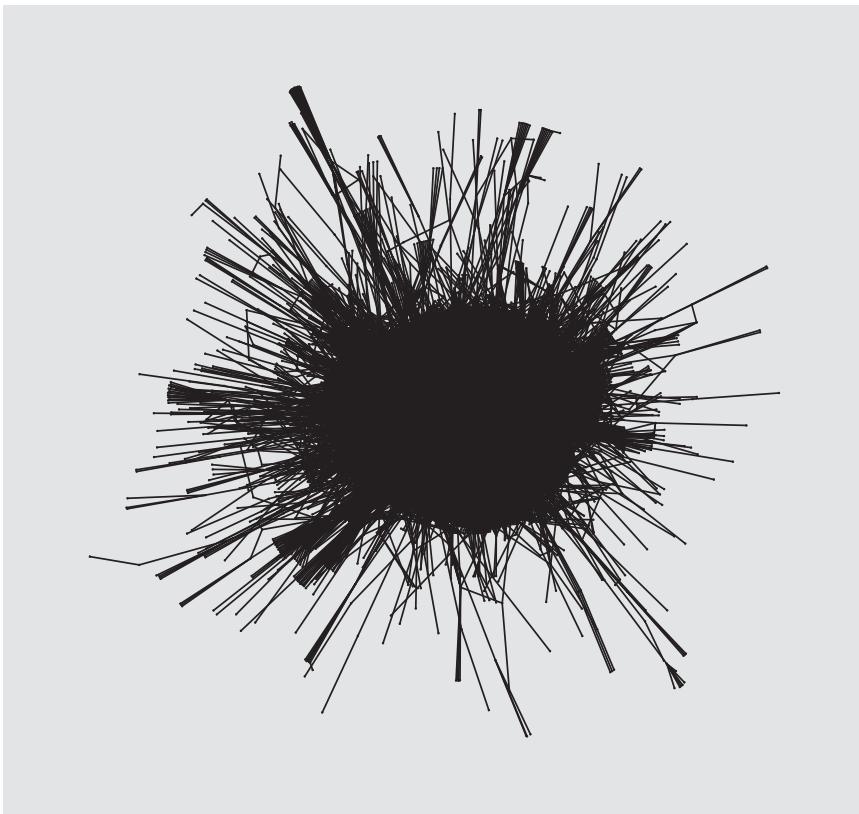


**Figure 4.10** A graph of 1284 human diseases and their genetic relationships, except that the 417 isolated diseases are not shown.

$b_{ij} = 0$ . This is another example of two-mode data that can be represented by a bipartite graph. Unlike the dolphin case, the disease matrix  $\mathbf{B}$  is sparse enough so that the  $N$  by  $N$  adjacency matrix  $\mathbf{A}$  can be defined by the rule that  $a_{ij} = 1$  if the corresponding entry in  $\mathbf{B}^T \mathbf{B}$  is positive, and otherwise  $a_{ij} = 0$ . That is, it is not necessary to choose a similarity measure and a cutoff to produce a graph that is not cluttered with an edge between almost all nodes. The graph induced by this matrix is drawn in Figure 4.10. Note that unlike the other two networks in this paper, there are a large number (417) of isolated nodes (not shown), as well as other, nontrivial, connected components. On the other hand, there are a few nodes of extremely high degree: colon cancer has the highest degree, 50, while breast cancer has the second highest degree, 30.

#### 4.2.2.4 Internet

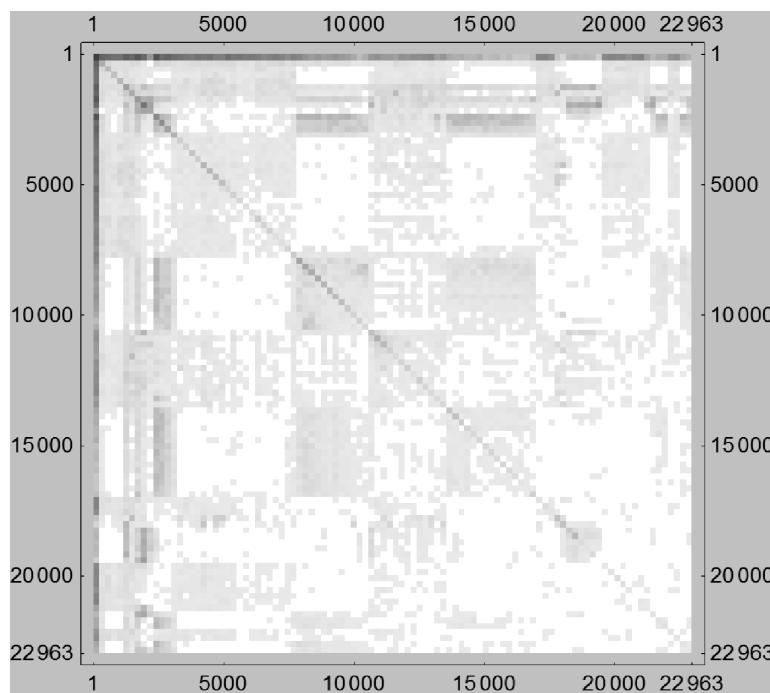
The fourth and last network example is the Internet, which potentially connects almost everyone in the world. To reduce the number of nodes we shall consider the Internet at a higher level of abstraction. Within the Internet an *autonomous system* (AS) is a collection of routing prefixes under the control of a single entity which



**Figure 4.11** A graph plot of the of Internet connections between autonomous systems.

has a unique autonomous system number (ASN). *Mathematica*<sup>®</sup> supplies a symmetrized snapshot of the structure of the Internet at the level of these autonomous systems, reconstructed from the University of Oregon Route Views Project. The resulting graph has 22,963 nodes and 48,436 edges, while the maximum node degree is 2390. Naturally, the graph of this version of the Internet is not too informative, except as a Rorschach test: is Figure 4.11 a hairy ball or a porcupine? Certainly, the pattern fits the “core–periphery” concept (Borgatti and Everett, 2000), where the “core” would be nodes in the developed countries and regions, while the “periphery” would be third world countries, small islands, and dependencies.

Another way to picture these Internet data is to show their adjacency matrix with gray scale indicating the number of connections from nodes near  $i$  to nodes near  $j$ . Naturally, this depends on having an ordering of the rows and columns that reflects the overall similarity. That is, if  $|i - i'|$  is small, then row  $i$  should be highly similar to row  $i'$ . A good choice is to order the rows and columns by the largest eigenvalue of the adjacency matrix, or by the second smallest eigenvalue of the Laplacian matrix, defined and discussed in Section 4.2.4. If instead the ordering



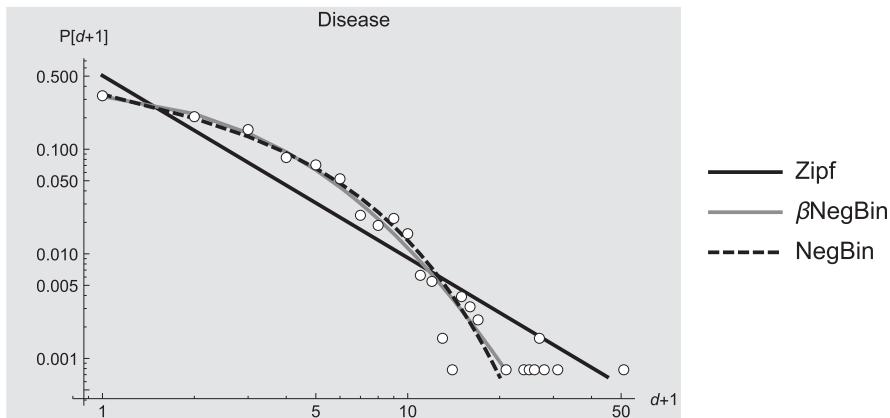
**Figure 4.12** A matrix plot of the Internet connections between autonomous systems.

is random, then the matrix plot will be uninformative, except that nodes of high degree will appear as lines in the corresponding rows and columns. Figure 4.12, however, shows that the internet ASes have 10 or 12 blocks, usually large countries or regions.

### 4.2.3 Probabilistic properties of degree sequences

The degree sequence of graphs is of great interest both mathematically, as in the Erdős–Galli condition (4.3), and empirically, especially for those networks with a large number of nodes. Sometimes the only information available about a large graph is the degree sequence. For example, in epidemiological studies of sexually transmitted diseases (Del Genio *et al.*, 2010), privacy concerns may lead to recording only the number of sexual partners, not their identity. In this case the data are just the degree sequence. Even if the total graph is given, many models partition the description of the graph into a popularity component (the degree sequence) and a structural component conditioned on the degree sequence. We shall examine several probability distributions that have been used to describe degree sequences, with the purpose of fitting the degrees in the disease graph of Figure 4.10.

The degree sequence for the disease data is plotted in Figure 4.13, where  $k$  refers to a particular degree value and  $p_k$  refers to the proportion of nodes in the graph



**Figure 4.13** A plot of the degree sequence of the disease graph. Note that the  $\beta$ -negative binomial gives a much better fit than does Zipf's power law.

with degree  $k$ . The log-log plot is usual for large data sets. This has the benefit of showing *power laws*, probabilities proportional to  $k^{-\alpha}$ , as straight lines. The reason why the log-log plot gives these straight lines is that if you take logs of the equation  $p_k = ck^{-\alpha}$ , where  $\alpha$  and  $c$  are constants, you get  $\log p_k = \log c - \alpha \log k$ , a linear equation relating  $\log p_k$  to  $\log k$ .

The first type of degree distribution discussed here is for the set of *Bernouilli random graphs*  $G(N, p)$ , where  $N$  is the number of nodes, and where  $p$  represents the probability that an edge is present. The Bernoulli process is to go through all  $N(N - 1)/2$  potential edges, placing an edge  $\{i, j\} \in [N]^{(2)}$  into the edge set with probability  $p$ . Any given node  $i$  is connected to any of the other  $N - 1$  nodes independently and with probability  $p$ . The probability  $p_k$  of this node being connected to exactly  $k$  other nodes is given by the *binomial* probability distribution:

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}, \quad (4.7)$$

where  $0 \leq k < N$ , which shows that  $G(N, p)$  has a binomial degree distribution. The mean and variance of this distribution are  $(N - 1)p$  and  $(N - 1)p(1 - p)$ , respectively.

If  $N$  is increased and  $p$  relatively small, as is the case in large sparse networks, then it is well-known that the binomial distribution in (4.7) converges in distribution to the *Poisson* distribution

$$p_k = \frac{e^{-\lambda} \lambda^k}{k!}, \quad (4.8)$$

where  $k \in \mathbb{Z}_{\geq 0}$  and  $\lambda = (N - 1)p$ . This means that the binomial distribution is well-approximated by a Poisson distribution when  $N$  is large,  $p$  is small, and  $Np = \lambda$  is moderate. However, in addition to small approximation errors to the binomial usually entailed with the Poisson, there is another restriction, namely that  $p_k$  must

be zero for  $k \geq N$ , because the largest possible degree of a graph with  $N$  nodes is  $N - 1$ . Fortunately for large sparse graphs, these logically impossible nonzero probabilities are usually negligible. In those few cases where these probabilities are not negligible, the corrected formula for (4.8) is the the *right truncated Poisson* distribution:

$$p_k = \frac{\lambda^k}{k!} \left( \sum_{j=0}^{N-1} \frac{\lambda^j}{j!} \right)^{-1}, \quad (4.9)$$

defined for  $k = 0, 1, \dots, N - 1$ .

A diagnostic property for a Poisson distribution is that the variance equals the mean. This allows us to immediately dismiss the Poisson distribution as a model for the disease network, because the mean and variance of its degrees are 2.378 and 12.405. Although small differences might be due to chance, in this case the maximum likelihood fitness test (as implemented in *Mathematica*<sup>®</sup>) finds that the  $P$ -value is  $2.7 \times 10^{-316}$ , making it highly unlikely the Poisson fits.

One difficulty with testing models for degree sequences like the binomial and Poisson models is that any particular degree sequence  $\mathbf{d}$  is not an an i.i.d. sample from the model. One obvious constraint is that the degree sequence must be graphical, as prescribed in the Erdős–Gallai theorem. Another such constraint in the uniform graph model  $G_{N,M}$  is that there is a negative correlation of  $-1/(N - 1)$  between any two degrees. For example, if  $d_1$  is greater than the mean degree,  $2M/N$ , then the mean of the other degrees must be less than  $2M/N$ . However, these constraints are less important for large  $N$ .

A more plausible and useful model for degree distributions (Rapoport, 1963), is the *negative binomial* distribution, which has parameters,  $n, p \in \mathbb{R}_{>0}$  such that  $p < 1$ , with the pdf

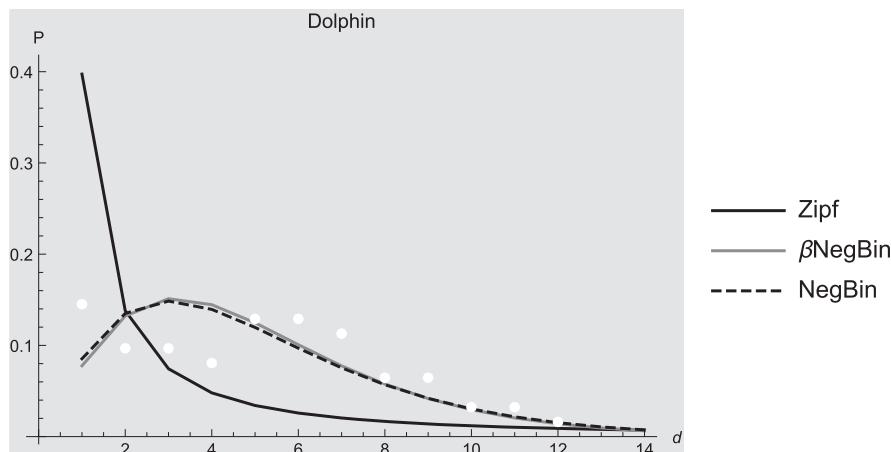
$$p_k = \binom{n+k-1}{n-1} p^n (1-p)^k \quad (4.10)$$

$$= \binom{-n}{k} p^n (p-1)^k, \quad (4.11)$$

where  $k \in \mathbb{Z}_{\geq 0}$ . Note that the binomial coefficient is defined by

$$\binom{n}{k} \stackrel{\text{def}}{=} \frac{n(n-1)\cdots(n-k+1)}{k!}, \quad (4.12)$$

for  $n \in \mathbb{Z}$  and  $k \in \mathbb{Z}_{\geq 0}$ , so that  $n$  can be negative, explaining the “negative” in negative binomial. If  $n$  is negative and  $k$  is odd, then (4.12) is negative, but then so is the term  $(p-1)^k$  from (4.11), resulting in a positive probability, as the reader can verify. If  $n$  is a positive integer, then the negative binomial distribution gives the number of failures before the  $n$ th success in a Bernoulli sequence with success probability  $p$ . Another process that leads to a negative binomial is given by a pure birth process with linear birthrate. This is also know as *preferential attachment*, where edges are added to a graph in sequence and the probability that a particular

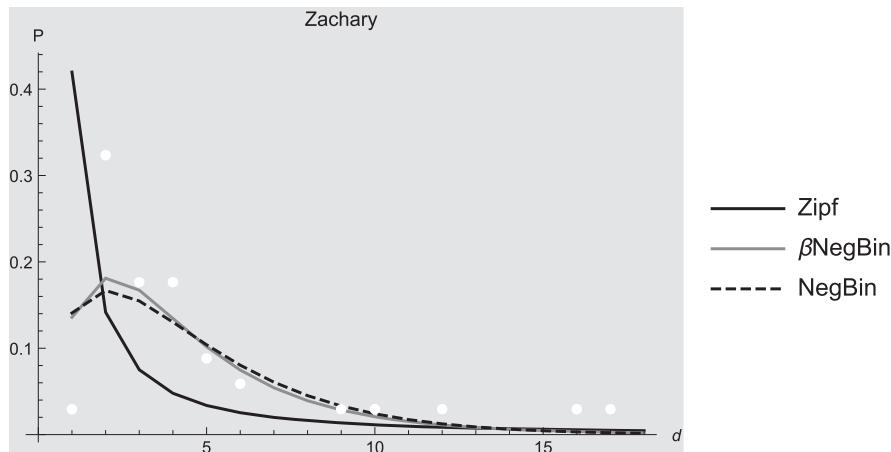


**Figure 4.14** A plot of the degree sequence of dolphin friendship. Note that the  $\beta$ -negative binomial and the negative binomial both fit the data equally well and better than does Zipf’s power law. However, the negative binomial has an extra parameter.

node receives a new edge is proportional to its current degree. Unlike the Poisson distribution, the negative binomial has a larger variance  $n(1 - p)/p^2$  than mean  $n(1 - p)/p$ , which is characteristic of most empirical degree sequences. The ratio of the variance to the mean  $D = \sigma^2/\mu$  is called the *index of dispersion* and can be estimated from data or from the theoretical distribution. If  $D > 1$  then the distribution is said to have *overdispersion*, which is typical of a lot of empirical networks where many nodes have a small degree while a few nodes called “hubs” have very large degrees. The other possibilities are called *underdispersion* when  $D < 1$  and *equidispersion* when  $D = 1$ . The Poisson distribution has equidispersion because  $D = 1$ , while for the negative binomial  $D = 1/p > 1$ , indicating overdispersion. Finally, the binomial distribution has underdispersion because  $D = 1 - p < 1$ .

Note that, like the Poisson distribution, the negative binomial has a positive probability for all nonnegative integers, whereas every degree of a graph must be less than the number of nodes,  $N$ . However, for large sparse graphs, this is not a practical problem because these probabilities are vanishingly small. Fitting the negative binomial distribution to the disease network degrees, we find that  $n = 0.78229$  and  $p = 0.24750$ , with a resulting  $P$ -value of 0.1033. This means the negative binomial cannot be rejected as a model for the disease degree sequence. See Figure 4.13 for the fitted negative binomial distribution plotted against the disease degree sequence. Similar plots for the degree sequences of the dolphin and karate graphs are shown in Figures 4.14 and 4.15. Although the numbers are smaller, again the negative binomial cannot be rejected.

Next, we consider the  $\beta$ -negative binomial distribution, which is a mixture distribution of the negative binomial with a beta distribution (with parameters  $\alpha$  and  $\beta$ ) on the parameter  $p$ . This results in three positive real parameters for the



**Figure 4.15** A plot of the degree sequence of the Zachary karate students.  
Note that Zipf's power law fits very poorly and while the  $\beta$ -negative binomial and negative binomial both fit the data much better, they still can be rejected at the 5% level.

$\beta$ -negative binomial,  $n$ ,  $\alpha$ , and  $\beta$ , with pdf

$$p_k = \frac{n^{(k)} \alpha^{(n)} \beta^{(k)}}{k! (\alpha + \beta)^{(n)} (n + \alpha + \beta)^{(k)}} \quad (4.13)$$

$$= \binom{-\beta}{k} \binom{\alpha + \beta - 1}{-n - k} / \binom{\alpha - 1}{-n}, \quad (4.14)$$

where  $k \in \mathbb{N}$  and  $n^{(k)}$  is the rising factorial. If  $k$  is a positive integer, then  $n^{(k)} = n(n+1)\cdots(n+k-1)$ , but if  $k$  is a positive real number, then the rising factorial can still be defined as  $\Gamma(n+k)/\Gamma(n)$ , where the gamma function  $\Gamma$  is defined by an integral. Similarly, the binomial coefficients in (4.13) are defined by gamma functions. The mean  $\mu$  and variance  $\sigma^2$  of the  $\beta$ -negative binomial in (4.13) are

$$\mu = \frac{n\beta}{\alpha - 1} \quad \text{if } \alpha > 1, \text{ else } \infty \quad (4.15)$$

$$\sigma^2 = \frac{n(n+\alpha-1)\beta(\alpha+\beta-1)}{(\alpha-2)(\alpha-1)^2} \quad \text{if } \alpha > 2, \text{ else } \infty. \quad (4.16)$$

The index of dispersion for the  $\beta$ -negative binomial for  $\alpha > 2$  is

$$D = \frac{(n+\alpha-1)(\alpha+\beta-1)}{(\alpha-2)(\alpha-1)} > 1,$$

which indicates overdispersion. If  $1 < \alpha \leq 2$ , then (4.13) implies that  $D = \infty$ , which is extreme overdispersion. However, if  $\alpha < 1$ , then both the mean and the variance are infinite, giving dispersion the indeterminate form  $\infty/\infty$ .

Fitting the  $\beta$ -negative binomial distribution to the disease network degrees, we find that  $n = 9.0702$ ,  $\alpha = 5.5135$ , and  $\beta = 1.1803$ , with a resulting  $P$ -value of

0.14835. This means the  $\beta$ -negative binomial cannot be rejected as a model for the disease degree sequence. Figure 4.13 displays the fitted  $\beta$ -negative binomial distribution plotted against the data. This distribution is only marginally better than the negative binomial, but it does seem to fit better in the right-hand tail.

The final distribution illustrated here is *Zipf's law*, also known as the *discrete Pareto law*. Its pdf is given by

$$p_k = \frac{k^{-\rho-1}}{\zeta(\rho+1)}, \quad (4.17)$$

where  $k \in \mathbb{Z}_{\geq 0}$  and  $\rho \in \mathbb{R}_{>0}$ , the positive integers and reals, respectively. The normalizing function  $\zeta(\rho)$  is known as the Riemann Zeta function. The mean  $\mu$  and variance  $\sigma^2$  of Zipf's law in (4.17) are

$$\mu = \frac{\zeta(\rho)}{\zeta(\rho+1)} \quad \text{if } \rho > 1, \text{ else } \infty \quad (4.18)$$

$$\sigma^2 = \frac{\zeta(\rho-1)}{\zeta(\rho+1)} - \frac{\zeta(\rho)^2}{\zeta(\rho+1)^2} \quad \text{if } \rho > 2, \text{ else } \infty. \quad (4.19)$$

The dispersion index, after subtracting 1 from the mean to compensate for Zipf's law not being defined at 0, is

$$D = \frac{\zeta(\rho+1)\zeta(\rho-1) - \zeta(\rho)^2}{\zeta(\rho)\zeta(\rho+1) - \zeta(\rho+1)^2} > 1, \quad (4.20)$$

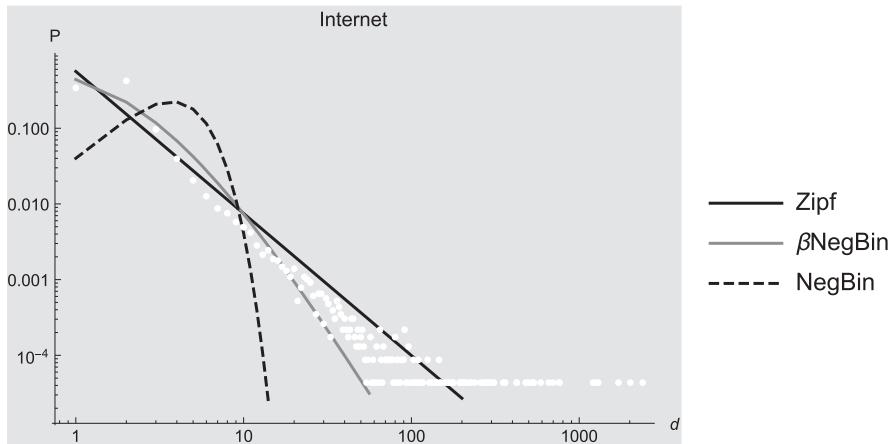
for  $\rho > 2$ , which is proven by noting that  $\zeta(\rho)$  is decreasing and convex for these values of the parameter.

Fitting Zipf's law to the disease network degrees, we find that  $\rho = 0.74190$  with a resulting  $P$ -value of  $4.1 \times 10^{-75}$ , which leads one to reject this model for the disease degree sequence. From the plot in Figure 4.13 it is also obvious that this power law is a poor fit at both ends of the distribution, but especially at the beginning. Some proponents of power law advocate choosing the minimum number  $k_{\min}$  such that the power law holds for  $k \geq k_{\min}$  and then fit only that part of the data (Newman, 2010). However, it would seem to be desirable to employ distributions that fit all the data without having to discard inconvenient points.

It is interesting to note that the tail of the  $\beta$ -negative binomial converges to a power law like the one found in the Zipf law with parameter  $\rho = \alpha$ . The limit ratio of the  $\beta$ -negative binomial (4.13) to Zipf's law (4.17), ignoring factors independent of  $k$ , can be shown to be

$$\lim_{k \rightarrow \infty} \frac{n^{(k)} \beta^{(k)}}{k! (n + \alpha + \beta)^{(k)} k^{-1-\alpha}} = \frac{\Gamma(n + \alpha + \beta)}{\Gamma(n) \Gamma(\beta)}. \quad (4.21)$$

In comparing the three statistical models in Figure 4.13, the Zipf gives a very poor fit, while the negative binomial and the  $\beta$ -negative binomial give adequate fits. The number of parameters in these models is 1, 2, and 3, respectively, but the computation of the  $P$ -values take this into account. In the choice between the last two models, however, the fact that the  $P$ -values are comparable for the disease data



**Figure 4.16** A plot of the degree sequence of Internet connections. Note that Zipf's power law and the  $\beta$ -negative binomial both fit the data better than does the negative binomial.

might tilt one towards the negative binomial because it eliminates one parameter. Still, the advantage of the  $\beta$ -negative is that in the limit it does converge to a power law, so that it might be more appropriate for other, larger, data sets.

The Internet data give another example of a graph with a degree sequence that does seem to fit the power law (Zipf's law). As plotted in Figure 4.16, it appears that both Zipf's law and the  $\beta$ -negative binomial fit the data much better than does the negative binomial distribution. Because Zipf's law has only one parameter, it would appear to be better than either of the other two distributions. Zipf's law seems to be most appropriate for extremely large data sets whose values range over several orders of magnitude.

## 4.2.4 Matrix methods

### 4.2.4.1 Matrix operations

It turns out that a lot of properties of graphs and digraphs can be computed using matrix methods. In fact, tools of analysis based on matrix methods have contributed greatly to major recent advances in network science. In this subsection we first define the essential matrix concepts and then show how they can be used to examine the properties of a network. Although adjacency matrices have only 0s and 1s as entries, they can be considered to be real-valued matrices, making available a wider range of methods and results.

If  $\mathbf{A}$  and  $\mathbf{B}$  are two matrices with dimensions  $L$  by  $M$  and  $M$  by  $N$ , respectively, then *matrix multiplication* is defined by

$$(\mathbf{AB})_{ij} \stackrel{\text{def}}{=} \sum_{k=1}^M a_{ik} b_{kj}, \quad (4.22)$$

where  $i \in [L]$  and  $j \in [N]$ . If  $\mathbf{A}$  is a square  $N$  by  $N$  matrix, it may have an *inverse*, a matrix denoted by  $\mathbf{A}^{-1}$  such that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}, \quad (4.23)$$

where  $\mathbf{I} = \mathbf{I}_N$  is the  $N$  by  $N$  *identity matrix*, consisting of 1s on the diagonal and 0s elsewhere. A square matrix that has no inverse is said to be *singular*, and is characterized by linear dependence among its rows (or columns).

The  $N$  by  $N$  identity matrix is said to be the *multiplicative identity* for all  $N$  by  $N$  matrices  $\mathbf{A}$  because the equation

$$\mathbf{A}\mathbf{I} = \mathbf{I}\mathbf{A} = \mathbf{A},$$

holds.

Because an  $M$  by 1 matrix  $\mathbf{A}$  has just one column, it can be considered to be a (*column*) *vector*,  $\mathbf{a} = (a_{i1})_{i \in [M]}$ . Usually we omit the redundant 1 index and refer to the  $i$ th component of this vector as  $a_i$ . The number  $N$  in this context is called the *dimension* of the vector. The vector consisting of  $N$  0s is called the *zero vector*  $\mathbf{0}$ . Another important special kind of vectors are the *standard basis* vectors for  $N$ -dimensional space. These vectors are denoted by  $\mathbf{e}_1, \dots, \mathbf{e}_N$ , where the  $i$ th standard basis vector,  $\mathbf{e}_i$ , is the vector of all 0s except for the  $i$ th component, which equals 1. In this chapter vectors, such as  $\mathbf{u}$  or  $\mathbf{v}$ , are by default *column* vectors, while row vectors are defined as transposed column vectors:  $\mathbf{u}^T$  or  $\mathbf{v}^T$ . Note that if a column vector  $\mathbf{v}$  is treated as an  $N$  by 1 matrix, then the corresponding row vector  $\mathbf{v}^T$  is the matrix transpose of  $\mathbf{v}$ . Therefore, (4.22) also defines vector–matrix, matrix–vector, and even vector–vector multiplication.

If we have two column vectors,  $\mathbf{u}$  and  $\mathbf{v}$ , both of dimension  $N$ , then one type of vector–vector multiplication has the special name, the *inner product*, defined (as in (4.22)) by

$$\mathbf{u}^T \mathbf{v} \stackrel{\text{def}}{=} \sum_{k=1}^N u_k v_k. \quad (4.24)$$

If the inner product between two vectors is 0, then they are said to be *orthogonal*. Note that the product  $\mathbf{u}\mathbf{v}^T$  is a matrix, instead of a number as in (4.24), and is defined even when the two vectors have different dimensions. The *length* or *magnitude* of a vector  $\mathbf{v}$  is given by

$$\|\mathbf{v}\| \stackrel{\text{def}}{=} \sqrt{\mathbf{v}^T \mathbf{v}}. \quad (4.25)$$

If the magnitude of a vector is 1, then it is said to be *normal*. To *normalize* a vector, just divide by its magnitude and denote the result with a “hat”:  $\hat{\mathbf{v}} \stackrel{\text{def}}{=} \mathbf{v}/\|\mathbf{v}\|$ . The *cosine* between two vectors  $\mathbf{u}$  and  $\mathbf{v}$  of the same dimension is defined with the previous concepts as

$$\cos(\mathbf{u}, \mathbf{v}) \stackrel{\text{def}}{=} \frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} = \hat{\mathbf{u}}^T \hat{\mathbf{v}}. \quad (4.26)$$

When the two vectors each have a mean of zero, the cosine reduces to the *correlation coefficient*. The cosine between a vector  $\mathbf{v}$  and the  $i$ th standard basis vector is  $v_i/\|\mathbf{v}\| = \hat{v}_i$  and is called the *i*th *directional cosine* for  $\mathbf{v}$ . The normalized vector  $\hat{\mathbf{v}}$  can be visualized as the point where  $\mathbf{v}$  intersects the  $(N-1)$ -hypersphere of radius 1.

If  $\mathbf{A}$  is an  $M$  by  $N$  matrix, then its  $j$ th column is the vector  $\mathbf{a}_j \stackrel{\text{def}}{=} (\mathbf{A}_{i,j})_{i \in [M]}$ . Similarly, the  $i$ th row vector of  $\mathbf{A}$  is defined as  $\mathbf{a}_i^T \stackrel{\text{def}}{=} (\mathbf{A}_{i,j})_{j \in [N]}$ . This means that the matrix can be treated as a row vector of its columns:  $\mathbf{A} = (\mathbf{a}_j)_{j \in [N]}$ . Similarly, a matrix can be treated as a column vector of its rows. The utility of this exercise is that it gives us a useful way to represent matrix multiplication equivalent to the definition (4.22) as a sum of the columns of the first matrix times the rows of the second:

$$\mathbf{AB} = \sum_{k=1}^N \mathbf{a}_k \mathbf{b}_k^T. \quad (4.27)$$

This concept is extremely useful in the eigenvalue decomposition of matrices in the next section.

As a special case of (4.27), the matrix–vector product  $\mathbf{Av}$  is a vector that is the sum of the columns of  $\mathbf{A}$  weighted by the components of  $\mathbf{v}$ . If  $\mathbf{Av} = \mathbf{0}$  for some nonzero vector  $\mathbf{v}$ , then the columns of  $\mathbf{A}$  are said to be *linearly dependent*; if no such vector exists, the columns are said to be *linearly independent*. If the columns of  $\mathbf{A}$  are linearly dependent, then it cannot have an inverse. For if there did exist an inverse,  $\mathbf{A}^{-1}$ , defined in (4.23), then

$$(\mathbf{A}^{-1}\mathbf{A})\mathbf{v} = \mathbf{I}\mathbf{v} = \mathbf{v} \neq \mathbf{0}, \text{ while} \quad (4.28)$$

$$\mathbf{A}^{-1}(\mathbf{Av}) = \mathbf{A}^{-1}\mathbf{0} = \mathbf{0}, \quad (4.29)$$

which contradicts the associative law. This shows that linear independence is necessary for the existence of a matrix inverse. The condition is also sufficient.

In network analysis the most common choices for the values of a matrix are  $\{0, 1\}$ , the nonnegative integers  $\mathbb{Z}_{\geq 0}$ , or the reals  $\mathbb{R}$ . Other kinds of structured values for matrices are possible, but the essential thing is that there be a well-defined addition and multiplication that will allow matrix multiplication to be defined as in (4.22). The nonnegative reals  $\mathbb{R}_{\geq 0}$ , or intervals thereof, can be used as matrix entries to indicate an objective “degree of association” or a probability of each edge or arc. This is a promising area for more accurate representations of relationships.

The entries for a square 0–1 matrix, such as in an adjacency matrix, can also be treated as nonnegative integers. Using the usual matrix multiplication (4.22),  $\mathbf{A}^2$  is a matrix of nonnegative integers, where the  $i, j$ th entry counts the number of walks of length two from  $i$  to  $j$ . Furthermore, the  $k$ th power of the adjacency matrix  $\mathbf{A}^k$ , where  $k \geq 0$ , counts the number of walks of length  $k$  between any two nodes. Matrices over the nonnegative integers  $\mathbb{Z}_{\geq 0}$  are also a useful way of representing so-called “multigraphs,” where  $a_{ij}$  is the number of edges or arcs from  $i$  to  $j$ .

#### 4.2.4.2 Eigenvalues and eigenvectors

Let  $\mathbf{A}$  be an  $N$  by  $N$  square matrix with entries in the reals. A nonzero vector  $\mathbf{v}$  is said to be an (*right*) *eigenvector* for the *eigenvalue*  $\lambda \in \mathbb{R}$  if

$$\mathbf{Av} = \mathbf{v}\lambda. \quad (4.30)$$

The set of all eigenvalues for a square matrix  $\mathbf{A}$  is called its *spectrum* and is denoted by  $\lambda(\mathbf{A})$ . The meaning of (4.30) is linked to the concept of vectors having both a direction and a magnitude. For example, the wind has both a direction, say north-east, and a magnitude, say 10 miles per hour. If we normalize the eigenvector  $\mathbf{v}$  in (4.30) and recall that the entries of  $\hat{\mathbf{v}}$  are the directional cosines, we obtain

$$\mathbf{A}\hat{\mathbf{v}} = \hat{\mathbf{v}}\lambda,$$

which shows that multiplying an eigenvector by its matrix preserves its direction (in the form of directional cosines) but multiplies its magnitude by its eigenvalue.

The following theorem shows that when you multiply a matrix by a constant or add a constant to the diagonal elements, then the eigenvalues are modified accordingly, but the eigenvectors remain invariant.

**Theorem 4.4** (Effects of scalars on eigenvalues) *Let  $(\mathbf{v}, \lambda)$  be an (eigenvector, eigenvalue) pair for the matrix  $\mathbf{A}$ . Then the following equations hold for all scalars  $c \in \mathbb{R}$ :*

$$(c\mathbf{A})\mathbf{v} = c(\mathbf{v}\lambda) = \mathbf{v}(c\lambda)$$

$$(\mathbf{A} + c\mathbf{I})\mathbf{v} = \mathbf{Av} + c\mathbf{v} = \mathbf{v}(\lambda + c),$$

where  $\mathbf{I}$  is the identity matrix of the same dimension as  $\mathbf{A}$ .

The proof uses only the definition of eigenvalues and vectors and the associative and distributive laws for matrices, vectors, and scalars. The impact of the first equation in Theorem 4.4 is that the entries of a matrix can be changed by a multiplicative constant, say by changing units from seconds to minutes, and the resulting eigenvalues are changed by the same constant, while the eigenvectors are changed not at all. The second equation shows that adding a constant to each diagonal entry adds the same constant to the eigenvalues and again does not change the corresponding eigenvectors.

However, adding a positive real constant  $c$  to every element of the matrix changes everything in complicated ways. One can say that as  $c \rightarrow \infty$ , then the largest eigenvalue approaches  $Nc$  with eigenvector the vector  $\mathbf{1}$  of all 1s. To see this, note that the  $N$  by  $N$  matrix  $\mathbf{C}$  whose entries are all equal to  $c$  satisfies the eigen equation,  $\mathbf{C}\mathbf{1} = Nc\mathbf{1}$ ,

Finding eigenvectors and values can be a difficult nonlinear problem. It is *nonlinear* because the right-hand side of (4.30) consists of products of the form  $v_i\lambda$  for the  $N + 1$  unknowns. But if we are just given an eigenvalue, then (4.30) becomes a linear problem easily solved by standard methods. An even simpler problem is to verify whether or not a given vector and number is an eigenvector–eigenvalue pair

for the matrix. For example, consider the eigenvalue problem for the adjacency matrix  $\mathbf{A}$  of the 4-cycle graph  $C_4$  from Figure 4.2:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}. \quad (4.31)$$

The reader can confirm that the vector  $\mathbf{v} = (-1, 1, -1, 1)^T$  and the number  $\lambda = -2$  satisfies (4.30) for the matrix  $\mathbf{A}$  in (4.31) because

$$\mathbf{Av} = (2, -2, 2, -2)^T = -2(-1, 1, -1, 1)^T.$$

Note that the eigenvector  $\mathbf{v}$  is not uniquely determined by the eigenvalue  $\lambda$ , because multiplying  $\mathbf{v}$  by any nonzero scalar  $c$  also is an eigenvector for  $\lambda$ , because it also satisfies the eigenvector equation (4.30):

$$\mathbf{A}(c\mathbf{v}) = c(\mathbf{Av}) = c(\lambda\mathbf{v}) = \lambda(c\mathbf{v}). \quad (4.32)$$

For example, the matrix  $\mathbf{A}$  in (4.31) also has the eigenvector  $(-10, 10, -10, 10)^T$  for the eigenvalue  $-2$ . In other words, any eigenvalue  $\lambda$  of a matrix  $\mathbf{A}$  determines a *subspace*,  $S_\lambda$ , a set of vectors closed under vector addition and scalar multiplication:

$$S_\lambda = \{\mathbf{v} \in \mathbb{R}^N \mid \mathbf{Av} = \lambda\mathbf{v}\}.$$

Note that the *zero vector*  $\mathbf{0} = 0\mathbf{v}$ , is included in the subspace  $S_\lambda$ , even though the zero vector itself is, by definition, not an eigenvector. The reader should verify that  $S_\lambda$  is closed under vector addition, meaning that if  $\mathbf{u}$  and  $\mathbf{v}$  are both eigenvectors for the same  $\lambda$ , then so is  $\mathbf{u} + \mathbf{v}$ .

If there are so many eigenvectors for each eigenvalue, then how do we decide which one to present? Sometimes we want the eigenvector to be normalized, so  $\hat{\mathbf{v}}$  is a logical choice, remembering that its negation  $-\hat{\mathbf{v}}$  is also normalized. The usual convention is to choose the normalized eigenvector with the fewest negative signs, although this does not help in (4.31) with the eigenvalue  $-2$  because the number of negative and positive entries are equal. Another criterion for eigenvectors is “simplicity,” avoiding fractions, roots, and negative signs whenever possible. This was the rational for choosing the eigenvector  $\mathbf{v} = (-1, 1, -1, 1)^T$  in (4.31) instead of the equally valid  $(-1.2, 1.2, -1.2, 1.2)^T$ .

Let  $\mathbf{V}$  be the matrix consisting of a maximal set of independent eigenvectors, and let  $\Lambda$  be the diagonal matrix of the corresponding eigenvalues. Then we can write (4.30) as a matrix equation:

$$\mathbf{AV} = \mathbf{V}\Lambda. \quad (4.33)$$

Note that  $\mathbf{V}$  is an  $N$  by  $M$  matrix with  $M \leq N$ . The reason for this caveat is that there are pathological examples of matrices  $\mathbf{A}$  that do not have  $N$  independent eigenvectors. For example, if  $\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 0 & 3 \end{pmatrix}$ , then the characteristic equation is  $(3 - \lambda)^2 = 0$ , which has the root 3 with algebraic multiplicity 2. However, all the

eigenvectors are multiples of  $(1, 0)^T$ , so its geometric multiplicity is 1. That is, there is only one independent eigenvector, so the matrix of independent eigenvectors  $\mathbf{V}$  is 2 by 1. Note, however, that  $\mathbf{A}$  does have an inverse,  $\frac{1}{9} \begin{pmatrix} 3 & -1 \\ 0 & 3 \end{pmatrix}$ .

Continuing our example with the adjacency matrix (4.31) for  $C_4$ , the cyclic graph of order 4, the eigenvector equation in matrix form from (4.33) would be

$$\begin{aligned} \mathbf{AV} &= \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} -1 & 1 & 0 & -1 \\ 1 & 1 & -1 & 0 \\ -1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} & (4.34) \\ &= \begin{pmatrix} -2 & 2 & 0 & 0 \\ 2 & 2 & 0 & 0 \\ -2 & 2 & 0 & 0 \\ 2 & 2 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} -1 & 1 & 0 & -1 \\ 1 & 1 & -1 & 0 \\ -1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} -2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \mathbf{V}\Lambda, \end{aligned}$$

where the first matrix is the adjacency matrix  $\mathbf{A}$ , the second is the matrix  $\mathbf{V}$  whose columns are a maximal set of independent eigenvectors, while the last matrix is  $\Lambda$ , the diagonal matrix of eigenvalues,  $-2, 2, 0, 0$ , respectively. When multiplied out, the first and third lines equal the middle line. Note that two new eigenvalues, 2 and 0, have been exhibited, and that there are two independent eigenvectors corresponding to 0.

Several warnings are in order here. First, observe that the the matrix  $\mathbf{V}$  of eigenvectors of (4.33) is to the right of  $\mathbf{A}$ , but is to the left of  $\Lambda$ . Because matrix multiplication is not commutative, it has to be this way. Can you see why? Second, the matrix  $\mathbf{V}$  will not have an inverse if there are not enough linearly independent eigenvectors (4.28). This means that one cannot multiply both sides on the right by  $\mathbf{V}^{-1}$  to decompose  $\mathbf{A}$  because the inverse simply does not exist. Finally, the eigenvalues and vectors could contain real or even *imaginary* numbers, numbers of the form  $a + b\sqrt{-1}$  for  $a, b \in \mathbb{R}$ ; this may arise even if the original matrix contains only 0s and 1s. As we shall see in the spectral theorem, however, the complication of imaginary eigenvalues or eigenvectors only occurs for nonsymmetric matrices, such as the adjacency matrices for digraphs.

In addition to having an inverse, another desirable property of the matrix of eigenvectors is that any two distinct column vectors,  $\mathbf{u}$  and  $\mathbf{v}$ , are *orthogonal*, meaning that  $\mathbf{u}^T \mathbf{v} = 0$ , and that each column vector is *normal*, meaning that  $\mathbf{u}^T \mathbf{u} = 1$ : such matrices are said to be *orthogonal* and are often denoted by  $\mathbf{Q}$ . Yes, it would be more logical to call such matrices *orthonormal*, but this term is rare. The beautiful thing about an orthogonal matrix is not only does it have an inverse, but the inverse equals the transpose:  $\mathbf{Q}^{-1} = \mathbf{Q}^T$ . The reader should verify that the columns of  $\mathbf{V}$  in (4.34) are orthogonal, but not normal. However, each column vector  $\mathbf{v}$  can

be normalized, resulting in the corresponding cosign vector  $\hat{\mathbf{v}}$ . Having done this in the example for  $C_4$ , we get the orthogonal matrix

$$\mathbf{Q} = \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} & 0 \end{pmatrix},$$

which still satisfies the matrix eigenvector equation (4.33).

Such orthogonal eigenvector matrices are critical in the celebrated spectral theorem:

**Theorem 4.5** (The Spectral Theorem) *A real  $N$  by  $N$  symmetric matrix  $\mathbf{A}$  can be factored into*

$$\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^T = \sum_k^N \mathbf{q}_k \lambda_k \mathbf{q}_k^T,$$

where  $\mathbf{Q}$  is an orthogonal matrix of column eigenvectors  $(\mathbf{q}_j)_{j \in [N]}$  and where  $\Lambda$  diagonal matrix of corresponding real eigenvalues  $(\lambda_i)_{i \in [N]}$ .

*Proof* The proof is routine, but a critical observation is that given the decomposition in the theorem, the matrix  $\mathbf{A}$  can be shown to be symmetric by the calculation

$$\begin{aligned} \mathbf{A}^T &= (\mathbf{Q}\Lambda\mathbf{Q}^T)^T \\ &= \mathbf{Q}^{TT} \Lambda^T \mathbf{Q}^T \\ &= \mathbf{Q}\Lambda\mathbf{Q}^T \\ &= \mathbf{A}. \end{aligned}$$

This uses two properties of the transpose operator that hold for all matrices (of compatible dimension),  $\mathbf{A}^{TT} = \mathbf{A}$  and  $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$ , plus the fact that diagonal matrices are symmetric.  $\square$

The spectral theorem is called “the principal axis theorem” in mechanics and is “principal component analysis” in statistics. In the statistical context, the spectral theorem is applied to a covariance matrix, which has the additional property of having nonnegative eigenvalues, not just real ones. Spectral theory has been generalized to infinite dimensional spaces, is the basis for Fourier analysis, and is a critical tool in quantum mechanics. Finally, in the study of sound, the spectrum corresponds to the fundamental frequency and its overtones, leading to the famous question “Can you hear the shape of a drum?” (Kac, 1966). The question for us is “Can you detect the structure of a network from its spectrum?”

Returning to the problem of finding eigenvectors and eigenvalues, we have noted that the hard part is finding the eigenvalues. To do this, we must express  $\mathbf{v}\lambda$  as a product of the identity matrix times a vector,  $\lambda\mathbf{I}\mathbf{v}$ . Only now is it legitimate to subtract the right-hand side of (4.30) and factor out the  $\mathbf{v}$  to get the equivalent

equation,

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{v} = \mathbf{0}. \quad (4.35)$$

This equation says that the linear combination (given by  $\mathbf{v}$ ) of the columns of the matrix  $(\mathbf{A} - \lambda \mathbf{I})$  is zero. This means that  $(\mathbf{A} - \lambda \mathbf{I})$  is nonsingular. A well-known result from linear algebra tells us that this can happen if and only if its determinant is zero. Recall that the *determinant* of a matrix  $\mathbf{B}$  is defined as

$$\det(\mathbf{B}) = \sum_{\pi \in S_N} \text{sign}(\pi) b_{1,\pi(1)} \cdots b_{N,\pi(N)}, \quad (4.36)$$

where  $S_N$  is the set of all permutations  $\pi$  of the set  $[N]$ , and where  $\text{sign}(\pi)$  is  $+1$  if the permutation  $\pi$  can be expressed as an even number of transpositions  $i \leftrightarrow j$  and otherwise is  $-1$ . If the matrix  $\mathbf{B}$  is 2 by 2, then the determinant is simply  $b_{11}b_{22} - b_{12}b_{21}$ , but for larger matrices the computations are best left to computers. The determinant of  $(\mathbf{A} - \lambda \mathbf{I})$  results in the *characteristic equation*, a polynomial in the variable  $\lambda$

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0,$$

which can be solved without involving the vector  $\mathbf{v}$ . Then, given an eigenvalue  $\lambda$ , equation (4.35) is now linear and can be easily solved for the eigenvector  $\mathbf{v}$  corresponding to  $\lambda$ .

Two important properties of eigenvalues can be deduced from the characteristic equation without even solving it.

**Theorem 4.6** *If a square matrix  $\mathbf{A}$  of size  $N$  has eigenvalues  $\lambda_1, \dots, \lambda_N$ , then*

$$\det \mathbf{A} = \prod_i \lambda_i,$$

$$\text{tr} \mathbf{A} = \sum_i \lambda_i.$$

*Proof* To show that the determinant equals the product of the eigenvalues, factor the characteristic polynomial into linear factors of its roots:

$$\det(\mathbf{A} - \lambda \mathbf{I}) = \prod_i (\lambda_i - \lambda),$$

which we can do because the fundamental theorem of algebra guarantees that any polynomial of degree  $N$  can be factored into  $N$  (possibly complex) linear factors. Then let  $\lambda = 0$ .

To show that the trace equals the sum of the eigenvalues, consider the coefficient of  $(-\lambda)^{N-1}$  in the equation above. On the right-hand side, the coefficient is  $\sum_i \lambda_i$ . On the left-hand side, the polynomial  $\det(\mathbf{A} - \lambda \mathbf{I})$ , a term which includes an off-diagonal element  $a_{ij}$  must also exclude  $a_{ii} - \lambda$  and  $a_{jj} - \lambda$ . Therefore, such a term does not have enough factors of  $\lambda$  to be a coefficient of  $(-\lambda)^{N-1}$ , which implies that the coefficient of  $(-\lambda)^{N-1}$  must come from the main diagonal and it is  $\sum_i a_{ii} = \sum_i \lambda_i$ .  $\square$

Returning to our example of the adjacency matrix for  $C_4$ , we see that the characteristic equation is

$$\det \begin{pmatrix} -\lambda & 1 & 0 & 1 \\ 1 & -\lambda & 1 & 0 \\ 0 & 1 & -\lambda & 1 \\ 1 & 0 & 1 & -\lambda \end{pmatrix} = \lambda^4 - 4\lambda^2 = \lambda^2(\lambda + 2)(\lambda - 2) = 0. \quad (4.37)$$

The four solutions to (4.37) are  $\lambda = -2, 2, 0, 0$ . Now it is a simple linear problem to find the eigenvectors in the columns of the matrix  $\mathbf{V}$  in (4.34). Note that when the characteristic equation is factored into (possibly complex) linear factors, the number of factors with the same eigenvalue is said to be the *algebraic multiplicity* of that eigenvalue. In (4.37), for example, the eigenvalue 0 has algebraic multiplicity 2. In another example, the  $N$  by  $N$  identity matrix has exactly one eigenvalue, 1, which is of algebraic multiplicity  $N$ .

For nonsymmetric matrices, such as for the adjacency matrices of digraphs, the eigenvalues may be complex and the spectral theorem does not apply. However, if we assume that the  $N$  eigenvectors in the columns of  $\mathbf{V}$  are independent, there exists an inverse  $\mathbf{V}^{-1}$ , so that we can multiply equation (4.33) on both the right and the left by  $\mathbf{V}^{-1}$ , resulting in

$$\mathbf{V}^{-1}\mathbf{A} = \mathbf{\Lambda}\mathbf{V}^{-1}.$$

This implies that the *rows* of  $\mathbf{V}^{-1}$  are the *left* eigenvectors for the corresponding eigenvalues in  $\mathbf{\Lambda}$ , showing that even though they may differ, there is an interesting relation between the left and right eigenvectors. It also confirms, for the case of a complete set of eigenvectors, that their eigenvalues are identical.

An even more important implication of a complete set of  $N$  eigenvalues for a matrix  $\mathbf{A}$  is that this makes the matrix *diagonalizable*. That is, multiplying both sides of (4.33) on the right by  $\mathbf{V}^{-1}$ , results in the equation,

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1} = \mathbf{v}_1\lambda_1\mathbf{v}_1^{-1} + \mathbf{v}_2\lambda_2\mathbf{v}_2^{-1} + \cdots + \mathbf{v}_N\lambda_N\mathbf{v}_N^{-1}, \quad (4.38)$$

and we say that  $\mathbf{A}$  is *similar* to the diagonal matrix  $\mathbf{\Lambda}$ . An important use of (4.38) is for the computation of matrix powers, where for all integers  $k$ , we have

$$\mathbf{A}^k = \mathbf{V}\mathbf{\Lambda}^k\mathbf{V}^{-1}. \quad (4.39)$$

As a further simplification, note that  $\mathbf{\Lambda}^k$  is a diagonal matrix with diagonal entries  $\lambda_1^k, \dots, \lambda_N^k$ . To prove formula (4.39) by induction for any positive integer  $k$ , the base case of  $k = 1$  is easy since  $\mathbf{A}^1 \stackrel{\text{def}}{=} \mathbf{A}$ . The induction step assumes that (4.39) holds for  $k$  and seeks to prove it for  $k + 1$ , which is done as follows:

$$\mathbf{A}^{k+1} = \mathbf{A}^k\mathbf{A} = \mathbf{V}\mathbf{\Lambda}^k\mathbf{V}^{-1}\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1} = \mathbf{V}\mathbf{\Lambda}^{k+1}\mathbf{V}^{-1}, \quad (4.40)$$

where  $\mathbf{V}$  and its inverse cancel out. This result is then extended to all integers by defining for  $k \in \mathbb{N}$ ,  $\mathbf{A}^{-k} \stackrel{\text{def}}{=} (\mathbf{A}^{-1})^k$  and noting that  $\mathbf{A}^0 = \mathbf{A}^1\mathbf{A}^{-1} = \mathbf{I}_N$ . An

interesting special case is when  $k = -1$ , which implies that for every (eigenvalue, eigenvalue) pair  $(\lambda, \mathbf{v})$  for  $\mathbf{A}$ , then  $(1/\lambda, \mathbf{v})$  is an eigen pair for  $\mathbf{A}^{-1}$ . That is, the eigenvalues for  $\mathbf{A}^{-1}$  are the reciprocals of the eigenvalues for  $\mathbf{A}$  while the eigenvectors remain invariant. Finally, it was noted that the  $i, j$ th entry  $\mathbf{A}^k$  is the number of walks of length  $k$  from  $i$  to  $j$ . The eigen expression in (4.40) is an efficient way to compute the number of such walks, and gives an important example of what the spectrum tells us about the structure of a graph.

An important kind of matrix in the social and behavioral sciences is a matrix with nonnegative values, since in many cases it is impossible for negative numbers to occur. The Perron–Frobenius theory is concerned with exploring the existence of a unique positive eigenvector for such nonnegative matrices. We shall examine a special case of this theory and its implications for networks. A nonnegative matrix  $\mathbf{A}$  is said to be *primitive* if there is a natural number  $k$  such that  $\mathbf{A}^k$  has all positive entries.

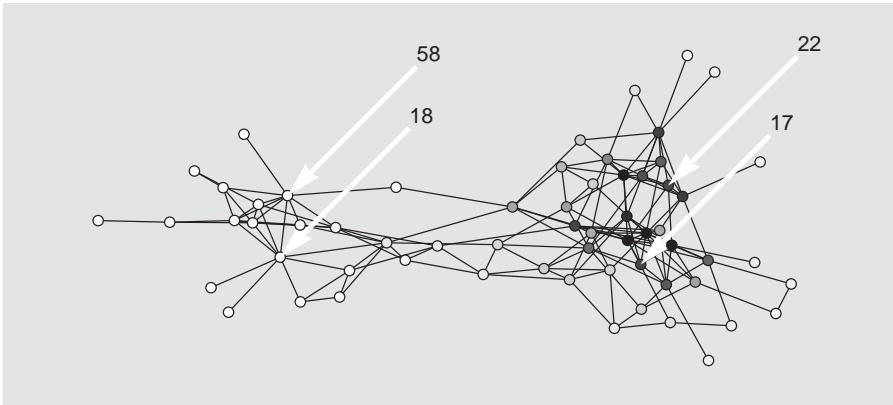
**Theorem 4.7** (Perron–Frobenius for primitive matrices) *Let  $\mathbf{A}$  be a nonnegative, primitive matrix. Then  $\mathbf{A}$  has a real eigenvalue  $\lambda_1$  and a corresponding real eigenvector  $\mathbf{v}_1$  with the following properties:*

1.  $\lambda_1 > 0$ .
2. *If  $\lambda$  is any other eigenvalue of  $\mathbf{A}$ , then  $\lambda_1 > |\lambda|$ .*
3.  $\lambda_1$  has algebraic multiplicity 1.
4.  $\mathbf{v}_1$  has only positive real values.
5. *All other eigenvectors that are not scalar multiples of  $\mathbf{v}_1$  must have at least one negative or complex value.*

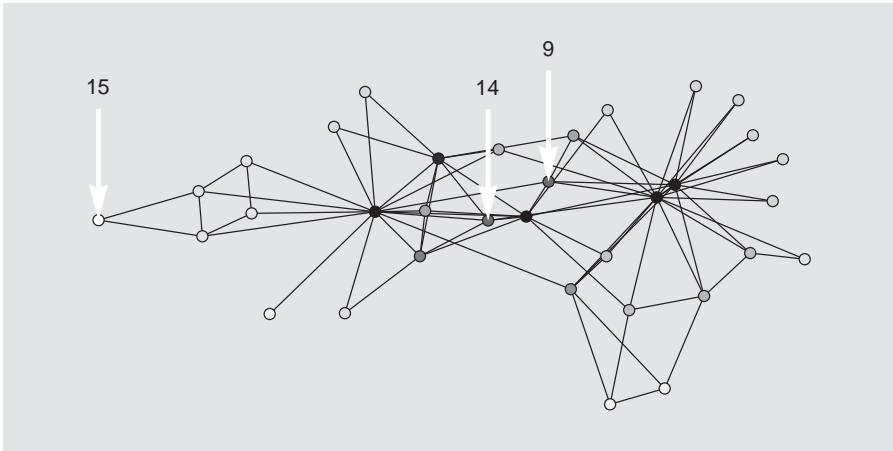
*Proof* See, for example, Minc (1985). □

The eigenvalue  $\lambda_1$  in Theorem 4.7 is sometimes called the *Perron eigenvalue* for  $\mathbf{A}$  or, more simply, its *largest*, or *dominant* eigenvalue, while its eigenvector is called the *Perron vector* for  $\mathbf{A}$ . The algebraic multiplicity being 1 implies that the Perron vector is unique up to a scalar multiple.

Suppose one wanted to develop a measure of “centrality” for nodes in a network that captures the intuitive notion that some people are “central” to a social network while others are “marginal.” In one of the most influential papers in networks (cited 7258 in Google Scholar as of October 22, 2014), Freeman (1979) discusses three concepts of centrality; the simplest is called *degree centrality*, which is merely the degree of each node. The problem with this measure is that it does not distinguish between two nodes of equal degree, but where one is adjacent to nodes of low degree while the other is adjacent to nodes of high degree. One could easily devise a measure of centrality that would take account of the degrees of nodes in the neighborhood, but then what about the neighborhood of the neighborhood, and so on? If the hypotheses of the Perron–Frobenius Theorem 4.7 is satisfied, as it is in most connected graphs, then the natural limit of this thought process is to use the components of the dominant eigenvector as a measure of centrality, called *eigenvector centrality* (Bonacich, 1987). If  $\mathbf{v}_1$  is such a dominant eigenvector, then



**Figure 4.17** The dolphin network with nodes shaded according to their eigenvalue centrality: darker nodes are more central. Nodes 58 and 18 have high degree but low centrality, while the opposite is true for nodes 22 and 17.



**Figure 4.18** The Zachary karate network with nodes shaded according to their eigenvalue centrality: darker nodes are more central. Node 15 has a higher degree (2) than predicted by its centrality, while nodes 9 and 14 have a lower degree than is predicted by their centralities.

the eigenvector centrality of the graph is  $\mathbf{c} \stackrel{\text{def}}{=} \mathbf{v}/\sum_i v_i$ . That is, the eigenvector centrality is the dominant eigenvector that sums to 1.

As an illustration of eigenvector centrality, Figures 4.17 and 4.18 show the dolphin and Zachary karate networks redrawn with the same spring-electrical embedding as before, but with new shading so that the darkness of each node is proportional to its eigenvector centrality. The correlations between degree and eigenvector centrality are 0.7196 and 0.9173 for the dolphin and Zachary networks, respectively. Of course, one would not want the correlations to be

perfect, because otherwise one would choose the degree centrality because of its simplicity. In the next two paragraphs we shall examine where the two centralities differ and in what sense eigenvector centrality is better.

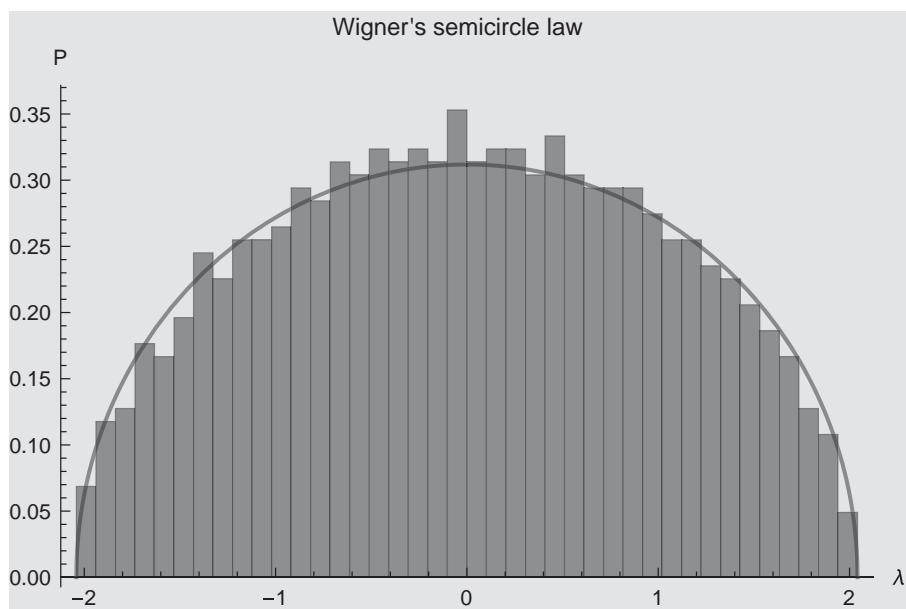
In order to compare the two centralities for the dolphin data, a linear regression gives the equation  $c = -0.00374 + 0.00387d$  as the best predictions for eigenvector centrality given the degrees. The thick white arrows point to nodes where there is most disagreement between the two centrality measures. For example, node 18 in the dolphin graph had a rather high degree of 9, while its eigenvector centrality was only 0.003087, whereas the linear regression predicted that a node of degree 9 would have an eigenvector centrality of 0.03112. As can be seen from Figure 4.17, node 18 is connected to other nodes of low eigenvector centrality, as indicated by their light shading. Node 58 also has a higher degree than predicted by its eigenvector centrality. The biggest deviation in the other direction from the regression equation was found in node 17, which has degree 6, which would predict an eigenvector centrality of 0.01950, whereas its actual eigenvector centrality is almost twice as high at 0.03696. Again returning to Figure 4.17, one can see that although its degree is not that large, it is connected to others of high eigenvector centrality, as indicated by their dark shading. Similar remarks apply to node 22.

Turning now to the Zachary karate data, another linear regression gives a slightly different equation  $c = 0.00965 + 0.00431d$ , which best predicts eigenvector centrality. The white arrow at the left of Figure 4.18 points to node 15 of degree 2, which has the highest residual in the direction of the predicted eigenvector centrality 0.01826, being much higher than the actual value of 0.004748. As with the dolphin data, this discrepancy can be explained by node 15 being connected to two nodes of low eigenvector centrality. Similar remarks apply to the two nodes, 9 and 14, with the highest discrepancy in the other direction.

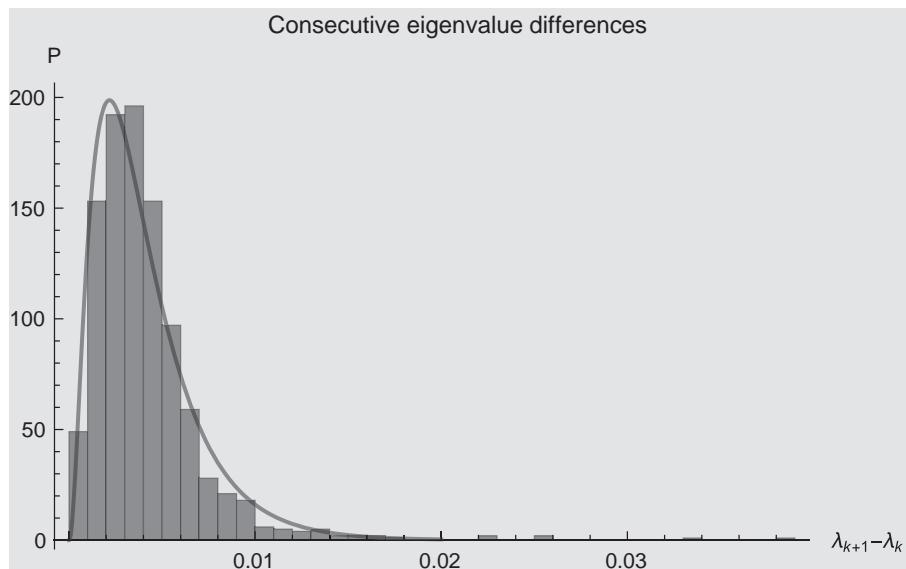
Although the spectrum of a real matrix may contain complex numbers, the spectral theorem 4.5 shows that the spectrum is real for symmetric matrices. The distribution of eigenvalues of random symmetric matrices follows *Wigner's semicircle law* (Wigner, 1958), where the pdf is approximately a semicircle as in Figure 4.19. Another special case is for a random *orthogonal* matrix (Diaconis, 2003), defined as matrix whose inverse equals its transpose. All eigenvalues of an orthogonal matrix are of unit length, distributed rather evenly around the complex circle. The distribution is too even for a Poisson process, which would predict that the distribution of the distances between points would be approximately exponentially distributed (ignoring the fact that the distances on the unit circle cannot be greater than  $\pi$ ). As can be seen in Figure 4.20, the empirical distribution of these distances has a maximum away from zero, suggesting a repulsive force keeping the eigenvalues from being too close, as if they were particles with the same electrostatic charge.

#### 4.2.4.3 Laplacians

In addition to the adjacency matrix  $\mathbf{A}$ , a graph leads to several other matrices whose eigenvalues and vectors reveal critical structural properties. The two most common



**Figure 4.19** A histogram of eigenvalues of a random orthogonal matrix. The continuous line is Wigner's semicircle law.



**Figure 4.20** A plot of the differences between adjacent eigenvalues of a random symmetric adjacency matrix. The plot is not decreasing, indicating that adjacent eigenvalues tend to repel each other.

such matrices are called the *Laplacian* (or *combinatorial Laplacian* or *Kirchhoff*)  $\mathbf{L}$  and the *analytic Laplacian* (or *normalized Laplacian*)  $\mathcal{L}$  (Chung, 1997; Bollobás, 1998). The Laplacian matrix is defined for a graph with  $N$  nodes by

$$l_{ij} \stackrel{\text{def}}{=} \begin{cases} d_i, & \text{if } i = j, \\ -1, & \text{if } i \sim j, \\ 0, & \text{otherwise,} \end{cases} \quad (4.41)$$

where  $0 \leq i, j < N$  and  $d_i$  is the degree of node  $i$ . If  $\mathbf{D}$  is the diagonal matrix of degrees, then the Laplacian matrix can be expressed in terms of the adjacency matrix:  $\mathbf{L} = \mathbf{D} - \mathbf{A}$ . Because the row and column sums of the Kirchhoff matrix are zero, the vector of ones is an eigenvector for the zero eigenvalue:  $\mathbf{L}\mathbf{1} = \mathbf{1}0 = \mathbf{0}$ .

The normalized Laplacian matrix is defined by

$$\mathcal{L} \stackrel{\text{def}}{=} \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2} \quad (4.42)$$

$$= \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}, \quad (4.43)$$

where  $\mathbf{D}^{-1/2}$  is the diagonal matrix of one over the square root of the degrees, with the convention that if  $v$  is an isolated node, then  $\mathbf{D}^{-1/2}$  is zero for the  $v$ th diagonal entry.

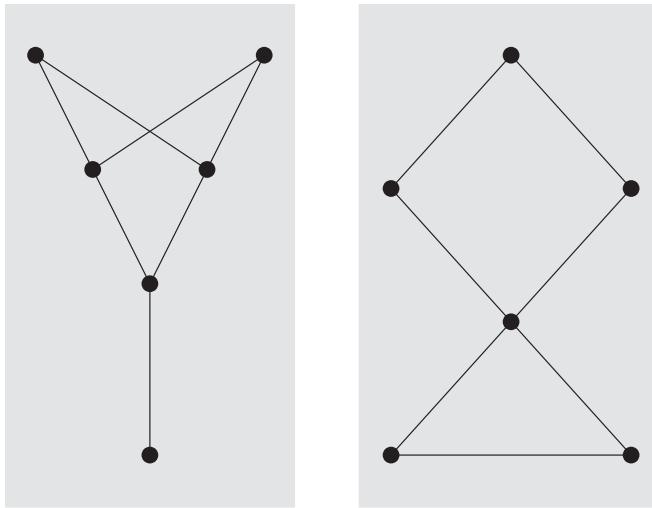
One significant property of the Laplacian of a graph  $G = (V, E)$  is that all of its eigenvalues are nonnegative. This is true for all matrices of the form  $\mathbf{S}\mathbf{S}^T$ , where  $\mathbf{S}$  is any real matrix. Such matrices are said to be *positive semidefinite*. One way to show that the Laplacian is positive semidefinite is to “orient” the edges of the graph  $G$ . To be precise, an *orientation* of a graph  $G = (V, E)$  is a function  $h : E \rightarrow V$  such that for all edges  $h(e) \in e$ . That is,  $h$  is a *choice function* that arbitrarily picks out one element of an edge, which we call its *head* (the other one being the *tail*, denoted by  $t(e)$ ). Obviously, there are  $2^{|E|}$  possible orientations of the graph. Next, an *orientated incidence matrix* of a graph  $G = (V, E)$  with orientation  $h$  is the  $|V|$  by  $|E|$  matrix  $\mathbf{S}$  defined by

$$s_{v,e} = \begin{cases} +1, & \text{if } v = h(e), \\ -1, & \text{if } v = t(e), \\ 0, & \text{otherwise.} \end{cases}$$

Each column of  $S$  has exactly two nonzero entries, summing to zero, while the sum of the absolute values in row  $v$  is the degree of  $v$ . The eigenvalues of the Laplacian are usually listed in nondecreasing order beginning with the zero eigenvalue:  $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$ .

Next, we list the Laplacian eigenvalues for some named graphs.

- For the complete graph  $K_N$ , the Laplacian eigenvalues are 0 and  $N$  (with multiplicity  $N - 1$ ).
- For the complete bipartite graph  $K_{M,N}$ , they are  $M$  (with multiplicity  $N - 1$ ),  $N$  (with multiplicity  $M - 1$ ), and  $M + N$ .
- For the star  $S_N$  on  $N + 1$  nodes, they are 0, 1 (with multiplicity  $N - 1$ ), and  $N + 1$ .



**Figure 4.21** The pair of Laplacian cospectral graphs with the fewest nodes.

- For the cycle  $C_N$ , they are  $2 - 2\cos \frac{2\pi k}{N}$  for  $k = 0, 1, \dots, N - 1$ .
- For the path  $P_N$ , they are  $2 - 2\cos \frac{\pi k}{N}$  for  $k = 0, 1, \dots, N - 1$ .

It is possible for two nonisomorphic graphs to have the same Laplacian eigenvalues. Such pairs are said to be (*Laplacian-*) *cospectral*. Figure 4.21 depicts a minimal example of two cospectral graphs (Cvetković *et al.*, 2010): both have the spectrum  $(0, 2, 3, 3, 3 - \sqrt{5}, 3 + \sqrt{5})$ .

A few basic facts about the Laplacian spectrum are listed in the next theorem.

**Theorem 4.8** *Let  $G$  be a graph with  $N$  nodes,  $M$  edges, and Laplacian eigenvalues  $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$ . Then the following statements hold:*

1.  $\sum_i \lambda_i = 2M$ .
2. Except for  $K_N$ , the second smallest eigenvalue satisfies  $\lambda_2 \leq \kappa$ , where  $\kappa$  is the connectivity of  $G$ .
3. If  $G$  has  $k$  connected components, then  $0 = \lambda_k < \lambda_{k+1}$ . I.e., it has exactly  $k$  zero eigenvalues. (This is not true of the eigenvalues of the adjacency matrix.)
4. The spectrum of a graph is equal to the disjoint union of the spectra of its individual connected components.
5. The largest eigenvalue equals  $N$  for  $G = K_N$ , otherwise, it is strictly less than  $N$ .

*Proof* The first result follows from Theorem 4.4: the diagonal entries of the Laplacian are the degrees, which add up to twice the number of edges.

For the other results, see Bollobás (1998). □

Recall that for the adjacency matrix, the largest eigenvalue and its vector are the focus of attention, while the smaller eigenvalues are progressively less important.

For the Laplacian matrix, however, the situation is essentially reversed because of the negative values in the off-diagonal entries. As mentioned before, if all nodes of a graph are of degree  $k$ , then the respective eigenvalues satisfy  $\lambda(\mathbf{L}) = k - \lambda(\mathbf{A})$ . The first Laplacian eigenvector is equal to any nonzero constant  $N$ -vector, such as  $\mathbf{1}$ . Therefore, the first interesting Laplacian eigenvalue and eigenvector pair is the second one,  $\lambda_2$  and  $\mathbf{v}_2$ . This second eigenvalue is called the *algebraic connectivity* and its eigenvector is called the *Fiedler vector* of the graph (Büyükoğlu *et al.*, 2007).

Looking at the examples (4.2.4) the algebraic connectivity ( $\lambda_2$ ) decreases from first to last in the list. For the complete graph  $K_N$ , it is  $N$ ; for the complete bipartite graph  $K_{M,N}$ , it is  $\min(M, N)$ ; for the star  $S_N$ , it is 1; for the cycle  $C_N$ , it is  $2(1 - \cos\frac{2\pi}{N})$ ; and for the path  $P_N$ , it is  $2(1 - \cos\frac{\pi}{N})$ . The last two are not obvious at first glance, but it is easy to show that  $\lambda(C_N)_2$  is greater than  $\lambda(P_N)_2$  for all  $N \geq 3$ . The sequence of the ratios,  $\lambda(C_N)_2/\lambda(P_N)_2 = 2(1 + \cos\frac{\pi}{N})$ , increases monotonically to 4 as  $N \rightarrow \infty$ . That is, on the same number of nodes, the algebraic connectivity of cycles approaches 4 times that of paths. In fact, for  $N = 3$ , the ratio already equals 3. Note, however, that the algebraic connectivity for both cycles and paths converges to zero as  $N$  increases.

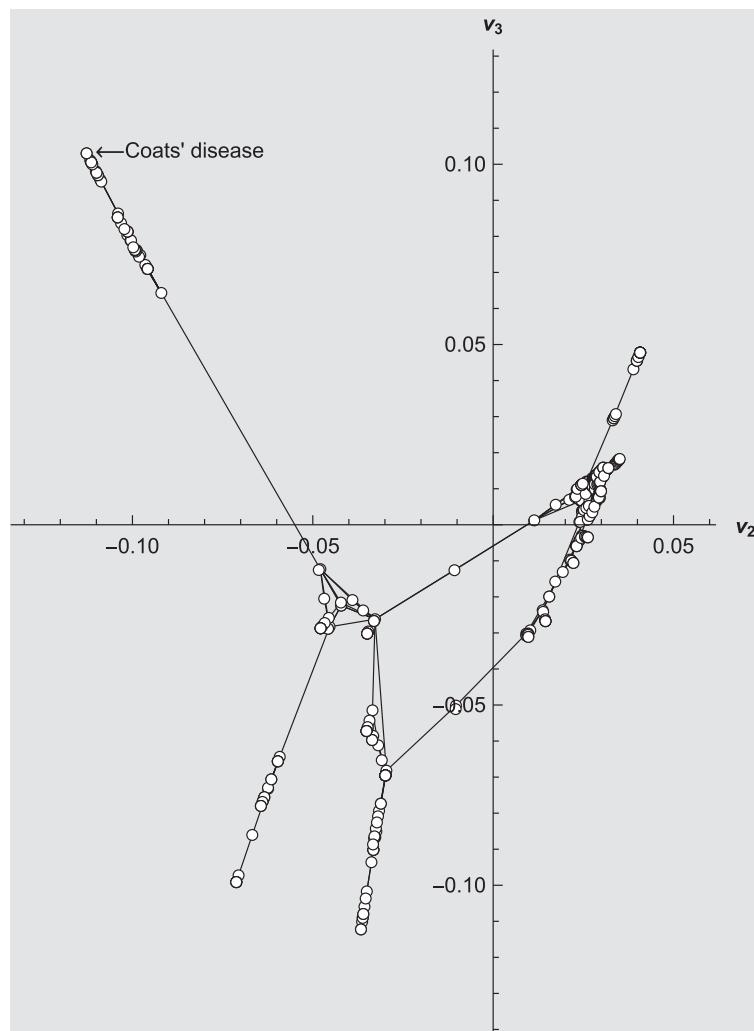
The eigenvectors of Laplacians are used in graph embeddings. These embeddings have a better theoretical underpinning than those based on “springs” and “electrical charges.” It can be shown (Bollobás, 1998) that the second Laplacian eigenvalue and its eigenvector satisfy the equation

$$\lambda_2 = \inf_{\mathbf{v} \perp \mathbf{1}} \frac{\sum_{i \sim j} (v_i - v_j)^2}{\sum_i v_i^2}. \quad (4.44)$$

In other words, the second smallest eigenvalue of the Laplacian is the minimum value of the *Rayleigh quotient* in (4.44) over all vectors  $\mathbf{v} \perp \mathbf{1}$ . That is, such that  $\mathbf{v}$  sums to zero. Similarly, the third eigenvector and value satisfy an equation similar to (4.44) except that the vector must also be perpendicular to  $\mathbf{v}_2$ .

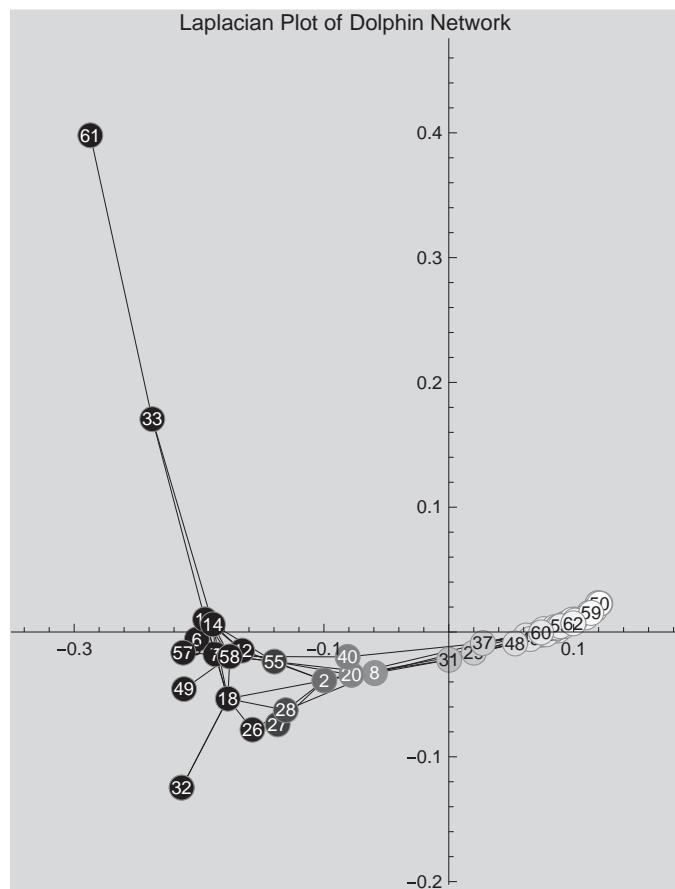
The Laplacian eigenvectors,  $\mathbf{v}_2$  (the Fiedler vector) and  $\mathbf{v}_3$ , lead to a more principled graph embedding in that the energy in (4.44) is minimized. This embedding can be seen in Figures 4.22, 4.23, and 4.24. Some of the nodes are placed far away from the others while some are on top of one another. While this may make the graph less aesthetically attractive than the spring/electrical embedding, it may more accurately reflect the relationships among the nodes.

One measure of difference between the Fiedler vector and the dominant eigenvector of the adjacency matrix is that their correlation for the dolphin data is only 0.121. The Fiedler vector also has a low correlation of 0.181 with the degrees, while the dominant eigenvector is highly correlated, 0.720, with them. This difference is illustrated in Figures 4.23 and 4.17. Another way to illuminate the meanings of these two vectors is with matrix plots, where these vectors determine the ordering of the rows and columns. Figure 4.25 orders the dolphin matrix by eigenvector centrality: note that this tends to place the 1s, indicated by black squares, close to the diagonal. The diagonal elements have been shaded by node degree: note



**Figure 4.22** A plot of the eigenvectors,  $\mathbf{v}_2$  and  $\mathbf{v}_3$ , corresponding to the smallest two eigenvalues of the Laplacian of the disease graph. For example, Coats' disease has the coordinates  $(-0.113, 0.103)$  corresponding to its components on  $\mathbf{v}_2$  and  $\mathbf{v}_3$ .

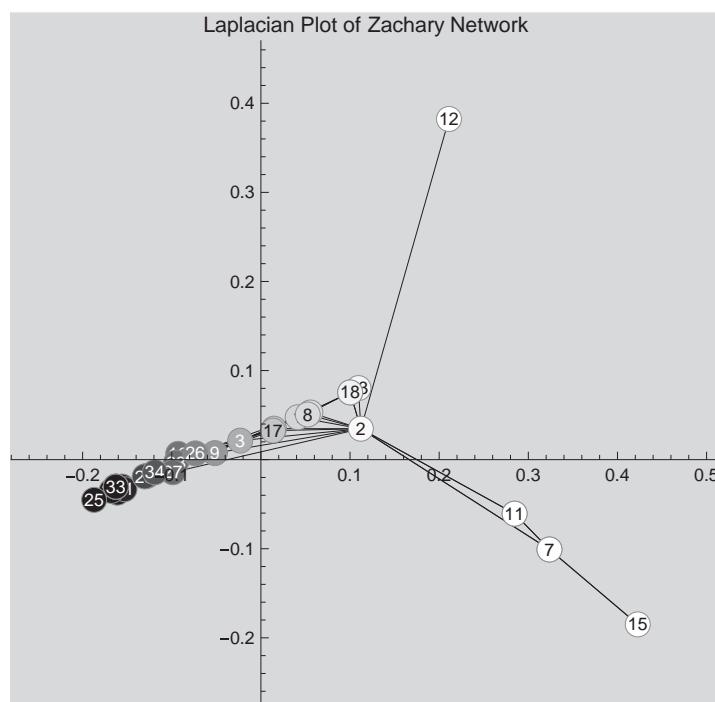
that the high-degree nodes tend to come first. On the other hand, the Laplacian-based Fiedler vector shown in the matrix plot 4.26 places the two nodes of highest degree at opposite ends of the vector, as indicated by the concentration of edges at the upper right and lower left of the matrix, far from the diagonal. This means that nodes have a similar value on the Fiedler vector not just because they are connected, but also because they are connected to the same other nodes. The reason that the Fiedler indicates more structural relationships than does the centrality vector is that the effect of node degree is largely eliminated.



**Figure 4.23** A plot of the eigenvectors corresponding to the smallest two eigenvalues of the Laplacian of the dolphin graph.

### 4.3 Probabilistic graph models

This section discusses probabilistic models for graphs and digraphs. So far we have seen the Bernoulli model for graphs and other models of degree sequences, but in this section we go well beyond these simple models. There is a vast literature in this area, and developments in the last few decades have greatly increased our knowledge. This section will not attempt to survey this large area; however, in the concluding section of the chapter several references will be provided that provide a more complete access. Probabilistic graph models are characterized by various approaches for dealing with the complex statistical dependencies that may exist among the random variables representing the presence or absence of various edges and arcs in graphs and digraphs. One general approach is to start with a particular specification of these dependencies and then use it to build up a parametric family of joint distributions over all the edge or arc random variables.

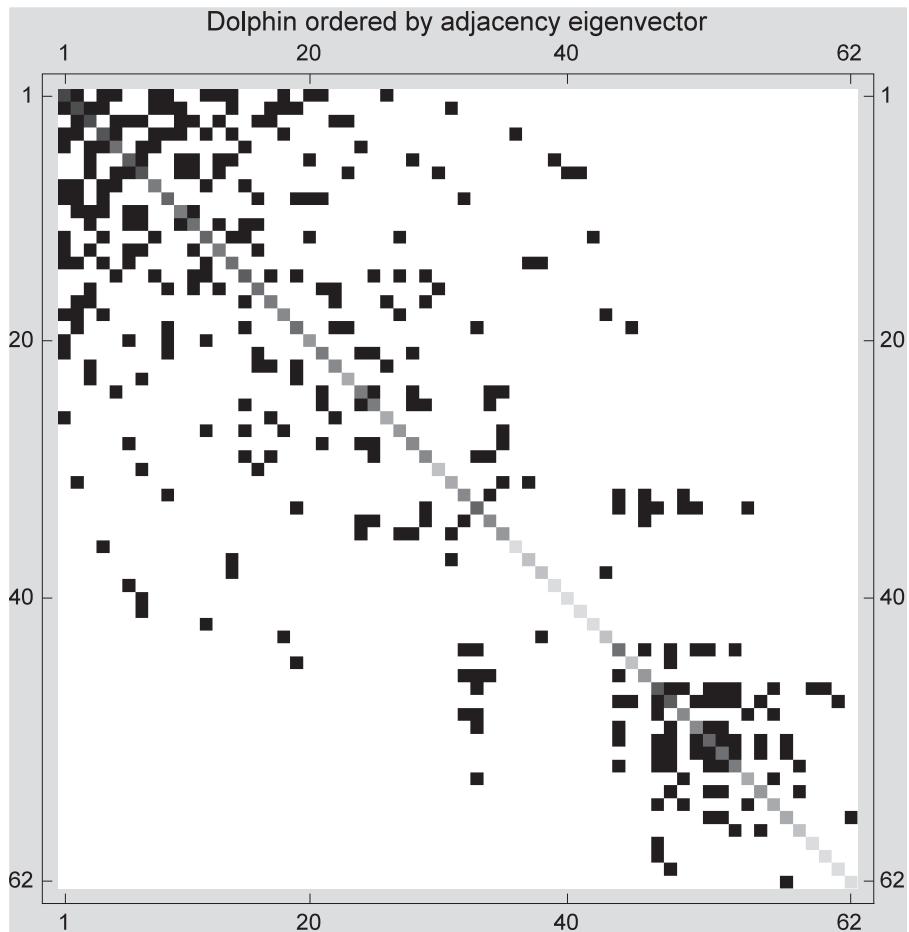


**Figure 4.24** A plot of the eigenvectors corresponding to the smallest two eigenvalues of the Laplacian of the karate graph.

This approach utilizes the so-called Hammersley–Clifford (H–C) theorem (e.g., Besag, 1974), and it leads to a variety of probabilistic graphical models including versions of spatial (rather than temporal) Markov models. This approach to dealing with dependencies between random variables has not been seen in other areas of modeling in the cognitive sciences, so part of this section will focus in detail on it.

Throughout, we assume a fixed finite set of nodes  $V = \{v_i \mid i \in [N]\}$ . It is easy to see that there are  $N(N - 1)/2$  possible edges and  $N(N - 1)$  possible arcs or directed edges from  $V$ . Each subset  $E$  of possible edges can be associated with a graph  $G = (V, E)$ , and each subset  $A$  of arcs can be associated with a digraph  $D = (V, A)$ . Thus there are  $2^{N(N-1)/2}$  distinct graphs and  $2^{N(N-1)}$  distinct digraphs on  $V$ . A *probabilistic graph model* (PGM) or a *probabilistic digraph model* (PDM) specifies, respectively, a family of one or more probability distributions on all possible graphs or digraphs on  $V$ . In the case that a PGM or PDM specifies only a single probability distribution, we will refer to the model, respectively, as a *probabilistic graph* (PG) or a *probabilistic digraph* (PD).

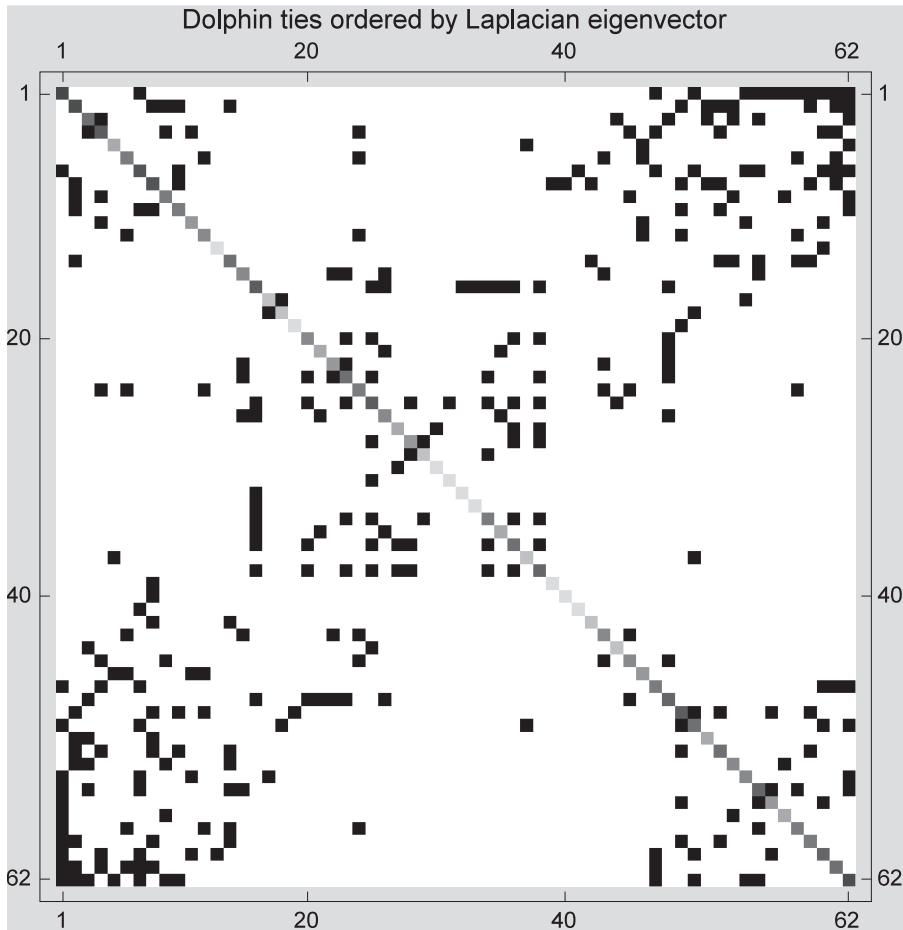
To describe probability distributions on graphs or digraphs, it is convenient to associate random variables with each possible edge or arc from  $V$ . A PGM associates a 0–1 valued random variable  $X_{ij}$  with each of the  $N(N - 1)/2$  possible edges. For example, the possible edge  $\{v_i, v_j\}$  is represented by a random variable



**Figure 4.25** A matrix plot of the dolphin friendship data, ordered by centrality, the dominant eigenvector of the adjacency matrix. The diagonals are shaded according the node degree, while a black off-diagonal entry indicates adjacency.

$X_{ij}$  with possible realizations  $R_{ij} = \{0, 1\}$ , where  $X_{ij} = 1$  means an edge  $\{v_i, v_j\}$  is present in the graph and  $X_{ij} = 0$  means that it is absent. Recall that edges are between nodes and are nondirected, so that for all  $i, j \in [N]$  we have  $X_{ij} = X_{ji}$ , which motivates the convention that we only need to consider edge random variables  $X_{ij}$  with  $i < j$ . Consequently, let  $\mathbf{X} = (X_{ij})_{i,j \in [N], i < j}$ , with possible realizations  $\Omega_{\mathbf{X}} \stackrel{\text{def}}{=} \{0, 1\}^{N(N-1)/2}$ , be the complete collection of random variables for a PGM on the node set  $V$ . The collection  $\mathbf{X}$  is called *complete* because its set of possible realizations,  $\mathbf{x} \in \Omega_{\mathbf{X}}$ , is in one-to-one correspondence with the set of possible graphs on  $V$ , under the correspondence  $\leftrightarrow$  defined for all  $\mathbf{x} \in \Omega_{\mathbf{X}}$  by

$$G(\mathbf{x}) \leftrightarrow (V, E = \{\{v_i, v_j\} \mid x_{ij} = 1\}).$$



**Figure 4.26** A matrix plot of the dolphin friendship data. The rows and columns are ordered by the vector corresponding to the second smallest Laplacian eigenvector, the Fiedler vector. The diagonals are shaded according to the node degree (blacker is larger), while a black off-diagonal entry means the nodes are adjacent.

In the case of a PDM, there is a complete collection of random variables,  $\mathbf{Y} = (Y_{ij})_{i,j \in [N], i \neq j}$ , one for each possible arc in the digraph. As before,  $Y_{ij} = 1$  if there is an arc from  $v_i$  to  $v_j$ , and otherwise  $Y_{ij} = 0$ . It is clear that the set of possible realizations for  $\mathbf{Y}$ , denoted by  $\Omega_{\mathbf{Y}} = \{0, 1\}^{N(N-1)}$ , has a one-to-one correspondence  $\leftrightarrow$  with the set of all digraphs on  $V$ , defined by

$$D(\mathbf{y}) \leftrightarrow (V, D = \{(v_i, v_j) \mid y_{ij} = 1\}).$$

A PG is specified by a probability distribution  $P(\mathbf{x})$  on  $\Omega_{\mathbf{X}}$ , where of course for all  $\mathbf{x} \in \Omega_{\mathbf{X}}$  we have  $P(\mathbf{x}) \geq 0$  and  $\sum_{\mathbf{x} \in \Omega_{\mathbf{X}}} P(\mathbf{x}) = 1$ . Similarly, a PD is specified by a probability distribution  $P(\mathbf{y})$  on  $\Omega_{\mathbf{Y}}$ . Probabilistic graph models (PGMs) and

probabilistic digraph models (PDGs) are, respectively, collections of PGs or PDs. These collections are usually specified in terms of a vector of  $S$  latent parameters denoted by  $\Theta = (\theta_s)_{s \in [S]}$ , with parameter space  $\Psi_\Theta \subseteq \mathbb{R}^S$ . Then corresponding to each  $\Theta \in \Psi_\Theta$ , there is a PG or PD given, respectively, by  $P(\mathbf{x} | \Theta)$  or  $P(\mathbf{y} | \Theta)$ .

The situation described above has the character of most parametric probability models where the data are described by a series of random variables,  $\mathbf{X} = (X_j)_{j \in [J]}$ , that may exhibit a dependency structure. The key to many parametric models in such a situation is to pick the latent parameters in such a way that the data random variables are *conditionally independent* (or *exchangeable*) given the parameters,

$$P(\mathbf{X} = \mathbf{x} | \Theta) = \prod_{j \in [J]} P(x_j | \Theta). \quad (4.45)$$

However, in the case of graphical models, apart from several very simple ones, this approach has not proven to be effective. The reason is that the dependencies between the edge or arc random variables may be complex leading to the need to decompose the joint edge or arc random variables into conditionally independent subsets of random variables rather than individual random variables, e.g., for a PGM such a subset would look like  $P((x_{ij})_{\{i,j\} \in A} | \Theta)$  for some edge set  $A \subseteq E$ .

It turns out that there is an approach to specifying graph models without using conditional independence in the form of equation (4.45) that arises in the area of spatial statistics. This approach is based on the H-C theorem mentioned earlier. The approach allows modelers to start with a specification of the sorts of dependencies between the random variables that they want the model to allow, and then from that specification alone a family of parametric graph models can be derived. The remainder of this section will consist of several subsections. First, in 4.3.1 the H-C theorem will be described in a form applicable to graphical models. It is hoped that the reader who follows the technical details in the subsection may consider the possibility of using this approach to construct models in other cognitive paradigms such as list-memory experiments, categorization tasks, and choice and decision-making settings. In the second subsection 4.3.2, it will be shown how to adapt the theorem to formulate probabilistic graph models. The result will lead to the so-called class of *exponential random graph models* (ERGMs) in subsection 4.3.3. Examples will be provided including ones based on Markovian assumptions. In the final subsection 4.3.4, we will discuss another approach to constructing probabilistic graphical models based on selecting a set of graph statistics, and making all graphs (or digraphs) that have the same values of these statistics equiprobable.

### 4.3.1 The Hammersley–Clifford theorem

In this subsection, we develop a way to represent the dependency structure in a collection of random variables. Once we establish this representation, then we can describe the possible joint probability distributions on the random variables. In doing this, we will make use of the so-called Hammersley–Clifford (H–C) theorem. This theorem uses the dependency structure among a set of random variables

to delimit their possible joint probability distributions. There are a number of applications of the H-C theorem in the statistical sciences, notably in spatial statistics, where the random variables are associated with locations in a metric space. Further, there are versions of the theorem that are applicable to general discrete and continuous random variables. However, our application in this subsection will be to PGMs and PDMs, so for that reason, we restrict ourselves to the special case of the H-C theorem for 0–1 valued random variables.

There are other papers that purport to provide proofs of related aspects of the H-C theorem (Besag, 1974), and others that provide the consequences of the H-C theorem for graph models while accepting the basic theory as given (Frank and Strauss, 1986; Koskinen *et al.*, 2008). However, in our view these papers often omit details that can confuse readers without advanced statistical and mathematical knowledge. So our goal in this subsection is not to cut corners, and instead lay out the H-C theorem as applied to 0–1-valued random variables in complete formal detail.

We start with a collection of random variables  $\mathbf{X} = (X_j)_{j \in [J]}$ , and when this will not introduce confusion we will refer to them by their subscript indices in  $[J]$ . We work with the case where each random variable is 0–1-valued with space (set of possible values)  $\{0, 1\}$ . Thus the space of  $\mathbf{X}$  (the set of possible realizations of  $X$ ) is  $\Omega_{\mathbf{X}} = \{0, 1\}^J$  with  $2^J$  possible values, and we denote the joint probability distribution of  $\mathbf{X}$  by  $P(\mathbf{x})$ . Throughout, we will assume that all possible realizations of  $X$  have positive probability, that is, for all  $\mathbf{x} \in \Omega_{\mathbf{X}}$ , we have  $P(\mathbf{x}) > 0$ .

There are three major steps that are needed to develop the H-C theorem under the assumptions above:

1. provide a special representation for arbitrary probability distributions on  $\mathbf{X}$  based on summing terms associated with the random variables in various subsets of  $[J]$ ,
2. develop the dependency structure of  $\mathbf{X}$  based on properties of the conditional distributions arising from  $P(\mathbf{x})$ , and
3. use properties of the dependency structure of  $\mathbf{X}$  to further simplify the representation in step 1.

The rest of this subsection will flesh out these steps in detail.

#### 4.3.1.1 Step 1: Special representation of probability distributions on $\mathbf{X}$

Let  $P(\mathbf{x})$  be an arbitrary joint probability distribution on  $\mathbf{X}$ , subject to the condition above and define for all  $\mathbf{x} \in \Omega_{\mathbf{X}}$

$$Q(\mathbf{x}) \stackrel{\text{def}}{=} \log\{P(\mathbf{x})/P(\mathbf{0})\}, \quad (4.46)$$

which is well-defined because  $P(\mathbf{x}) > 0$  for all  $\mathbf{x} \in \Omega_{\mathbf{X}}$ . Clearly, knowledge of  $Q(\mathbf{x})$  leads to a unique  $P(\mathbf{x})$  and vice versa, as can be noted by computing for all  $\mathbf{x} \in \Omega_{\mathbf{X}}$

$$P(\mathbf{x}) = P(\mathbf{0}) \exp\{Q(\mathbf{x})\}, \quad (4.47)$$

and noting of course that  $P(\mathbf{0}) = 1 - \sum_{\mathbf{x} \in \Omega_{\mathbf{X}} \setminus \mathbf{0}} P(\mathbf{x})$ .

Before preceding, some notation is needed concerning the joint random variables associated with subsets of  $[J]$ . Consequently, define for every  $A \subseteq [J]$  and  $\mathbf{x} \in \Omega_{\mathbf{X}}$  the vector  $\mathbf{x}_A \stackrel{\text{def}}{=} (x_i)_{i \in A}$  and let  $A^c \stackrel{\text{def}}{=} [J] \setminus A$  denote *complementation*. Thus, for any  $A \subseteq [J]$ , we can represent the realization  $\mathbf{x} \in \Omega_{\mathbf{X}}$  by permuting the random variables into the form  $\mathbf{x}^* = (\mathbf{x}_A, \mathbf{x}_{A^c})$ , where the sequence of subscripts of the random variables in  $\mathbf{x}$  will be a reordering of the variables so that the ones in  $A$  come first. Because the random variables in  $\mathbf{x}$  and  $\mathbf{x}^*$  retain their subscripts, it is always the case that  $P(\mathbf{x}) = P(\mathbf{x}^*)$ , so hereafter we will use  $P(\mathbf{x})$  to stand for any division of the random variables indexed by  $[J]$ . The main result of Step 1 is given next.

**Theorem 4.9** *Let the random vector  $\mathbf{X}$  with probability distribution  $P(\mathbf{x})$  be subject to the conditions described above. Then for each  $A \subseteq [J]$ , there is a real number  $\eta_A$  such that for all  $\mathbf{x} \in \Omega_{\mathbf{X}}$*

$$Q(\mathbf{x}) = \sum_{A \subseteq [J]} \eta_A \prod_{i \in A} x_i. \quad (4.48)$$

*Proof* First note that for  $\mathbf{x} = \mathbf{0}$ , definition (4.48) shows that  $Q(\mathbf{0}) = \log 1 = 0$ , which is what (4.48) computes if  $\eta_{\emptyset} = 0$ . Second, consider any  $\mathbf{x}$  with a realization such that  $x_i = 1$  for exactly one  $i \in [J]$ , while  $x_j = 0$  for the rest of the indices. That is,  $\mathbf{x}_{\{i\}^c} = \mathbf{0}$ , where  $\mathbf{0}$  stands for a vector of zeros of the appropriate dimension, in this case  $J - 1$ . From (4.48) it is easily seen that for such choices of  $\mathbf{x}$ , the equation

$$Q(x_i, \mathbf{0}) = \eta_{\{i\}} \quad (4.49)$$

satisfies (4.48), where of course in terms of the original probability distribution,  $\eta_{\{i\}} = \log P(x_i, \mathbf{0})/P(\mathbf{0})$ .

Next, suppose  $\mathbf{x}$  has two distinct random variables with realizations of 1,  $x_i = x_j = 1$ , and the rest set to zero,  $\mathbf{x}_{\{i,j\}^c} = \mathbf{0}$ . Then (4.48) requires that such an  $\mathbf{x}$  has the form

$$Q(\mathbf{x}) = \eta_{\{i\}} + \eta_{\{j\}} + \eta_{\{i,j\}} = Q(x_i, x_j, \mathbf{0}). \quad (4.50)$$

From (4.50) and the already computed values  $\eta_{\{i\}}$  and  $\eta_{\{j\}}$  in (4.49), we can solve for  $\eta_{\{i,j\}}$  yielding

$$\eta_{\{i,j\}} = Q(x_i, x_j, \mathbf{0}) - \eta_{\{i\}} - \eta_{\{j\}}. \quad (4.51)$$

The ideas leading to (4.51) can be applied to any  $\mathbf{x} \in \Omega_{\mathbf{X}}$ . Let  $\mathbf{x} \in \Omega_{\mathbf{X}}$ , and define  $A \subseteq [J]$  as that subset of  $[J]$ , where for all  $i \in A$ ,  $x_i = 1$  and for all  $j \in A^c$ ,  $x_j = 0$ . Then it is clear from (4.48) that

$$Q(\mathbf{x}) = Q(\mathbf{x}_A, \mathbf{x}_{A^c}) = \sum_{B \subseteq A} \eta_B,$$

and this requires

$$\eta_A = Q(\mathbf{x}_A, \mathbf{x}_{A^c}) - \sum_{B \subset A} \eta_B, \quad (4.52)$$

noting that  $B \subset A$  means that  $B$  is a *proper* subset of  $A$ . Thus by proceeding in steps from  $|A| = 0$  to  $|A| = J$ , as in (4.52), all the  $\eta_A$  are well defined in terms of  $Q(\mathbf{x})$  and thus (4.48) is satisfied for all  $\mathbf{x} \in \Omega_{\mathbf{X}}$ .  $\square$

#### 4.3.1.2 Step 2: The dependency structure of $\mathbf{X}$

Let  $\mathbf{X}$  with probability distribution  $P(\mathbf{x})$  be defined as above. Following Besag (1974), the set of *neighbors* of  $X_i$  are all the random variables  $X_j$ , where  $j \neq i$ , such that the conditional distribution of  $X_i$  given all the other random variables, depends on  $X_j$ . To state this formally, the notation in the previous section is useful. First use the rules of conditional probability to rewrite

$$P(x_i | x_j, \mathbf{x}_{\{i,j\}^c}) = P(x_i, x_j | \mathbf{x}_{\{i,j\}^c}) / P(x_j | \mathbf{x}_{\{i,j\}^c}).$$

Then from the above, if  $X_j$  is not a neighbor of  $X_i$ , we can drop  $x_j$  in the condition on the left hand side yielding for all  $\mathbf{x} \in \Omega_{\mathbf{X}}$

$$P(x_i | \mathbf{x}_{\{i,j\}^c})P(x_j | \mathbf{x}_{\{i,j\}^c}) = P(x_i, x_j | \mathbf{x}_{\{i,j\}^c}). \quad (4.53)$$

Thus, from (4.53), the requirement that  $X_j$  is a neighbor of  $X_i$  means that there is at least one  $\mathbf{x} \in \Omega_{\mathbf{X}}$  such that

$$P(x_i, x_j | \mathbf{x}_{\{i,j\}^c}) \neq P(x_i | \mathbf{x}_{\{i,j\}^c})P(x_j | \mathbf{x}_{\{i,j\}^c}), \quad (4.54)$$

meaning that  $X_i$  and  $X_j$  are not conditionally independent given the rest.

Denote the neighbors of  $X_i$  by  $N(i) = \{j \mid X_j \text{ is a neighbor of } X_i\}$ . It is clear from (4.54) that being neighbors is a symmetric relation, that is for all  $i, j \in [J]$

$$j \in N(i) \Leftrightarrow i \in N(j).$$

Thus the dependency structure of  $\mathbf{X}$  with probability distribution  $P(\mathbf{x})$  can be represented by a dependency graph,  $G_{\mathbf{X}} = ([J], E_{\mathbf{X}})$ , where  $E_{\mathbf{X}}$  is the set of all pairs  $\{i, j\}$  such that  $X_i$  and  $X_j$  are neighbors.

#### 4.3.1.3 Step 3: The H-C theorem for 0–1-valued random variables

The final step in our representation of  $\mathbf{X}$  with probability distribution  $P(\mathbf{x})$  is to present the H-C theorem for 0–1-valued random variables. The theorem provides a simplification of the representation of  $Q(\mathbf{x})$  in (4.48). In order to state the theorem, we need to use the cliques of the dependency graph  $G_{\mathbf{X}} = ([J], E_{\mathbf{X}})$ . Recall from earlier definitions that the *cliques* of a graph consist of all subsets of nodes where every pair of nodes in the subset has an edge. Notice that every singleton node is (vacuously) a clique. Consequently, denote the set of cliques of  $G_{\mathbf{X}} = ([J], E_{\mathbf{X}})$  by

$$\text{Cl}_{\mathbf{X}} \stackrel{\text{def}}{=} \{A \subseteq [J] \mid (\forall i, j)(i \neq j \ \& \ i, j \in A \Rightarrow i \in N(j))\}.$$

Note that the definition makes use of the fact that being neighbors is a symmetric relation and that it implies that the relation is irreflexive. The crucial consequence of the H-C theorem is that it allows us to set some of the  $\eta_A$ , where  $A \subseteq [J]$ , in the representation (4.48) to zero, as determined by the dependency structure of  $\mathbf{X}$ .

**Theorem 4.10 (H-C)** Let  $\mathbf{X} = (X_j)_{j \in [J]}$  be a vector of 0–1-valued random variables with probability distribution  $P(\mathbf{x})$  subject to the condition that for all  $\mathbf{x} \in \Omega_{\mathbf{X}}$  we have  $P(\mathbf{x}) > 0$ , and let  $G_{\mathbf{X}} = ([J], E_{\mathbf{X}})$  be the dependency graph of  $\mathbf{X}$  with cliques  $\text{Cl}_{\mathbf{X}}$ . Consider the expansion of  $Q(\mathbf{x})$  in (4.48). Then for all  $A \subseteq [J]$

$$A \notin \text{Cl}_{\mathbf{X}} \Rightarrow \eta_A = 0.$$

*Proof* Suppose  $A \subseteq [J]$  is not a clique of  $G_{\mathbf{X}} = ([J], E_{\mathbf{X}})$ . Then there are distinct  $i, j \in A$  such that  $j \notin N(i)$  and  $i \notin N(j)$ . Next, select a realization  $\mathbf{x} \in \Omega_{\mathbf{X}}$  subject to the following conditions on the values in  $\mathbf{x}_{\{i,j\}^c}$ : if  $k \in A \setminus \{i, j\}$  then  $x_k = 1$ , else if  $k \notin A$  then  $x_k = 0$ , leaving  $x_i$  and  $x_j$  undetermined. Denote such a realization by  $(x_i, x_j, \mathbf{x}_{\{i,j\}^c})$ . Considering the four possible values of  $(x_i, x_j)$ , we then get from conditional independence in (4.53) that

$$P(1, 1 | \mathbf{x}_{\{i,j\}^c}) P(0, 0 | \mathbf{x}_{\{i,j\}^c}) = P(1, 0 | \mathbf{x}_{\{i,j\}^c}) P(0, 1 | \mathbf{x}_{\{i,j\}^c}). \quad (4.55)$$

Because the conditional in each term is the same, simple probability calculations on (4.55) lead to

$$P(1, 1, \mathbf{x}_{\{i,j\}^c}) P(0, 0, \mathbf{x}_{\{i,j\}^c}) = P(1, 0, \mathbf{x}_{\{i,j\}^c}) P(0, 1, \mathbf{x}_{\{i,j\}^c}),$$

from which it easily follows that

$$Q(1, 1, \mathbf{x}_{\{i,j\}^c}) + Q(0, 0, \mathbf{x}_{\{i,j\}^c}) = Q(1, 0, \mathbf{x}_{\{i,j\}^c}) + Q(0, 1, \mathbf{x}_{\{i,j\}^c}). \quad (4.56)$$

Next apply the expansion in (4.48) to each term in (4.56) yielding

$$\sum_{B \subseteq A} \eta_B + \sum_{B \subseteq A \setminus \{i, j\}} \eta_B = \sum_{B \subseteq A \setminus \{j\}} \eta_B + \sum_{B \subseteq A \setminus \{i\}} \eta_B. \quad (4.57)$$

It is easy to derive from (4.57) that

$$\eta_A + \eta_{A \setminus \{i\}} + \eta_{A \setminus \{j\}} + 2 \sum_{B \subseteq A \setminus \{i, j\}} \eta_B = \eta_{A \setminus \{i\}} + \eta_{A \setminus \{j\}} + 2 \sum_{B \subseteq A \setminus \{i, j\}} \eta_B,$$

from which it follows that  $\eta_A = 0$ .  $\square$

The consequence of the H-C theorem is that the expansion in (4.48) can be simplified by dropping all subsets of  $[J]$  that do not correspond to cliques in the dependency graph. The result is that

$$Q(\mathbf{x}) = \sum_{A \in \text{Cl}_{\mathbf{X}}} \eta_A \prod_{i \in A} x_i. \quad (4.58)$$

### 4.3.2 The H-C theorem and joint probability distributions for models

It is of course possible to be given a specified joint pdf  $P(\mathbf{x})$  over  $J$  0–1 random variables, construct the dependency graph, and express  $Q(\mathbf{x})$  as a sum over the cliques of the dependency graph as in (4.58). Such an undertaking could easily be tedious, and if one already had  $P(\mathbf{x})$ , there would be little to gain by such a representation. Instead, the H-C theorem is used in a different way to construct a parameterized class of joint pdfs. The idea is to decide *a priori* what dependency structure

one wants to specify in a model. From that dependency structure one determines the cliques,  $\text{Cl}_X$ , and then develops the expansion in (4.58). From the expansion one has  $Q(\mathbf{x})$ , and from that one can derive the nature of  $P(\mathbf{x})$  from (4.47). The coefficients in (4.58) can be interpreted as model parameters rather than numbers, and when this is done one has a parametric family of pdfs over  $\mathbf{X} = (X_j)_{j \in [J]}$ , with parameter  $\Theta = (\eta_A)_{A \in \text{Cl}_X}$  with space  $\Psi_\Theta = \mathbb{R}^{|\text{Cl}_X|}$ . This process will be illustrated in the next two examples.

**Example 4.1** (Bernoulli model) Suppose one chooses the simplest possible dependency structure of the  $J$  0–1 random variables, which is the case where for all  $j \in [J]$ ,  $N(j) = \emptyset$ . In such a case

$$\text{Cl}_X = \{\{X_j\} \mid j \in [J]\},$$

and the expansion in (4.58) becomes

$$Q(\mathbf{x}) = \sum_{j \in [J]} \eta_j x_j, \quad (4.59)$$

with parameters  $\Theta = (\eta_j)_{j \in [J]}$  having space  $\Psi_\Theta = \mathbb{R}^J$ . Such a parameterization is easily seen to produce a model equivalent to the independent Bernoulli trials model. Recall that the Bernoulli model has parameters  $\mathbf{P} = (p_j)_{j \in [J]}$ , with space  $\Psi_\mathbf{P} = (0, 1)^J$ , and probability distribution

$$P(\mathbf{x}) = \prod_{j \in [J]} p_j^{x_j} (1 - p_j)^{1-x_j} = \prod_{j \in [J]} \left( \frac{p_j}{1 - p_j} \right)^{x_j} (1 - p_j). \quad (4.60)$$

If we divide the Bernoulli expression by  $P(\mathbf{0}) = \prod_{j \in [J]} (1 - p_j)$  and take logs to get  $Q(\mathbf{x})$  in (4.48), it is easily seen that the two models are statistically equivalent with the parameter  $\eta_j = \log(\frac{p_j}{1-p_j})$  that reduces (4.59) to (4.60). If for all  $j \in [J]$  we have  $\eta_j = \eta$ , then the model expressed by (4.59) is equivalent to the Bernoulli trials model, with  $P(\mathbf{x}) = p^{k(\mathbf{x})} (1 - p)^{J-k(\mathbf{x})}$ , where  $k(\mathbf{x}) = \sum_{j \in [J]} x_j$  and  $p \in (0, 1)$ .

**Example 4.2** (Markov chain) The goal of this example is to interrelate two representations of a simple homogeneous Markov chain on  $M$  0–1 random variables,  $\mathbf{X} = (X_i)_{i \in [M]}$ , whose space is  $\Omega_X = \{0, 1\}^M$  and where  $M > 1$ . The first representation will be in terms of the ordinary transition matrix and start vector of a homogeneous Markov chain, and the second will come from the H-C theorem. Consider the following Markov chain representation:

$$\begin{array}{c|cc} & 1 & 0 \\ \mathbf{T} = 1 & t_1 & 1 - t_1 \\ 0 & 1 - t_0 & t_0 \\ \hline \end{array} \quad \mathbf{P} = (p, 1 - p). \quad (4.61)$$

The states of the Markov chain are denoted by “1” and “0,” and the parameters are  $\Theta = (p, t_0, t_1) \in \Psi_\Theta = (0, 1)^3$ . The start vector is  $\mathbf{P}$ , and  $p = P(X_1 = 1)$  is the

probability that the chain starts in state 1, while  $\mathbf{T}$  is the transition matrix of the chain. The parameters  $t_0, t_1$  are state-to-state transition probabilities. For example,  $t_0$  is the probability that the chain stays in state 0 for the next trial following any trial that it is in state 0, and  $t_1$  plays the same role for state 1. Below are some useful probabilities for a string of  $M$  trials computed from the Markov model (4.61) (as usual letting  $\mathbf{0}$  denote a string of one or more zeros):

$$\begin{aligned} P(\mathbf{0}) &= (1-p)t_0^{M-1}, \\ P(\mathbf{10}) &= p(1-t_1)t_0^{M-2}, \\ P(\mathbf{01}) &= (1-p)t_0^{M-2}(1-t_0), \\ P(\mathbf{010}) &= (1-p)t_0^{i-2}(1-t_0)(1-t_1)t_0^{M-i-1}, \quad 1 < i < M, \\ P(\mathbf{110}) &= pt_1(1-t_1)t_0^{M-3}, \\ P(\mathbf{011}) &= (1-p)t_0^{M-3}(1-t_0)t_1, \\ P(\mathbf{0110}) &= (1-p)t_0^{i-2}(1-t_0)t_1(1-t_1)t_0^{M-i-2}, \quad 1 < i < M-1. \end{aligned} \quad (4.62)$$

Now let's model a homogeneous Markov chain using the H-C theorem. The cliques of the dependency graph are given by

$$\text{Cl}_\mathbf{x} = \{\{X_i\} \mid 1 \leq i \leq M\} \cup \{\{X_i, X_{i+1}\} \mid 1 \leq i \leq M-1\}. \quad (4.63)$$

Now for our homogeneous Markov chain using these cliques we have the H-C representation given from (4.58) by

$$\begin{aligned} Q(\mathbf{x}) &= \eta_1 x_1 + \eta_M x_M + \eta \sum_{i=2}^{M-1} x_i \\ &\quad + \eta_{12} x_1 x_2 + \eta_{(M-1)M} x_{(M-1)} x_M + \eta^* \sum_{i=2}^{M-1} x_i x_{i+1}, \end{aligned} \quad (4.64)$$

and each of these six  $\eta$ s has a space that is the full real line. Notice in (4.64) we have assigned the same parameter  $\eta$  to all cliques of the form  $\{\{X_i\} \mid 1 < i < M\}$  and the same parameter  $\eta^*$  to all cliques of the form  $\{\{X_i, X_{i+1}\} \mid 1 < i < M-1\}$ . The reason why these six clique types are needed for a homogeneous Markov chain comes from the different probabilities in (4.62) for the usual representation of a homogeneous Markov chain. We see that the three cases where  $\mathbf{x}$  has exactly one random variable with a value of 1 have different probabilities and hence for such  $\mathbf{x}$ ,  $Q(\mathbf{x})$  has three possible values. The same is true for the case where  $\mathbf{x}$  has exactly two adjacent random variables with values of 1 and the rest have values of 0.

We can compute the  $\eta$ s from the H-C expansion  $Q(\mathbf{x})$  in (4.48) by using (4.52). The results are given in the next six equations, where the middle term of each equation is the H-C value of each  $\eta$ , and the right term is its value in terms of the

traditional representation in (4.61).

$$\begin{aligned}
 \eta_1 &= Q(1\mathbf{0}) = \log \frac{p(1-t_1)}{(1-p)t_0}, \\
 \eta_M &= Q(\mathbf{0}1) = \log \frac{(1-t_0)}{t_0}, \\
 \eta &= Q(\mathbf{0}1\mathbf{0}) = \log \frac{(1-t_0)(1-t_1)}{t_0^2}, \\
 \eta_{12} &= Q(11\mathbf{0}) - (\eta_1 + \eta) = \log \frac{pt_1(1-t_1)}{(1-p)t_0^2} - (\eta_1 + \eta), \\
 \eta_{(M-1)M} &= Q(\mathbf{0}11) - (\eta + \eta_M) = \log \frac{(1-t_0)t_1}{t_0^2} - (\eta + \eta_1), \\
 \eta^* &= Q(\mathbf{0}11\mathbf{0}) - 2\eta = \log \frac{(1-t_0)t_1(1-t_1)}{t_0^3} - 2\eta.
 \end{aligned} \tag{4.65}$$

Now, according to the H-C theorem and our knowledge of the usual homogeneous Markov chain theory, these two representations must be statistically equivalent. The first concern in showing this is that there are six  $\eta$ s in the H-C representation and only three functionally independent parameters in the usual definition in (4.61), so the six  $\eta$ s cannot be functionally independent. Notice that we can solve for the three parameters in (4.61) from the first three of the equations in (4.65). The results are:

$$\begin{aligned}
 t_0 &= \frac{1}{1+e^{\eta_M}}, \\
 t_1 &= \frac{1+e^{\eta_M}-e^{\eta-\eta_M}}{1+e^{\eta_M}}, \\
 p &= \frac{e^{\eta_1-\eta+\eta_M}}{1+e^{\eta_1-\eta+\eta_M}}.
 \end{aligned} \tag{4.66}$$

The equations in (4.66) show that as the H-C parameters  $(\eta_1, \eta_M, \eta)$  range independently through the real numbers, the three usual Markov parameters  $(p, t_1, t_0)$  range independently in  $(0, 1)$ . Further, these three equations are invertible, as can be seen in the first three equations of (4.65), and in that form these three are functionally independent and free to vary in the reals. Finally, it is clear from the last three equations of (4.65) that the values of  $(\eta_{12}, \eta_{(M-1)M}, \eta^*)$  are functionally determined from the values of  $(\eta_1, \eta_M, \eta)$ .

In order to complete the demonstration that the two specifications are statistically equivalent, we need to show that for all  $\mathbf{x} \in \Omega_{\mathbf{X}}$ ,  $P(\mathbf{x})$  generated from (4.61) match the values of  $P(\mathbf{x})$  that are obtained from (4.64) under the parameter equivalences between the two models. For example, consider  $x = (0011110001)$ , for  $M = 10$ . From (4.61) we have

$$Q(\mathbf{x}) = \log \frac{P(\mathbf{x})}{P(\mathbf{0})} = \log \frac{t_1^3(1-t_0)^2(1-t_1)}{t_0^6},$$

and from (4.64), replacing the  $\eta$  by their values in (4.65), we have

$$\begin{aligned} Q(\mathbf{x}) &= 4\eta + 3\eta^* + \eta_M \\ &= 4\eta + 3 \log \frac{(1-t_0)t_1(1-t_1)}{t_0^3} - 6\eta + \log \frac{1-t_0}{t_0} \\ &= \log \frac{t_1^3(1-t_0)^2(1-t_1)}{t_0^6}. \end{aligned}$$

This establishes the equivalence of the two representations for the selected  $\mathbf{x}$ , and we leave it as an exercise for the reader to see that arbitrary probabilities  $P(\mathbf{x})$  match in the two representations when parameter equivalences are used.

### 4.3.3 The H-C theorem for PGMs and PDMs

In the case of PGMs and PDMs, the random variables are, respectively, the edge and arc random variables, which as described earlier are 0–1 random variables. To employ the H-C theorem for probabilistic graphical models, one does not start with a particular joint probability distribution on the edge or arc random variables, and then determine the dependence graph, and then create a representation of the distribution as in (4.48). Instead, as with the examples presented so far, one reverses this process by starting with a particular specification of the dependence structure of the random variables in the form of a dependency graph, and then the H-C theorem is used to derive the form of the PGM or PDM. While this procedure can be done for any dependency graph defined on the node set  $V$ , it is often of interest to consider a special case where isomorphic graphs or digraphs have identical probabilities. This condition is stated in subsection 4.2.1 on graphs. This condition assumes that the tendencies to participate in various substructures are the same for all nodes in the graph, which of course is at best an approximation. However, the homogeneity assumption allows a considerable reduction in model parameters and it arguably serves as a representation of the overall tendencies exhibited in the graph as a whole. However, the assumption would be inappropriate in a case with covariates that applied to each node in the graph. We illustrate this process in a series of examples to follow.

**Example 4.3** (Bernoulli graph models) This example is set up using the H-C theorem with cliques only of size 1, and then evoking homogeneity, i.e., that all isomorphic configurations are equally likely. We show that the Bernoulli model derives from the H-C theorem and homogeneity.

The Bernoulli PGM on  $N$  nodes has a single parameter  $p$  with space  $\Psi_p = (0, 1)$ . The model assumes that edges occur between pairs of nodes independently with probability  $p$ . Recall that there are  $2^{N(N-1)/2}$  possible edges, and the presence or absence of edges is encoded by the family of random variables  $\mathbf{X} = (X_{ij})_{i,j \in [N], i < j}$ . Therefore, the Bernoulli PGM has probability distribution for

all  $\mathbf{x} \in \Omega_{\mathbf{X}}$ ,

$$P(\mathbf{x}, p) = p^{t(\mathbf{x})}(1 - p)^{[2^{N(N-1)/2} - t(\mathbf{x})]}, \quad (4.67)$$

where  $t(\mathbf{x}) = \sum_{i < j} x_{ij}$  is the number of edges in the realization  $\mathbf{x}$ . Now to derive the Bernoulli PGM from the H-C theorem, assume that the cliques consist of only single edge random variables, that is

$$\text{Cl}_{\mathbf{X}} = \{\{X_{ij}\} \mid 1 \leq i < j \leq N\}.$$

Then by the H-C theorem (Theorem 4.10) we have for all  $\mathbf{x} \in \Omega_{\mathbf{X}}$ ,

$$Q(\mathbf{x}) = \sum_{i, j \in [N], i < j} \eta_{ij} x_{ij}. \quad (4.68)$$

Now, evoking the homogeneity condition that all isomorphic graphs are equally likely, we can set  $\eta_{ij} \stackrel{\text{def}}{=} \eta$ , and this single parameter has space  $\Psi_\eta = \mathbb{R}$ . In Example 4.1 we derived the usual Bernoulli model for a sequence of binary random variables from (4.59) by equating  $\eta_j = \log \frac{p_j}{1-p_j}$ . Then it is easy to see that for all  $\mathbf{x} \in \Omega_{\mathbf{X}}$ ,

$$Q(\mathbf{x}) = \log \left( \frac{p}{1-p} \right)^{t(\mathbf{x})} = \log \frac{P(\mathbf{x})}{P(\mathbf{0})}. \quad (4.69)$$

Using the same logic as in Example 4.1, Equation (4.69) is equivalent to (4.67), and thus we see how the Bernoulli PGM derives from the H-C theorem.

The Bernoulli PDM is easily derived from the H-C theorem using single arc random variables as the cliques of the dependency graph,

$$\text{Cl}_{\mathbf{Y}} = \{\{Y_{ij}\} \mid 1 \leq i, j \leq N, i \neq j\}. \quad (4.70)$$

We leave this derivation as an exercise for the reader. The most salient problem with the Bernoulli PDM is that it assumes that reciprocal arcs  $\{Y_{ij}, Y_{ji}\}$  are statistically independent. It is well established that in many situations in networks there is dependency between reciprocal arcs. For example, in friendship networks there is a tendency toward reciprocation, i.e.,  $Y_{ij} = Y_{ji}$ , and in networks representing authority relations there is a tendency toward asymmetry, i.e.,  $Y_{ij} \neq Y_{ji}$ . It is easy to develop a model that has dependence between reciprocal arcs using the H-C theorem while retaining the assumption that dyads  $\{Y_{ij}, Y_{ji}\}$  are conditionally independent. In addition to the cliques in the dependency graph for the Bernoulli PDM (4.70), one adds reciprocal random variables. Then the dependency graph becomes

$$\text{Cl}_{\mathbf{Y}} = \{\{Y_{ij}, \{Y_{ij}, Y_{ji}\} \mid 1 \leq i, j \leq N, i \neq j\}\}. \quad (4.71)$$

Then employing the H-C theorem 4.10, we have a parameterized PDM that includes possible tendencies toward or against reciprocity given by for all  $\mathbf{y} \in \Omega_{\mathbf{Y}}$ ,

$$Q(\mathbf{y}) = \sum_{i, j \in [N], i \neq j} \eta_{ij} y_{ij} + \sum_{i, j \in [N], i < j} \tilde{\eta}_{ij} y_{ij} y_{ji}, \quad (4.72)$$

where  $\eta_{ij}$ ,  $\tilde{\eta}_{ij} \in \mathbb{R}$ . Now invoking the homogeneity condition discussed earlier, set  $\eta_{ij} \stackrel{\text{def}}{=} \eta$  and  $\tilde{\eta}_{ij} \stackrel{\text{def}}{=} \tilde{\eta}$ , with parameter space  $\Psi_{(\eta, \tilde{\eta})} = \mathbb{R}^2$ . Then (4.72) simplifies to

$$\mathcal{Q}(\mathbf{y}) = s(\mathbf{y})\eta + m(\mathbf{y})\tilde{\eta}, \quad (4.73)$$

where

$$\begin{aligned} s(\mathbf{y}) &= \sum_{i,j \in [N], i \neq j} y_{ij} && \text{(sum of all arcs)} \\ m(\mathbf{y}) &= \sum_{i,j \in [N], i < j} y_{ij} y_{ji} && \text{(mutual pairs).} \end{aligned} \quad (4.74)$$

From (4.73), we can derive the probability distributions of the new PDM. They are given by

$$P(\mathbf{y}) = P(\mathbf{0}) \exp[s(\mathbf{y})\eta + m(\mathbf{y})\tilde{\eta}]. \quad (4.75)$$

The form of the *dyad independence model* in (4.75) is not immediately useful unless one has  $P(\mathbf{0})$ ; however, this quantity is a constant for each value of the parameter  $(\eta, \tilde{\eta})$ , so we can see that for all  $\mathbf{y} \in \Omega_{\mathbf{Y}}$ ,

$$P(\mathbf{y}) = k(\eta, \tilde{\eta}) \exp[s(\mathbf{y})\eta + m(\mathbf{y})\tilde{\eta}]. \quad (4.76)$$

A more traditional way to specify a model for dyad independence of the sort represented in (4.76) would be to assume that dyads can be classified into three classes: *mutual* ( $Y_{ij} = Y_{ji} = 1$ ), where arcs are reciprocated; *asymmetric* ( $Y_{ij} \neq Y_{ji}$ ), where exactly one of the two arcs is present, and *null*, where neither arc is present. Then one could assume that dyads are classified i.i.d. into these three classes with probabilities, respectively,  $p$ ,  $q$ , and  $r = 1 - p - q$ . Such a model would have parameter  $\Theta = (p, q)$ , with parameter space  $\Psi_{\Theta} = \{\Theta = (p, q) \mid p, q > 0, p + q < 1\}$ . This version of the homogeneous dyad independence model is defined for all  $\mathbf{y} \in \Omega_{\mathbf{Y}}$  by

$$P(\mathbf{y}) = p^{m(\mathbf{y})} q^{a(\mathbf{y})} (1 - p - q)^{n(\mathbf{y})}, \quad (4.77)$$

where  $a(\mathbf{y}) = \sum_{i,j \in [N], i \neq j} y_{ij}(1 - y_{ji})$  (asymmetric ordered pairs),  $m(\mathbf{y})$  was defined in (4.74), and  $n(\mathbf{y}) = N(N - 1)/2 - [m(\mathbf{y}) + a(\mathbf{y})]$  (the null unordered pairs are the complement of mutual plus asymmetric pairs).

By the logic of the H-C theorem, the model in (4.76) should be statistically equivalent to the model in (4.77), but it is not immediately obvious that this is so. To see their relationship, note that  $a(\mathbf{y}) = s(\mathbf{y}) - 2m(\mathbf{y})$  implies that (4.77) can be rewritten as

$$\begin{aligned} P(\mathbf{y}) &= p^{m(\mathbf{y})} q^{s(\mathbf{y}) - 2m(\mathbf{y})} (1 - p - q)^{N(N-1)/2 - m(\mathbf{y}) - s(\mathbf{y}) + 2m(\mathbf{y})} \\ &= \left( \frac{p(1 - p - q)}{q^2} \right)^{m(\mathbf{y})} \left( \frac{q}{1 - p - q} \right)^{s(\mathbf{y})} (1 - p - q)^{N(N-1)/2}. \end{aligned} \quad (4.78)$$

Next, note that if

$$\begin{aligned}\eta &= \log \frac{q}{1-p-q}, \\ \tilde{\eta} &= \log \frac{p(1-p-q)}{q^2}, \\ k(\eta, \tilde{\eta}) &= (1-p-q)^{N(N-1)/2},\end{aligned}\tag{4.79}$$

then (4.76) is identical to (4.78), and this establishes that the models defined in (4.76) and (4.77) are statistically equivalent, where (4.79) converts the parameters of (4.77) into  $(\eta, \tilde{\eta})$ , and the transformation is invertible with

$$p = \frac{e^{2\eta+\tilde{\eta}}}{1 + e^\eta + e^{2\eta+\tilde{\eta}}}, \quad q = \frac{e^\eta}{1 + e^\eta + e^{2\eta+\tilde{\eta}}}.$$

So far in all our examples of parameterized models using the H-C theorem, we have stated a conventional probability model for the joint distribution of the random variables, and then have shown that a statistically equivalent version of the model can be derived from the theorem applied to the appropriate dependency graph. For example, for the dyad independence model in the previous example we stated the model in terms of a trinomial distribution in (4.77) and then we showed that it was equivalent to the model derived from the H-C theorem from the dependency graph in (4.71) along with the homogeneity condition leading to (4.76). The real advantage of using the H-C theorem to create PGMs and PDMs is to create new models when the dependency graph is more complex and there are no corresponding established conventional models to fall back on. The next example is historically interesting because it is the first model in the networks area to be built up from the H-C theorem, and it has proved to lead the way to other useful models.

**Example 4.4** (Markov graphs) In this example we show how one can define a class of PGMs that incorporate Markov properties in a spatial, rather than a temporal, sense. Consider the collection of graphs on  $N$  nodes, with edge random variables  $\mathbf{X} = (X_{ij})_{i,j \in [N], i < j}$ . Markov graphs were introduced by Frank and Strauss (1986), who developed the idea from work in the area of spatial statistics, e.g., Besag (1974). These Markov PGMs are developed from the idea that only “adjacent” edges are conditionally dependent. Specifically, assume that the edge random variables  $X_{ij}$  and  $X_{kl}$  are conditionally dependent only if  $\{i, j\} \cap \{k, l\} \neq \emptyset$ . From this restriction it is possible to state the cliques of the dependency graph.

Such cliques are easily seen to consist exactly of all singleton edge random variables (also called 1-stars),  $A = \{\{X_{ij}\} \mid 1 \leq i < j \leq N\}$ , all possible triangles,  $T = \{\{X_{ij}, X_{jk}, X_{ik}\} \mid 1 \leq i < j < k \leq N\}$ , and all possible  $k$ -stars  $S_k = \{\{X_{i_1j_1}, X_{i_2j_2}, \dots, X_{i_kj_k}\} \mid \text{for distinct } i, j_1, \dots, j_k \in [N]\}$ . (Note in this representation of  $S_k$  we have made use of the fact that for graphs  $X_{ij} = X_{ji}$ , and this harmlessly violates the convention that  $X_{ij}$  implies  $i < j$ .) There are  $\binom{N}{2}$  single edge cliques,  $\binom{N}{3}$  distinct triangles, and  $N \binom{N-1}{k} = (k+1) \binom{N}{k+1}$  distinct  $k$ -stars in the dependency graph  $\text{Cl}_{\mathbf{X}}$ . The two alternative expressions for the number of distinct  $k$ -stars

corresponding to two methods of constructing a star: (1) pick out the  $N$  possible star centers and then pick out  $k$  endpoints, or (2) pick out the  $k + 1$  nodes for the whole star and then designate one node as the center. Note that any  $k$ -star contains a number of smaller stars, e.g., any 4-star contains four 3-stars and six 2-stars.

We leave it as an exercise to show that if the graph has  $N$  nodes, the dependency graph for the Markov PGM has cardinality

$$N2^{N-1} + \binom{N}{3} - \binom{N+1}{2}.$$

The H-C expansion for Markov PGMs would include a prohibitive number of terms if each of these cliques were represented by a different parameter; however, invoking homogeneity allows us to have a single parameter for each type of clique in the H-C expansion in (4.58). The result is for all  $\mathbf{x} \in \Omega_{\mathbf{x}}$ ,

$$Q(\mathbf{x}) = \eta_A a(\mathbf{x}) + \eta_T t(\mathbf{x}) + \sum_{k=2}^{N-1} \eta_{S_k} s_k(\mathbf{x}), \quad (4.80)$$

where  $a(\mathbf{x})$  is the number of edges present in  $\mathbf{x}$ ,  $t(\mathbf{x})$  is the number of triangles present in  $\mathbf{x}$ , and  $s_k(\mathbf{x})$  is the number of  $k$ -stars present in  $\mathbf{x}$ . The Markov graph model on  $N$  nodes now has  $N$  parameters,

$$\Theta = (\eta_A, \eta_T, \eta_{S_2}, \dots, \eta_{S_{N-1}}),$$

with space  $\Psi_\Theta = \mathbb{R}^N$ .

From (4.80) we can write the parameterized probability distribution of the model as for all  $\mathbf{x} \in \Omega_{\mathbf{x}}$ ,

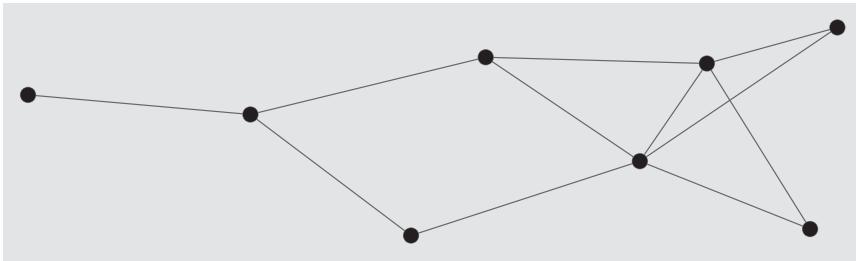
$$P(\mathbf{x}) = P_\Theta(\mathbf{0}) \exp[\eta_A a(\mathbf{x}) + \eta_T t(\mathbf{x}) + \sum_{k=2}^{N-1} \eta_{S_k} s_k(\mathbf{x})]. \quad (4.81)$$

The term  $P_\Theta(\mathbf{0})$  is needed to completely specify the probabilistic structure of the model, and of course it is determined as a function of  $\Theta$  by the equation

$$P_\Theta(\mathbf{0}) = 1 - \sum_{\mathbf{x} \in \Omega_{\mathbf{x}} \setminus \mathbf{0}} \exp[\eta_A a(\mathbf{x}) + \eta_T t(\mathbf{x}) + \sum_{k=2}^{N-1} \eta_{S_k} s_k(\mathbf{x})]. \quad (4.82)$$

Unfortunately it is not easy to obtain (4.82) except for very small graphs, and modelers generally regard  $P_\Theta(\mathbf{0})$  as a normalizing constant, and even without its value one can easily see that ratios of probabilities of the form  $P(\mathbf{x})/P(\mathbf{y})$  are easily computed from (4.82) because the normalizing constant cancels. More generally, there are Markov Chain Monte Carlo (MCMC) methods that can be used to compute with the model, e.g., Snijders (2002). Further discussion and relevant references are in section 4.4.

The full Markov PGM in (4.81) has parameters that can increase or decrease the probability of a particular graph as a function of the number of edges, triangles, and the various types of stars. Despite evoking homogeneity to reduce the number of parameters, there are still  $N$  parameters for a graph with  $N$  nodes. However,



**Figure 4.27** Graph illustrating PGM counts.

there are interesting nested models that can be constructed by setting some of these parameters to zero. For example, suppose one wanted only to control the tendency for the graph to have or not have high levels of transitivity. In this case one could retain as free parameters  $\eta_A$  and  $\eta_T$  and set the rest of the parameters in (4.81) to zero. Then the resulting PGM would have parameter  $\Theta = (\eta_A, \eta_T)$  with space  $\Psi_\Theta = \mathbb{R}^2$ , and probability function, for all  $\mathbf{x} \in \Omega_{\mathbf{X}}$ ,

$$P(\mathbf{x}) = P_\Theta(\mathbf{0}) \exp[\eta_A a(\mathbf{x}) + \eta_T t(\mathbf{x})]. \quad (4.83)$$

To illustrate the PGMs in (4.81) and (4.83), consider the graph  $\mathbf{x}$  in Figure 4.27. There are  $N = 8$  nodes, and it is easy to count the eight quantities needed for (4.83). The results are:  $a(\mathbf{x}) = 11$ ,  $t(\mathbf{x}) = 3$ ,  $s_2 = 25$ ,  $s_3 = 16$ ,  $s_4 = 6$ ,  $s_5 = 1$ , and the last two stars are not present in the graph. For example, the transitive model in (4.83) would generate the probability of the graph in Figure 4.27 as

$$P(\mathbf{x} | \Theta) = p_\theta(\mathbf{0}) \exp(11\eta_A + 3\eta_T),$$

where  $p_\theta(\mathbf{0})$  is the probability of a graph on eight nodes with no edges, given a particular value of the parameter  $\Theta$ , which could in principle be calculated from (4.81) by setting all the  $\eta_{S_k} = 0$ .

Thus far in this example we have developed Markov PGMs. In the case of PDMs (Frank and Strauss, 1986), define a Markov PDM as one where two arcs,  $Y_{ij}$  and  $Y_{kl}$ , are conditionally dependent given the remaining arcs only if  $\{i, j\} \cap \{k, l\} \neq \emptyset$ . This means that the dependency graph has edges only between such pairs of arcs. The cliques of the dependency graph include all possible arcs, all mutual arcs, for example  $\{Y_{ij}, Y_{ji}\}$ , stars of order  $k > 2$ , and triangles of various kinds. In the case of digraphs, stars and triangles can take different forms: for example,  $\{Y_{ij}, Y_{jk}, Y_{ik}\}$ ,  $\{Y_{ij}, Y_{jk}, Y_{ki}\}$ , and  $\{Y_{ij}, Y_{ji}, Y_{ik}, Y_{jk}\}$  are all distinct triangles and therefore cliques in the dependency graph of the Markov PDM. We leave it as an exercise to show that for digraphs on  $N$  nodes, there are  $27\binom{N}{3}$  distinct triangles and  $N\binom{N-1}{k}3^k$  distinct  $k$ -stars (each has  $k + 1$  nodes) in the dependency graph.

The Markov PDG on  $N$  nodes, even with the homogeneity assumption leading to (4.81), has a large number of parameters, and in general one would examine various sub models analogous to the one in (4.83) to control for the tendencies of the digraph to have various structures.

**Example 4.5** (Exponential random graph models (ERGMs)) The Markov models of Example 4.4 in this section represented a paradigm shift in probabilistic models for networks, especially social networks. The accepted idea was that large networks are built up of the interactions between small Platonic structures of the sort that were described in subsection 4.2.1. In particular, the structure of (4.81) suggests that one could select a set of simple graph substructures like triangles,  $k$ -stars,  $P_4$ , or  $C_4$  (see Figure 4.2), use their counts as sufficient statistics, and then write a model directly without having to derive it from the H-C theorem. Suppose  $S$  simple graph structures are selected, and let  $t_s(\mathbf{x})$  be the count of the number of occurrences of structure  $s$  in a graph  $\mathbf{x}$ . Then by analogy to (4.81), one could write the probability distribution in the form, for all  $\mathbf{x} \in \Omega_{\mathbf{x}}$ ,

$$P(\mathbf{x} | \Theta) = k(\Theta) \exp\left\{\sum_{s=1}^S \theta_s t_s(\mathbf{x})\right\}, \quad (4.84)$$

where the parameter of the model is  $\Theta \in \Psi_{\Theta} = \mathbb{R}^S$ , and  $k(\Theta)$  is a normalizing constant for each  $\Theta$  to assure that

$$\sum_{\mathbf{x} \in \Omega_{\mathbf{x}}} P(\mathbf{x} | \Theta) = 1.$$

Clearly (4.84) defines a valid probability distribution for every value of  $\Theta$ ; however, in many cases it would be prohibitive to compute the value of  $k(\Theta)$ . This has led to MCMC methods for dealing with (4.84), both for estimating the parameters from a given observation of the graph as well as to simulate graphs drawn from a particular setting of the parameters (Lusher *et al.*, 2012). As a consequence, statisticians began to take a closer look at the Markov models as special cases of ERGMs, and things turned out not to be as clear as might have been hoped. The basic problem with the Markov models was that for certain reasonable parameter values most of the probability mass was distributed, sometimes over a small subset of possible graphs with multiple modes. As a consequence, special forms of parameterizing the substructures were developed that removed some of these problems. Section 4.4 will provide references for further study of these points.

Before ending the discussion of ERGMs, it should be noted that there are some substructures that could be used in (4.84) that would appear on the surface not to be ones that would emerge from cliques of the dependency graph in (4.81). For example, suppose one was interested in using 2-stars that did not form a triangle. Such a special 2-star would have the form  $\{X_{ij} = 1, X_{ik} = 1, X_{jk} = 0\}$ , and clearly  $\{X_{ij}, X_{ik}, X_{jk}\}$  would be a clique in the dependency graph. Assuming homogeneity with a single parameter  $\eta_W$  for these special 2-stars, then in the H-C expansion of (4.84) one would have

$$Q(\mathbf{x}) = \eta_A \sum_{1 \leq i < j \leq N} x_{ij} + \eta_W \sum_{1 \leq i < j < k \leq N} x_{ij}x_{ik}x_{jk}.$$

However, unfortunately, each product for a special 2-star that does not form a triangle  $x_{ij}x_{ik}x_{jk}$  is zero, so such structures would not appear in the H-C form in (4.58).

There is a happy solution to the problem by computing the Boolean conjunctive normal form for the special two star structure, namely

$$X_{ij}X_{ik}(1 - X_{jk}) = X_{ij}X_{ik} - X_{ij}X_{ik}X_{jk}.$$

This expression has the value 1 if and only if the realizations of these three random variables form a special 2-star. Let  $w(\mathbf{x})$  be the number of special 2-stars,  $s_2(\mathbf{x})$  be the number of (ordinary) 2-stars, and  $t(\mathbf{x})$  be the number of triangles in a realization  $\mathbf{x}$ . It is clear that  $w(\mathbf{x}) = s_2(\mathbf{x}) - 3t(\mathbf{x})$ . The way to represent this ERGM as deriving from the H-C theorem is now possible. The cliques of the dependency graph include all edge random variables and all triangles as in the PGM in (4.83) and all two edge cliques of the form  $\{X_{ij}, X_{ij}\}$ . Then the H-C expansion for the special 2-star ERGM has the form

$$P(\mathbf{x} \mid \Theta = (\eta_A, \eta_W)) = P_\Theta(\mathbf{0}) \exp\{\eta_A a(\mathbf{x}) + \eta_W s_2(\mathbf{x}) - 3\eta_W t(\mathbf{x})\}, \quad (4.85)$$

with parameter  $\Theta = (\eta_A, \eta_W) \in \Omega_\Theta = \mathbb{R}^2$ .

It is a reasonable conjecture that other substructures that are defined by presence/absence patterns like the one for special 2-stars in (4.85) could be handled by computing the disjunctive terms of the conjunctive normal form and basing cliques on them as in (4.58), equating parameters when necessary.

### 4.3.4 Conditionally uniform models

In the previous section, probabilistic network (graph or digraph) models were developed using the H-C theorem, and this development gave rise to the class of models known as exponential random graph models (ERGMs). Recall that the way that the number of parameters in an ERGM was reduced to manageable proportions was to impose the regularity condition, namely, the requirement that all isomorphic graphs or digraphs have the same probability. In this section, a related restriction is placed on graphs leading to a class of conditionally uniform models.

Conditionally uniform models start with selecting a set of graph statistics just as in the case of ERGMs; however, instead of using a given empirical network to estimate the parameters of an ERGM, one works directly with the observations of the selected graph statistics for the empirical network. The condition that is imposed on conditionally uniform models is that all graphs that have the same observations of the selected statistics as does the empirical network are equally probable, and graphs that do not share these same values have probability zero. Given this condition, it is obvious that graphs that are isomorphic to a given empirical graph are equally probable; however, many more graphs that are not isomorphic may also have the same probability as the given graph. The reason is that, as discussed in

section 4.2.1, for two graphs or digraphs to be isomorphic they must have identical values of all structural statistics, not just the ones that one might select for a particular conditionally uniform model.

A simple example of a conditionally uniform model due to Erdős and Rényi (1959) is to fix the number of edges for a graph to be  $m^*$ , where  $0 \leq m^* \leq \binom{N}{2}$ , and require that all such graphs be equally likely, while graphs with a different number of edges are assigned zero probability. From the previous work with the Bernoulli PGM, denoted by  $G(N, p)$  in Example 4.3, it is easy to see that there are  $\binom{N(N-1)/2}{m}$  graphs on  $N$  nodes with exactly  $m$  edges, and each of these graphs has a degree sum of  $d = 2m$ . Consequently, we have the probability distribution for this model given for each  $\mathbf{x} \in \Omega_{\mathbf{X}}$  by

$$P(G(\mathbf{x}) \mid m) = \begin{cases} \left( \frac{N(N-1)/2}{m^*} \right)^{-1}, & \text{if } m = m^*, \\ 0, & \text{otherwise.} \end{cases} \quad (4.86)$$

This probability model is called the *uniform random graph model*, and is denoted by  $G(N, m^*)$ .

Another PGM would be to fix a given graphical  $N$ -vector of nodal degrees  $\mathbf{d}$  and require that all graphs with this degree sequence to be equally likely. This random graph model is denoted by  $G(\mathbf{d})$ . Note that this is a further restriction than merely fixing the number of nodes and edges as in  $G(N, m^*)$ . This is because, given the degree sequence  $\mathbf{d}$ , the number of nodes is the length of  $\mathbf{d}$ , and the number of edges is half the sum of  $\mathbf{d}$ . While it may be difficult to determine the exact number of graphs with the same degree sequence, there are asymptotic approximations (McKay and Wormald, 2013). Given the extreme departures from the binomial distribution of the degree sequences of the empirical networks examined in Section 4.2.3, it might be important to control for degree sequence. Later examples will confirm this suspicion.

In the case of digraphs, Example 4.5 provides a multinomial model in (4.78) that classifies dyads into mutual, asymmetric, or null. Suppose a given empirical digraph  $D(\mathbf{y}^*)$  on  $N$  nodes has  $m^*$  mutual dyads,  $a^*$  asymmetric dyads, and  $n^* = N(N - 1)/2 - (m^* + a^*)$  null dyads. The triad of these counts is called the MAN-statistic. In this case,  $\text{MAN}(\mathbf{y}^*) = (m^*, a^*, n^*)$ . Then a conditionally uniform PDM based on  $D(\mathbf{y}^*)$  is given by

$$P[D(\mathbf{y}) \mid \text{MAN}(\mathbf{y}^*)] = \begin{cases} \left( \frac{N(N-1)/2}{m^* a^* n^*} \right)^{-1}, & \text{if } \text{MAN}(\mathbf{y}) = \text{MAN}(\mathbf{y}^*) \\ 0, & \text{otherwise.} \end{cases} \quad (4.87)$$

The model in (4.87) was proposed originally by Holland and Leinhardt (1976).

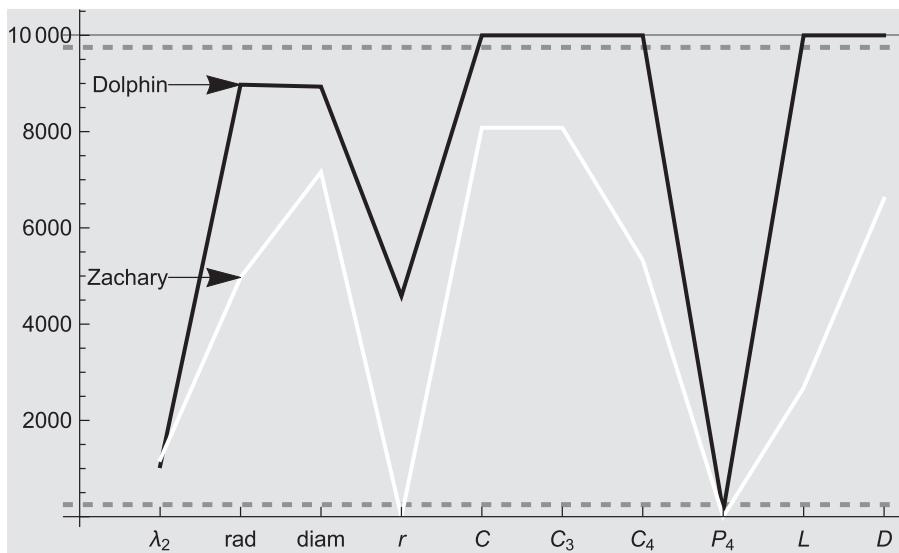
One use of conditionally uniform models is to develop nonparametric, null hypotheses for testing hypotheses about the frequency of occurrence of some particular type of subgraph structure in a given empirical graph. For example, suppose that one was interested in whether or not the frequency of triangles in an empirical graph was unusually high or low. Clearly, other things being equal, graphs with a

large number of edges  $M$  are likely to have more triangles than are graphs with small  $M$ . Thus, one might want to control for number of edges by using the conditionally uniform model  $G(N, m^*)$ . Here are the steps that could be employed. First, one would take an empirical graph  $G(\mathbf{x}^*)$  on  $N$  nodes and calculate the number of edges  $m^*$  and the number of triangles  $t^*$ . Then one would randomly simulate many graphs from the model in (4.86), and for each such simulated graph count the number of triangles. From this sample, one could construct an empirical cumulative distribution of triangle counts  $F(t \mid m^*)$ , the proportion of samples each with  $m^*$  edges and that have a  $t$ -statistic (triangle count) less than or equal to  $t$ , for  $0 \leq t \leq N(N - 1)(N - 2)/6$ . Finally, one could see where in this random distribution the observed value  $t^*$  falls. This would yield a  $P$ -value, namely,  $F(t^* \mid m^*)$ , and a usual two-tailed hypothesis test could be performed.

The key to the test above is the ability to simulate graphs from the model in (4.86). This would be computationally quite easy, as one could order the  $N(N - 1)/2$  potential edges, create a 0–1 vector by placing a 1 in the slots corresponding to the edges present in the given graph  $G(\mathbf{x}^*)$  and a 0 in the other slots. Then one would select a series of random permutations of the 0–1 vector, and for each calculate the  $t$ -statistic. The speed of this method depends on the best algorithms for permutations, which are  $O(N^2)$ . Alternatively, one could select random subsets of  $M$  edges, which can be done in  $O(M)$  time. Since most networks are rather sparse,  $M \ll N(N - 1)/2$ , the subset selection method is much faster than the permutation method. Algorithms for obtaining random permutations and random subsets are readily available (Knuth, 1973; Nijenhuis and Wilf, 1978). Software for generating uniform random graphs for large  $N$  and  $M$  are available in most software packages relating to graphs (Knuth, 1994; Batagelj and Mrvar, 2011; Wolfram Research, 2014).

Unfortunately, unlike the simple example above, the choice of other statistics to define a conditionally uniform model can lead to quite complex situations where explicit derivations are not possible. In some cases, as described in the next example, sophisticated computational techniques can be developed to handle sets of interesting statistics. This case is where the statistics that are controlled for are the degree sequences for each node in a graph.

Modern computer programs like *Mathematica*<sup>®</sup> can directly generate random graphs that preserve certain important properties, such as the number of edges or the degree sequence. This is desirable because one wants to compute graph measures or motifs and then determine if these values are more frequent (or less frequent) than is found in random samples of graphs of the same general type as the graph observed in a given network. As a minimum, one would like to preserve the number of nodes and edges. This is easily done, but leads to a binomial distribution of degrees, which is not characteristic of most observed graphs. The next step would be to control for the degree sequence itself. Algorithms to compute uniform random samples with a given degree sequence are well-known (Del Genio *et al.*, 2010) and implemented, for example, in *Mathematica*<sup>®</sup>.



**Figure 4.28** Plots of dolphin and Zachary karate measures and motifs. In each case, 10,000 random graphs were generated preserving the degree sequences. The height represents the number of random graphs that had measures or motif counts that were less than or equal to those of the dolphin or Zachary graphs. The upper and lower dotted gray lines represent the 0.975 and 0.025 two-sided *p*-values, respectively.

**Example 4.6** (Karate and dolphin profiles) Figure 4.28 shows the result of 10,000 random samples of graphs preserving the dolphin and karate degree sequences. The plot represents the number of random graphs with the given degree sequence with measures or motifs less than the given graph. The five measures are:

1. The second smallest Laplacian eigenvalue  $\lambda_2$ , also called the algebraic connectivity.
2. The graph radius,  $\text{rad}$ .
3. The graph diameter,  $\text{diam}$ .
4. The *Global Clustering Coefficient*,  $C$ , defined as the fraction of paths  $P_3$  whose nodes induce a triangle  $C_3$  divided by the total number of paths on three nodes. It is a measure of the degree of transitivity of the graph.
5. *Assortativity*, denoted by  $r$ , is in fact the Pearson correlation coefficient between the degrees of adjacent nodes. A positive  $r$  suggests, for example, a network where popular people tend to be friends with other popular people. On the other hand, a negative  $r$  suggests a network where popular people tend to associate with unpopular people, implying an asymmetry as is seen, for example, with the relation between gurus and their acolytes.

Next, the five motifs are the cycles  $C_3$  and  $C_4$ , the path  $P_4$ , the lollipop  $L$ , and the diamond  $D$ , all shown in Figure 4.2. Why did we not include the path  $P_3$ ? The

reason is that the number of subgraphs (the general type of subgraphs, not *induced* subgraphs) isomorphic to  $P_3$  is  $\sum_i^N \binom{d_i}{2}$ , so if the degree sequence is fixed, then, as the reader should try to show, the number of such subgraphs is also fixed. Similarly, the star with three edges  $S_3$  was not included because its count is given by  $\sum_i^N \binom{d_i}{3}$ .

In Figure 4.28 we see the results of two runs of 10,000 random graphs preserving the degree sequences of the dolphin and Zachary networks. The 20 points in the plot are connected by white lines for the dolphin data, and black lines for the Zachary data. Each point represents the number of random graphs (preserving the degree sequence) that are less than or equal to the observed data on the given measure. Division by 100 gives its empirical percentile. For the first measure,  $\lambda_2$  (the second smallest Laplacian eigenvalue), note that both the dolphin and the Zachary plots have the same value of about 1000. This means that of the 10,000 random graphs, 10%, had smaller or equal values of  $\lambda_2$ . Note that the actual values of  $\lambda_2$  are 0.17297 and 0.46852, respectively, so we are not plotting the values themselves, but rather their percentile position in their respective samples. However, neither the dolphin nor the Zachary  $p$ -values of  $\lambda_2$  are outside the two-sided range for  $p = 0.95$ , indicated by the two dotted lines. In the same way the plot indicates that neither the graph radius,  $\text{rad}$ , nor the graph diameter,  $\text{diam}$ , differs from what you would expect from random graphs with the same degree sequences.

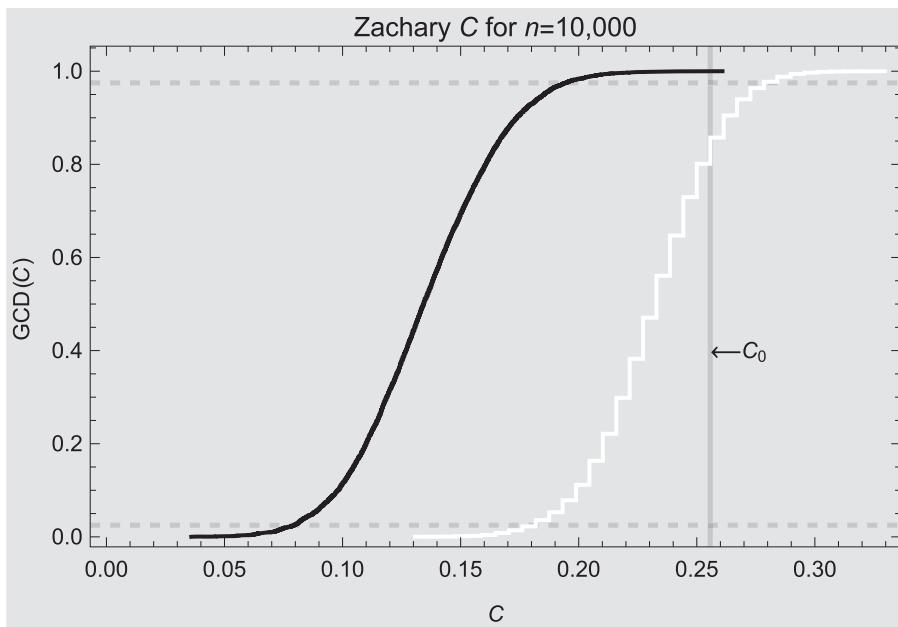
The next measure, the assortativity index  $r$ , is in the middle of the range for the dolphins but is outside the 95% range for Zachary's karate students. In fact, there are no cases out of 10,000 random graphs with the karate degree sequence that had a lower  $r$ , so it really blows past the 95% barriers. Looking at Figure 4.9 you can see that there is a negative correlation between one end of an edge and the other: the high degree nodes are linked with low degree nodes, the "acolytes."

On the other hand, the global clustering index  $C$  is extremely high for dolphins, meaning that most of the paths  $P_3$  are subgraphs of a cycle  $C_3$ . That is, for dolphins, a friend of a friend usually is a friend. Again, all the 10,000 random graphs have a lower  $C$ , so this result is way more significant than the modest 95% often used in social science. This conclusion is reinforced by just counting the number of triangles  $C_3$ : the dolphins are again at the 100th percentile. Dolphins are also at the 100th percentile on cycles  $C_4$ .

However, both dolphins and karate students are near zero on paths on four nodes: the dolphins do have 108 random graphs with a lower numbers of  $P_4$ , while the Karate students have none. While the number of such paths might seem rather large, 5023 and 2371, respectively, these numbers are small compared to the randomly generated graphs.

Finally, the dolphins are near to the 100th percentile in the lollipop and diamond graphs. Perhaps dolphins prefer motifs containing closed cycles? That is, the lollipop graph contains the subgraph  $C_3$ , while the diamond graph contains  $C_4$  as well as two copies of  $C_3$ .

**Example 4.7** (Conditionally uniform models with  $C$  and  $r$  on all our network examples) We will present examples of two PGMs applied to our four previous



**Figure 4.29** Empirical cdfs of the global clustering coefficient  $C$  for random graphs with the uniform (black line), and degree sequence (white line) models for the Zachary karate network. The white line is jagged because there are fewer distinct graphs with the given degree sequence.

network examples using two important statistics for the study of networks. The two statistics are the global clustering coefficient  $C$  and the assortativity index  $r$ . The two PGMs are the uniform graph model and the model preserving the degree sequences of each of the four networks. We generated 10,000 random graphs preserving each of these two properties. In Figure 4.29 we examine in detail only  $C$  for only the Zachary karate student data. We plot the entire empirical distribution function for the 10,000 values of  $C$  under the two conditionally uniform models, the uniform graph model, which preserves only the number of edges, and the degree sequence model, which preserves  $\mathbf{d}$ . The black line is the cdf for the uniform graph model. Note that all but one or two of the random graphs have a value of  $C$  that is greater than the observed value,  $C_0 = 0.256$ . This value is indicated by the vertical gray line. The conclusion under this model is that the global clustering coefficient for the Zachary data has a two sided  $p$ -value less than 0.05, indicated by the two dotted lines.

However, the degree sequence model gives a different conclusion. If we look at the white line in Figure 4.29, we see that it intersects the vertical gray line at about 0.86, indicating that the global clustering coefficient does not differ from chance. Because there are fewer graphs that match a given degree sequence than match the number of edges, the white line is jagged because of the presence of more tied values of  $C$ .

Table 4.1 *Measures C and r for network examples.*

Data	Measure	Value	Uniform	Degree
Zachary	$C$	0.256	100	86
	$r$	-0.476	0	0
Dolphin	$C$	0.309	100	100
	$r$	-0.044	50	45
Disease	$C$	0.850	100	100
	$r$	0.652	45	83
Internet	$C$	0.011	100	100
	$r$	-0.198	0	0

Next, the  $p$ -values for both  $C$  and  $r$  are tabulated for all four data sets in Table 4.1. The results are 10,000 random graphs for each of the four data sets. The two measures are  $C$ , the global clustering coefficient, and  $r$ , the graph assortativity measure. The “Value” column is the value of these measures. The “Uniform” column gives the percent of uniform random graphs that have a measure less than or equal to the value. Finally, the “Degree” column gives the percentile result for random graphs with the same degree sequence as the data. The global clustering coefficients  $C$  are significant for all four data sets under the uniform graph model, and for all but the Zachary data (as discussed in the previous paragraph). That is, transitivity holds in all but one case. Looking at the  $p$ -values for the assortativity index  $r$ , however, a more complex situation is revealed. For both the Zachary and Internet data,  $r$  significantly *less* than one would be expected under both random models. As mentioned above, the interpretation for the low  $r$  for the Zachary data is that high-status individuals (as measured by their degree) tend to have ties with lower-, rather than higher-, status individuals. Similarly, the disease gene network connects diseases of high degree with mutated versions of low degree. Concerning the other examples, dolphins and the Internet have a tendency toward homophily: like attracting like.

#### 4.4 Further topics

One type of network that has recently received much attention in the literature is called a *small-world* network, originally suggested by Milgram’s experiment (Milgram, 1967). This type of network is characterized by having a high global clustering coefficient and a small diameter. This means that it is highly transitive while making all nodes relatively close to each other. The intuition behind this model is that in many large networks, say people and their friendship networks, there seem to be two seemingly inconsistent tendencies: first, a friend of a friend is usually a friend, and second, any two people are usually joined by a path of length six or less, the so-called “six-degrees of separation” phenomenon.

At least three ways of randomly generating small-world graphs have been proposed, all of which are implemented in *Mathematica*<sup>®</sup>. The idea is to generate many such random graphs and then examine how closely they capture the essence of a given network. These small-world models, and many others, are well-covered in survey books (Newman, 2010; Estrada, 2011).

The first of these three models was introduced by Price (1976) and has three parameters:  $N$ , the number of nodes;  $k$ , the number of arcs added (if possible) at each step, and  $a \in \mathbb{R}_{>0}$ , an attractiveness parameter. Step 1 starts with the digraph with a single node, 1. Step  $i$ , for  $1 < i \leq k + 1$ , adds node  $i$  and arcs from  $i$  to each of the preceding nodes. The digraph after step  $k + 1$ , consists of all arcs  $i \rightsquigarrow j$ , where  $k + 1 \geq i > j \geq 1$ . Each of the next steps, step  $i$ , for  $k + 1 < i \leq N$ , adds node  $i$  and exactly  $k$  arcs from  $i$  to the previous nodes. These  $k$  arcs are randomly sampled without replacement (to avoid duplicating arcs) from the possible arcs with probability proportional to weights  $q_j + a$ , where  $q_j$  is the in-degree of node  $j$ , where  $j < i$  and  $j$  has not been previously sampled, and where  $a$  is an attractiveness parameter. The resulting digraph is asymmetric and has  $kN - \frac{k(k+1)}{2}$  arcs. Often this digraph is transformed into a graph by replacing each arc  $i \rightsquigarrow j$  with edge  $i \sim j$ .

Let us examine the Price model for two extreme values of the *attractiveness* parameter. In both cases let  $N = 100$  and  $k = 3$ : if  $a = 0.0001$ , then the first 3 nodes receive almost all the arcs, usually resulting in an in-degree distribution of  $(99, 98, 97, 0, \dots, 0)$ , while if  $a = 1000$ , the effect of the  $q_j$  are swamped out and the resulting in-degree distribution looks more like a negative binomial distribution with parameters  $n \approx 1$  and  $p \approx 0.25$ . However, this model captures both a general attractiveness parameter  $a$  and a node-specific parameter  $q_i$  proportional to its degree. It has been useful in creating power law degree distributions.

The next such model was proposed by Watts and Strogatz (1998). A graph is constructed starting from a circular embedding of  $N$  nodes. These nodes are then connected to, say,  $2k$  of their nearest neighbors in the plane. Then each edge is randomly “rewired,” one endpoint is changed to another with probability  $p$ , making sure that no loop or multiple edge is created. The vertex count is preserved, but it is possible to create isolated nodes. However, with a relatively small  $p$  it is possible to generate graphs that satisfy the criteria for small-world graphs.

Soon after the Watts–Strogatz model appeared, a third small-world model was proposed by Barabási and Albert (1999). This construction starts from the triangle  $C_3$ , and a vertex with, say,  $k$  edges is added at each step. The  $k$  edges are attached to nodes at random, following a distribution proportional to the node degree. This model also generates small-world type graphs, but has the additional virtue of creating preferential attachment resulting in a degree sequence that follows a power law for large  $N$ .

There are several obvious limitations of the network models presented in this chapter and which predominate in the field. The first limitation is that the nodes and links, once generated, do not change with time. The dynamics of networks,

once elucidated, may shed light on their statics. The second limitation is that nodes have intrinsic, non-network, properties that may affect their role in the network. These properties may range from relatively static “traits,” such as sex or age, to variable “states,” like being “depressed” or “high on heroin.” In many applications of network science, these nodal covariates are incorporated in the graphical model. The third limitation is that ties as presented here are either “on” or “off.” However, many ties can be described on some sort of scale. These nodal and link properties may influence and be influenced by network variables such as centrality or by the nodal properties of the graph neighborhood. A leader in this expansion of the scope of network analysis is Snijders (2014), who has laid out a series of rigorous statistical models of longitudinal network data including nodal variables such as demographics and behavior. In these models the network may influence the dynamics of behavior, and the behavior may influence the dynamics of the network. In other words, this describes the co-evolution of networks and behavior. These models are freely available in a program called RSIENA or SIENA 4 (Ripley *et al.*, 2014).

The problem of finding graph isomorphisms (GI), discussed briefly in section 4.2.1, is very difficult in general and is the subject of intense competition to find the best methods. The fastest proven running time for GI is  $e^{O(\sqrt{N \log N})}$  (Babai *et al.*, 1983), “despite many published claims” of linear time algorithms (McKay and Piperno, 2013). On the other hand, polynomial time algorithms are known for many special classes of graphs, such as connected graphs with bounded degree (Luks, 1982). For example, the best algorithm in Luks (1982) for *trivalent* graphs, where every node has degree three, is  $O(N^5)$ . Two of the best computer programs for the general graph isomorphism problem are *nauty* and *Traces*, both developed by McKay and Piperno (2013), which can be downloaded for free at <http://cs.anu.edu.au/bdm/nauty/>.

The emphasis in this chapter has been on graphs, but many empirical relational structures fail to be symmetric. For example, a relation may occasionally fail to be symmetric, as in unrequited love, or it may never be symmetric, as in the outcome of tournaments that do not allow ties. These structures may be modeled with digraphs if they are irreflexive. An excellent introduction to the theory of digraphs is found in Bang-Jensen and Gregory (2001).

A difficulty in working with the adjacency matrix of a digraph is that many of the eigenvalues may be complex. Recall that a complex number is of the form  $a + bi$ , where  $a, b \in \mathbb{R}$ , and where  $i = \sqrt{-1}$ . While complex eigenvalues and vectors are very useful in such fields as electrical engineering and quantum mechanics, some of the nice properties, such as those found in the Spectral Theorem 4.5 and the Perron–Forbenius Theorem 4.7, of symmetric matrices are lost. A surprising solution to this problem is to recode a real-valued adjacency matrix  $\mathbf{A}$  into a complex-valued matrix  $\mathbf{H}$  such that the eigenvalues are nonnegative (Hoser and Geyer-Schulz, 2005; Hoser and Bierhance, 2007). The complex-valued matrix is given by  $\mathbf{H} \stackrel{\text{def}}{=} [\mathbf{A} + \mathbf{A}^T - (\mathbf{A} - \mathbf{A}^T)i]/\sqrt{2}$ . The beauty of this transformation

is that  $\mathbf{H}$  is a *Hermitian*, meaning that if  $h_{jk} = a + bi$ , then  $h_{kj} = a - bi$ . The Hermitian matrix  $\mathbf{H}$  has real (positive and negative) eigenvalues, although the eigenvectors may contain complex numbers. However, this approach has been found to be useful in clustering of digraphs.

In Section 4.3, probabilistic graphs and digraphs were discussed. This is a very large topic in network science, and our presentation only covered one facet of the topic. In particular, we assumed a graph or digraph with a fixed, finite set of nodes. Then the possible edges or arcs were treated as dichotomous random variables. The stress was placed on using the Hammersley–Clifford theorem to allow one to construct probabilistic graph or digraph models with preselected types of dependency structures among the random variables. This approach originated first in network science with the paper by Frank and Strauss (1986), and it immediately motivated generalizations that led to the family of exponential random graph models (ERGMs) presented in Example 4.5 and Equation (4.84). In addition, it was soon discovered that the estimation theory for ERGMs was more complicated than originally thought, as briefly described after the probability distribution for an ERGM was presented in Equation (4.84). These problems led to some new approaches to estimation as well as some modifications in the models, e.g., Hunter and Handcock (2006) and Snijders *et al.* (2006). A good source for additional references and software packages for ERMGs is Lusher *et al.* (2012).

There are several good books that discuss probabilistic models with an underlying graph-theoretic basis, e.g., Grimmett (2010) and Koller and Friedman (2009). Some of these models depart from assumptions we imposed as described above. For example, in many cases the number of nodes can be infinite or even governed by a random variable. One example of the latter case is the Bernoulli random graph  $G(N, p)$ , where edges between pairs of nodes are independent and identically distributed with probability  $p$ , and the number of nodes is treated as a random variable. While this assumption is unrealistic for naturally occurring networks, interesting results are nevertheless found for these models in probability theory. For example, as a function of  $N$  and  $p$ , one is interested in the probability that the graph satisfies certain properties, such as being connected. In addition, results are known for run-time complexity to detect various graph properties for the Bernoulli graph with a variable number of nodes that help shed some light on the complexity of certain related algorithms for more general graphs.

Another topic concerns random walks on the nodes of a fixed graph or digraph, where the next step from a particular node in the walk is a random choice among the accessible nodes that follow from edges or out-arcs. This area is a generalization from the usual study of random walks on a multidimensional checkerboard design. The assumption about next moves makes the random walk a Markov chain, and this results in a desire to seek conditions under which the chain is reversible. This is an important topic behind Markov Chain Monte Carlo simulation discussed briefly for ERGMs, and more generally it is behind most approaches to Bayesian inference of complex parametric models both inside and outside of network theory.

Another generalization of probabilistic graphs is to associate random variables with the nodes of a digraph. This assumption is the underpinning of Bayesian networks, e.g., Neapolitan (2004). The random variables can be discrete or continuous, and a directed arc goes from one random variable to another if the second has a probability distribution that is conditionally dependent on the first. The random variables are organized as a directed acyclic graph (DAG), namely, a directed graph with no directed cycles. For example, in a medical application the random variables could refer to symptoms and diseases. Then the probabilistic structure could compute the probability distribution over diseases given any pattern of symptoms.

It turns out if the probabilistic structure of the random variables in the DAG satisfies a Markov condition, a lot of computational power is achieved. To state the condition we need some terminology. Suppose  $u$  and  $v$  are two nodes in the DAG; if there is an arc from  $u$  to  $v$ , then  $u$  is called a *parent* of  $v$  and  $v$  a *child* of  $u$ . If instead there is a path of length two or more from  $u$  to  $v$ , then  $u$  is an *ancestor* of  $v$  and  $v$  a *descendent* of  $u$ . Now the *Markov condition* holds on the DAG,  $D = (V, A)$ , if for each random variable  $X$  in  $V$ ,  $\{X\}$  is conditionally independent of the set of all of its nondescendents given the set of all its parents. Many theorems and computational shortcuts are available if a Bayesian network DAG satisfies the Markov condition, e.g., Neapolitan (2004).

For further reading about networks we suggest perusing articles in *The Journal of Social Networks*. This journal, together with the professional association INSNA: *International Network for Social Network Analysis* and the annual *International Sunbelt Social Network Conference*, forms the intellectual hub of social networks. Also, the listserv, SOCNET, provides a forum for information and discussion about social networks. A more recent journal, *Network Science*, also has outstanding articles and in addition has the virtue of free online access at <http://journals.cambridge.org/action/displayJournal?jid=NWS>.

Another source for additional information is the special issue of the *Journal of Mathematical Psychology*, especially the methods tutorial by Robins (2013). According to the *Web of Science*, as of March 31, 2015, the most cited paper with the topic “Social networks” is the Watts and Strogatz (1998) article “Collective dynamics of ‘small-world’ network” in *Nature*. This paper is cited 10,116 times, while the second-most cited paper is “only” cited 3118 times. Journals whose network articles emphasize mathematical models of large networks include the *Physical Review Letters*, the *Journal of Physics A: Mathematical and Theoretical*, the *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, the *Physica A: Statistical Mechanics and its Applications*, and the journal *Chaos*.

One of the top journals of graph theory is, well, the *Journal of Graph Theory*, as well as the *Journal of Combinatorial Theory, Series B*. The journal/magazine *IEEE Computer* emphasizes computational complexity and other aspects of theoretical computer science.

Wasserman and Faust (1994) is still the best introductory book on social networks, while more general networks with more advanced methods are covered by books by the Newman (2010) and Estrada (2011).

## Notation

Expression	Meaning
$\mathbb{R}, \mathbb{R}_{\geq 0}, \mathbb{R}_{> 0}$	reals, nonnegative reals, positive reals
$\mathbb{Z}, \mathbb{Z}_{\geq 0}, \mathbb{N}$	integers, nonnegative integers, naturals
$\cap, \cup, \setminus$	set intersection, union, relative complement
$\Delta, S^c$	symmetric difference, complement
$\emptyset$	the empty set
$[k]$	$\{i \in \mathbb{N} \mid i \leq k\}$ N.B. $ [k]  = k$
$I_S$	indicator function: $I_S(x) = 1$ if $x \in S$ else 0
$f: X \rightarrow Y: x \mapsto y$	function from $X$ to $Y$ : $x$ maps to $y$
$\{A_i\}_{i \in I}$	$A: I \rightarrow U$ a family of sets indexed by $I$
$x \sim y, x \rightsquigarrow y$	$x$ is related to $y$ in a graph or digraph
$N(v), N^-(v), N^+(v)$	neighborhood, in-, out-, of node $v$
$d(v), d^-(v), d^+(v)$	degree, in-, out- of node $v$
$f(x) = O(g(x))$	$g$ is an asymptotic upper bound for $f$
$\mathbf{v}, \mathbf{v}^T, \mathbf{A}$	column vector, row vector, matrix
$\mathbf{A}\mathbf{v}$	matrix times a column vector
$\mathbf{AB}$	matrix multiplication
$\mathbf{A}^T$	matrix transpose
$\mathbf{u}^T \mathbf{v}$	inner product of vectors
$\Omega$	sample space
$\mathcal{X}$	a collection of random variables
$P_X(x) \stackrel{\text{def}}{=} P(X = x)$	$P(X^{-1}\{x\}) \stackrel{\text{def}}{=} P(\{\omega \in \Omega \mid X(\omega) = x\})$
$P_X(E) \stackrel{\text{def}}{=} P(X^{-1}E)$	probability of an event $E \subseteq \mathbb{R}$
$F_X(x) \stackrel{\text{def}}{=} P(X \leq x)$	cumulative distribution function (cdf)
$f_X(x) \stackrel{\text{def}}{=} \frac{d}{dx}F_X(x)$	density function (pdf) for a continuous r.v.
$p_X(x) \stackrel{\text{def}}{=} P_X(x)$	prob. mass function (pmf) for a discrete r.v.
$P(X = x \mid Y)$	conditional probability that $X = x$ given $Y$
$X \perp Y \mid Z$	conditional independence of $X$ and $Y$ given $Z$
$\text{Cl}_X$	set of cliques of a dependency graph $X$
$\square$	QED

## References

- Babai, L., Kantor, W. M. and Luks, E. M. (1983). Computational complexity and the classification of finite simple groups. In *Foundations of Computer Science, 1983, 24th Annual Symposium on*. New York, NY: IEEE, pp. 162–171.
- Bang-Jensen, J. and Gregory, G. (2001). *Digraphs: Theory, Algorithms and Applications*. London: Springer-Verlag.
- Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**, 509–512.

- Batagelj, V. and Mrvar, A. (2011). *Pajek*.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B (Methodological)*, **36**, 192–236.
- Büyükoğlu, T., Leydold, J. and Stadler, P. F. (2007). *Laplacian Eigenvectors of Graphs*. Berlin: Springer-Verlag.
- Bollobás, B. (1998). *Modern Graph Theory*. New York, NY: Springer-Verlag.
- Bonacich, P. (1987). Power and centrality: a family of measures. *American Journal of Sociology*, **92**, 1170–1182.
- Borgatti, S. P. and Everett, M. G. (2000). Models of core/periphery structures. *Social Networks*, **21**, 375–395.
- Choudum, S. A. (1986). A simple proof of the Erdos–Gallai theorem on graph sequences. *Bulletin of the Australian Mathematical Society*, **33**, 67–70.
- Chung, F. R. K. (1997). *Spectral Graph Theory*. Providence, RI: American Mathematical Society.
- Cvetković, D., Rowlinson, P. and Slobodan, S. (2010). *An Introduction to the Theory of Graph Spectra*. London: Mathematical Society.
- Del Genio, C. I., Kim, H., Toroczkai, Z. and Bassler, K. E. (2010). Efficient and exact sampling of simple graphs with given arbitrary degree sequence. *PLoS ONE*, **5**(4).
- Diaconis, P. (2003). Patterns in eigenvalues: the 70th Josiah Willard Gibbs Lecture. *Bulletin of the American Mathematical Society*, **40**, 155–178.
- Erdős, P. and Gallai, T. (1960). Graphs with prescribed degree of vertices. *Mat. Lapok*, **11**, 264–274.
- Erdős, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae Debrecen*, **6**, 290–297.
- Estrada, E. (2011). *The Structure of Complex Networks: Theory and Applications*. Oxford: Oxford University Press.
- Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, **81**, 832–842.
- Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, **1**, 215–239.
- Freeman, L. C. (2004). *The Development of Social Network Analysis*. Vancouver: ΣP Empirical Press.
- Goh, K., Cusick, M. E., Valle, D., et al. (2007). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 8685–8690.
- Greenacre, M. J. (2007). *Correspondence Analysis in Practice*. Second edn. London: Chapman & Hall/CRC Interdisciplinary Statistics.
- Grimmett, G. (2010). *Probability on Graphs: Random Processes on Graphs and Lattices*. Cambridge: Cambridge University Press.
- Harary, F. (1969). *Graph Theory*. Reading, MA: Addison-Wesley.
- Holland, P. W. and Leinhardt, S. (1976). Conditions for eliminating intransitivities in binary digraphs. *Journal of Mathematical Sociology*, **4**, 315–318.
- Hoser, B. and Bierhance, T. (2007). Finding cliques in directed weighted graphs using complex hermitian adjacency matrices. In Decker, R. (ed.), *Advances in Data Analysis*. Berlin: Springer, pp. 83–90.
- Hoser, B. and Geyer-Schulz, A. (2005). Eigenspectral analysis of hermitian adjacency matrices for the analysis of group substructures. *Journal of Mathematical Sociology*, **29**, 265–294.

- Hunter, D. R. and Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, **15**, 565–583.
- Kac, M. (1966). Can one hear the shape of a drum? *The American Mathematical Monthly*, **73**(4), 1–23.
- Knuth, D. E. (1973). *The Art of Computer Programming. 1. Fundamental Algorithms*. Reading, MA: Addison-Wesley.
- Knuth, D. E. (1994). *The Stanford GraphBase: a platform for combinatorial computing*. Reading, MA: Addison-Wesley.
- Kocay, W. and Kreher, D. P. (2005). *Graphs, Algorithms, and Optimization*. London: Chapman & Hall/CRC.
- Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.
- Koskinen, J. H., Wang, P., Robins, G. L. and Pattison, P. E. (2008). Extreme Actors-Outliers and Influential Observations in Exponential Random Graph (p-star) Models. *MelNet Social Networks Laboratory Technical Report*, 08–05.
- Krackhardt, D. (1987). Cognitive social structures. *Social Networks*, **9**, 109–134.
- Luks, E. M. (1982). Isomorphism of graphs of bounded valence can be tested in polynomial time. *Journal of Computer and System Sciences*, **25**, 42–65.
- Lusher, D., Koskinen, J. and Robins, G. (eds). (2012). *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge: Cambridge University Press.
- Lusseau, D., Schneider, K., Boisseau, O. J., et al. (2003). The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, **54**, 396–405.
- McKay, B. D. and Piperno, A. (2013). Practical graph isomorphism, II. *arXiv preprint arXiv:1301.1493*.
- McKay, B. D. and Wormald, N. C. (2013). Asymptotic enumeration by degree sequence of graphs of high degree. *European Journal of Combinatorics*, **11**, 565–580.
- Milgram, S. (1967). The small world problem. *Psychology Today*, **2**, 60–67.
- Milo, R., Shen-Orr, S., Itzkovitz, S., et al. (2002). Network motifs: simple building blocks of complex networks. *Science*, **298**(5594), 824–827.
- Minc, H. (1985). *Nonnegative Matrices*. Chichester: Wiley.
- Moreno, J. L. (1934). *Who Shall Survive? Foundations of Sociometry, Group Psychotherapy, and Sociodrama*. Nervous and Mental Disease.
- Morgan, L. H. (1851/1997). *Systems of Consanguinity and Affinity of the Human Family*. Lincoln, NE: University of Nebraska.
- Neapolitan, R. E. (2004). *Learning Bayesian Networks*. Upper Saddle River, NJ: Prentice Hall.
- Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford: Oxford University Press.
- Nijenhuis, A. and Wilf, H. S. (1978). *Combinatorial Algorithms*. Second edn. New York, NY: Academic Press.
- Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, **27**, 292–306.
- Rapoport, A. (1963). Mathematical models of social interaction. In Luce, R. D., Bush, R. R., and Galanter, E. (eds), *Handbook of Mathematical Psychology*, vol. II. Chichester: Wiley, pp. 495–579.

- 
- Ripley, R., Boitmanis, K. and Snijders, T. A. B. (2014). *RSiena: Siena-simulation investigation for empirical network analysis. R package version 1.1-274/r274.*
- Robins, G. (2013). A tutorial on methods for the modeling and analysis of social network data. *Journal of Mathematical Psychology*, **57**, 261–274.
- Snijders, T. A. B. (2002). Markov Chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, **3**(2), 1–40.
- Snijders, T. A. B. (2014). Siena algorithms.
- Snijders, T. A. B., Pattison, P. E. and Robins, G. L. (2006). New specifications for exponential random graph models. *Sociological Methodology*, **36**, 99–153.
- Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.
- Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of “small-world” networks. *Nature*, **393**, 440–442.
- Wigner, E. P. (1958). On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, **67**, 325–327.
- Wolfram Research, Inc. (2014). *Mathematica*.
- Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, **33**, 452–473.

# 5 Knowledge spaces and learning spaces

Jean-Paul Doignon and Jean-Claude Falmagne

5.1	Origin and motivation	274
5.2	Knowledge structures and learning spaces	276
5.3	Knowledge spaces and wellgradedness	279
5.4	The base and the atoms	283
5.5	Surmise systems	285
5.6	The fringe theorem	287
5.7	Learning words and learning strings	288
5.8	The projection theorem	294
5.9	Probabilistic knowledge structures	299
5.10	The stochastic assessment algorithm	300
5.10.1	Sketch of the algorithm in the straightforward situation	301
5.10.2	The parallel algorithm for large domains	307
5.11	About building knowledge spaces or learning spaces	308
5.12	Some applications – the ALEKS system	313
5.12.1	The extra problem method	313
5.12.2	Correcting for careless errors	315
5.12.3	Learning readiness	316
5.12.4	ALEKS based placement at the University of Illinois	316
5.13	Bibliographical notes	317
	Acknowledgments	318
	References	318

## 5.1 Origin and motivation

Knowledge space theory (abbreviated as KST) originated with a paper by Doignon and Falmagne (1985). This work was motivated by the shortcomings of the psychometric approach to the assessment of competence. The psychometric models are based on the notion that competence can be measured, which the two authors thought was at least debatable. Moreover, a typical application of a psychometric model in the form of a standardized test results in placing an individual

in one of a few dozen ordered categories, which is far too coarse a classification to be useful. In the case of the SAT,<sup>1</sup> for example, the result of the test is a number between 200 and 800 with only multiples of 10 being possible scores.

In the cited paper, Doignon and Falmagne proposed a fundamentally different theory. The paper was followed by many others, written by them and other researchers (see the bibliographical notes in Section 5.13).

The basic idea is that an assessment in a scholarly subject should uncover the individual’s “knowledge state,” that is, the exact set of concepts mastered by the individual. Here, “concept” means a type of problem that the individual has learned to master, such as, in Beginning Algebra:

*solving a quadratic equation with integer coefficients;*

or, in Basic Chemistry

*balance a chemical equation using the smallest whole number stoichiometric coefficients.<sup>2</sup>*

In KST, a problem type is referred to as an “item.” Note that this usage differs to that in psychometric, where an item is a particular problem, such as: *Solve the quadratic equation  $x^2 - x - 12 = 0$ .* In our case, the examples of an item are called *instances*.<sup>3</sup>

The items or problem types form a possibly quite large set, which we call the “domain” of the body of knowledge. A knowledge state is a subset of the domain, but not any subset is a state: the knowledge states form a particular collection of subsets, which is called the “knowledge structure” or more specifically (when certain requirements are satisfied) the “knowledge space” or the “learning space.” The collection of states captures the whole structure of the domain. As in Beginning Algebra, which will be our lead example in this chapter, the domain may contain as many as 650 items, and the learning space may contain many millions of states, in sharp contrast with the few dozen scoring categories of a psychometric test. Despite this large number of possible knowledge states, an efficient assessment is feasible in the course of 25–35 questions.

In Sections 5.2 to 5.8, we review the fundamental combinatorial concepts and the main axiomatizations. We introduce the important special case of KST, learning space theory (LST). As the collection of all the feasible, realistic knowledge states may be very large, it is essential to find efficient summaries. We describe two such summaries in Sections 5.4 and 5.7. The theory discussed in this chapter has been extensively applied in schools and universities. The success of the applications is due in large part to a feature of the knowledge state produced by the

<sup>1</sup> A test for college admission and placement in the USA. The acronym SAT used to mean “Scholastic Aptitude Test.” A few years ago the meaning of SAT was changed into “Scholastic Assessment Test.” Today this acronym stands alone, without any associated meaning.

<sup>2</sup> For example:  $\text{Fe}_2\text{O}_3(s) \rightarrow \text{Fe}(s) + \text{O}_2(g)$ , which is not balanced. The correct response is:  $2\text{Fe}_2\text{O}_3(s) \rightarrow 4\text{Fe}(s) + 3\text{O}_2(g)$ .

<sup>3</sup> So, the instances of learning space theory are the items of psychometrics.

assessment: the state is predictive of what a student is ready to learn. The reason lies in a formal result, the “fringe theorem” (see Section 5.6). In some situations, it is important to focus on a part of a knowledge structure. We call the relevant concept a “projection” of a knowledge structure on a subset of the items. This is the subject of Section 5.8.

A subsequent section is devoted to the description of the Markovian type stochastic assessment procedure (Section 5.10). It relies on the notion of a probabilistic knowledge structure, introduced in Section 5.9. In Section 5.11, we give an outline of our methods for building the fundamental structure of states for a particular scholarly domain, such as Beginning Algebra, Pre-Calculus, or Statistics. Such constructions are enormously demanding and time-consuming. They rely not only on dedicated mathematical algorithms, but also on huge bases of assessment data. The most extensive applications of KST are in the form of the web-based system called ALEKS,<sup>4</sup> which includes a teaching component. Millions of students have used the system, either at home, or in schools and universities. Section 5.12 reports some results of these applications.

This chapter summarizes key concepts and results from two books. One is the monograph of Falmagne and Doignon (2011).<sup>5</sup> The other book is the edited volume of Falmagne *et al.* (2013), which contains recent data on the applications of the theory and also some new theoretical results. A few additional results appear here in Sections 5.7 and 5.11.

## 5.2 Knowledge structures and learning spaces

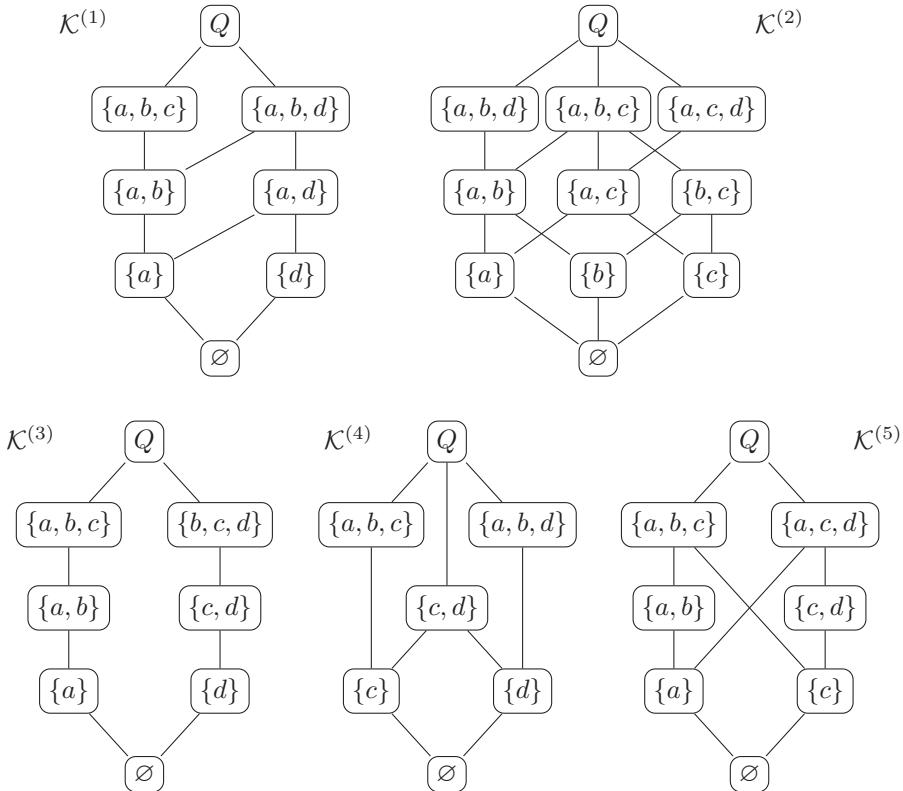
We formalize the cognitive structure of a scholarly subject as a collection  $\mathcal{K}$  of subsets of a basic set  $Q$  of items. In the case of Beginning Algebra, the items forming  $Q$  are the types of problems a student must master to be fully conversant in the subject. We suppose that the collection  $\mathcal{K}$  contains at least two subsets: the empty set, which is that of a student knowing nothing at all in the subject, and the full set  $Q$  of problems. The next definition cast these basic notions in set-theoretic terms. We illustrate the definition by a few examples.

**Definition 5.1** A *knowledge structure* is a pair  $(Q, \mathcal{K})$  consisting of a nonempty set  $Q$  and a collection  $\mathcal{K}$  of subsets of  $Q$ ; we assume  $\emptyset \in \mathcal{K}$  and  $Q \in \mathcal{K}$ . The set  $Q$  is called the *domain* of the knowledge structure  $(Q, \mathcal{K})$ . The elements of  $Q$  are the *items*, and the elements of  $\mathcal{K}$  are the *knowledge states*, or just the *states*. A knowledge structure  $(Q, \mathcal{K})$  is *finite* when its domain  $Q$  is a finite set.

For any item  $q$  in  $Q$ , we write  $\mathcal{K}_q$  for  $\{K \in \mathcal{K} \mid q \in K\}$ , the subcollection of  $\mathcal{K}$  consisting of all the states containing  $q$ . A knowledge structure  $(Q, \mathcal{K})$  is *discriminative* when for any two items  $q$  and  $r$  in the domain, we have  $\mathcal{K}_q = \mathcal{K}_r$  only if  $q = r$ .

<sup>4</sup> ALEKS is an acronym for “Assessment and LEarning in Knowledge Spaces.”

<sup>5</sup> This is a much expanded re-edition of Doignon and Falmagne (1999).



**Figure 5.1** The five examples of knowledge structures in Example 5.1.

We often abbreviate  $(Q, \mathcal{K})$  into  $\mathcal{K}$  (with no loss of information because  $Q = \cup \mathcal{K}$ ).

**Example 5.1** Here are five examples of knowledge structures all on the same domain  $Q = \{a, b, c, d\}$ :

$$\begin{aligned}\mathcal{K}^{(1)} &= \{\emptyset, \{a\}, \{d\}, \{a, b\}, \{a, d\}, \{a, b, c\}, \{a, b, d\}, Q\}, \\ \mathcal{K}^{(2)} &= \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, d\}, \{a, b, c\}, \{a, c, d\}, Q\}, \\ \mathcal{K}^{(3)} &= \{\emptyset, \{a\}, \{d\}, \{a, b\}, \{c, d\}, \{a, b, c\}, \{b, c, d\}, Q\}, \\ \mathcal{K}^{(4)} &= \{\emptyset, \{c\}, \{d\}, \{c, d\}, \{a, b, c\}, \{a, b, d\}, Q\}, \\ \mathcal{K}^{(5)} &= \{\emptyset, \{a\}, \{c\}, \{a, b\}, \{c, d\}, \{a, b, c\}, \{a, c, d\}, Q\}.\end{aligned}$$

The five knowledge structures are finite, and all but  $\mathcal{K}^{(4)}$  are discriminative: we have

$$\mathcal{K}_a^{(4)} = \mathcal{K}_b^{(4)} = \{\{a, b, c\}, \{a, b, d\}, Q\} \quad \text{with } a \neq b. \quad (5.1)$$

In the graphs of these knowledge structures displayed in Figure 5.1, the ascending lines show the covering relation of the states; that is, we have an ascending line

from the point representing the state  $K$  to the one representing the state  $L$  exactly when  $K$  is *covered* by  $L$ , that is when  $K \subset L$  and moreover there is no state  $A$  in  $\mathcal{K}$  such that  $K \subset A \subset L$ .

We call a representation of a knowledge structure as exemplified in Figure 5.1 a *covering diagram* of the structure. Note in passing two extreme cases of knowledge structures on a given domain  $Q$ . One is  $(Q, 2^Q)$ , where  $2^Q$  denotes the *power set* of  $Q$ , that is the collection of all the subsets of  $Q$ . The other one is  $(Q, \{\emptyset, Q\})$ , in which the knowledge structure contains only the two required states  $\emptyset$  and  $Q$ . These two examples are trivial and uninteresting because they entail a complete lack of organization in the body of information covered by the items in  $Q$ .

Two requirements on a knowledge structure make good pedagogical sense. One is that there should be no gaps in the organization of the material: the student should be able to master the items one by one. Also, there should be some consistency in the items: an advanced student should have less trouble learning a new item than another, less-competent student has. The two axioms incorporated in the next definition formalize the two ideas.

**Definition 5.2** A *learning space*  $(Q, \mathcal{K})$  is a knowledge structure which satisfies the two following conditions:

[L1] LEARNING SMOOTHNESS. For any two states  $K, L$  with  $K \subset L$ , there exists a finite chain of states

$$K = K_0 \subset K_1 \subset \cdots \subset K_p = L \tag{5.2}$$

such that  $|K_i \setminus K_{i-1}| = 1$  for  $1 \leq i \leq p$  (thus we have  $|L \setminus K| = p$ ).

In words: *If the learner is in some state  $K$  included in some state  $L$ , then the learner can reach state  $L$  by mastering items one by one.*

[L2] LEARNING CONSISTENCY. For any two states  $K, L$  with  $K \subset L$ , if  $q$  is an item such that  $K \cup \{q\}$  is a state, then  $L \cup \{q\}$  is also a state.

In words: *Knowing more does not prevent learning something new.*

Notice that any learning space is finite. Indeed, Condition [L1] applied to the two states  $\emptyset$  and  $Q$  yields a finite chain of states from  $\emptyset$  to  $Q$ . In Example 5.1 (see also Figure 5.1), only the two structures  $\mathcal{K}^{(1)}$  and  $\mathcal{K}^{(2)}$  are learning spaces. The knowledge structure  $\mathcal{K}^{(3)}$  satisfies Condition [L1] in Definition 5.2 but not Condition [L2] (take  $K = \emptyset, L = \{a\}$  and  $q = d$ ). As for  $\mathcal{K}^{(4)}$ , it satisfies [L2] but not [L1]. Finally,  $\mathcal{K}^{(5)}$  does not satisfy either condition. A simpler way to check whether a covering diagram as in Figure 5.1 represents a learning space is provided in the next section, just after Theorem 5.2.

We give one more example of a learning space, which is realistic in that its 10 items belong to the domain of the very large learning space of Beginning Algebra used in the ALEKS system. A remarkable feature of a learning space is that any subset of the items of a learning space also defines a learning space (see below Definition 5.11 and Theorem 5.11). Thus we have a learning space on these

Table 5.1 *The items of the 10-item example of Figure 5.2.*

a. Quotients of expressions involving exponents	b. Multiplying two binomials
c. Plotting a point in the coordinate plane using a virtual pencil on a Cartesian graph	d. Writing the equation of a line given the slope and a point on the line
e. Solving a word problem using a system of linear equations (advanced problem)	f. Graphing a line given its equation
g. Multiplication of a decimal by a whole number	h. Integer addition (introductory problem)
i. Equivalent fractions: fill the blank in the equation $\frac{a}{b} = \frac{\square}{c}$ , where $a, b$ and $c$ are whole numbers	j. Graphing integer functions

10 items. This example will be used repeatedly later on, and in particular to illustrate the assessment mechanism, that is, the questioning algorithm uncovering the knowledge state of a student.

**Example 5.2 Ten items in Beginning Algebra.** Table 5.1 lists 10 items. Remember that items are types of problems, and not particular cases (which are called instances). Here is an instance of item **d**: *a line passes through the point  $(x, y) = (-3, 2)$  and has a slope of 6. Write an equation for this line.* Figure 5.2 shows 34 knowledge states in  $Q = \{a, b, \dots, j\}$  which together form a learning space.

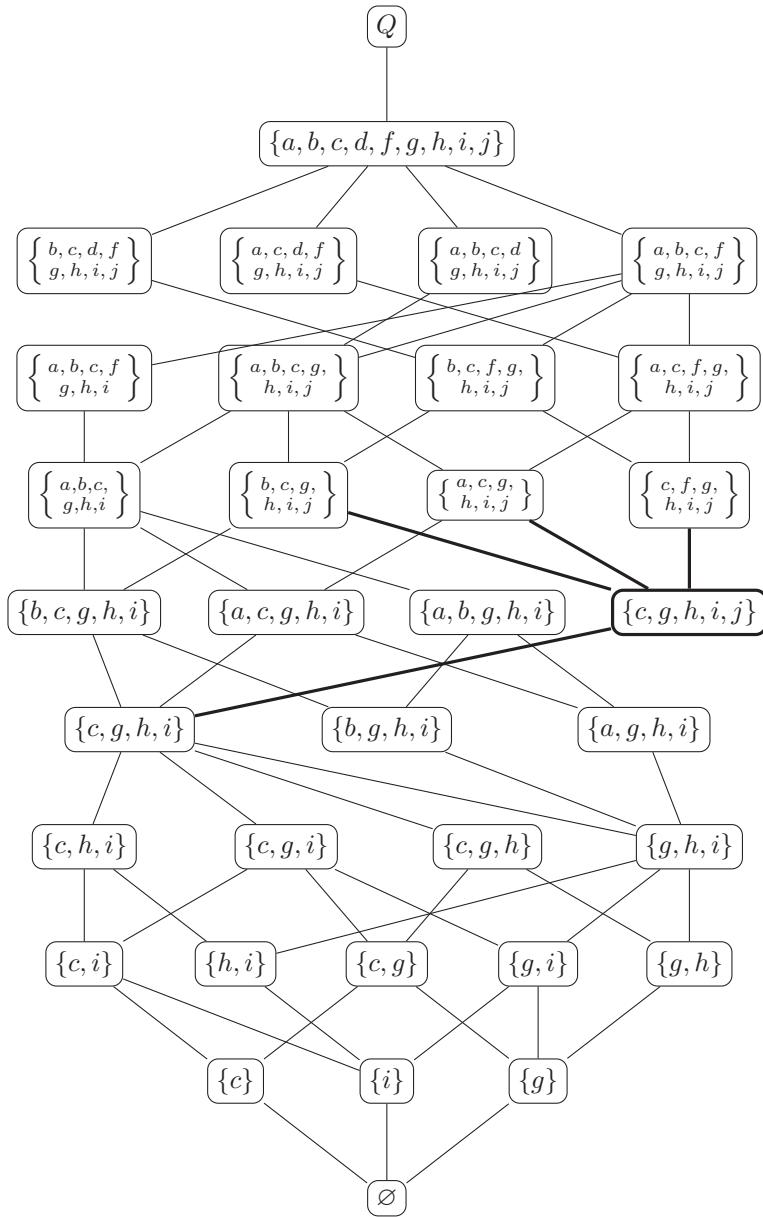
### 5.3 Knowledge spaces and wellgradedness

The name “learning space” specifically refers to Axioms [L1] and [L2] of Definition 5.2. As mentioned earlier, the two axioms have an interesting, pedagogical interpretation. Other characterizations of the same combinatorial concept focus on some other key concepts, which we review in the present section. The *symmetric difference* between two sets  $K$  and  $L$  is defined by  $K \Delta L = (K \setminus L) \cup (L \setminus K)$ .

**Definition 5.3** A *knowledge space*  $\mathcal{K}$  is a knowledge structure which is closed under union, or  *$\cup$ -closed*, that is,  $\cup \mathcal{C} \in \mathcal{K}$  for any subcollection  $\mathcal{C}$  of  $\mathcal{K}$ . The knowledge structure  $\mathcal{K}$  is *well-graded* if for any two states  $K$  and  $L$  in  $\mathcal{K}$ , there exists a natural number  $h$  such that  $|K \Delta L| = h$  and a finite sequence of states  $K = K_0, K_1, \dots, K_h = L$  such that  $|K_{i-1} \Delta K_i| = 1$  for  $1 \leq i \leq h$ . The knowledge structure  $\mathcal{K}$  is *accessible* or *downgradable*<sup>6</sup> if for any nonempty state  $K$  in  $\mathcal{K}$ , there is some item  $q$  in  $K$  such that  $K \setminus \{q\} \in \mathcal{K}$ . A downgradable, finite knowledge space is called an *antimatroid*.<sup>7</sup>

<sup>6</sup> See Doble *et al.* (2001) for the latter term.

<sup>7</sup> See Korte *et al.* (1991) for this use of the word. For other authors, an antimatroid is closed under intersection rather than under union (and “upgradable” rather than downgradable), see, for example, Edelman and Jamison (1985).



**Figure 5.2** The covering diagram of the 10-item learning space  $\mathcal{L}$  of Example 5.2. The meaning of the four bold lines joining the state  $\{c, g, h, i, j\}$  to four other states is explained in Section 5.6.

The closure under union is a critical property because it enables a (sometimes highly efficient) summary of a knowledge space by a minimal subcollection of its states. The subcollection is called the “base” of the knowledge space and is one of the topics of our next section. Note that a well-graded knowledge

structure  $(Q, \mathcal{K})$  is necessarily downgradable (given the state  $K$ , take  $L = \emptyset$  in Definition 5.2 [L1]). However, a downgradable knowledge structure is not necessarily finite nor well-graded.

**Example 5.3** Take as items all the natural numbers, thus the domain is  $\mathbb{N}$ . As states, take the empty set plus all the subsets of  $\mathbb{N}$  whose complement is finite. We denote by  $\mathcal{G}$  the collection of states:

$$\mathcal{G} = \{\emptyset\} \cup \{K \in 2^{\mathbb{N}} \mid |\mathbb{N} \setminus K| < +\infty\}. \quad (5.3)$$

The resulting structure  $(\mathbb{N}, \mathcal{G})$  is downgradable. It is infinite, and not well-graded (consider, for instance, the two states  $\emptyset$  and  $\mathbb{N}$ ).

**Theorem 5.1** *For any knowledge structure  $(Q, \mathcal{K})$ , the following three statements are equivalent.*

- (i)  $(Q, \mathcal{K})$  is a learning space.
- (ii)  $(Q, \mathcal{K})$  is an antimatroid.
- (iii)  $(Q, \mathcal{K})$  is a well-graded knowledge space.

Cosyn and Uzun (2009) proved the equivalence of Conditions (i) and (ii) in Theorem 5.1, while Korte *et al.* (1991) established still another characterization of antimatroids (or learning spaces): Theorem 5.2 below is Lemma 1.2 of their chapter 3. We provide a combined proof of Theorems 5.1 and 5.2 below.

**Theorem 5.2** *A knowledge structure  $(Q, \mathcal{K})$  is a learning space if and only if its collection  $\mathcal{K}$  of states satisfies the following three conditions:*

- (a)  $Q$  is finite;
- (b)  $\mathcal{K}$  is downgradable, that is: any state  $K$  contains some item  $q$  such that  $K \setminus \{q\} \in \mathcal{K}$ ;
- (c) for any state  $K$  and any items  $q, r$ , if  $K \cup \{q\}, K \cup \{r\} \in \mathcal{K}$ , then  $K \cup \{q, r\} \in \mathcal{K}$ .

Theorem 5.2 makes it easy to check whether a (finite) covering diagram such as that pictured in Figure 5.1 represents a learning space. Assume that the points representing two states are at the same level (or height) if and only if they have the same number of items; then it suffices to check that: (i) any ascending line connects points at two successive levels; (ii) any point representing a nonempty state is the end of at least one ascending line; (iii) if two ascending lines start from the same point, then their endpoints are the origins of ascending lines having the same endpoint.

*Proofs of Theorems 5.1 and 5.2* For a given knowledge structure  $(Q, \mathcal{K})$ , we show that

$$(i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow ((a), (b) \text{ and } (c)) \Rightarrow (i).$$

First notice that any of the three conditions (i), (ii), (iii) entails the finiteness of  $Q$ , that is, Condition (a).

(i)  $\Rightarrow$  (ii). Let  $(Q, \mathcal{K})$  be a learning space as in Definition 5.2. We prove that  $(Q, \mathcal{K})$  is an antimatroid as in Definition 5.3; in other words, that  $(Q, \mathcal{K})$  is a finite knowledge space which is moreover downgradable. Suppose that  $K, L$  are states. We first apply Learning Smoothness to  $\emptyset$  and  $L$  and derive a sequence  $L_0 = \emptyset, L_1, \dots, L_\ell = L$  of states such that  $|L_i \setminus L_{i-1}| = 1$  for  $1 \leq i \leq \ell$ . Then applying Learning Consistency to the states  $\emptyset$  and  $K$  and the item forming  $L_1$ , we derive  $K \cup L_1 \in \mathcal{K}$ . Next, we apply Learning Consistency to the states  $L_1$  and  $K \cup L_1$  and the item forming  $L_2 \setminus L_1$  to derive  $K \cup L_2 \in \mathcal{K}$ . The general step, for  $i = 1, 2, \dots, \ell$ , applies Learning Consistency to  $L_{i-1}$  and  $K \cup L_{i-1}$  and the item forming  $L_i \setminus L_{i-1}$  to derive  $K \cup L_i \in \mathcal{K}$ . At the last step ( $i = \ell$ ), we get  $K \cup L \in \mathcal{K}$ . On the other hand, downgradability of  $\mathcal{K}$  at the state  $K$  is just a particular case of Learning Smoothness at the states  $\emptyset$  and  $K$ .

(ii)  $\Rightarrow$  (iii). If  $(Q, \mathcal{K})$  is an antimatroid, then  $\mathcal{K}$  is closed under union by definition. To prove the wellgradedness of  $\mathcal{K}$  (Definition 5.3), we take two states  $K$  and  $L$ . By assumption,  $K \cup L$  is also a state, and moreover by downgradability there exists a sequence of states  $M_0 = \emptyset, M_1, \dots, M_h = K \cup L$  with  $|M_i \setminus M_{i-1}| = 1$  for  $1 \leq i \leq h$ . Then  $K \cup M_0, K \cup M_1, \dots, K \cup M_p$ , after deletion of repetitions, becomes an increasing sequence  $K_0, K_1, \dots, K_k$  from  $K$  to  $K \cup L$  with increments consisting of one item. We derive a similar sequence  $L_0, L_1, \dots, L_\ell$  from  $L$  to  $K \cup L$ . Finally,  $K_0, K_1, \dots, K_k = L_\ell, L_{\ell-1}, \dots, L_0$  is the required sequence from  $K$  to  $L$  (indeed,  $k + \ell = |K \Delta L|$ ).

(iii)  $\Rightarrow$  ((a), (b) and (c)). Downgradability (b) is a direct consequence of wellgradeness. To prove (c), we only need to notice  $K \cup \{q, r\} = (K \cup \{q\}) \cup (K \cup \{r\})$  and apply the assumed closure under union.

((a), (b) and (c))  $\Rightarrow$  (i). To prove Learning Smoothness, consider states  $K$  and  $L$  with  $K \subset L$ . By downgradability, there exist sequences  $K_0 = \emptyset, K_1, \dots, K_k = K$  and  $L_0 = \emptyset, L_1, \dots, L_\ell = L$  of states such that  $|K_i \setminus K_{i-1}| = 1$  for  $1 \leq i \leq k$  and  $|L_j \setminus L_{j-1}| = 1$  for  $1 \leq j \leq \ell$ . Repeated applications of (c) show  $K_i \cup L_j \in \mathcal{K}$ . After deleting repetitions in  $K \cup L_0 = K, K \cup L_1, \dots, K \cup L_{\ell-1}, L$ , we obtain the desired sequence from  $K$  to  $L$ .

To prove Learning Consistency, we again consider two states  $K$  and  $L$  with  $K \subset L$  together with an item  $q$  such that  $K \cup \{q\} \in \mathcal{K}$ . In the previous paragraph, we proved the existence of a sequence  $M_0 = K, M_1, \dots, M_h = L$  of states such that  $|M_i \setminus M_{i-1}| = 1$  for  $1 \leq i \leq h$ . Applying (c) repeatedly, we obtain  $M_1 \cup \{q\} \in \mathcal{K}, M_2 \cup \{q\} \in \mathcal{K}, \dots, M_h \cup \{q\} \in \mathcal{K}$ , the last one being  $L \cup \{q\} \in \mathcal{K}$  as desired.  $\square$

A simple case of a learning space arises when the collection of states is closed under both union and intersection.

**Definition 5.4** A *quasi ordinal space* is a knowledge space closed under intersection. A *(partially) ordinal space* is a quasi ordinal space which is discriminative.

In Example 5.1, only the structure  $\mathcal{K}^{(1)}$  is a quasi ordinal space, and it is even an ordinal space. The reason for the terminology in Definition 5.4 lies in Theorem 5.3 below, due to Birkhoff (1937). We recall that a *quasi order* on  $Q$  is a reflexive

and transitive relation on  $Q$ . A *partial order* on  $Q$  is a quasi order on  $Q$  which is an *antisymmetric relation* (that is, for all  $q$  and  $r$  in  $Q$ , it holds that  $qRr$  and  $rRq$  implies  $q = r$ ).

**Theorem 5.3** (Birkhoff, 1937) *There exists a one-to-one correspondence between the collection of all quasi ordinal spaces  $\mathcal{K}$  on a set  $Q$  and the collection of all quasi orders  $\mathcal{Q}$  on  $Q$ . One such correspondence is specified by the two equivalences<sup>8</sup>*

$$\text{for all } q, r \text{ in } Q: \quad q\mathcal{Q}r \iff \mathcal{K}_q \supseteq \mathcal{K}_r; \quad (5.4)$$

$$\text{for all } K \subseteq Q: \quad K \in \mathcal{K} \iff (\forall (q, r) \in \mathcal{Q}: r \in K \Rightarrow q \in K). \quad (5.5)$$

*Its restriction to discriminative spaces links partially ordinal spaces to partial orders.*

Note in passing that the closure under intersection does not make good pedagogical sense. A variant of Theorem 5.3 for knowledge spaces appears below as Theorem 5.6; a variant for learning spaces follows from Theorem 5.7.

## 5.4 The base and the atoms

In practice, learning spaces tend to be very large, counting millions of states. For various purposes – for example, to store the structure in a computer’s memory – such huge structures need to be summarized. One such summary is the “base” of the structure, which we define below.

**Definition 5.5** The *span* of a collection of sets  $\mathcal{F}$  is the collection  $\mathbb{S}(\mathcal{F})$  containing exactly those sets that are unions of sets in  $\mathcal{F}$ . We then say that  $\mathcal{F}$  spans  $\mathbb{S}(\mathcal{F})$ . So,  $\mathbb{S}(\mathcal{F})$  is necessarily  $\cup$ -closed. A *base* of a  $\cup$ -closed collection  $\mathcal{S}$  of sets is a minimal subcollection  $\mathcal{B}$  of  $\mathcal{S}$  spanning  $\mathcal{S}$  – where “minimal” refers to inclusion, that is, if  $\mathbb{S}(\mathcal{H}) = \mathcal{S}$  for some  $\mathcal{H} \subseteq \mathcal{B}$ , then necessarily  $\mathcal{B} \subseteq \mathcal{H}$ .

Note that by a common convention, the empty set is the union of the zero set from  $\mathcal{B}$ . Accordingly, the empty set never belongs to a base. It is easily shown that the base of a knowledge space is unique when it exists. Also, any finite knowledge space has a base (see theorems 3.4.2 and 3.4.4 in Falmagne and Doignon, 2011). However, some  $\cup$ -closed collections of sets have no base; an example is the collection of all the open subsets of the set of real numbers. Any learning space has a base because it is finite and  $\cup$ -closed (cf. Theorem 5.1, (i)  $\Leftrightarrow$  (iii)).

For example, the base of the learning space  $\mathcal{K}^{(2)}$  displayed in Figure 5.1 is

$$\{ \{a\}, \{b\}, \{c\}, \{a, b, d\}, \{a, c, d\} \}. \quad (5.6)$$

The economy is not great in this little example, but may become spectacular in the case of the very large structures encountered in practice.

<sup>8</sup> We recall the formula  $\mathcal{K}_q = \{K \in \mathcal{K} \mid q \in K\}$ .

Table 5.2 *The atoms at all the items of the 10-item learning space from Example 5.2 (see Example 5.4).*

Items	Atoms
$a$	$\{a, g, h, i\}$
$b$	$\{b, g, h, i\}$
$c$	$\{c\}$
$d$	$\{b, c, d, f, g, h, i, j\}, \{a, c, d, f, g, h, i, j\}, \{a, b, c, d, g, h, i, j\}$
$e$	$Q$
$f$	$\{c, f, g, h, i, j\}, \{a, b, c, f, g, h, i\}$
$g$	$\{g\}$
$h$	$\{h, i\}, \{g, h\}$
$i$	$\{i\}$
$j$	$\{c, g, h, i, j\}$

Another reason for the importance of the base stems from a pedagogical concept. The relevant question is: “*Given some item  $q$ , which minimal set, or sets, of items must be mastered for  $q$  to be learnable?*” Or restated in set-theoretical terms: “*what are the minimal states containing a given item  $q$ ?*” As one might guess, these minimal sets coincide with the elements of the base.

**Definition 5.6** Let  $\mathcal{K}$  be a knowledge space. For any item  $q$ , an *atom at  $q$*  is a minimal state of  $\mathcal{K}$  containing  $q$ . A state  $K$  is called an *atom* if  $K$  is an atom at  $q$  for some item  $q$ . A knowledge space is *granular* if for any item  $q$  and any state  $K$  containing  $q$ , there is an atom at  $q$  which is included in  $K$ .

Clearly, any finite knowledge space is granular. On the other hand, a state  $K$  is an atom in a knowledge space  $\mathcal{K}$  if and only if any subcollection of states  $\mathcal{F}$  such that  $K = \cup \mathcal{F}$  contains  $K$  (cf. theorem 3.4.7 in Falmagne and Doignon, 2011). Note also that any granular knowledge space has a base (cf. proposition 3.6.6 in Falmagne and Doignon, 2011).

**Example 5.4** For the 10-item learning space pictured in Figure 5.2, there are two atoms at item  $f$ , namely

$$\{c, f, g, h, i, j\}, \{a, b, c, f, g, h, i\}. \quad (5.7)$$

You can check from the figure that these two sets are indeed minimal states containing  $f$  and that they are the only ones with that property. Note that there is just one atom at the item  $b$ , which is  $\{b, g, h, i\}$ , while there are three atoms at  $d$ . Table 5.2 displays the full information on the atoms.

We conclude this section with the expected result (a proof is given in Falmagne and Doignon, 2011<sup>9</sup>).

9 The reader should refer to that monograph for most of the proofs omitted in this chapter.

**Theorem 5.4** Suppose that a knowledge space has a base. Then this base is exactly the collection of all the atoms.

A simple algorithm, due to Dowling (1993) and grounded on the concept of an atom, constructs the base of a finite knowledge space given the states. In the same paper, she also describes a more elaborate algorithm for efficiently building the span of a collection of subsets of a finite set. Both algorithms are sketched in Falmagne and Doignon (2011, pp. 49–50) (another algorithm for the second task, in another context, is due to Ganter; see Ganter and Reuter, 1991). The concept of an atom is closely related to that of the “surmise system,” which is the topic of our next section. We complete the present section with a characterization of learning spaces through their atoms (Koppen, 1998).

**Theorem 5.5** For any finite knowledge space  $(Q, \mathcal{K})$ , the following three statements are equivalent:

- (i)  $(Q, \mathcal{K})$  is a learning space;
- (ii) for any atom  $A$  at item  $q$ , the set  $A \setminus \{q\}$  is a state;
- (iii) any atom is an atom at only one item.

## 5.5 Surmise systems

In a finite knowledge space, a student masters an item  $q$  only when his state includes some atom  $C$  at  $q$ . So, the collection of all the atoms at the various items may provide a new way to specify a knowledge space. We illustrate this idea by the following example of a knowledge space.

**Example 5.5** Consider the knowledge space

$$\begin{aligned} \mathcal{H} = \{ &\emptyset, \{a\}, \{b, d\}, \{a, b, c\}, \{a, b, d\}, \{b, c, e\}, \\ &\{a, b, c, d\}, \{a, b, c, e\}, \{b, c, d, e\}, \{a, b, c, d, e\} \} \end{aligned} \quad (5.8)$$

on the domain  $Q = \{a, b, c, d, e\}$ . Figure 5.3 provides its covering diagram, while Table 5.3 lists its atoms. Table 5.3 links each of items  $a, d$ , and  $e$  to a single atom of  $\mathcal{H}$ , and items  $b$  and  $c$  to three and two atoms, respectively. So, to master item  $b$ , one must first master either item  $d$ , or items  $a$  and  $c$ , or items  $c$  and  $e$ .

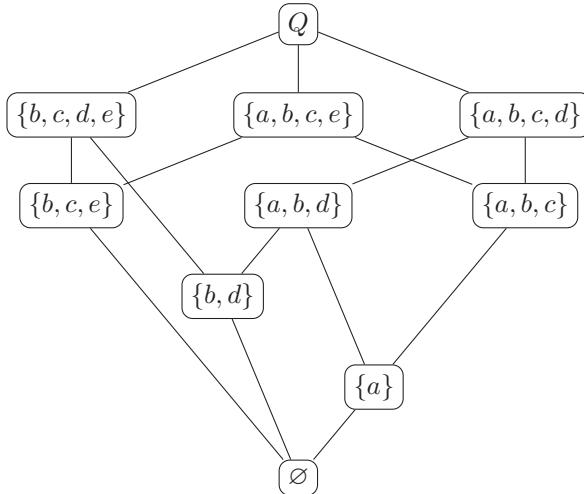
Example 5.5 illustrates the following definition.

**Definition 5.7** Let  $Q$  be a nonempty set of items. A function  $\sigma : Q \rightarrow 2^{\mathcal{P}}$  mapping each item  $q$  in  $Q$  to a nonempty collection  $\sigma(q)$  of subsets of  $Q$  (so,  $\sigma(q) \neq \emptyset$ ) is called an *attribution function* on the set  $Q$ . For each  $q$  in  $Q$ , any  $C$  in  $\sigma(q)$  is called a *clause for  $q$*  (in  $\sigma$ ). A *surmise function*  $\sigma$  on  $Q$  is an attribution function on  $Q$  which satisfies the three additional conditions, for all  $q, q' \in Q$ , and  $C, C' \subseteq Q$ :

- (i) if  $C \in \sigma(q)$ , then  $q \in C$ ;
- (ii) if  $q' \in C \in \sigma(q)$ , then  $C' \subseteq C$  for some  $C' \in \sigma(q')$ ;
- (iii) if  $C, C' \in \sigma(q)$  and  $C' \subseteq C$ , then  $C = C'$ .

Table 5.3 *Items and their atoms in the knowledge space of Equation (5.8) (see also Figure 5.3).*

Items	Atoms
$a$	$\{a\}$
$b$	$\{b, d\}, \{a, b, c\}, \{b, c, e\}$
$c$	$\{a, b, c\}, \{b, c, e\}$
$d$	$\{b, d\}$
$e$	$\{b, c, e\}$



**Figure 5.3** The covering diagram of the knowledge space in Example 5.5.

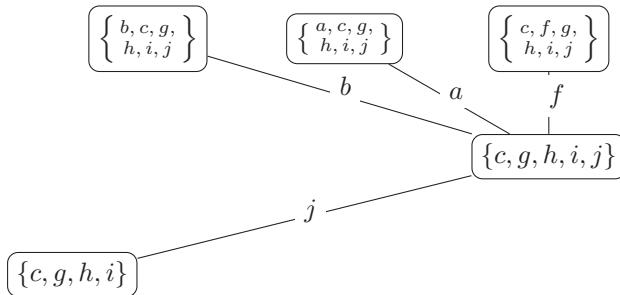
In such a case, the pair  $(Q, \sigma)$  is a *surmise system*. A surmise system  $(Q, \sigma)$  is *discriminative* if  $\sigma$  is injective (that is: whenever  $\sigma(q) = \sigma(q')$  for some  $q, q' \in Q$ , then  $q = q'$ ). Then the surmise function  $\sigma$  is also called *discriminative*.

It is easily shown that any attribution function  $\sigma$  on a set  $Q$  defines a knowledge space  $(Q, \mathcal{K})$  via the equivalence

$$K \in \mathcal{K} \iff \forall q \in K, \exists C \in \sigma(q) : C \subseteq K. \quad (5.9)$$

In fact, we have the following extension of Birkhoff's Theorem 5.3 (it is an extension in the sense that the one-to-one correspondence we obtain extends the correspondence in Birkhoff's Theorem). The result is due to Doignon and Falmagne (1985), who derive it from an appropriate “Galois connection.”

**Theorem 5.6** *There exists a one-to-one correspondence between the collection of all granular knowledge spaces  $\mathcal{K}$  on a set  $Q$  and the collection of all the surmise functions  $\sigma$  on  $Q$ . One such correspondence is specified by the equivalence, for all*



**Figure 5.4** Part of Figure 5.2 showing the items in the inner fringe and outer fringe of the state  $\{c, g, h, i, j\}$  (see Definition 5.8).

$q$  in  $Q$  and  $A$  in  $2^Q$ ,

$$A \text{ is an atom at } q \text{ in } \mathcal{K} \iff A \in \sigma(q). \quad (5.10)$$

This one-to-one correspondence links discriminative knowledge spaces to discriminative surmise functions.

The correspondence between knowledge spaces and surmise functions is suggestive of a practical method for building a knowledge space or even a learning space, based on analyzing large sets of learning data. We describe such a method in Section 5.11.

A characterization of learning spaces through their surmise functions derives directly from Theorem 5.5.

**Theorem 5.7** *A finite knowledge space  $(Q, \mathcal{K})$  is a learning space if and only if in the corresponding surmise system any clause is a clause for only one item.*

In the case of finite, partially ordinal spaces, a highly efficient summary of the space takes the form of the “Hasse diagram” of the partial order. Attempts to extend the notion of a Hasse diagram from partially ordered sets to surmise systems are reported in Doignon and Falmagne (1999) and Falmagne and Doignon (2011).

## 5.6 The fringe theorem

The final result of a standardized test is a numerical score.<sup>10</sup> In the case of an assessment in the framework of a learning space, the result is a knowledge state which may contain hundreds of items. Fortunately, a meaningful summary of that state can be given in the form of its “inner fringe” and “outer fringe.”

In the 10-item Example 5.2, consider the state  $\{c, g, h, i, j\}$ , which is printed in the bold frame of Figure 5.2. Figure 5.4 reproduces the relevant part of the graph,

<sup>10</sup> Or a couple of such scores, in the case of a multidimensional model.

in particular all the adjacent states. From the state  $\{c, g, h, i, j\}$ , only three items are learnable,<sup>11</sup> which are  $a$ ,  $b$ , and  $f$  (we mark them on their respective lines). On the other hand, the only way to reach state  $\{c, g, h, i, j\}$  is to learn item  $j$  from the state  $\{c, g, h, i\}$  (which is the unique state giving access to  $\{c, g, h, i, j\}$ ). The two sets of items  $\{j\}$  and  $\{a, b, f\}$  completely specify the state  $\{c, g, h, i, j\}$  among all the states in the structure; this is a remarkable property of learning spaces which we now formalize.

**Definition 5.8** Let  $(Q, \mathcal{K})$  be a knowledge structure. The *inner fringe* of a state  $K$  in  $\mathcal{K}$  is the set of items

$$K^{\mathcal{T}} = \{q \in K \mid K \setminus \{q\} \in \mathcal{K}\}. \quad (5.11)$$

The *outer fringe* of a state  $K$  is the set of items

$$K^O = \{q \in Q \setminus K \mid K \cup \{q\} \in \mathcal{K}\}. \quad (5.12)$$

Note that the empty state  $\emptyset$  always has an empty inner fringe, and that the whole domain  $Q$  has an empty outer fringe.

**Theorem 5.8** *In a learning space, any state is specified by the pair formed of its inner fringe and its outer fringe.*

Theorem 5.8 has the following important consequence. In any learning space, the knowledge state uncovered by an assessment can be reported as two sets of items: those in its inner fringe, and those in its outer fringe. The outer fringe is especially important because, assuming that the learning space is a faithful representation of the cognitive organization of the material, it tells us exactly what the student is ready to learn. We will see in Section 5.12 that this information is supported by real-life data: the probability that a student actually succeeds in learning an item picked in the outer fringe of his or her state is about .93 estimated on the basis of hundreds of thousand ALEKS assessments (see Subsection 5.12.2).

## 5.7 Learning words and learning strings

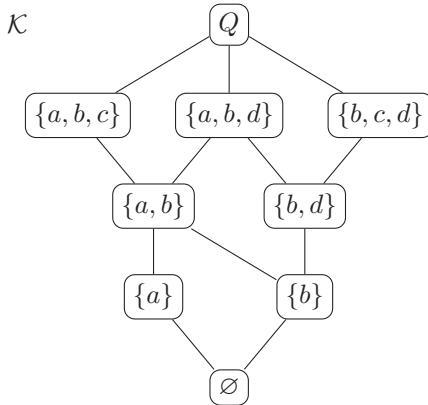
In a learning space, a learner can reach any state by learning its items one at a time – but not in any order. Let us look at an example.

**Example 5.6** A learning space on the domain  $Q = \{a, b, c, d\}$  is described by its covering diagram in Figure 5.5.

The state  $\{a, b, d\}$  can be reached by mastering the items in three possible successions, which we also call “words” (see Definition 5.9):

$a b d,$   
 $b a d,$   
 $b d a.$

<sup>11</sup> We mean directly learnable without requiring the mastery of any other item outside  $\{c, g, h, i, j\}$ .



**Figure 5.5** The covering diagram of the learning space in Example 5.6.

For the mastery of the whole domain  $Q$ , there are six “strings” in all:

$a b c d,$   
 $a b d c,$   
 $b a c d,$   
 $b a d c,$   
 $b d a c,$   
 $b d c a.$

All of the words and strings in Example 5.6 share a self-explanatory property: for any of their “prefixes,” the items appearing in the prefix form a knowledge state. Let us define the new terminology.

**Definition 5.9** Given some finite set  $Q$ , a *word* on  $Q$  is any injective mapping  $f$  from  $\{1, 2, \dots, k\}$  to  $Q$ , for some  $k$  with  $0 \leq k \leq |Q|$ ; the case  $k = 0$  produces the *empty word*. With  $f(i) = w_i$  for  $1 \leq i \leq k$ , we write the word  $f$  as  $w = w_1 w_2 \cdots w_k$ , and we call  $k$  the *length* of the word  $w$ . A *prefix* of  $w$  is a word  $w_1 w_2 \cdots w_i$ , where  $0 \leq i \leq k$  (thus any word is a prefix of itself). If an item  $q$  does not appear in  $w$ , the *concatenation* of  $w$  with  $q$  is the word  $w q = w_1 w_2 \cdots w_k q$ . A *string* on  $Q$  is a word of length  $|Q|$ . Notice that words (and strings) do not involve repetitions (because of the required injectivity of  $f$ ), a reason for which some authors rather speak of “simple words” (as, for example, Boyd and Faigle, 1990). Any word  $w = w_1 w_2 \cdots w_k$  determines the set  $\tilde{w} = \{w_1, w_2, \dots, w_k\}$  (if all words are counted,  $k!$  of them determine the same set of size  $k$ ).

Let  $(Q, \mathcal{K})$  be a finite knowledge structure with  $|Q| = m$ . A *learning word* (in  $(Q, \mathcal{K})$ ) is a word  $w$  on  $Q$  such that for each of its prefixes  $v = w_1 w_2 \cdots w_i$  the subset  $\tilde{v} = \{w_1, w_2, \dots, w_i\}$  is a state in  $\mathcal{K}$ ; here  $0 \leq i \leq k$  if  $k$  is the length of  $w$ . A *learning string* is such a learning word with  $k = m$ .

Korte *et al.* (1991) use the expression “shelling sequence” for our “learning word,” and the expression “basic word” for our “learning string,” while Eppstein (2013a) uses “learning sequence” for our “learning string.” General knowledge spaces can be without any learning string (for instance, it is the case when  $\mathcal{K} = \{\emptyset, Q\}$  as soon as  $|Q| \geq 2$ ). The axioms of learning spaces are typically consistent with the existence of (many) learning strings and words. In fact, learning spaces can be recognized from properties of the collection of their learning strings (see next theorem) or the collection of their words (see Theorem 5.10).

**Theorem 5.9** *Let  $Q$  be a finite set, with  $|Q| = m$ . A nonempty collection  $\mathcal{S}$  of strings on  $Q$  is the collection of all learning strings of some learning space on  $Q$  if and only if  $\mathcal{S}$  satisfies the three conditions below:*

- (i) *any item appears in some string of  $\mathcal{S}$ ;*
- (ii) *if  $u$  and  $v$  are two strings in  $\mathcal{S}$  such that for some  $k$  in  $\{1, 2, \dots, m - 1\}$  we have*

$$\{u_1, u_2, \dots, u_{k-1}\} = \{v_1, v_2, \dots, v_{k-1}\} \quad \text{and} \quad u_k \neq v_k, \quad (5.13)$$

*then*

$$u_1 u_2 \cdots u_{k-1} u_k v_k \quad (5.14)$$

*is a prefix of some string in  $\mathcal{S}$ ;*

- (iii) *if  $u$  and  $v$  are two strings in  $\mathcal{S}$  such that for some  $k$  in  $\{0, 1, \dots, m - 1\}$  and some item  $q$  we have*

$$\{v_1, v_2, \dots, v_{k-1}, v_k, v_{k+1}\} \setminus \{u_1, u_2, \dots, u_k\} = \{q\}, \quad (5.15)$$

*then  $u_1 u_2 \cdots u_k q$  is a prefix of some string in  $\mathcal{S}$ .*

*Proof* (Necessity.) Assume  $(Q, \mathcal{L})$  is a learning space with  $|Q| = m$ , and denote by  $\mathcal{S}$  the collection of its learning strings. By Learning Smoothness,  $\mathcal{S}$  is nonempty and all the items of  $Q$  appear in any string in  $\mathcal{S}$ , so Condition (i) is true. If the hypothesis of Condition (ii) holds, then  $\{u_1, u_2, \dots, u_{k-1}\}$ ,  $\{u_1, u_2, \dots, u_{k-1}, u_k\}$  and  $\{u_1, u_2, \dots, u_{k-1}, v_k\}$  are all states of  $\mathcal{L}$ . Hence by Learning Consistency  $\{u_1, u_2, \dots, u_{k-1}, u_k, v_k\}$  is also a state, which we denote by  $L$ . On the other hand, by Learning Smoothness, there is a string  $L_{k+1}, L_{k+2}, \dots, L_m$  of states with  $L_{k+1} = L$ ,  $L_m = Q$ , and  $|L_i \setminus L_{i-1}| = 1$  for  $i = k + 2, k + 3, \dots, m$ . Taking  $w_i$  as the item in  $L_i \setminus L_{i-1}$ , we obtain the learning string  $u_1 u_2 \cdots u_{k-1} u_k v_k w_{k+2} w_{k+3} \cdots w_m$ . Thus Condition (ii) holds.

Now suppose the strings  $u$  and  $v$  fulfill the assumption in Condition (iii). Thus  $\{v_1, v_2, \dots, v_{k+1}\}$  is a state, that we call  $L$ . By Learning Smoothness, there is a string  $L_{k+1}, L_{k+2}, \dots, L_m$  of states in  $\mathcal{L}$  with  $L_{k+1} = L$ ,  $L_m = Q$  and  $|L_i \setminus L_{i-1}| = 1$  for  $i = k + 2, k + 3, \dots, m$ . With  $\{w_i\} = L_i \setminus L_{i-1}$ , we obtain the learning string

$$u_1 u_2 \cdots u_{k-1} u_k q w_{k+2} w_{k+3} \cdots w_m.$$

Hence Condition (iii) holds.

(Sufficiency.) Given a collection  $\mathcal{S}$  of strings on  $Q$  satisfying (i)–(iii), we call  $\mathcal{L}$  the collection of all prefixes of strings in  $\mathcal{S}$ . Then  $\emptyset$  and  $Q$  are in  $\mathcal{L}$ . Moreover,  $\mathcal{L}$  clearly satisfies downgradability, that is Condition (b) in Theorem 5.2. To establish that  $\mathcal{L}$  satisfies Condition (c), let  $K, K \cup \{q\}$  and  $K \cup \{r\}$  be in  $\mathcal{L}$ . There exist strings  $u$  and  $v$  such that  $\{u_1, u_2, \dots, u_{|K|}\} = K$  and  $\{v_1, v_2, \dots, v_{|K|+1}\} = K \cup \{q\}$ . Then by Condition (iii)  $u_1 u_2 \cdots u_{|K|} q$  is a prefix of some string in  $\mathcal{S}$ . A similar argument shows that  $u_1 u_2 \cdots u_{|K|} r$  is a string prefix. Then by Condition (ii),  $u_1 u_2 \cdots u_{|K|} q r$  is also a prefix of some string. So  $K \cup \{q, r\} \in \mathcal{L}$ . Hence by Theorem 5.2  $(Q, \mathcal{L})$  is a learning space. Moreover, the learning strings of  $(Q, \mathcal{L})$  constitute exactly  $\mathcal{S}$  (this derives again from the definition of  $\mathcal{S}$  together with Condition (iii)).  $\square$

Here is an example showing that Conditions (ii) and (iii) in Theorem 5.9 are independent.

**Example 5.7** On the domain  $Q = \{a, b, c, d\}$ , the two strings

$$abdc,$$

$$acdb$$

form a collection which satisfies Condition (iii) in Theorem 5.9 but not Condition (ii) (take  $k = 2$ ,  $u_1 u_2 = ab$  and  $v_1 v_2 = ac$ ). Conversely, the two strings on the same domain  $Q = \{a, b, c, d\}$

$$abcd,$$

$$badc$$

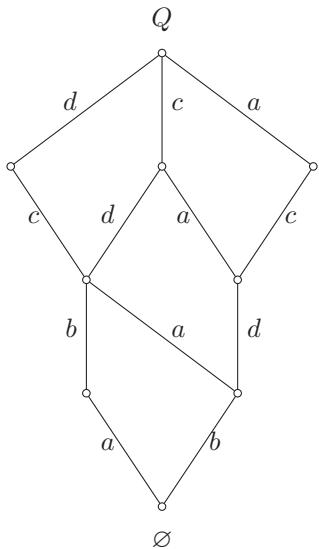
form a collection which satisfies Condition (ii) in Theorem 5.9 but not Condition (iii) (take  $k = 2$ ,  $u_1 u_2 = ba$  and  $v_1 v_2 v_3 = abc$ ).

The next result is Theorem 2.1 in Boyd and Faigle (1990) (compare with Theorem 1.4 in Korte *et al.*, 1991).

**Theorem 5.10** *Let  $Q$  be a finite domain. A collection  $\mathcal{W}$  of words on  $Q$  is the collection of all learning words of some learning space on  $Q$  if and only if  $\mathcal{W}$  satisfies the following three conditions:*

- (i) *any item from  $Q$  appears in at least one word of  $\mathcal{W}$ ;*
- (ii) *any prefix of a word in  $\mathcal{W}$  also belongs to  $\mathcal{W}$ ;*
- (iii) *if  $v$  and  $w$  are two words of  $\mathcal{W}$  with  $\tilde{v} \not\subseteq \tilde{w}$ , then for some item  $q$  in  $\tilde{v} \setminus \tilde{w}$  the concatenation  $wq$  is a word again in  $\mathcal{W}$ .*

*Proof* (Necessity.) Assume  $(Q, \mathcal{L})$  is a learning space, and denote by  $\mathcal{W}$  the collection of all its learning words. Then by downgradability of  $\mathcal{L}$ , the collection  $\mathcal{W}$  contains some string, so Condition (i) is true. By the definition of a learning word  $w$ , any prefix of  $w$  is also a learning word, so Condition (ii) holds. Now take two words  $v$  and  $w$  as in Condition (iii). Then  $\tilde{w} \cup \tilde{v} \in \mathcal{L}$  (by Theorem 5.1,  $\mathcal{L}$  is  $\cup$ -closed). Because of Learning Smoothness, there is a string  $L_0 = \tilde{w}, L_1, \dots, L_\ell = \tilde{w} \cup \tilde{v}$  with  $|L_i \setminus L_{i-1}| = 1$  for  $i = 1, 2, \dots, \ell$ . Let  $\{q\} = L_1 \setminus L_0$ . Then  $q \in \tilde{v} \setminus \tilde{w}$  and  $wq \in \mathcal{W}$ .



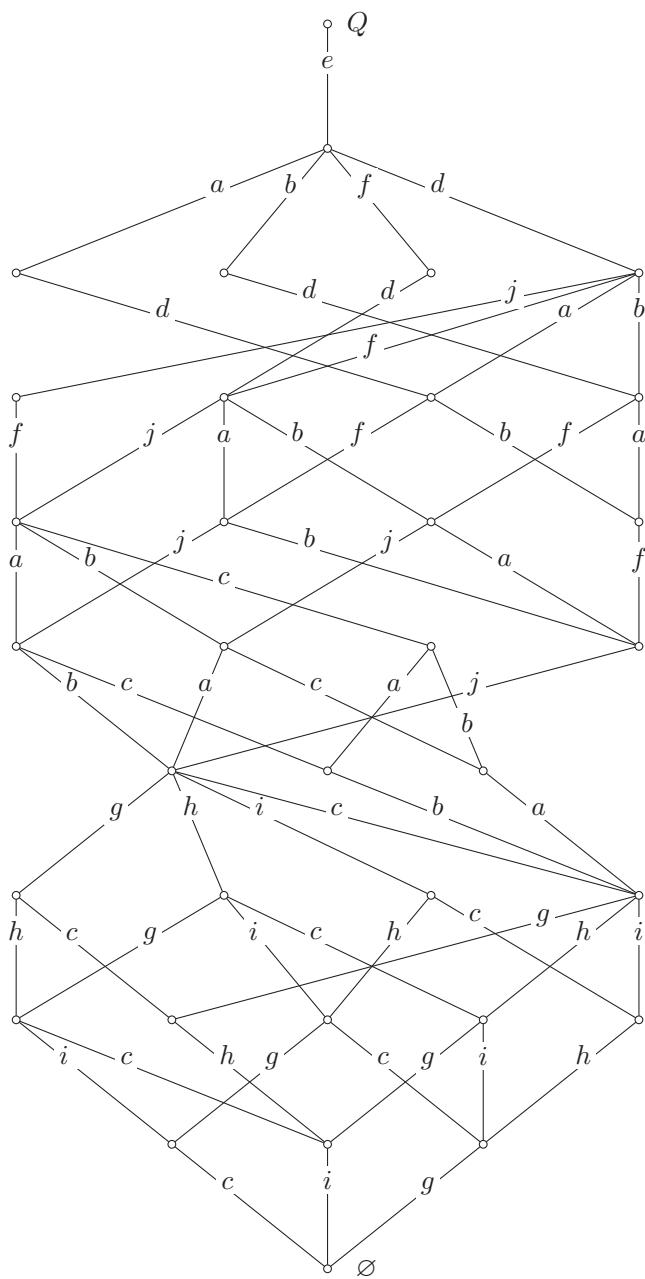
**Figure 5.6** The learning diagram representing the learning space in Example 5.6 (see also Figure 5.5).

(Sufficiency.) Given a collection  $\mathcal{W}$  of words satisfying (i)–(iii), set  $\mathcal{L} = \{\tilde{w} \mid w \in \mathcal{W}\}$ . Then  $\emptyset \in \mathcal{L}$ . Repeatedly applying Conditions (i) and (iii), we infer that there is a string in  $\mathcal{W}$ , and so  $Q \in \mathcal{L}$ . By (ii),  $\mathcal{L}$  is downgradable. To conclude that  $(Q, \mathcal{L})$  is a learning space it now suffices to prove that  $\mathcal{L}$  satisfies Condition (iii) in Theorem 5.2. Let  $K, K \cup \{q\}$  and  $K \cup \{r\}$  be states in  $\mathcal{L}$  with  $q \neq r$ . There are then some words  $v$  and  $w$  in  $\mathcal{W}$  such that  $\tilde{v} = K \cup \{q\}$  and  $\tilde{w} = K \cup \{r\}$ . Because  $\tilde{v} \setminus \tilde{w} = \{q\}$ , Condition (iii) implies  $w q \in \mathcal{W}$ . Now  $\tilde{w} q = K \cup \{q, r\}$ , and so  $K \cup \{q, r\} \in \mathcal{L}$ . Finally, it is easily checked that  $\mathcal{W}$  consists of all learning words of  $\mathcal{L}$ .  $\square$

Learning words and strings form a useful tool for the handling of large learning spaces. For instance, they are implicit in the new representation in Figure 5.6 of the learning space  $\mathcal{L}$  from Example 5.6: a learning string consists of the letters (representing items) on a path from the vertex representing  $\emptyset$  to the vertex representing  $Q$ . We call such a representation (with letters displayed only to show addition of a single item) a *learning diagram*.

Figure 5.7 shows a similar learning diagram for our 10-item example from Example 5.2.

Theorem 5.9 characterizes learning spaces through their complete collections of learning strings. In the same vein as the base which, containing only a relatively small number of states of the knowledge structure, gives us access to the whole collection, we might want to summarize in a similar way the collection of learning strings in a subcollection. The following definition comes from Eppstein (2013a).



**Figure 5.7** The learning diagram of the ten-item learning space  $\mathcal{L}$  of Example 5.2 (see also Figure 5.2).

**Definition 5.10** Let  $\mathcal{S}$  be a collection of strings on a finite domain  $Q$ . Form the collection  $\mathcal{L}_{\mathcal{S}}$  containing all possible unions of the sets determined by prefixes of strings in  $\mathcal{S}$ . Then  $(Q, \mathcal{L}_{\mathcal{S}})$  is the learning space *encoded* by  $\mathcal{S}$ .

That  $(Q, \mathcal{L}_{\mathcal{S}})$  indeed forms a learning space is easy to verify (for instance, apply Theorem 5.1(ii)). Now, conversely, given a learning space  $(Q, \mathcal{L})$ , we may select any nonempty subset  $\mathcal{S}$  of its collection of learning strings; then, in general,  $\mathcal{L}_{\mathcal{S}} \subseteq \mathcal{L}$  holds, but there is no reason to have equality here (see theorem 13.5.7 in Eppstein, 2013a, for a criterion).

The case in which we have the equality  $\mathcal{L}_{\mathcal{S}} = \mathcal{L}$  is interesting for algorithmic work on the learning space  $\mathcal{L}$ . Let us denote as  $S_1, S_2, \dots, S_k$  the strings forming  $\mathcal{S}$ . Then any state in  $\mathcal{L}$  is univocally encoded by a list of natural numbers  $n_1, n_2, \dots, n_k$ : each of the numbers specifies the length of the prefix we need to extract from the corresponding string in  $\mathcal{S}$  to get the state at hand as the union of the prefixes. Eppstein (2013a) shows how to exploit the new state encoding for various tasks. The present context generates the following problem: given a learning space, how do we compute the smallest number of learning strings needed to encode it? The ensuing invariant was dubbed *convex dimension* by Edelman and Jamison (1985) (see also Korte *et al.*, 1991). Eppstein (2013b) gives an algorithm to compute it.

**Example 5.8** The learning space of Example 5.6 (see also Figure 5.5) is encoded by the following three of its learning strings:

$$\begin{aligned} &a b d c, \\ &b a c d, \\ &b d c a. \end{aligned}$$

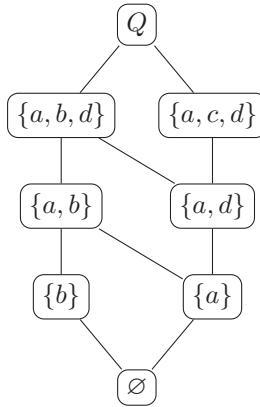
It is also encoded by the two strings

$$\begin{aligned} &a b c d, \\ &b d c a, \end{aligned}$$

but never by just one string.

## 5.8 The projection theorem

How large is the structure of a real-life learning space? For instance, what is the ratio of the number of knowledge states to the number of possible subsets of the domain? In the 10-item example of Table 5.1 and Figure 5.2 we have 34 knowledge states, which gives the ratio  $34/2^{10} \approx .03$ . However, this example may be misleading. In real-life learning spaces, the ratio may become considerably smaller as soon as there are a couple of dozen items. As another illustration, we again take the 37 items example in Beginning Algebra. There are 4615 knowledge states in the corresponding (induced) learning space. With just 37 items, the ratio of the number of states to the number of subsets is  $4615/2^{37} \approx .03 \times 10^{-6}$ . We mention in passing that there are 217 different knowledge states containing exactly 25 items.



**Figure 5.8** The covering diagram of the learning space in Example 5.9.

Presumably all these knowledge states would be assigned the same psychometric score in classical test theory, while KST treats them as different.

As mentioned in Section 5.1, the full domain of Beginning Algebra in the ALEKS system contains about 650 items. The complexity of the resulting learning space is daunting. It calls for ways of parsing such huge learning spaces into meaningful components. One of the goals could be a placement test for which only part of the full collection of items is needed. A more important reason arises when we need an assessment on the full structure. In such a case, the number of knowledge states is so large that a straightforward approach becomes infeasible.

A practical solution has been worked out which consists in suitably partitioning the domain and then carrying on simultaneous, parallel, mutually informative assessments of the resulting substructures, ultimately followed by the computation of the final state. We outline this technique in Section 5.10. Here, we define two useful concepts, that of a “projection” and that of “children” of a learning space, given a subset of the domain. We show – without proof – that such a projection remains a learning space, while in general children satisfy only some of the requirements in Definition 5.2.

We begin with an illustration based on a small learning space.

**Example 5.9** Let  $(Q, \mathcal{K})$  be the learning space on the domain  $\{a, b, c, d\}$  whose covering diagram is provided in Figure 5.8. Consider the subset  $Q' = \{c, d\}$  of the domain  $Q$  and form all the “traces”  $K \cap Q'$ , for  $K$  a state in  $\mathcal{K}$ . The resulting collection, the “projection” of  $\mathcal{K}$  on  $Q'$ ,

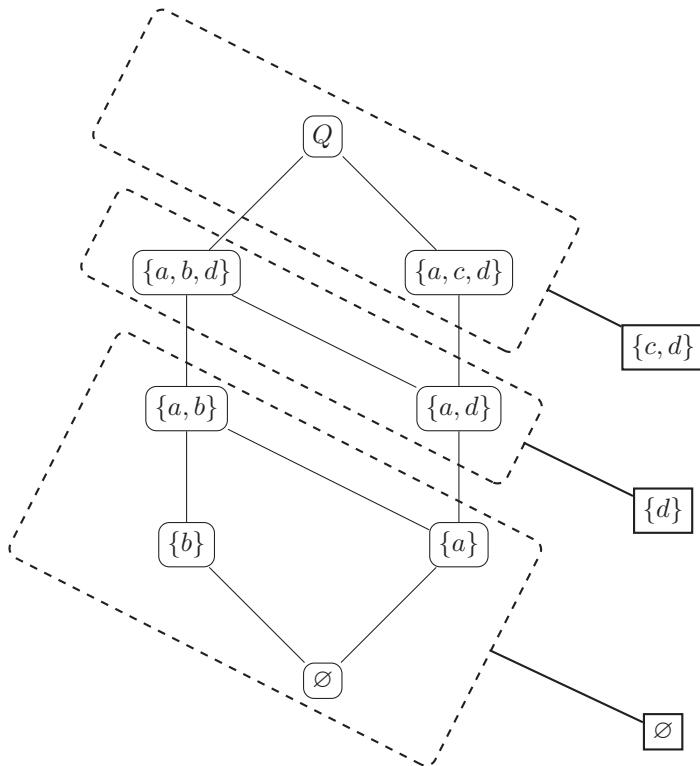
$$\{\emptyset, \{d\}, \{c, d\}\}, \quad (5.16)$$

forms again a learning space. The general result appears in Theorem 5.11(i) below. We summarize the construction in Figure 5.9.

With the same space we now illustrate another construction. This time, we sort out the states in  $\mathcal{K}$  according to their intersections with  $Q' = \{c, d\}$ . The resulting

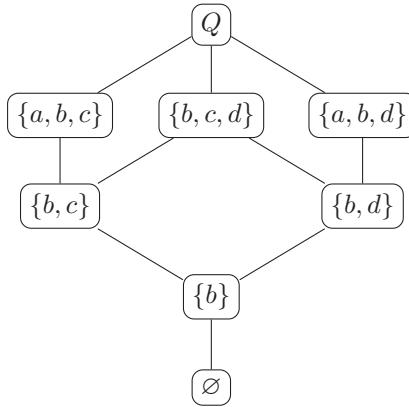
Table 5.4 *The states of the learning space  $(Q, \mathcal{K})$  in Example 5.9 are sorted according to their intersection with  $\{c, d\}$ . The third column provides the corresponding children.*

Classes of states	Intersections with $\{c, d\}$	Children
$\{\{a, c, d\}, Q\}$	$\{c, d\}$	$\{\emptyset, \{b\}\}$
$\{\{a, d\}, \{a, b, d\}\}$	$\{d\}$	$\{\emptyset, \{b\}\}$
$\{\emptyset, \{a\}, \{b\}, \{a, b\}\}$	$\emptyset$	$\{\emptyset, \{a\}, \{b\}, \{a, b\}\}$



**Figure 5.9** An illustration of the two constructions in Example 5.9 (with  $Q' = \{c, d\}$ ): the three equivalence classes are in the rounded, dashed rectangles; the traces are displayed on the right.

equivalence classes are displayed row by row in Table 5.4. Thus the states of  $\mathcal{K}$  in a same row all have the same trace on  $Q' = \{c, d\}$  (shown in the second column). It happens that they always form a “union-stable,” well-graded knowledge collection (see Theorem 5.11(ii) below). However, in view of the absence of  $\emptyset$ , the two first collections do not constitute a learning space. In the third column, we show the “children”; they are obtained by subtracting from the states (of  $\mathcal{K}$  shown in that row) their common intersection.



**Figure 5.10** The covering diagram of the learning space in Example 5.10.

In the second row of Table 5.4, all the states (of the original learning space) have in common the items  $d$  (by construction) and moreover  $a$ . Removing the two common items gives the two sets  $\emptyset$  and  $\{b\}$  which altogether form a learning space on the new domain  $\{b\}$ . The last assertion is not true in general.

**Example 5.10** Let  $(Q, \mathcal{K})$  be the learning space on the domain  $\{a, b, c, d\}$  whose covering diagram is provided in Figure 5.10. For  $Q' = \{a\}$ , one of the two children equals  $\{\{c\}, \{d\}, \{c, d\}\}$ ; it does not contain the empty set.

We now define the concept of a “projection,” and then that of “children.”

**Definition 5.11** Suppose that  $(Q, \mathcal{K})$  is a knowledge structure with  $|Q| \geq 2$ , and let  $Q'$  be any proper nonempty subset of  $Q$ . For  $K$  in  $\mathcal{K}$ , the subset  $K \cap Q'$  of  $Q'$  is the *trace* of  $K$  on  $Q'$ . The collection of all traces

$$\mathcal{K}_{|Q'} = \{K \cap Q' \mid K \in \mathcal{K}\} \quad (5.17)$$

is the *projection* of  $\mathcal{K}$  on  $Q'$ .

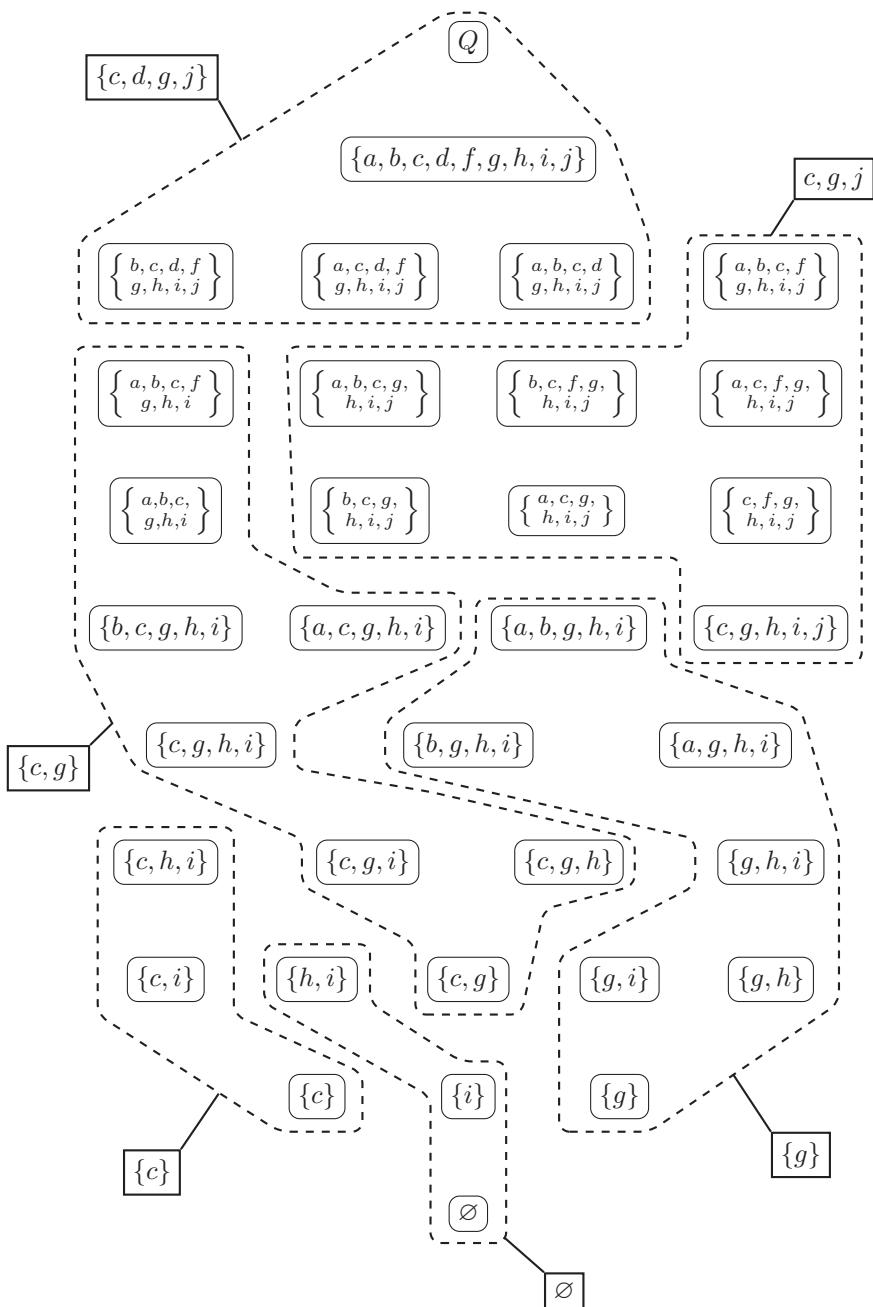
Figure 5.11 shows the trace of the 10-item learning space on  $\{c, d, g, j\}$ . Note that the sets in  $\mathcal{K}_{|Q'}$  may be or not be states of  $\mathcal{K}$ .

**Definition 5.12** Suppose again that  $(Q, \mathcal{K})$  is a knowledge structure with  $|Q| \geq 2$ , and let  $Q'$  be any proper nonempty subset of  $Q$ . Define the relation  $\sim_{Q'}$  on  $\mathcal{K}$  by

$$K \sim_{Q'} L \iff K \cap Q' = L \cap Q' \quad (5.18)$$

$$\iff K \Delta L \subseteq Q \setminus Q'. \quad (5.19)$$

(The equivalence between the right-hand sides of (5.18) and (5.19) is easily verified.) Then  $\sim_{Q'}$  is an equivalence relation on  $\mathcal{K}$ . When the context specifies the subset  $Q'$ , we may simply write  $\sim$  for  $\sim_{Q'}$ . We denote by  $[K]$  the equivalence class of  $\sim$  containing  $K$ , and by  $\mathcal{K}_\sim = \{[K] \mid K \in \mathcal{K}\}$  the partition of  $\mathcal{K}$  induced



**Figure 5.11** Projection of the 10-item example on  $\{c, d, g, j\}$ : the dashed lines delineate the equivalence classes, the little black rectangles show the traces.

by  $\sim$  (or by  $Q'$ ). For any state  $K$  in  $\mathcal{K}$ , we define the collection

$$\mathcal{K}_{[K]} = \{L \setminus \cap[K] \mid L \in [K]\}. \quad (5.20)$$

The collection  $\mathcal{K}_{[K]}$  is a  $Q'$ -child of  $\mathcal{K}$ , or simply a child of  $\mathcal{K}$  when the set  $Q'$  is made clear by the context. A child of  $\mathcal{K}$  may take the form of the singleton  $\{\emptyset\}$ ; it is then the trivial child. We refer to  $\mathcal{K}$  as the parent structure.

Because  $\emptyset \in \mathcal{K}$  we have  $\mathcal{K}_{[\emptyset]} = [\emptyset]$ . We may have  $\mathcal{K}_{[K]} = \mathcal{K}_{[L]}$  even when  $K \not\sim L$ . (Examples are easily built: see Table 5.4.)

**Theorem 5.11** *Let  $(Q, \mathcal{K})$  be a learning space, with  $|Q| \geq 2$ . The following two properties hold for any proper nonempty subset  $Q'$  of  $Q$ .*

- (i) *The projection  $\mathcal{K}_{|Q'}$  of  $\mathcal{K}$  on  $Q'$  is a learning space.*
- (ii) *The children of  $\mathcal{K}$  are well-graded and  $\cup$ -stable collections. This means that the union of any nonempty subcollection of the collection also belongs to the collection.<sup>12</sup> Example 5.9 shows that the children are not necessarily learning spaces.*

We can impose some restricting conditions on the set  $Q'$  that guarantee any child to be a learning space, provided that the empty state is added to the collection if necessary (see Falmagne and Doignon, 2011, definition 2.4.11 and theorem 2.4.12).

The concept of a projection plays an essential role in designing assessment algorithms for realistic learning spaces. In applications, the size of the collection of states may be so prohibitively large that the obvious strategy of gradually narrowing down, by some method or other, the class of states consistent with the assessment results is not practical. The solution discussed in Section 5.10.1 is to first design a suitable partition of the domain into  $N$  manageable classes. Second, one builds the  $N$  projections on these classes. The assessment procedure then operates in parallel on the  $N$  projections. A combination of the results of the  $N$  simultaneous assessments delivers a final knowledge state.

## 5.9 Probabilistic knowledge structures

The concept of a learning space is deterministic. As such, it does not provide realistic predictions of subjects' responses to the problems of a test. Probabilities must enter in at least two ways in a realistic model. For one, the knowledge states will certainly occur with different frequencies in the population of reference. So, it makes sense to postulate the existence of a probability distribution on the collection of states. For another, a subject's knowledge state does not necessarily specify the observed responses. A subject having mastered an item may be careless in responding, and make an error. Also, in some situations, a subject may

<sup>12</sup> Notice a slight difference between “union-stable” and “union-closed”; only the second condition entails that the empty set belongs to the collection.

be able to guess the correct response to a question not yet mastered.<sup>13</sup> In general, it makes sense to introduce conditional probabilities of responses, given the states.

**Definition 5.13** A *probabilistic (knowledge structure)*  $(Q, \mathcal{K}, p)$  consists of a finite knowledge structure  $(Q, \mathcal{K})$  with a probability distribution  $p$  on the collection  $\mathcal{K}$  of states. Thus  $p(K)$  is a real number with  $0 \leq p(K) \leq 1$ , for any state  $K$ , and moreover  $\sum_{K \in \mathcal{K}} p(K) = 1$ . A *parametrized (probabilistic knowledge structure)*  $(Q, \mathcal{K}, p, \beta, \eta)$  is a probabilistic knowledge structure  $(Q, \mathcal{K}, p)$  equipped with two functions  $\beta : Q \rightarrow [0, 1] : q \mapsto \beta_q$  and  $\eta : Q \rightarrow [0, 1] : q \mapsto \eta_q$ . The number  $\beta_q$  represents the *careless error probability* to item  $q$ , and the number  $\eta_q$  is the *lucky guess probability* for item  $q$ .

**Definition 5.14** A parametrized probabilistic knowledge structure  $(Q, \mathcal{K}, p, \beta, \eta)$  is *straight* if  $\beta_q = \eta_q = 0$  for all items  $q$ .

We now describe two types of construction of probability distributions on a set of states. In the first type, the set of states results from a projection.

**Definition 5.15** Let  $(Q, \mathcal{K}, p)$  be a probabilistic knowledge structure, and let  $\emptyset \neq Q' \subset Q$ . On the projection  $\mathcal{K}_{|Q'}$ , we define the *projected distribution*  $p'$  by setting, for  $K'$  a state in  $\mathcal{K}_{|Q'}$ ,

$$p'(K') = \sum \{p(K) \mid K \in \mathcal{K} \text{ and } K \cap Q' = K'\}.$$

Then  $(Q', \mathcal{K}_{|Q'}, p')$  is the *probabilistic projection* on  $Q'$  of the probabilistic knowledge structure  $(Q, \mathcal{K}, p)$ .

In the second case, we start with a probabilistic knowledge structure and extend it on a larger set of states.

**Definition 5.16** Let  $(Q, \mathcal{K})$  be a knowledge structure, and let  $\emptyset \neq Q' \subset Q$ . Assume  $p'$  is a probability distribution on the projection  $\mathcal{K}_{|Q'}$ . We define the *extended distribution*  $p^+$  to  $\mathcal{K}$  of  $p$  by setting, for  $K$  a state in  $\mathcal{K}$ ,

$$p^+(K) = \frac{p'(K \cap Q')}{\sum |\{L \in \mathcal{K} \mid L \cap Q' = K \cap Q'\}|}.$$

Then  $(Q, \mathcal{K}, p^+)$  is the *uniform extension* to  $(Q, \mathcal{K})$  of the probabilistic knowledge structure  $(Q', \mathcal{K}', p')$ .

## 5.10 The stochastic assessment algorithm

The general idea of the assessment algorithm is to gradually update, after each response of the subject, the distribution of probabilities on the collection of states. On each step of the assessment, the system selects an item, and presents a randomly chosen instance of that item to the student. The student's

<sup>13</sup> Such lucky guesses have probability zero or are very unlikely in assessment systems requiring open responses to the items, instead of multiple-choice. ALEKS is one of those systems.

response is evaluated and classified as “correct” or “false.” The result serves to update the probability distribution on the set of states. The new distribution is the starting point of the next step. Ultimately, only one or a few states will remain with a high probability. The system then chooses the final state.

The assessment algorithm we just sketched is applicable in the “straightforward situation,” that is, when the learning space (or the knowledge structure for that matter) is moderately large, with a domain not exceeding 50 items. Such learning spaces can serve in the design of some placement tests, for example. In Section 5.10.2, we deal with the more usual case of domains having hundreds of items. On such large domains, the application of the algorithm requires considerably more sophistication because the number of states becomes so large that operating on the probability distribution on the set of states is unmanageable. The solution outlined in Section 5.10.2 is to build a suitable partition of the domain and to perform parallel (simultaneous, mutually informative) assessments on the projections on all the subdomains. The final state is constructed by combining the outcomes obtained in each of the parallel assessments.

### 5.10.1 Sketch of the algorithm in the straightforward situation

We suppose that, at the outset of the algorithm execution, there exists some probability distribution on the collection of states.<sup>14</sup> Such a probability distribution may be inferred from some information on the population that the testee belongs to. Failing such information, we may simply assume that the distribution is uniform. The assessment is adaptive. On each of its main steps, the algorithm applies a Bayesian type updating operator of the current probability distribution, producing thus for each knowledge state an estimate of the probability that the student be in this state.

We keep illustrating our discussion by our example of the 10 items in Beginning Algebra and give in Figure 5.12 a picture of the learning space at the beginning of the assessment. Notice that the values of states probabilities are represented by shades of gray (the lighter the disk is, the higher the probability value is).

Figure 5.13 summarizes the sequence of events on each step. We call such steps *trials*. At the beginning of each trial, the algorithm picks the “most informative item”<sup>15</sup> to present to the student; the item selection relies on the current, estimated probabilities of the states. Next, it randomly chooses an instance of the item. For example, if the item selected is

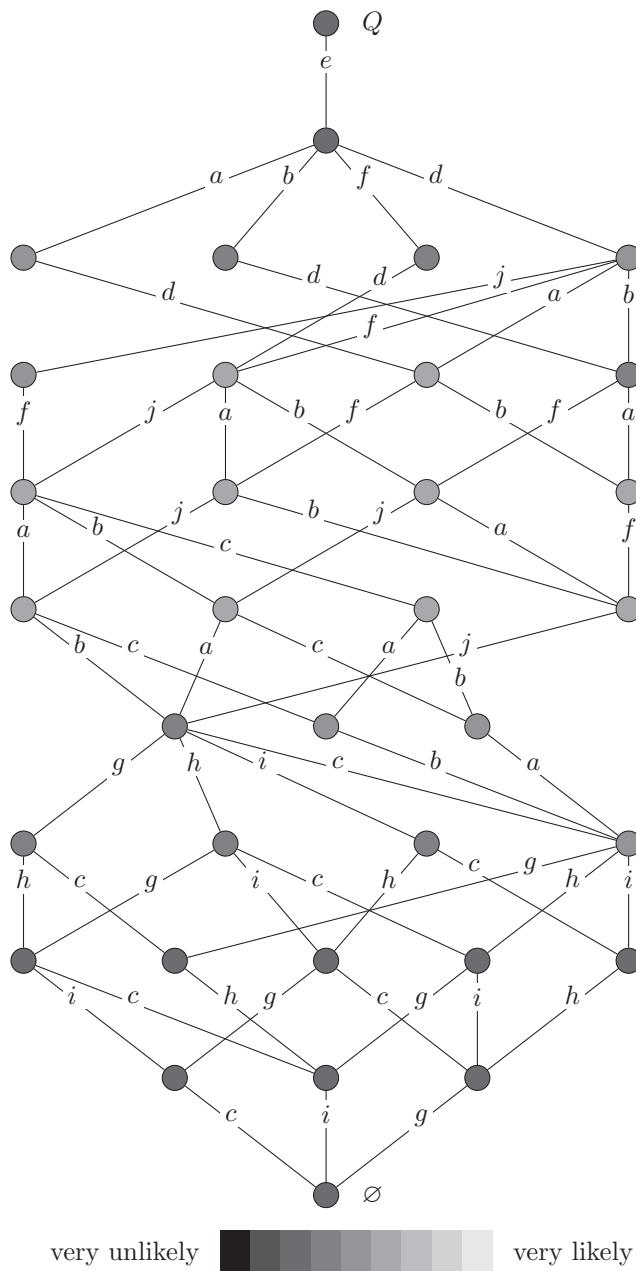
a: Quotients of expressions involving exponents

the instance could be

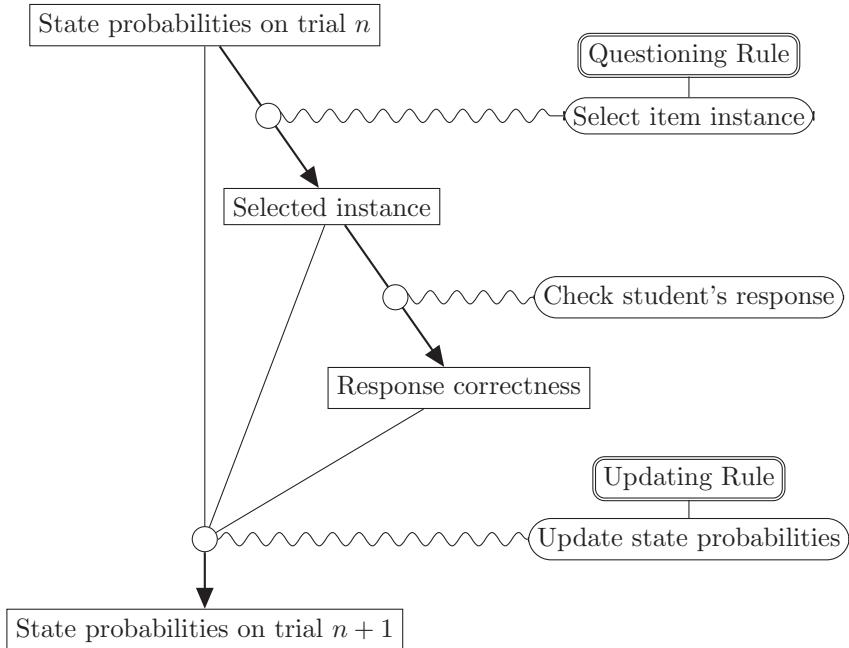
Simplify the expression  $\frac{a^4 b^5}{5a^6 b}$  as much as possible.

<sup>14</sup> This is one of the reasons why the algorithm cannot be readily applied in the case of large domains: we cannot manage a probability distribution on a set containing billions of states.

<sup>15</sup> For the technical meaning of this term, see the Questioning Rule.



**Figure 5.12** Initial probabilities of all the states at the beginning of the assessment. The probabilities are marked by the shading of the circles representing the states. Dark grey means very unlikely, and bright grey very likely.



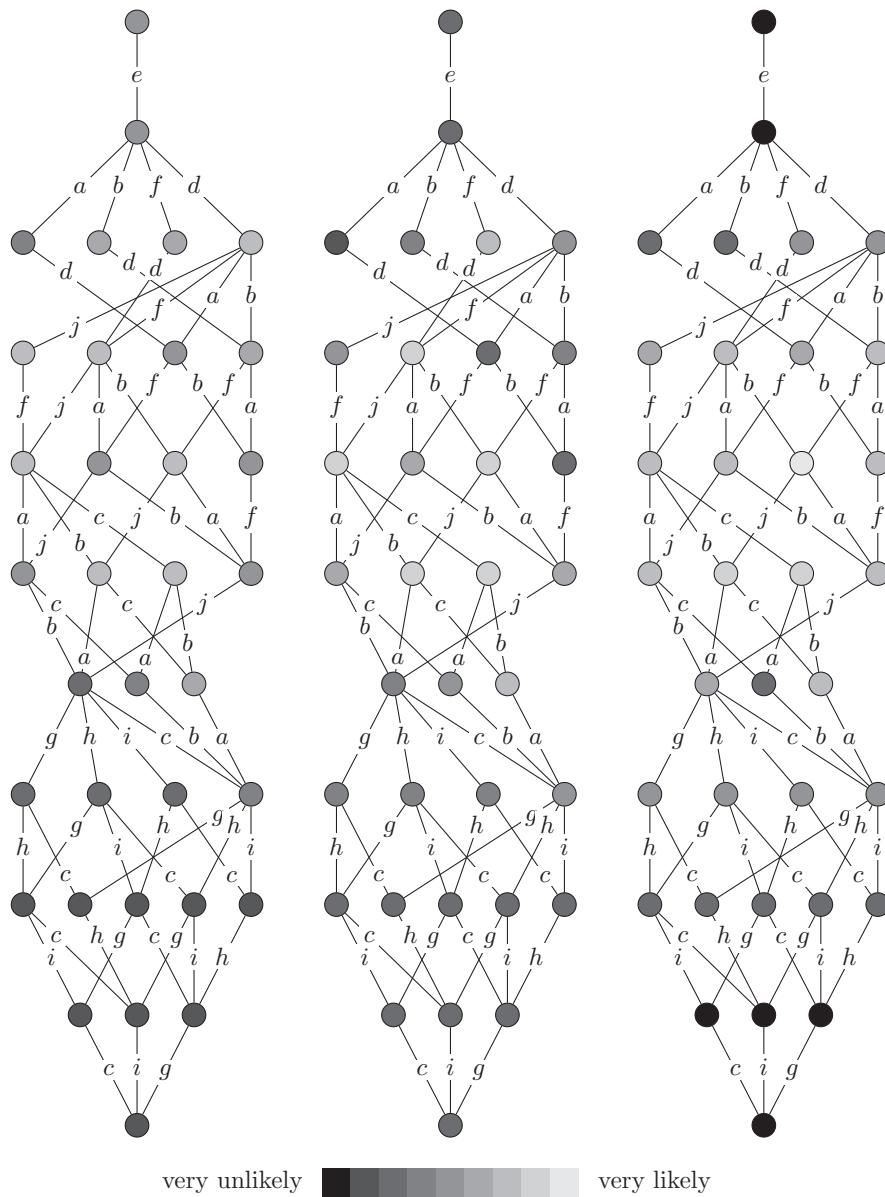
**Figure 5.13** Top to bottom, a schematic representation of the assessment from trial  $n$  to trial  $n + 1$ : on the left the data acted upon, on the right the general instructions executed according to specific rules.

The algorithm then proposes the instance to the student. It records the response and checks the correctness. To complete the trial, the algorithm uses the information just collected in order to update the probability distribution on the collection of states. Of course, we still need to specify how the item selection and the probability updating are performed (see the Questioning Rule and the Updating Rule, below).

The basic idea of the assessment algorithm is to ensure that, in the course of the assessment, the probability distribution becomes gradually concentrated on a knowledge state, or on a few knowledge states which are close together.<sup>16</sup> If several states end up with the same high probability, the system chooses randomly between them. In practical situations, the assessment terminates in about 25–35 trials with the algorithm we will describe. We first provide an illustration and then introduce the required notation.

Figure 5.12 pictured a (hypothetical) probability distribution on the set of states at the beginning of the assessment. The brightest shades indicate the high probability states, the darkest ones represent the lowest probability states. We infer from the picture that the student may have mastered about 4–7 items. Figure 5.14 shows the graphs of three later situations. Suppose during the first trial the student gives a correct response to item  $a$ . The graph on the left shows the updated probability values: note the increase of the probability values for states containing  $a$ ,

16 From the standpoint of their symmetric difference.



**Figure 5.14** Left to right, three successive situations along the assessment:  
(i) after a correct response to  $a$ ; (ii) next, after a false answer to  $f$ ; (iii) later,  
at the end of the assessment.

and the decrease of the values for the states not containing  $a$ . The middle graph sketches the estimated distribution at the end of trial 2, after the student gave next a false response to item  $f$ . The last graph represents the typical situation at the end of the assessment: only one state remains with a high probability, which is

$\{a, c, g, h, i\}$ . The algorithm would choose that state as representing the competence of the student in that part of Beginning Algebra.

We now give a precise mathematical meaning to the concepts outlined above, starting with a list of notations based on a knowledge structure  $(Q, \mathcal{K})$ .

## Mathematical notation

$n$	the step number, or <i>trial number</i> , with $n = 1, 2, \dots$ ;
$\mathcal{K}_q$	the subcollection of $\mathcal{K}$ formed by the states containing $q$ ;
$\Lambda_+$	the set of all positive probability distributions on $\mathcal{K}$ ;
$\mathbf{L}_n$	a positive random probability distribution on $\mathcal{K}$ : we have $\mathbf{L}_n = L_n \in \Lambda_+$ if $L_n$ is the probability distribution on $\mathcal{K}$ at the beginning of trial $n$ (so $L_n > 0$ );
$\mathbf{L}_n(K)$	a random variable (r.v.) evaluating the probability of state $K$ on trial $n$ ;
$\mathbf{Q}_n$	a r.v. representing the question asked on trial $n$ : we have $\mathbf{Q}_n = q$ if $q$ in $Q$ is the question asked on trial $n$ ;
$\mathbf{R}_n$	a r.v. coding the response on trial $n$ : $\mathbf{R}_n = \begin{cases} 1 & \text{if the response is correct,} \\ 0 & \text{otherwise;} \end{cases}$
$\mathbf{W}_n$	the random history of the process from trial 1 to trial $n$ ;
$\iota_A$	the indicator function of a set $A$ : $\iota_A(q) = \begin{cases} 1 & \text{if } q \in A, \\ 0 & \text{if } q \notin A; \end{cases}$
$\zeta_{q,r}$	a collection of parameters defined for $q \in Q$ , $r \in \{0, 1\}$ and satisfying $1 < \zeta_{q,r}$ (see the Updating Rule [U]).

The formal rules governing the assessment algorithm are as follows (where  $n$  is the trial number).

Using the “Questioning Rule,” the algorithm picks a most informative item, that is, an item which, on the basis of the current probability distribution on the set of states, has a probability of being responded to correctly as close to .5 as possible. There may be more than one such item, in which case a uniform, random selection is made. Formally, in terms of the actual probability distribution  $L_n$ :

[Q] **Questioning Rule.** For all  $q \in Q$  and all positive integers,

$$\mathbb{P}(\mathbf{Q}_n = q \mid \mathbf{L}_n, \mathbf{W}_{n-1}) = \frac{\iota_{S(L_n)}(q)}{|S(L_n)|} \quad (5.21)$$

where  $S(L_n)$  is the subset of  $Q$  containing all those items  $q$  minimizing  $|2L_n(\mathcal{K}_q) - 1|$ . Under this questioning rule, which is called *half-split*, we must have  $\mathbf{Q}_n \in S(L_n)$  with a probability equal to one. The questions in the set  $S(L_n)$  are then chosen with equal probability.

The “Response Rule” formalized below states that the student’s response to an instance of an item is correct with probability 1 if the item belongs to the student’s knowledge state  $K_0$ , and false with probability 1 otherwise. While this rule is

unrealistic and plays no role in the assessment algorithm per se, it is essential in some simulations. The rule used in practice is the “Modified Response Rule” (see below). These rules are used, in particular, in the “Extra problem method” (see Subsection 5.12.1).

**[R] Response Rule.** There is some state  $K_0$  such that, for all positive integers  $n$ ,

$$\mathbb{P}(\mathbf{R}_n = \iota_{K_0}(q) \mid \mathbf{Q}_n = q, \mathbf{L}_n, \mathbf{W}_{n-1}) = 1. \quad (5.22)$$

The state  $K_0$  is the *latent state* representing the set of all the items currently mastered by the student. It is the state that the assessment algorithm aims to uncover.

**[U] Updating Rule.** We have  $\mathbb{P}(\mathbf{L}_1 = L) = 1$  (with  $L$  the initial probability distribution). Moreover, there are real parameters  $\zeta_{q,r}$  (where  $q \in Q$ ,  $r \in \{0, 1\}$  and  $1 < \zeta_{q,r}$ ) such that for any positive integer  $n$ , if  $\mathbf{L}_n = L_n$ ,  $\mathbf{Q}_n = q$ ,  $\mathbf{R}_n = r$  and

$$\zeta_{q,r}^K = \begin{cases} 1 & \text{if } \iota_K(q) \neq r, \\ \zeta_{q,r} & \text{if } \iota_K(q) = r, \end{cases} \quad (5.23)$$

then

$$L_{n+1}(K) = \frac{\zeta_{q,r}^K L_n(K)}{\sum_{K' \in \mathcal{K}} \zeta_{q,r}^{K'} L_n(K')}. \quad (5.24)$$

This updating rule is called *multiplicative with parameters  $\zeta_{q,r}$* . The operator mapping  $L_n$  to  $L_{n+1}$  has two important, related properties. First, it is commutative: the order of the successive items of the assessment has no effect on the result, that is, on the final probability distribution on the set of states. Second, the operator is essentially Bayesian (this was proved by Mathieu Koppen<sup>17</sup>).

It can be shown that, under the above Rules [Q], [R] and [U], the latent state is recoverable in the sense that:

$$\mathbf{L}_n(K_0) \xrightarrow{\text{a.s.}} 1. \quad (5.25)$$

We recall that “a.s.” stands for “almost surely.” For a proof, see theorem 1.6.7 in Falmagne and Doignon (2011). Note that this theorem assumes that the careless errors and the lucky guesses have probability zero (the straight case as in Definition 5.14).

While Rules [U], [Q], and [R] are fundamental to the assessment, a straightforward implementation of these rules in an assessment engine is not feasible for two reasons. One is that students often make careless errors. A solution is to amend Rule [R] by introducing a “careless error parameter.” This would result, for example, in the modified rule:

<sup>17</sup> Personnal communication; see subsection 13.4.5 in Falmagne and Doignon (2011).

[R'] **Modified Response Rule.** For all positive integers  $n$ ,

$$\mathbb{P}(\mathbf{R}_n = \iota_{K_0}(q) \mid \mathbf{Q}_n = q, \mathbf{L}_n, \mathbf{W}_{n-1}) = \begin{cases} 1 - \beta_q & \text{if } q \in K_0 \\ 0 & \text{if } q \notin K_0, \end{cases} \quad (5.26)$$

in which  $\beta_q$  is the probability of making an error in responding to the item  $q$  which lies in the latent state  $K_0$ . The parameters  $\beta_q$  can be estimated from the data (see Section 5.12.1). In some cases, “lucky guess” parameters may also be used, for example to deal with cases of the multiple choice paradigm. However, in real-life situations, the occurrence of lucky guesses would render the result of the assessment so unreliable as to render it practically useless. Multiple choice paradigms are cheap, but their results are questionable and they should be avoided. The ALEKS assessment system, which is based on the three rules [U], [Q], and [R'], rarely use multiple choice. When it does, the number of possible responses is so large that the probability of a lucky guess is negligible.

The second reason is that, in many real life applications, the domain formalizing an actual curriculum is typically very large, containing several hundred items. The learning space can still be constructed, but its collection of states is huge, counting many million states. It means that the assessment algorithm cannot proceed in the straightforward manner outlined above: the probability distribution on such a large collection of states cannot be managed. We now expose a solution.

### 5.10.2 The parallel algorithm for large domains

In case the domain  $Q$  of the “parent learning space”  $\mathcal{L}$  is large, the algorithm first partitions  $Q$  into  $N$  subdomains  $Q^1, Q^2, \dots, Q^N$ . Here, the subdomains are approximately of equal sizes, and they are manageable because  $N$  is chosen sufficiently large. They are representatives of the parent domain  $Q$ , in the sense, for example, that each one of them could serve as a placement test. By the projection theorem 5.11(i), these  $N$  subdomains determine  $N$  projections  $\mathcal{L}^1, \mathcal{L}^2, \dots, \mathcal{L}^N$  which are all learning spaces. The algorithm then manages  $N$  parallel, alternating, mutually informative assessments on the  $N$  projected learning spaces  $(Q^i, \mathcal{L}^i)$  (where  $i = 1, 2, \dots, N$ ). This means that if the chosen item  $q$  belongs to, say, the subdomain  $Q^i$ , the updating of the probability distribution on  $\mathcal{L}^i$  is performed. Next, the information provided by the response to  $q$  is transferred to all the other  $N - 1$  learning spaces with corresponding updating of their state probabilities.

The main features of the algorithm are described in Falmagne *et al.* (2013, see section 8.8, pp. 143–144). The basic ideas are as follows.

- (1) On each trial, scan the  $N$  learning spaces and choose the “most informative” item to be presented to the student. Suppose that this item is  $q$ , belonging to the subdomain  $Q^i$  of the projected learning space  $(Q^i, \mathcal{L}^i)$ .
- (2) Record the student’s response to item  $q$ . Update the state probabilities of the learning space  $\mathcal{L}^i$ .

- (3) Update the state probabilities of all the other learning subspaces  $\mathcal{L}^j$ ,  $j \neq i$ . This is achieved by temporarily adding item  $q$  to the other learning spaces  $\mathcal{L}^j$  and working with extended distributions (cf. Definition 5.16).
- (4) At the end of the assessment, combine the information gathered in all the  $N$  learning subspaces into a final knowledge state in  $\mathcal{L}$ . For details, see Falmagne *et al.* (2013, subsection 8.8.1, p. 144).

## 5.11 About building knowledge spaces or learning spaces

Building a knowledge space on a given domain is a highly demanding, data-driven enterprise. Building a learning space is an even more intricate process. Chapters 15 and 16 of Falmagne and Doignon (2011) describe a method in detail. We only sketch the main ideas here, proceeding in two steps. We first describe some theoretical constructions which are instrumental for building knowledge spaces. We then show how these tools can be amended in the case of learning spaces.

The basic algorithm is the “QUERY” routine, which is due to Koppen (1993) and Müller (1989) (see also Dowling, 1994). The routine manages the following type of queries, which are either put to expert teachers or have responses inferred from assessment data.

[Q] *Suppose that a student under examination has just provided wrong responses to all the items in some set A. Is it practically certain that this student will also fail item q? Assume that the conditions are ideal in the sense that errors and lucky guesses are excluded.*

Such a query is summarized as the pair  $(A, q)$ , with  $\emptyset \neq A \subset Q$  and  $q \in Q$ . The QUERY routine gradually builds the relation  $\mathcal{P}$  from  $2^Q$  to  $Q$  consisting of all the queries  $(A, q)$  which receive a positive response. For  $m = |Q|$ , there are  $m(2^{m-1} - 1)$  useful queries (asking a query  $(A, q)$  with  $q \in A$  is useless: the answer must be affirmative). So their number is superexponential in  $m$ . Fortunately, in most cases, only a small fraction of them need to be asked: as we explain below, the responses to some queries may be inferred from the response to previously asked queries. The importance of the queries regarding the problem at hand – building a knowledge space or a learning space – lies in the following theorem (Koppen and Doignon, 1990).

**Theorem 5.12** *For a finite domain  $Q$ , there is a one-to-one correspondence between the collection of knowledge spaces  $\mathcal{K}$  on  $Q$  and the set of relations  $\mathcal{P}$  from  $2^Q \setminus \{\emptyset\}$  to  $Q$  satisfying for all  $q$  in  $Q$  and  $A, B$  in  $2^Q \setminus \{\emptyset\}$ :*

- (i) *if  $q \in A$ , then  $A\mathcal{P}q$ ;*
- (ii) *if  $A\mathcal{P}b$  for all  $b$  in  $B$  and  $B\mathcal{P}q$ , then  $A\mathcal{P}q$ .*

One such correspondence is defined by the two equivalences, where  $r \in Q$ ,  $K, B \subseteq Q$ ,

$$K \in \mathcal{K} \iff (\forall (A, q) \in \mathcal{P} : A \cap K = \emptyset \implies q \notin K), \quad (5.27)$$

$$(B, r) \in \mathcal{P} \iff (\forall K \in \mathcal{K} : B \cap K = \emptyset \implies r \notin K). \quad (5.28)$$

In view of Formula (5.27), a positive answer to the query  $(A, q)$  may lead to remove some subsets from the collection of potential states (namely, all the subsets  $K$  satisfying both  $A \cap K = \emptyset$  and  $q \in K$ ). Negating both sides of Formula (5.27) gives us the equivalent formula

$$K \notin \mathcal{K} \iff (\exists (A, q) \in \mathcal{P} : A \cap K = \emptyset \text{ and } q \in K), \quad (5.29)$$

which clearly shows that any non-state  $K$  is ruled out by a positive answer to some query. Thus, Theorem 5.12 is instrumental for building a knowledge space – which may or may not be a learning space. We outline below how we can adapt the implementation of Theorem 5.12 so that learning spaces are generated.

In the QUERY procedure, Block  $j$  consists of all the queries  $(A, q)$  with  $j = |A|$ . The procedure starts with Block 1, followed by Block 2, etc. In principle an expert teacher is able to provide the answer to any query in Block 1, because such a query takes the following simple form.

[Q1] Suppose that a student under examination has just provided a false response to question  $q$ . Is it practically certain that this student will also fail item  $r$ ? Assume that the conditions are ideal in the sense that errors and lucky guesses are excluded.

Collecting the responses to such queries  $(\{r\}, q)$ , or  $(r, q)$  for short, with  $r, q$  in  $Q$  and  $r \neq q$  amounts to constructing a relation  $R$  on the set  $Q$ . However, not all such queries need to be asked. Assuming that the expert is consistent, we can infer the responses of some queries from other queries. For instance if the responses to the queries  $(r, t)$  and  $(t, q)$  are positive, then the response to  $(r, q)$  should also be positive. In general, we only need to ask the expert a relatively small set of queries of the type [Q1]. Their responses yield some relation  $R$ . The *transitive closure*<sup>18</sup>  $t(R)$  of the relation  $R$  is a quasi order on  $Q$ . By Birkhoff's theorem 5.3, the quasi order  $t(R)$  uniquely defines a quasi ordinal knowledge space on  $Q$ , that is, a knowledge structure closed under union and intersection. The design of the QUERY algorithm ensures that this quasi order is in fact a partial order. Thus, the quasi ordinal structure is an ordinal space  $\mathcal{L}_1$ , namely, a learning space closed under intersection. The learning space  $\mathcal{L}_1$  contains all the actual (true) knowledge states of the learning space under construction. However, it also typically contains a possibly very large number of false states, which are due to the closure of intersection of  $\mathcal{L}^1$ .

While human experts are capable of providing useful responses to queries of the type [Q1], their responses to queries of higher blocks are less reliable. Fortunately,

<sup>18</sup> That is, the smallest transitive relation on  $Q$  that includes  $R$ .

despite the presence of false states, the learning space  $\mathcal{L}_1$  is sufficiently informative to be used in the schools and colleges.<sup>19</sup> The data collected by assessments using  $\mathcal{L}_1$  can then be used to simulate human expert responses to queries of higher block numbers, such as Block 2 or Block 3.

Here is how, taking Block 2 as our example. Assuming that we have a very large collection of assessment data,<sup>20</sup> we can estimate the conditional probabilities

$$P(q; r, t) \quad \text{of failing item } q \text{ given that both items } r \text{ and } t \text{ are failed.}$$

Choosing a suitably large threshold value  $\theta$ , with  $0 < \theta < 1$ , we estimate the relation  $\mathcal{P}$  for Block 2 by the rule

$$\{r, t\}\mathcal{P}q \quad \text{exactly when} \quad P(q; r, t) > \theta. \quad (5.30)$$

Using such an estimate and Formulas (5.29) and (5.30) would implement Block 2 of the QUERY routine. However, the resulting knowledge space would not necessarily be a learning space. Consequently, not all positive responses  $\{r, t\}\mathcal{P}q$  should be implemented and result in the elimination of states. This remark also applies in the general case, to the queries of any block.

The QUERY routine has been adapted for the construction of learning spaces. As suggested by our example of Block 2, the general idea is to verify that the removal of a state by the application of Theorem 5.12 would not result in a violation of the learning space axioms. The key result is theorem 16.1.6 in Falmagne and Doignon (2011), which is restated as Theorem 5.14 here. It relies on the concept of a “hanging state.”

**Definition 5.17** A nonempty state  $K$  in a knowledge structure  $(Q, \mathcal{K})$  is *hanging* if its inner fringe  $K^{\mathcal{T}} = \{q \in Q \mid K \setminus \{q\} \in \mathcal{K}\}$  (cf. Definition 5.8) is empty. The state  $K$  is *almost hanging* if it contains more than one item, but its inner fringe consists of a single item.

**Example 5.11** The knowledge space

$$\mathcal{L} = \{\emptyset, \{a\}, \{b\}, \{a, b\}, \{a, c\}, \{a, d\}, \{a, b, c\}, \{a, b, d\}, \{a, c, d\}, Q\},$$

depicted in Figure 5.15 has no hanging states, and has two almost hanging states, which are  $\{a, c\}$  and  $\{a, d\}$ . According to Theorem 5.13,  $\mathcal{L}$  must be a learning space.

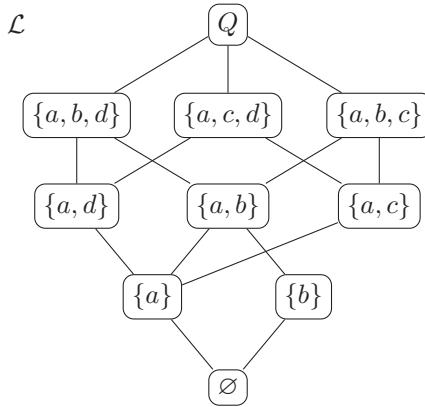
The following result is straightforward.

**Theorem 5.13** *A finite knowledge space is a learning space if and only if it has no hanging state.*

Theorem 5.13 indicates that, in the application of the QUERY routine, we must avoid the creation of hanging states (which could result from the removal of states

<sup>19</sup> The likely reason is that any false state  $L$  may be close to some true state  $K$  (in the sense of the symmetric difference distance  $|K \Delta L| = |(L \setminus K) \cup (K \setminus L)|$ ).

<sup>20</sup> Which is the typical case in the ALEKS system, for example.



**Figure 5.15** The covering diagram of the knowledge space in Example 5.11.

by a positive response to a query). We need one more tool to analyze the effect of such a positive response.

**Definition 5.18** Let  $(Q, \mathcal{K})$  be a knowledge space and let  $(A, q)$  be any query such that  $\emptyset \neq A \subset Q$  and  $q \in Q \setminus A$ . For any subfamily  $\mathcal{F}$  of  $\mathcal{K}$ , we define

$$\mathcal{D}_{\mathcal{F}}(A, q) = \{K \in \mathcal{F} \mid A \cap K = \emptyset \text{ and } q \in K\}. \quad (5.31)$$

Thus,  $\mathcal{D}_{\mathcal{K}}(A, q)$  is the subfamily of all those states of  $\mathcal{K}$  that would be removed by a positive response  $A\mathcal{P}q$  to the query  $(A, q)$  in the framework of the `QUERY` routine.

**Theorem 5.14** For any knowledge space  $\mathcal{K}$  and any query  $(A, q)$ , the family of sets  $\mathcal{K} \setminus \mathcal{D}_{\mathcal{K}}(A, q)$  is a knowledge space. If  $\mathcal{K}$  is a learning space, then  $\mathcal{K} \setminus \mathcal{D}_{\mathcal{K}}(A, q)$  is a learning space if and only if there is no almost hanging state  $L$  in  $\mathcal{K}$  such that  $A \cap L = L^{\mathcal{I}}$  and  $q \in L$ .

The *adapted* `QUERY` routine for building a learning space is based on Theorem 5.14. As we described above, we obtain a learning space (an ordinal space)  $\mathcal{L}_1$  at the end of Block 1. The next query from Block 2 is of the form  $(\{r, t\}, q)$ . A positive response  $\{r, t\}\mathcal{P}q$  induced from the estimated conditional probabilities of Statement (5.30) would lead to the removal from  $\mathcal{L}_1$  of any state  $K$  such that

$$\{r, t\} \cap K = \emptyset \quad \text{and} \quad q \in K, \quad (5.32)$$

provided  $\mathcal{L}_1 \setminus \{K\}$  does not contain any hanging state (cf. Formula (5.29)). However, there is one more subtlety. The fact that  $\mathcal{L}_1 \setminus \{K\}$  contains some hanging state does not necessarily mean that the query  $(\{r, t\}, q)$  must be discarded forever. Indeed, it may be that, at a later stage of the execution of the algorithm, after the removal of some other states, implementing  $\{r, t\}\mathcal{P}q$  no longer creates hanging states. The adapted `QUERY` routine takes care of such details (see Falmagne and Doignon, 2011, chapters 15 and 16).

So far, only Block 2 of the adapted *QUERY* routine has been implemented to build the learning space  $\mathcal{L}_2$  in realistic situations involving several hundred items, such as Beginning Algebra as the subject is taught in the USA. Remarkably, in these cases, it was observed that 99% of the false states in  $\mathcal{L}_1$  are removed to form  $\mathcal{L}_2$ . It is highly plausible that even better, smaller learning spaces would result from implementing Block 3 to  $\mathcal{L}_2$ , and maybe even higher blocks later on. We have been told by the ALEKS team<sup>21</sup> that this demanding enterprise is in progress.

Doignon (2014) recently proposed another modification of the *QUERY* routine to build a learning space. He uses a fundamental property of the collection of all the knowledge structures on the domain  $Q$ , a property which follows from results in Caspard and Monjardet (2004). Here, a knowledge structure  $\mathcal{K}$  on the domain  $Q$  is included in a knowledge structure  $\mathcal{L}$  also on  $Q$  when any state in  $\mathcal{K}$  is also a state in  $\mathcal{L}$  (that is,  $\mathcal{K} \subseteq \mathcal{L}$ ).

**Theorem 5.15** *For a given finite knowledge space  $(Q, \mathcal{K})$ , there are two cases. Either there is no learning space on  $Q$  included in  $\mathcal{K}$ , or among all the learning spaces on  $Q$  included in  $\mathcal{K}$ , there is one which includes any other one.*

When the second case in Theorem 5.15 occurs, we denote by  $\mathcal{K}^\Delta$  the largest learning space included in the knowledge space  $\mathcal{K}$ . Theorem 5.16 provides a description of  $\mathcal{K}^\Delta$  in terms of “gradations” (Theorems 5.15 and 5.16 first appeared in Doignon, 2014).

**Definition 5.19** Let  $(Q, \mathcal{K})$  be a finite knowledge structure. A *learning path* in  $\mathcal{K}$  is a maximal chain of states. A *gradation*  $\mathcal{C}$  is a learning path which is downgradable (Definition 5.3), that is: for any nonempty state  $K$  in  $\mathcal{C}$  there is some item  $q$  in  $K$  for which  $K \setminus \{q\} \in \mathcal{K}$ .

Notice that a learning path necessarily contains both states  $\emptyset$  and  $Q$ . A gradation always consists of  $1 + |Q|$  states of respective sizes  $0, 1, \dots, |Q|$ , and it forms itself a learning space on  $Q$ . There is an obvious correspondence between gradations and learning strings (Definition 5.9).

**Theorem 5.16** *Let  $\mathcal{K}$  be a knowledge space on the finite domain  $Q$ . If  $\mathcal{K}$  includes at least one learning space on  $Q$ , then the largest learning space  $\mathcal{K}^\Delta$  on  $Q$  included in  $\mathcal{K}$  is formed by all the states in all gradations in  $\mathcal{K}$ .*

As the classical *QUERY* routine, the *adjusted QUERY routine* proposed in Doignon (2014) repeatedly asks queries and maintains a collection  $\mathcal{L}$  of subsets of  $Q$  which, although it is decreasing, always forms a learning space on  $Q$ . When the query  $(A, q)$  prompts a positive answer from the expert (or the database), the routine builds the collection  $\mathcal{K} = \mathcal{L} \setminus \mathcal{D}_{\mathcal{L}}(A, q)$ . By Theorem 5.14, the collection  $\mathcal{K}$  is again a space. Now there are two cases:

- (i)  $\mathcal{K}$  includes some learning space (in other words: there is at least one gradation in  $\mathcal{K}$ , cf. Theorem 5.16). Then, in view of Theorem 5.15,  $\mathcal{K}$  must include for

21 Personal communication.

sure the learning space  $\mathcal{K}^\Delta$ . The routine then replaces  $\mathcal{L}$  with the latter learning space and asks a new query, if any is left unanswered.

- (ii)  $\mathcal{K}$  does not include any gradation. Then the routine exits execution, delivering the actual collection  $\mathcal{L}$ .

Killing the routine in Case (ii) makes sense, because there is no learning space that complies with the answers collected to queries. On the contrary, when the query answers are coherent in the sense that they reflect some latent learning space, Case (ii) does not occur. Even more, the routine uncovers the latent learning space.

**Theorem 5.17** *If  $\mathcal{L}$  is a latent learning space on the finite domain  $Q$  and the query answers are truthful with respect to  $\mathcal{L}$ , then the adjusted QUERY routine will ultimately uncover  $\mathcal{L}$ .*

In any case, the adjusted QUERY routine always produces a learning space (even in case it reached Case (ii) and therefore exited execution). We will not go into the details of the implementation of the adjusted QUERY routine, nor report any comparison of its performances with those of the adapted QUERY routine.

## 5.12 Some applications – the ALEKS system

In psychometrics, the concepts of “validity” and “reliability” are distinct<sup>22</sup> and dealt with separately, which is justified because a psychometric test is a measurement instrument which is not automatically predictive of a criterion. In the applications of learning space theory (LST) such as ALEKS, however, the items potentially<sup>23</sup> used in any assessment are, by construction, a fully comprehensive coverage of a curriculum, typically based on the consultation of numerous standard textbooks. This implies that if an LST-type assessment is reliable, it is presumably also valid, and vice versa (cf. section 17.1 in Falmagne and Doignon, 2011).

In our Introduction, we have presented LST as a more predictive, or valid, alternative to psychometrics. We describe here four analyses of data which all vindicate that position. The material is taken from Cosyn *et al.* (2013). However, the statistical analysis is so complex and detailed that we can only provide a summary here.

### 5.12.1 The extra problem method

The first inquiry investigates whether the knowledge state uncovered at the end of an assessment is predictive of the responses to questions that were not asked. For example, there are 416 items in the elementary school mathematics domain used in the ALEKS system, of which at most 35 are used in any assessment. Can we reliably predict the student’s responses to the  $381 = 416 - 35$  remaining questions?

<sup>22</sup> See Anastasi and Urbina (1997) for explanations.

<sup>23</sup> Potentially, in the sense that any item of the domain could be part of the test.

Table 5.5 *Basic data matrix for the computation of the correlation between the cases “in or out of the final state” and the student’s response coded as 1/0 for correct/false. So, theoretically, z stands for the number of lucky guesses, and y for the number of careless errors.*

		Response	
		1 (correct)	0 (false)
State	In	$x$	$y$
	Out	$z$	$w$

Table 5.6 *Values of the coefficients correlating the 1/0 variables coding the in/out of the final state and the correct/false response in the two cases: median value of the correlation distribution and grouped data.*

	Tetrachoric	Phi
Median	.68	.43
Grouped data	.80	.58

To find out, a randomly chosen *extra problem* is asked in any assessment, the response to which is not used in uncovering the state. The authors of Cosyn *et al.* (2013) have evaluated the correlation between the response to the extra problem – coded as 1/0 for correct/false – and a couple of predictive indices obtained from or during the assessment.

One predictive index is the dichotomic variable coding the observation that the extra problem is either in or out of the final state. Accordingly, the data takes the form of the 2 by 2 matrix in Table 5.5.

Because no correlation coefficient is available that would be fully adequate for the situation, two potentially useful ones were computed by Cosyn and his colleagues for these data: the tetrachoric and the Phi coefficient. The results are given in Table 5.6 for the median values of the coefficients and the grouped data. These data pertain to 125,786 assessments using 324 problems out of the 370 problems mentioned above.<sup>24</sup> The tetrachoric values are in the left column of the table. The median of the distribution is around .68. The grouped data, obtained from gathering the 324 individual  $2 \times 2$  matrices into one, yields a higher correlation of about .80. These are high values, but the tetrachoric coefficient is regarded as biased upward (however, see Greer *et al.*, 2003). The right column of the table contains similar values for the phi coefficient. These values are much lower, yielding a median of .43 (contrasting with the .68 value obtained for the tetrachoric) and a grouped data value of .58 (instead of .80).

24 Forty-six problems were discarded because the relevant data were not sufficient to provide reliable estimates of the coefficients.

### 5.12.2 Correcting for careless errors

However, the relatively low correlation values obtained for the phi coefficient in the above analysis may be due in part to the occurrence of careless errors. However, the basic data matrix of Table 5.6 may be revised to include such careless errors. Instead of the dichotomic *in/out (of the final state)* variable, Cosyn *et al.* (2013) define the new variable

$$S_a = \begin{cases} 1 - \epsilon_a & \text{if the final state contains the extra question } a, \\ 0 & \text{otherwise,} \end{cases} \quad (5.33)$$

in which  $\epsilon_a$  stands for the probability of committing a careless error to item **a**. So, if the extra question **a** is contained in the uncovered state,  $S_a$  is the probability of not committing a careless error to item **a**. The careless errors probabilities  $\epsilon_a$  were estimated from those cases in which, by chance, the same item **a** appears twice in an assessment, once as the extra problem and the other as one of the other items, and, moreover, the response in at least one of the two instances of **a** is correct. The relative frequency of cases in which a false response is given to the other instance of **a** provides an estimate of  $\epsilon_a$ .

The variable  $S_a$  is neither exactly continuous nor exactly discrete.<sup>25</sup> Nevertheless, for the purpose of comparison with similar analyses performed in psychometric situations, the authors have used the point biserial coefficient  $r_{pbis}$  to compute the correlation between the variables  $S_a$  and the 1/0 variable coding the correct/incorrect responses to the extra problem. The value reported for  $r_{pbis}$  was .67, noticeably higher than the .58 obtained for the phi coefficient for the same grouped data.

The authors compare this .67 value for the point biserial coefficient with those disclosed for the same coefficient in the Educational Testing Services (ETS) (2008) report<sup>26</sup> for the Algebra I California Standards Test (CST), which covers approximately the same curriculum as the ALEKS assessment for elementary algebra and is given to more than 100,000 students each year. This test consists of 65 multiple choice questions (items) and is built and scored in the framework of Item Response Theory (IRT), for which the point biserial coefficient is a standard measure. In particular, for each of the 65 items, a point biserial item–test correlation was computed, which measured the relationship between the dichotomous variable giving the 1/0 item score (correct/incorrect) and the continuous variable giving the total test score (see ETS 2008, p. 397). For the 2007 administration of the Algebra I CST, the mean point biserial coefficient for the 65 items was .36, and the median was .38 (see table 7.2, p. 397 of the ETS report). The minimum coefficient obtained for an item was .10 and the maximum was .53 (table 7.A.4, pp. 408–409, of the ETS report). The averages for preceding years on the test were similar, namely, the

<sup>25</sup> For example, the distribution of  $S_a$  vanishes in a positive neighborhood of 0, but is positive at the point 0 itself.

<sup>26</sup> Produced for the California Department of Education (Test and Assessment Division). See <http://www.cde.ca.gov/ta/tg/sr/documents/csttech rpt07.pdf>.

mean point biserial coefficients were .38 in 2005 and .36 in 2006 (see table 10.B.3, p. 553, of the same report).<sup>27</sup>

### 5.12.3 Learning readiness

A student using the ALEKS system is given regular assessments. At the end of each assessment,<sup>28</sup> the system gives the student the choice of the next item to learn, such items being located in the outer fringe of the student's knowledge state.<sup>29</sup> This makes sense because the items in the outer fringe are exactly those that the student is ready to learn. So, if the validity of the assessment is high, the probability of successfully learning an item chosen in the outer fringe should also be high. Cosyn *et al.* (2013) have estimated the probability that a student successfully learns an item chosen in the outer fringe of his or her state. For elementary school mathematics, the median of the distribution of the estimated (conditional) success probabilities was .93. This estimate was based on 1,940,473 learning occasions.

### 5.12.4 ALEKS based placement at the University of Illinois

Until 2006, students entering the University of Illinois had to take a mathematics placement test based on the ACT.<sup>30</sup> The results were not satisfactory because many students lack an adequate preparation for the course they were advised to take and ended up withdrawing. Beginning in 2007, the ACT was replaced by ALEKS. Ahlgren and Harper (2013) report a comparison of the two situations, which we briefly sum up here (see also Ahlgren and Harper, 2011).

The consequence of a withdrawal from a course may have dire consequences. For example, we read in the introductory section of Ahlgren and Harper (2013):

*At the University of Illinois, the standard introductory calculus course (**Calc: Calculus I**) is a five credit course, the withdrawal from which beyond the add-deadline may reduce students to a credit total below full-time status, resulting in the loss of tuition benefits, health benefits, scholarships, and athletic eligibility.*

The placement program of the Department of Mathematics at the University of Illinois deals with four courses: Preparation for Calculus (**Pre-Calc**), Calculus 1 (**Calc**), Calculus 1 for student with experience (**CalcExp**), and Business Calculus (**BusCalc**). Because the **Pre-Calc** course was only offered for the first time in 2007, it is not included in the comparison statistics below.

The new placement exam is an ALEKS assessment focused on the module<sup>31</sup> “*Prep for Calculus*,” with some items removed that are not part of the relevant course at the University of Illinois. The students take the non-proctored assessment either at home or on campus. Students failing to reach the required

<sup>27</sup> These low correlation values were obtained even though some items with low item-test correlations were removed from the test in a preliminary analysis.

<sup>28</sup> Except the final one, at the end of the course.

<sup>29</sup> See Section 5.6 for the concept of outer fringe.

<sup>30</sup> Originally, an acronym for American College Testing.

<sup>31</sup> A particular learning space used in the ALEKS system.

Table 5.7 *Percentages of decrease in withdrawals from three courses between Fall 2006 (before ALEKS) and Fall 2007, Fall 2008.*

Course	Percentage of decrease in withdrawals		
	<b>BusCalc</b>	<b>Calc</b>	<b>CalcExp</b>
2006–2007	↓ 57%	↓ 49%	↓ 67%
2006–2008	↓ 19%	↓ 81%	↓ 42%

score for placement in one of the four courses have the options of taking the placement test again, or using the ALEKS learning module (or some other method) to improve their results. The students can keep retaking assessments until the course add-deadline.

Table 5.7 gives the percentages of the decrease in the numbers of withdrawals observed in 2007 and 2008 for the **BusCalc**, **Calc** and **CalcExp** courses, in comparison with the numbers of withdrawals from the same courses in 2006.

The decrease in the percentages of withdrawals is remarkable. The authors also note a shift in the enrollment – not reported here – from the least advanced course, **BusCalc**, to the more advanced one, which is **CalcExp**. Presumably, this may be due to the combination of two factors: the accuracy of the assessment in the ALEKS system, which may improve the placement, and the fact the students were given the possibility of using the system to bridge some gaps in their expertise.

### 5.13 Bibliographical notes

The first paper on knowledge spaces was Doignon and Falmagne (1985), which contains some of the combinatorial basis of the theory, and in particular one of two key properties, namely, the closure under union: a knowledge space is a knowledge structure closed under union. Doignon and Falmagne (1988) was a technical follow up. The stochastic part of the theory was presented in the two papers by Falmagne and Doignon (1988a,b). The second paper also introduces a second key property, wellgradedness. A comprehensive, nontechnical description of these basic ideas is contained in Falmagne *et al.* (1990). (See also Doignon and Falmagne, 1987; Falmagne, 1989; Doignon, 1994, for other introductory papers.)

Even though, taken together, the closure under union and wellgradedness form a legitimate basis of the theory, it is not obvious that these axioms are pedagogically sound. Around 2002, Falmagne proposed a reaxiomatization in the form of the conditions labelled here as [L1] and [L2] (see Section 5.3), which he called a *learning space*. These axioms were unpublished at the time, but they were communicated to Eric Cosyn and Hasan Uzun, who then proved that a knowledge structure is a learning space if and only if it is a well-graded knowledge space (see Cosyn and Uzun, 2009, and our Theorem 5.1). A similar result is contained in Korte *et al.* (1991). In fact, learning spaces are exactly the same structures as antimatroids or convex geometries. The origin of the term antimatroid (in the

setting of intersection-closed families of subsets) goes back to Jamison (1980) and Jamison (1982); see also Edelman (1986) for a lattice-theoretic approach, and Edelman and Jamison (1985).

In time, other scientists became interested in knowledge space theory, notably Cornelia Dowling,<sup>32</sup> Jürgen Heller, and Reinhard Suck in Germany; Dietrich Albert and Cord Hockemeyer in Austria; Mathieu Koppen in the Netherlands; and Luca Stefanutti in Italy. The literature on the subject grew and now contains several hundred titles. The database at <http://liinwww.ira.uka.de/bibliography/Ai/knowledge.spaces.html> contains an extensive list of reference on knowledge spaces. It is maintained by Cord Hockemeyer at the University of Graz (see also Hockemeyer, 2001).

The monograph by Doignon and Falmagne (1999) gives a comprehensive description of knowledge space theory at that time. Beginning in 1994, an Internet-based educational software based on learning space theory was developed at the University of California, Irvine (UCI), by a team of software engineers including (among many others) Eric Cosyn, Damien Lauly, David Lemoine, and Nicolas Thiéry, under the supervision of Falmagne. This was supported by a large NSF grant to UCI. The end product of this software development was called ALEKS, which is also the name of a company created in 1996 by Falmagne and his graduate students.<sup>33</sup> Around 1999, it was thought that the educational software had matured enough for a fruitful application in the schools and universities. This led to further developments in the form of statistical analysis of the assessment data and new mathematical results. Falmagne *et al.* (2006) was a not-too-technical introduction to the topic at that time. The monograph by Falmagne and Doignon (2011, a much expanded re-edition of Doignon and Falmagne, 1999) is a comprehensive and (almost) up-to-date technical presentation of the theory.

Several assessment systems are founded on knowledge space theory and learning space theory, the most prominent ones being ALEKS and RATH. (The RATH system was developed by a team of researchers at the University of Graz, in Austria; see Hockemeyer, 1997, and Hockemeyer *et al.*, 1998.)

## Acknowledgments

The authors thank Laurent Fourny and Keno Merckx for their careful reading of a preliminary draft.

## References

- Ahlgren, A. and Harper, M. (2011). Assessment and placement through Calculus I at the University of Illinois. *Notices of the American Mathematical Society*, **58**, 1460–1461.

<sup>32</sup> Formerly Cornelia Müller; see this name also for references.

<sup>33</sup> In particular Cosyn, Lauly, and Thiéry. ALEKS Corporation was sold to McGraw-Hill Education in 2013.

- Ahlgren, A. and Harper, M. (2013). ALEKS-based placement at the University of Illinois. In: Falmagne, J.-Cl., Albert, D., Doble, C. W., Eppstein, D. and Hu, X. (eds), *Knowledge Spaces: Applications in Education*. Berlin: Springer, pp. 51–68.
- Anastasi, A. and Urbina, S. (1997). *Psychological Testing* (seventh edition). Upper Saddle River, NJ: Prentice Hall.
- Birkhoff, G. (1937). Rings of sets. *Duke Mathematical Journal*, **3**, 443–454.
- Boyd, E. A. and Faigle, U. (1990). An algorithmic characterization of antimatroids. *Discrete Applied Mathematics*, **28**, 197–205.
- Caspard, N. and Monjardet, B. (2004). Some lattices of closure systems on a finite set. *Discrete Mathematics and Theoretical Computer Science*, **6**, 163–190 (electronic).
- Cosyn, E. and Uzun, H. B. (2009). Note on two necessary and sufficient axioms for a well-graded knowledge space. *Journal of Mathematical Psychology*, **53**, 40–42.
- Cosyn, E., Doble, C. W., Falmagne, J.-Cl., et al. (2013). Assessing mathematical knowledge in a learning space. In: Falmagne, J.-Cl., Albert, D., Doble, C. W., Eppstein, D. and Hu, X. (eds), *Knowledge Spaces: Applications in Education*. Berlin: Springer, pp. 27–50.
- Doble, C. W., Doignon, J.-P., Falmagne, J.-Cl. and Fishburn, P. C. (2001). Almost connected orders. *Order*, **18**, 295–311.
- Doignon, J.-P. (1994). Probabilistic assessment of knowledge. In: Albert, D. (ed), *Knowledge Structures*. New York, NY: Springer Verlag, pp. 1–56.
- Doignon, J.-P. (2014). Learning spaces, and how to build them. In: Glodeanu, C. V., Kaytoue, M. and Sacarea, Chr. (eds), *Formal Concept Analysis, 12th International Conference, ICFCA 2014, Cluj-Napoca, Romania, June 10–13, 2014*. Lecture Notes in Artificial Intelligence, vol. 8478. Berlin, Heidelberg, and New York: Springer-Verlag, pp. 1–14.
- Doignon, J.-P. and Falmagne, J.-Cl. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, **23**, 175 – 196.
- Doignon, J.-P. and Falmagne, J.-Cl. (1987). Knowledge assessment: a set theoretical framework. In: Ganter, B., Wille, R., and Wolfe, K. E. (eds), *Beiträge zur Begriffsanalyse: Vorträge der Arbeitstagung Begriffsanalyse, Darmstadt 1986*. Mannheim: BI Wissenschaftsverlag, pp. 129–140.
- Doignon, J.-P. and Falmagne, J.-Cl. (1988). Parametrization of knowledge structures. *Discrete Applied Mathematics*, **21**, 87–100.
- Doignon, J.-P. and Falmagne, J.-Cl. (1999). *Knowledge Spaces*. Berlin: Springer-Verlag.
- Dowling, C. (1993). On the irredundant construction of knowledge spaces. *Journal of Mathematical Psychology*, **37**, 49–62.
- Dowling, C. E. (1994). Integrating different knowledge spaces. In: Fischer, G. H., and Laming, D. (eds), *Contributions to Mathematical Psychology, Psychometrics, and Methodology*. Berlin, Heidelberg, and New York: Springer-Verlag, pp. 149–158.
- Edelman, P. H. (1986). Abstract convexity and meet-distributive lattices. In: *Combinatorics and ordered sets (Arcata, Calif., 1985)*. Contemp. Math., vol. 57. Providence, RI: American Mathematical Society, pp. 127–150.
- Edelman, P. H. and Jamison, R. E. (1985). The theory of convex geometries. *Geometriae Dedicata*, **19**, 247–270.

- Educational Testing Services (ETS) (2008). *California Standard Tests (CSTs) Technical Report, Spring 2007 Administration*. Tech. rept. California Department of Education – Standard and Assessment Division. Contract 5417, <http://www.cde.ca.gov/ta/tg/sr/documents/csttech rpt07.pdf>.
- Eppstein, D. (2013a). Learning sequences: an efficient data structure for learning spaces. In: Falmagne, J.-Cl., Albert, D., Doble, C. W., Eppstein, D., and Hu, X. (eds), *Knowledge Spaces: Applications in Education*. Berlin: Springer, pp. 287–304.
- Eppstein, D. (2013b). Projection, decomposition, and adaption of learning spaces. In: Falmagne, J.-Cl., Albert, D., Doble, C. W., Eppstein, D. and Hu, X. (eds), *Knowledge Spaces: Applications in Education*. Berlin: Springer, pp. 305–322.
- Falmagne, J.-C. and Doignon, J.-P. (2011). *Learning Spaces*. Berlin: Springer-Verlag.
- Falmagne, J.-C., Albert, D., Doble, C. W., Eppstein, D. and Hu, X. (eds) (2013). *Knowledge Spaces: Applications in Education*. Berlin: Springer-Verlag.
- Falmagne, J.-Cl. (1989). Probabilistic knowledge spaces: a review. In: Roberts, F. (ed.), *Applications of Combinatorics and Graph Theory to the Biological and Social Sciences, IMA Volume 17*. New York, NY: Springer Verlag.
- Falmagne, J.-Cl. and Doignon, J.-P. (1988a). A class of stochastic procedures for the assessment of knowledge. *British Journal of Mathematical and Statistical Psychology*, **41**, 1–23.
- Falmagne, J.-Cl. and Doignon, J.-P. (1988b). A Markovian procedure for assessing the state of a system. *Journal of Mathematical Psychology*, **32**, 232–258.
- Falmagne, J.-Cl., Koppen, M., Villano, M., Doignon, J.-P. and Johanessen, L. (1990). Introduction to knowledge spaces: how to build, test and search them. *Psychological Review*, **97**, 204–224.
- Falmagne, J.-Cl., Cosyn, E., Doignon, J.-P. and Thiéry, N. (2006). The assessment of knowledge, in theory and in practice. In: Ganter, B. and Kwuida, L. (eds), *Formal Concept Analysis, 4th International Conference, ICFCA 2006, Dresden, Germany, February 13–17, 2006*. Lecture Notes in Artificial Intelligence. Berlin, Heidelberg, and New York: Springer-Verlag, pp. 61–79.
- Ganter, B. and Reuter, K. (1991). Finding all closed sets: a general approach. *Order*, **8**, 283–290.
- Greer, T., Dunlap, W. P. and Beaty, G. O. (2003). Intensity discrimination with gated and continuous sinusoids. *Educational and Psychological Measurement*, **63**, 931–950.
- Hockemeyer, C. (1997). *RATH – a relational adaptive tutoring hypertext WWW-environment*. Technical report 1997/3. Institut für Psychologie, Karl-Franzens-Universität Graz, Austria. Available online at <http://css.uni-graz.at/rath/publications/documentation.pdf>.
- Hockemeyer, C. (2001). *KST Tools User Manual*. Unpublished Technical Report. Institut für Psychologie, Karl-Franzens-Universität Graz, Austria. Available: <http://wundt.uni-graz.at/kst.html>.
- Hockemeyer, C., Held, Th. and Albert, D. (1998). RATH – a relational adaptive tutoring hypertext WWW-environment based on knowledge space theory. In: Alvegard, Chr. (ed), *CALISCE'98: Proceedings of the Fourth International Conference on Computer Aided Learning in Science and Engineering*. Göteborg, Sweden: Chalmers University of Technology, pp. 417–423.

- Jamison, R. E. (1980). Copoints in antimatroids. In: *Proceedings of the Eleventh Southeastern Conference on Combinatorics, Graph Theory and Computing (Florida Atlantic Univ., Boca Raton, Fla., 1980), Vol. II*, vol. 29, pp. 535–544.
- Jamison, R. E. (1982). A perspective on abstract convexity: classifying alignments by varieties. In: *Convexity and Related Combinatorial Geometry (Norman, Okla., 1980)*. Lecture Notes in Pure and Appl. Math., vol. 76. New York, NY: Dekker, pp. 113–150.
- Koppen, M. (1993). Extracting human expertise for constructing knowledge spaces: an algorithm. *Journal of Mathematical Psychology*, **37**, 1–20.
- Koppen, M. (1998). On alternative representations for knowledge spaces. *Mathematics and Social Science*, **36**, 127–143.
- Koppen, M. and Doignon, J.-P. (1990). How to build a knowledge space by querying an expert. *Journal of Mathematical Psychology*, **34**, 311–331.
- Korte, B., Lovász, L. and Schrader, R. (1991). *Greedoids*. Algorithms and Combinatorics, vol. 4. Berlin: Springer-Verlag.
- Müller, C. E. (1989). A procedure for facilitating an expert's judgments on a set of rules. In: Roskam, E. E. (ed.), *Mathematical Psychology in Progress. Recent Research in Psychology*. Berlin, Heidelberg, and New York: Springer-Verlag, pp. 157–170.

# 6 Evolutionary game theory

J. McKenzie Alexander

6.1 Introduction	322
6.2 Nash equilibria and evolutionarily stable strategies	323
6.3 Set-based evolutionary stability concepts	333
6.4 Evolutionary dynamics	336
6.4.1 The replicator dynamics	337
6.4.1.1 Discrete replicator dynamics	359
6.4.1.2 Multipopulation models	360
6.4.2 Replicator–mutator dynamics	361
6.4.3 Finite population models	363
6.5 Local interaction models	366
6.6 Further reading	370
References	371

## 6.1 Introduction

Evolutionary game theory made its way into the social sciences through a curiously circuitous path. Its origin lay in the realization that the mathematical theory of games developed by von Neumann and Morgenstern (1944) could be used to analyze problems of population biology. Because the fitness of an organism (or a trait) in a population often depends on the relative frequency of other organisms (or traits) present, natural selection can take on a strategic character even when none of the organisms are “rational” in any standard sense.

Although modern evolutionary game theory is typically considered to have begun in the work of Maynard Smith and Price (1973), important precursors exist in the work of R. A. Fisher as early as 1930. In *The Genetical Theory of Natural Selection*, Fisher sought to explain the approximate equality of the sex ratio in mammals using arguments which can be readily understood in game-theoretic terms. The watershed moment for evolutionary game theory, though, was the publication of Maynard Smith’s seminal work *Evolution and the Theory of Games* in 1982. In that text, Maynard Smith drew together a number of results, both from his own work and that of a number of mathematical biologists, presenting them in a clear, coherent format with minimal mathematical prerequisites.

Over time, economists and other social scientists became interested in evolutionary game theory for a variety of reasons, some stemming from long-standing issues in the traditional theory of games. Of these, perhaps the two most important were as follows: first, as von Neumann and Morgenstern clearly recognized, their theory lacked any underlying dynamics. Early on in *Theory of Games and Economic Behaviour*, they wrote:

We repeat most emphatically that our theory is thoroughly static. A dynamic theory would unquestionably be more complete and therefore preferable. But there is ample evidence from other branches of science that it is futile to try to build one as long as the static side is not thoroughly understood. (§4.8.2)

Second, the traditional solution concept – that of a Nash equilibrium (which shall be defined below) – suffered from the problem that it was not necessarily unique: games could, and frequently did, have multiple solutions, each of which had some claim to be the “rational” solution. As attempts were made to provide more subtle characterizations of what was to count as a solution (the “refinements programme”), it soon became apparent that a plethora of competing equilibrium refinements existed, each of which had some intuitive force supporting it. Evolutionary game theory, with its concept of an evolutionary stable strategy, not only provided another equilibrium refinement, but it also provided an underlying dynamical theory via the replicator dynamics of Taylor and Jonker (1978). The fact that the replicator dynamics can be interpreted as a model of *cultural* evolution, as well as biological evolution, provided additional motivation, and thus brought evolutionary game theory into the remit of the social sciences.

Since its inception, the topics falling under evolutionary game theory, broadly construed, have expanded well beyond its original biological remit, and now include areas such as games played on social networks, strategic learning in the presence of limited information, and many others. In the social sciences, evolutionary game theory has been used to study problems relating to markets, bargaining and resource allocation, the evolution of social norms, the provisioning of public goods, the emergence of language, signaling and cheap talk, the evolution of personal preferences, racial segregation, and the evolution of morality. For a list of references covering some of these areas of application, see Sandholm (2010, p. 16).

## 6.2 Nash equilibria and evolutionarily stable strategies

Let us begin with some game-theoretic preliminaries. Consider a game played by a group of  $N$  individuals, where each player  $i$  has a set of *pure strategies* denoted  $S_i$ .<sup>1</sup> It often proves useful to allow individuals to use an indeterminate

<sup>1</sup> A *strategy* is nothing more than an action available to a player. For example, *left* and *right*, when deciding what side of the road to drive on. The set of strategies available to each player will be clear whenever a particular game is being discussed; including explicit notation for each possible pure strategy for a player into the formalization becomes unwieldy quickly. For example, consider a simple two-player game consisting of a driver and a policeman: the driver may drive *fast*, *moderate*, or *slow*, and the police officer may *arrest* or *ignore*. The two sets of strategies have no actions in

strategy; such a strategy, corresponding to a probability distribution over the set of a player's pure strategies, is known as a *mixed strategy*. One common notation for mixed strategies expresses them as  $p_1s_1 + p_2s_2 + \dots + p_ns_n$ , where  $p_i$  is the probability of playing the pure strategy  $s_i$ . (The use of “+” is an abuse of notation.) The set of all mixed strategies for player  $i$  is denoted by  $\Delta_i$ , and is simply the set of all probability distributions over  $S_i$ . A specification of strategies to all players is a *strategy profile*, and the set of all strategy profiles is the Cartesian product  $\mathcal{S} = \Delta_1 \times \dots \times \Delta_N$ .

The *payoff function* for player  $i$ , denoted  $\pi_i$ , maps strategy profiles to real numbers, the real number being the payoff received by player  $i$ . This is typically done by first defining the payoff function for the set of pure strategy profiles, and then extending the definition of  $\pi_i$  to the set of mixed strategy profiles in the natural way.<sup>2</sup> A *strategic form game*  $\Gamma$  maps pure strategy profiles to payoffs for all the players, i.e.,  $\Gamma : S_1 \times \dots \times S_N \rightarrow \mathbb{R}^N$ . In the two-player case, a strategic form game conveniently corresponds to an  $n \times m$  matrix, each cell being of the form  $(r, c)$ , where  $r$  denotes the payoff for Row, and  $c$  denotes the payoff for Column. When the payoff function is the same for all players, the subscript is typically omitted.

The fundamental solution concept of traditional game theory is that of the *Nash equilibrium*<sup>3</sup>: a strategy profile where each player uses a best-response, given the strategies of the others. The following notational convention permits a more precise statement of this concept: if  $\sigma = (\sigma_1, \dots, \sigma_N) \in \mathcal{S}$  is a strategy profile and  $\xi_i \in \Delta_i$ , then  $(\xi_i, \sigma_{-i}) =_{\text{df}} (\sigma_1, \dots, \sigma_{i-1}, \xi_i, \sigma_{i+1}, \dots, \sigma_N)$ . Intuitively,  $\sigma_{-i}$  can be thought of as an incomplete strategy profile, specifying strategies for all players other than  $i$ .<sup>4</sup> Given this,  $\sigma^* = (\sigma_1^*, \dots, \sigma_N^*) \in \mathcal{S}$  is a Nash equilibrium if and only if, for all players  $i$ ,

$$\pi_i(\sigma_i^*, \sigma_{-i}^*) \geq \pi_i(\sigma_i, \sigma_{-i}^*), \text{ for all } \sigma_i \in \Delta_i.$$

(This illustrates what people mean when they say game theory is “notationally challenged”: any sufficiently general formulation requires notation disproportionate to its conceptual difficulty; see Gintis, 2000, p. xxii.)

common, and even differ in the number of actions available. To accommodate for this, in general, the formalization would look like this:  $S_i = \{s_1^i, \dots, s_{n_i}^i\}$ , where the superscript indicates to which player the strategy belongs to, the subscript serves to index the strategy, and the range of subscripts goes up to  $n_i$  to allow for the number of strategies to vary across players.

- 2 If players can use mixed strategies, the payoff function for player  $i$  is the weighted sum of payoffs conferred to player  $i$  by every possible pure strategy profile, where each term in the sum is weighted by the probability of that pure strategy profile occurring. See Binmore (1992, pp. 232–233) for a discussion of the two-player case, and Luce and Raiffa (1957, pp. 157–158) for the general  $N$ -player case.
- 3 The concept of a Nash equilibrium derives its name from the seminal work of John Nash, who proved in 1951 that any game with a finite number of players and a finite number of strategies has at least one Nash equilibrium if players can adopt mixed strategies. (It's also worth knowing that the characterization of a Nash equilibrium given in the film *A Beautiful Mind* is incorrect.)
- 4 Note that  $(\xi_i, \sigma_{-i})$  is another abuse of notation. Strictly speaking,  $\sigma_{-i}$  is a function from  $\Delta_i$  to  $\mathcal{S}$ , defined as  $\xi_i \mapsto (\sigma_1, \dots, \sigma_{i-1}, \xi_i, \sigma_{i+1}, \dots, \sigma_N)$ , and so it would be more accurate to write  $\sigma_{-i}(\xi_i)$  instead of  $(\xi_i, \sigma_{-i})$ . However, convention dictates one write  $(\xi_i, \sigma_{-i})$  instead.

Although we have just defined a Nash equilibrium in considerable generality, allowing for a number of players, each with a different strategy set, for the remainder of this section we shall only consider two-player symmetric games. That is, two-player games where the players share a common set of strategies, and the payoffs do not depend on the identity of the players. (For example, the payoff I receive when I play  $A$  and you play  $B$  is the same as the payoff you receive if you play  $A$  and I play  $B$ .)

Given this, let  $S = \{s_1, \dots, s_n\}$  denote the common set of pure strategies, and let  $\Delta$  be the set of all mixed strategies. The payoff received by a player who uses the pure strategy  $s_i$  against a player using the pure strategy  $s_j$  is denoted  $\pi(s_i|s_j)$ . If  $\sigma = p_1s_1 + \dots + p_ns_n$  and  $\mu = q_1s_1 + \dots + q_ns_n$  are mixed strategies, then

$$\begin{aligned}\pi(\sigma|\mu) &= \sum_{i=1}^n \sum_{j=1}^n p_i q_j \pi(s_i|s_j) \\ &= p_1 \pi(s_1|\mu) + p_2 \pi(s_2|\mu) + \dots + p_n \pi(s_n|\mu) \\ &= q_1 \pi(\sigma|s_1) + q_2 \pi(\sigma|s_2) + \dots + q_n \pi(\sigma|s_n).\end{aligned}$$

The bilinear nature of the payoff function will often be invoked in proofs.

An alternate notation, which in some cases proves to be more convenient, uses matrices. The fact that the underlying game is symmetric with  $n$  strategies enables us to represent it using an  $n \times n$  square matrix  $\mathbf{A}$ . Because mixed strategies are nothing more than vectors  $\mathbf{p} = (p_1, \dots, p_n) \in \mathbb{R}^n$ , the payoff to a player using strategy  $\mathbf{p}$  against someone using strategy  $\mathbf{q}$  is simply  $\mathbf{p} \cdot \mathbf{Aq}$ . (I do not distinguish between row or column vectors because it is clear from the context what the intended interpretation of  $\mathbf{p}$  should be.)

How should we use this game-theoretic machinery to analyze evolutionary stability? One point to note is that, although evolution does not provide globally optimal solutions to adaptive problems (think of the appendix or the human birth canal), it seems plausible that evolution would yield outcomes which are, in some sense, mutual best-responses. In a predator-prey system, those prey who fail to adopt a best-response behavior when pursued by a predator would, presumably, be selected against more heavily (i.e., eaten) than those prey which did adopt a best-response. Yet the following example shows that the concept of a Nash equilibrium is not, on its own, adequate for analyzing evolutionary phenomena.

**Example 6.1** Consider the following game:

	$A$	$B$
$A$	(2, 2)	(1, 1)
$B$	(1, 1)	(1, 1)

It is easy to verify that both  $(A, A)$  and  $(B, B)$  are Nash equilibria. However, if we imagine this game played in a population, where individuals are paired at random,

a population where everyone used the strategy  $B$  is not, in an informal sense, evolutionarily stable. (We will give a formal definition of an evolutionarily stable strategy below.) The reason why is that a Nash equilibrium does not require deviation to produce a *worse* outcome for a player, but merely an outcome that is no better. (A Nash equilibrium where a deviating player always obtains a worse outcome is a *strict* Nash equilibrium.) Hence, an  $A$ -mutant can invade an all- $B$  population: because an  $A$ -player paired against a  $B$ -player receives the exact same payoff as a  $B$ -player paired against a  $B$ -player, there is no selection against  $A$ . Furthermore, when two  $A$  players are paired with each other, each receives a payoff twice that of any  $B$ -player. Hence, not only is it possible for an  $A$ -mutant to invade an all- $B$  population and not be driven out, but the  $A$ -mutant might even spread to take over the entire population.  $\square$

One way of strengthening the concept of a Nash equilibrium to handle the concern of Example 6.1 is to incorporate a second-order stability condition; indeed, this is exactly the approach taken by Maynard Smith and Price (1973). Suppose we have a pure population of  $\sigma$ -players which are invaded by a small group of individuals using the strategy  $\mu$ . If  $p \in (0, 1)$  denotes the proportion of the population following the mutant strategy  $\mu$ , then the expected fitness of each strategy in the population is as follows:

$$\begin{aligned} W(\sigma) &= (1 - p)\pi(\sigma|\sigma) + p\pi(\sigma|\mu) \\ W(\mu) &= (1 - p)\pi(\mu|\sigma) + p\pi(\mu|\mu). \end{aligned}$$

Intuitively, if  $\sigma$  is evolutionarily stable, it will repel the invasion attempt. One way for this to happen is for  $\sigma$  to have a higher expected fitness:  $W(\sigma) > W(\mu)$ . Because  $p \ll 1$ , we should not need to worry about the magnitude of the fitness contributions of  $\pi(\sigma|\mu)$  and  $\pi(\mu|\mu)$  unless  $\pi(\sigma|\sigma) = \pi(\mu|\sigma)$ . This is the motivation behind Maynard Smith and Price's original definition of an *evolutionarily stable strategy*:

**Definition 6.1** A strategy  $\sigma$  is an *evolutionarily stable strategy* (ESS) if, for all strategies  $\mu \neq \sigma$ , it is the case that

$$\begin{aligned} &\text{either } \pi(\sigma|\sigma) > \pi(\mu|\sigma) \\ &\text{or } \pi(\sigma|\sigma) = \pi(\mu|\sigma) \text{ and } \pi(\sigma|\mu) > \pi(\mu|\mu). \end{aligned}$$

This, of course, is logically equivalent to requiring, for all  $\mu \neq \sigma$ ,

1.  $\pi(\sigma|\sigma) \geq \pi(\mu|\sigma)$ ,
2. if  $\pi(\sigma|\sigma) = \pi(\mu|\sigma)$ , then  $\pi(\sigma|\mu) > \pi(\mu|\mu)$ .

The second formulation has the advantage of making explicit the fact that an ESS consists of a Nash equilibrium (condition 1) augmented by a stability requirement (condition 2). The first formulation has the advantage of making it obvious that every strict Nash equilibrium is an ESS.

It is now perhaps more common to find another definition of an ESS, originally due to Taylor and Jonker (1978). Suppose that  $\sigma$  and  $\mu$  are two strategies, possibly mixed. If  $\epsilon \in [0, 1]$ , then  $\epsilon\mu + (1 - \epsilon)\sigma$  denotes the strategy which plays  $\mu$  with probability  $\epsilon$  and plays  $\sigma$  with probability  $1 - \epsilon$ . That is, if  $\mu = q_1s_1 + \dots + q_ns_n$  and  $\sigma = p_1s_1 + \dots + p_ns_n$ , then

$$\epsilon\mu + (1 - \epsilon)\sigma = \sum_{i=1}^n (\epsilon q_i + (1 - \epsilon)p_i)s_i.$$

**Definition 6.2** A strategy  $\sigma$  is an *evolutionarily stable strategy* if for every  $\mu \neq \sigma$  there exists a threshold  $\epsilon_\mu > 0$  such that, for every  $\epsilon \in (0, \epsilon_\mu)$ ,

$$\pi(\sigma|\epsilon\mu + (1 - \epsilon)\sigma) > \pi(\mu|\epsilon\mu + (1 - \epsilon)\sigma).$$

The equivalence of these two definitions can readily be seen by rewriting the above inequality as:

$$(1 - \epsilon)(\pi(\sigma|\sigma) - \pi(\mu|\sigma)) + \epsilon(\pi(\sigma|\mu) - \pi(\mu|\mu)) > 0.$$

As  $\epsilon \rightarrow 0$ , that requires either  $\pi(\sigma|\sigma) - \pi(\mu|\sigma) > 0$  or  $\pi(\sigma|\sigma) = \pi(\mu|\sigma)$  and  $\pi(\sigma|\mu) - \pi(\mu|\mu) > 0$ .

At this point, we can show that the ESS criterion excludes some Nash equilibrium strategies from being evolutionarily stable strategies. Consider the game below. One can verify that it has two Nash equilibria in pure strategies:  $(A, A)$  and  $(B, B)$ . Even though  $(B, B)$  is a Nash equilibrium, strategy  $B$  stands in an odd relation to strategy  $A$ . If my opponent plays  $B$ , I do no worse by playing  $A$  instead of  $B$ , as I receive 100 in both cases; but if my opponent plays  $A$ , I do strictly better, receiving a payoff of 1 instead of 0. That is, strategy  $A$  *weakly dominates* strategy  $B$ . Given that the  $(B, B)$  outcome is optimal, and the likely outcome when played with another person given the natural salience of the outcome, the following result may seem counterintuitive.

	A	B
A	(1, 1)	(100, 0)
B	(0, 100)	(100, 100)

**Theorem 6.1** *No weakly dominated strategy is an ESS.*

*Proof* Suppose  $\sigma$  is a Nash equilibrium strategy weakly dominated by  $\mu$ . That is, for all strategies  $\tau$ , it is the case that  $\pi(\mu|\tau) \geq \pi(\sigma|\tau)$ , and there exists at least one strategy  $\tau^*$  such that  $\pi(\mu|\tau^*) > \pi(\sigma|\tau^*)$ . (This is just the precise definition of a weakly dominated strategy.) Because  $\sigma$  is a Nash equilibrium, it must be the case that  $\pi(\mu|\sigma) = \pi(\sigma|\sigma)$ . But then, when we check the second criterion in the Maynard Smith definition of an ESS, we find that it fails: the fact that  $\mu$  weakly dominates  $\sigma$  means that  $\pi(\mu|\mu) \geq \pi(\sigma|\mu)$ .  $\square$

A strategy  $\sigma$  is said to be *strictly dominated* if there exists another strategy  $\sigma^*$  such that  $\pi(\sigma^*|\mu) > \pi(\sigma|\mu)$  for all strategies  $\mu$ . Because a strictly dominated

strategy cannot be a Nash equilibrium, it follows trivially that no strictly dominated strategy is an ESS.

In a finite game, with finitely many players, at least one Nash equilibrium always exists when players can use mixed strategies. However, the additional stability requirement imposed by an ESS is sufficiently strong that some games may not have any ESS at all, as the next example illustrates.

**Example 6.2** Consider the well-known game of Rock–Paper–Scissors:

	Rock	Paper	Scissors
Rock	(0, 0)	(−1, 1)	(1, −1)
Paper	(1, −1)	(0, 0)	(−1, 1)
Scissors	(−1, 1)	(1, −1)	(0, 0)

Because the only Nash equilibrium of this game is the mixed strategy  $\sigma = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ , we can show this game has no ESS by constructing another strategy  $\mu$  which shows  $\sigma$  does not satisfy the definition. (In this case, we shall use the Maynard Smith and Price definition of an ESS, as it is easier to work with.) Consider the strategy which always plays Rock. Because both  $\pi(\text{Rock}|\sigma) = 0$  and  $\pi(\sigma|\sigma) = 0$ , we need to check whether  $\pi(\sigma|\text{Rock}) > \pi(\text{Rock}|\text{Rock})$ . By symmetry,  $\pi(\sigma|\text{Rock}) = 0$ , and, trivially,  $\pi(\text{Rock}|\text{Rock}) = 0$  as well. Hence  $\sigma$  does not satisfy the definition and so the game has no ESS.  $\square$

Note that an ESS may not be unique, need not be a strict Nash equilibrium, and may not be globally optimal, as the following example shows.

**Example 6.3** Consider the following game:

	A	B	C
A	(3, 3)	(0, 0)	(0, 0)
B	(0, 0)	(1, 1)	(1, 2)
C	(0, 0)	(2, 1)	(1, 1)

Because  $\pi(A|A) > \pi(B|A)$  and  $\pi(A|A) > \pi(C|A)$ , strategy A is evolutionarily stable. In addition,  $\pi(C|C) > \pi(A|C)$  and, although  $\pi(C|C) = \pi(B|C)$ , because  $\pi(C|B) > \pi(B|B)$ , strategy C is also evolutionarily stable. Note that C is not a strict Nash equilibrium and is also suboptimal.  $\square$

These two examples raise the question of what mathematical properties characterize evolutionarily stable strategies, because ESS need not correspond to a strict Nash equilibrium, and do not always exist. Before addressing this question in detail, we first establish a few useful properties of ESS which enable us to place an upper bound on the number of ESS a game may have.

**Definition 6.3** Let  $\sigma$  be a mixed strategy. The *support* of  $\sigma$ , denoted  $\text{supp}(\sigma)$ , is the set of all pure strategies played with positive probability by  $\sigma$ .

**Theorem 6.2** Suppose  $\sigma$  is an ESS. If  $\mu$  is a strategy and  $\text{supp}(\mu) \subset \text{supp}(\sigma)$ , then  $\mu$  is not an ESS. If  $\mu$  is an ESS with  $\text{supp}(\mu) = \text{supp}(\sigma)$ , then  $\mu = \sigma$ .

*Proof* Suppose that  $\sigma$  is an ESS. We will first show that the expected payoff of any two pure strategies in the support of  $\sigma$  are equal, when played against  $\sigma$ . To see why, suppose that, to the contrary, there are  $x, y \in \text{supp}(\sigma)$  such that  $\pi(x|\sigma) > \pi(y|\sigma)$ . For each pure strategy  $s \in \text{supp}(\sigma)$ , let  $p_s$  denote the probability assigned to  $s$  by  $\sigma$ ; in particular,  $p_x$  and  $p_y$  are the probabilities assigned to  $x$  and  $y$ , respectively.

Now consider the strategy  $\sigma'$  which plays  $x$  with probability  $p_x + p_y$ , plays  $y$  with probability 0, and plays all other pure strategies  $s \in \text{supp}(\sigma)$  with probability  $p_s$ . Then

$$\pi(\sigma'|\sigma) = p_x\pi(x|\sigma) + p_y\pi(y|\sigma) + \sum_{\substack{s \in \text{supp}(\sigma) \\ s \neq x, y}} p_s\pi(s|\sigma)$$

and

$$\pi(\sigma'|\sigma) = (p_x + p_y)\pi(x|\sigma) + 0 \cdot \pi(y|\sigma) + \sum_{\substack{s \in \text{supp}(\sigma) \\ s \neq x, y}} p_s\pi(s|\sigma)$$

which means  $\pi(\sigma'|\sigma) > \pi(\sigma|\sigma)$ , because  $\pi(x|\sigma) > \pi(y|\sigma)$ . However, that contradicts the assumption that  $\sigma$  is an ESS. Hence  $\pi(x|\sigma) = \pi(y|\sigma)$  for all  $x, y \in \text{supp}(\sigma)$ .

Let  $\mu$  be a strategy such that  $\text{supp}(\mu) \subset \text{supp}(\sigma)$ , where  $\mu = \sum_{s \in \text{supp}(\mu)} q_s s$ . Then

$$\pi(\sigma|\sigma) = \sum_{s \in \text{supp}(\sigma)} p_s\pi(s|\sigma) = \sum_{s \in \text{supp}(\mu)} q_s\pi(s|\sigma) = \pi(\mu|\sigma) \quad (6.1)$$

as the two middle expressions are merely different convex combinations of equal quantities. Because  $\sigma$  is an ESS, the second clause of definition 6.1 requires that  $\pi(\sigma|\mu) > \pi(\mu|\mu)$ , and so  $\mu$  is not an ESS.

Finally, suppose  $\mu$  is an ESS with  $\text{supp}(\mu) = \text{supp}(\sigma)$ , but  $\mu \neq \sigma$ . This yields a contradiction with  $\mu$  being an ESS (via Equation (6.1), as before), and so  $\mu = \sigma$ .  $\square$

There are several interesting, albeit trivial, consequences of the above theorem.

**Corollary 6.1** (Bishop and Cannings 1978) Suppose that  $\sigma$  is an ESS with  $\text{supp}(\sigma) = \{s_1, \dots, s_n\}$ . Then

$$\pi(s_1|\sigma) = \pi(s_2|\sigma) = \dots = \pi(s_n|\sigma) = \pi(\sigma|\sigma). \quad (6.2)$$

The Bishop–Cannings theorem provides one means of searching for an ESS: let  $S' = \{s_{i_1}, \dots, s_{i_k}\} \subseteq S$  be a subset of pure strategies, and consider the mixed strategy  $\sigma = p_{i_1}s_{i_1} + \dots + p_{i_{k-1}}s_{i_{k-1}} + (1 - p_{i_1} - \dots - p_{i_{k-1}})s_{i_k}$ . If one cannot find values for  $p_{i_k}$  which satisfy equation (6.2), then no mixed strategy with support  $S'$  exists. If a solution does exist, one still must check that  $\sigma$  satisfies the definition of an ESS, because the Bishop–Cannings theorem provides a necessary, not a sufficient, condition.

A *completely mixed* strategy is one which assigns positive probability to all of the pure strategies. Theorem 6.2 immediately implies the following:

**Corollary 6.2** *A completely mixed ESS is the only ESS of the game.*

Because all of the games we are considering have only finitely many pure strategies, there are only a finite number of possible support sets for ESS. Hence:

**Corollary 6.3** *The number of ESS is finite (possibly zero).*

In the Taylor and Jonker definition of an ESS (6.2), the invasion threshold  $\epsilon_\mu$  may depend on the strategy attempting to invade. One might wonder whether this dependency can be removed. Hofbauer *et al.* (1979) showed that it can. This result means that an ESS can be understood as a strategy capable of driving out any single type from the population, provided that the frequency of that type in the population is below a common threshold, known as a “uniform invasion barrier.”

**Definition 6.4** A strategy  $\sigma$  is said to have a *uniform invasion barrier* if there exists an  $\bar{\epsilon} > 0$  such that, for all  $\mu \neq \sigma$  and every  $\epsilon < \bar{\epsilon}$ ,

$$\pi(\sigma|\epsilon\mu + (1 - \epsilon)\sigma) > \pi(\mu|\epsilon\mu + (1 - \epsilon)\sigma).$$

(Note that the difference between the Taylor and Jonker definition of an ESS and the definition of a uniform invasion barrier is the order of the quantifiers.)

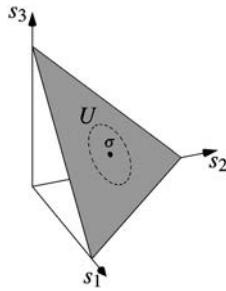
**Theorem 6.3** (Hofbauer *et al.*, 1979) *A strategy  $\sigma$  is an ESS if and only if  $\sigma$  has a uniform invasion barrier.*

*Proof* ( $\Rightarrow$ ) Assume that  $\sigma$  is an ESS. According to the Taylor and Jonker definition, for every strategy  $\mu$  there exists a  $\epsilon_\mu > 0$  such that, for all  $\epsilon \in (0, \epsilon_\mu)$ ,

$$\pi(\sigma|\epsilon\mu + (1 - \epsilon)\sigma) > \pi(\mu|\epsilon\mu + (1 - \epsilon)\sigma).$$

Because the payoff function  $\pi$  is bilinear, it is completely determined once the values of  $\pi$  for all combinations of pure strategies have been specified. Thus, although the particular value of  $\epsilon_\mu$  is a function of  $\mu$ , it can be chosen so as to vary continuously with  $\mu$ . As the set of all mixed strategies  $\Delta$  is closed and bounded, by the extreme value theorem there exists a minimum  $\bar{\epsilon} = \min_{\mu \in \Delta} \epsilon_\mu$  which serves as a uniform invasion barrier.

( $\Leftarrow$ ) If  $\sigma$  has a uniform invasion barrier  $\bar{\epsilon}$ , then the definition of an ESS is trivially satisfied by taking  $\epsilon_\mu = \bar{\epsilon}$  for all  $\mu \in \Delta$ .  $\square$



**Figure 6.1** The 2-simplex, showing the mixed strategy  $\sigma = (.3, .4, .3)$  and the neighbourhood  $U$  of all strategies  $\mu$  such that  $\|\mu - \sigma\| < \frac{1}{5}$ .

Another characterization of an ESS can be given in terms of how well it does when played against strategies which are “close” to it. We can make this notion precise as follows. If there are  $N$  pure strategies, the set of all mixed strategies is simply the  $N - 1$  unit simplex, typically denoted  $\Delta^{N-1}$ . Because  $\Delta^{N-1} \subset \mathbb{R}^N$ , we can measure the distance between any two mixed strategies  $\sigma_1, \sigma_2 \in \Delta^{N-1}$  via the standard Euclidean metric on  $\mathbb{R}^N$ . Thus we can talk about the *neighbourhood* of a mixed strategy, as illustrated in Figure 6.1. Note that the neighborhood of a strategy is that portion of an  $\epsilon$ -neighborhood about a point in  $\mathbb{R}^N$  restricted to the simplex, as points off the simplex have no natural game theoretic interpretation.

In Example 6.3, we saw that a strategy can be evolutionarily stable even if the payoffs received when it plays against itself are not globally optimal. However, what does turn out to be true is that an ESS is the optimal response against strategies sufficiently close.

**Definition 6.5** A strategy  $\sigma$  is said to be *locally superior* if it has a neighborhood  $U$  such that  $\pi(\sigma|\mu) > \pi(\mu|\mu)$  for all  $\mu \in U$  where  $\mu \neq \sigma$ .

**Theorem 6.4** (Hofbauer *et al.*, 1979) A strategy  $\sigma$  is an ESS if and only if  $\sigma$  is locally superior.

*Proof* ( $\Rightarrow$ ) Assume that  $\sigma$  is locally superior, and let  $U$  be a neighborhood such that  $\pi(\sigma|\mu) > \pi(\mu|\mu)$  for all  $\mu \in U$ . For any strategy  $\tau$ , there is a threshold  $\epsilon_\tau > 0$  such that

$$\mu' = \epsilon\tau + (1 - \epsilon)\sigma \in U \text{ whenever } \epsilon \in (0, \epsilon_\tau).$$

The local superiority of  $\sigma$  implies  $\pi(\sigma|\mu') > \pi(\mu'|\mu')$ , where

$$\begin{aligned} \pi(\mu'|\mu') &= \pi(\epsilon\tau + (1 - \epsilon)\sigma|\mu') \\ &= \epsilon\pi(\tau|\mu') + (1 - \epsilon)\pi(\sigma|\mu'). \end{aligned} \tag{6.3}$$

Because Equation (6.3) holds for all  $\epsilon \in (0, \epsilon_\tau)$ , it follows that  $\pi(\sigma|\mu') > \pi(\mu'|\mu')$  if and only if  $\pi(\sigma|\mu') > \pi(\tau|\mu')$ . Expanding the latter inequality, we

see that

$$\begin{aligned}\pi(\sigma|\epsilon\tau + (1-\epsilon)\sigma) &> \pi(\tau|\epsilon\tau + (1-\epsilon)\sigma) \\ \epsilon\pi(\sigma|\tau) + (1-\epsilon)\pi(\sigma|\sigma) &> \epsilon\pi(\tau|\tau) + (1-\epsilon)\pi(\tau|\sigma)\end{aligned}$$

and so either  $\pi(\sigma|\sigma) > \pi(\tau|\sigma)$  or else  $\pi(\sigma|\sigma) = \pi(\tau|\sigma)$  and  $\pi(\sigma|\tau) > \pi(\tau|\tau)$ , which means  $\sigma$  is an evolutionarily stable strategy.

( $\Leftarrow$ ) Assume that  $\sigma = p_1s_1 + \dots + p_ns_n$  is an ESS. To begin, notice that we only need to consider strategies  $\mu$  such that  $\text{supp}(\mu) \not\subseteq \text{supp}(\sigma)$ . The Bishop–Cannings theorem implies, for any  $\mu$  with  $\text{supp}(\mu) \subseteq \text{supp}(\sigma)$ , that  $\pi(\mu|\sigma) = \pi(\sigma|\sigma)$ , which means, from the Maynard Smith and Price definition of an ESS,  $\pi(\sigma|\mu) > \pi(\mu|\mu)$ .

Now consider the set of strategies

$$C = \{q_1s_1 + \dots + q_ns_n | q_i = 0 \text{ and } p_i > 0 \text{ for some } i\}$$

and let  $\bar{\epsilon}$  denote the uniform invasion boundary for  $\sigma$ . For all  $\tau \in C$  and every  $\epsilon \in (0, \bar{\epsilon})$ , we know that

$$\pi(\sigma|\epsilon\tau + (1-\epsilon)\sigma) > \pi(\tau|\epsilon\tau + (1-\epsilon)\sigma).$$

For convenience, denote the strategy  $\epsilon\tau + (1-\epsilon)\sigma$  by  $\mu$ . Multiplying the above inequality by  $\epsilon$  and adding  $(1-\epsilon)\pi(\sigma|\mu)$  to both sides yields

$$\begin{aligned}\epsilon\pi(\sigma|\mu) + (1-\epsilon)\pi(\sigma|\mu) &> \epsilon\pi(\tau|\mu) + (1-\epsilon)\pi(\sigma|\mu) \\ \pi(\sigma|\mu) &> \pi(\epsilon\tau + (1-\epsilon)\sigma|\mu) \\ \pi(\sigma|\mu) &> \pi(\mu|\mu).\end{aligned}$$

□

If a game has a completely mixed evolutionarily stable strategy, it turns out that we can state of necessary and sufficient conditions for it in terms of properties of a certain class of vectors. Suppose that  $\mathbf{p} = (p_1, \dots, p_n)$  and  $\mathbf{q} = (q_1, \dots, q_n)$  are two mixed strategies for a game. Then notice that  $\boldsymbol{\xi} = \mathbf{p} - \mathbf{q}$  has the property that  $\sum \xi_i = 0$ , because  $\boldsymbol{\xi}$  just tells how to redistribute the probability in  $\mathbf{q}$  in order to get  $\mathbf{p}$ , and probability is conserved. We will call a vector like  $\boldsymbol{\xi}$  a *probability redistribution vector*. The next theorem characterizes a completely mixed evolutionarily stable strategy in terms of properties of probability redistribution vectors. This is a case when the matrix notation introduced earlier proves more convenient in stating and proving the theorem, largely because  $\boldsymbol{\xi}$  has no natural game theoretic interpretation.

Hofbauer and Sigmund (2002) prove the following theorem.

**Theorem 6.5** *Let  $\mathbf{p}$  be a completely mixed Nash equilibrium for a game with payoff matrix  $\mathbf{A}$ . Then  $\mathbf{p}$  is an ESS if and only if*

$$\boldsymbol{\xi} \cdot \mathbf{A}\boldsymbol{\xi} < 0 \text{ for all } \boldsymbol{\xi} \neq \mathbf{0} \text{ with } \sum_{i=1}^N \xi_i = 0.$$

*Proof* ( $\Rightarrow$ ) Assume that  $\mathbf{p}$  is a completely mixed ESS. From the proof of 6.4, we know that  $\mathbf{p} \cdot \mathbf{Aq} > \mathbf{q} \cdot \mathbf{Aq}$  for all  $\mathbf{q} \in \Delta$  with  $\mathbf{p} \neq \mathbf{q}$ . So fix  $\mathbf{q} = \mathbf{p} + \boldsymbol{\zeta}$ , where  $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_n)$  and  $\sum \zeta_i = 0$ . Then

$$\begin{aligned}\mathbf{p} \cdot \mathbf{A}(\mathbf{p} + \boldsymbol{\zeta}) &> \mathbf{q} \cdot \mathbf{A}(\mathbf{p} + \boldsymbol{\zeta}) \\ \mathbf{p} \cdot \mathbf{Ap} + \mathbf{p} \cdot \mathbf{A}\boldsymbol{\zeta} &> \mathbf{q} \cdot \mathbf{Ap} + \mathbf{q} \cdot \mathbf{A}\boldsymbol{\zeta}.\end{aligned}$$

Because  $\mathbf{p}$  is a completely mixed Nash equilibrium, from the Bishop–Cannings theorem it follows that  $\mathbf{p} \cdot \mathbf{Ap} = \mathbf{q} \cdot \mathbf{Ap}$ , hence

$$\begin{aligned}\mathbf{p} \cdot \mathbf{A}\boldsymbol{\zeta} &> \mathbf{q} \cdot \mathbf{A}\boldsymbol{\zeta} \\ \mathbf{p} \cdot \mathbf{A}\boldsymbol{\zeta} &> (\mathbf{p} + \boldsymbol{\zeta}) \cdot \mathbf{A}\boldsymbol{\zeta} \\ 0 &> \boldsymbol{\zeta} \cdot \mathbf{A}\boldsymbol{\zeta}.\end{aligned}$$

Now consider an arbitrary  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n) \neq 0$  such that  $\sum \xi_i = 0$ . Then the vector  $\frac{\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|}$  is a probability redistribution vector, and so taking  $\boldsymbol{\zeta} = \frac{\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|}$  in the above calculations gives  $\frac{\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|} \cdot \mathbf{A} \frac{\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|} < 0$ . Multiplying both sides by  $\|\boldsymbol{\xi}\|^2$  yields  $\boldsymbol{\xi} \cdot \mathbf{A}\boldsymbol{\xi} < 0$ , as desired.

( $\Leftarrow$ ) Assume that  $\mathbf{p}$  is a completely mixed Nash equilibrium, and that the payoff matrix  $\mathbf{A}$  is such that  $\boldsymbol{\xi} \cdot \mathbf{A}\boldsymbol{\xi} < 0$  for all  $\boldsymbol{\xi} \neq \mathbf{0}$  with  $\sum \xi_i = 0$ . Fix some  $\epsilon > 0$  and consider the set of all mixed strategies  $\mathbf{q}$  within  $\epsilon$  distance of  $\mathbf{p}$ . Now, any such  $\mathbf{q}$  can be written as  $\mathbf{p} + \boldsymbol{\xi}$  for some vector  $\boldsymbol{\xi}$  which serves to redistribute the probability mass. By simply following the inequality manipulations above in reverse, we obtain  $\mathbf{p} \cdot \mathbf{A} > \mathbf{q} \cdot \mathbf{Aq}$ , which shows that  $\mathbf{p}$  is a locally superior strategy and thus an ESS.  $\square$

### 6.3 Set-based evolutionary stability concepts

The definition of an evolutionarily stable strategy only considers stability from the point of view of an individual strategy. One interesting question is how to extend the concept of evolutionary stability to a *set* of strategies. As motivation, consider a four-strategy game, with payoffs as follows: all strategies receive a payoff of 1 when played against themselves, strategies  $S_1$  and  $S_2$  both receive a payoff of 1 when played against any other strategy, but strategies  $S_3$  and  $S_4$  both receive a payoff of  $\frac{1}{2}$  when played against any other strategy. In this setting, neither  $S_1$  nor  $S_2$  satisfy the definition of an ESS because a pure population of  $S_1$  can be invaded by  $S_2$ , and vice versa. Yet, a population consisting of any mix of  $S_1$  and  $S_2$  is stable in the sense that any  $S_3$  or  $S_4$  mutants will be driven out. This, essentially, extends the idea of an ESS from a single strategy to a set of strategies.

We first need to introduce some notation. If  $\mu \in \Delta$  is a strategy (possibly mixed), let  $\beta(\mu)$  denote the set of best-replies to  $\mu$ ; that is,

$$\beta(\mu) = \{\sigma \in \Delta : \pi(\sigma|\mu) \geq \pi(\varphi|\mu), \text{ for all } \varphi \in \Delta\}.$$

In addition, let  $\Delta^{\text{NE}}$  denote the set of all Nash equilibrium strategies for the game under consideration.

Using this, we can now state one set-based definition of evolutionary stability, due to Thomas (1985), as follows:

**Definition 6.6** A set  $S \subset \Delta^{\text{NE}}$  is said to be an *evolutionarily stable set* if  $S$  is nonempty and closed, and each  $\mu \in S$  has a neighborhood  $N$  such that  $\pi(\mu|\sigma) \geq \pi(\sigma|\sigma)$  for all strategies  $\sigma \in N \cap \beta(\mu)$ , with strict inequality when  $\sigma \notin N$ .

There is a close relationship between evolutionarily stable sets and a weaker concept of evolutionarily stability, known as neutral stability. This is defined as follows:

**Definition 6.7** A strategy  $\sigma \in \Delta$  is a *neutrally stable strategy* if for every strategy  $\mu \in \Delta$  there is some threshold  $\bar{\epsilon}_\mu \in (0, 1)$  such that

$$\pi(\sigma|\epsilon\mu + (1 - \epsilon)\sigma) \geq \pi(\mu|\epsilon\mu + (1 - \epsilon)\sigma)$$

holds for all  $0 < \epsilon < \bar{\epsilon}_\mu$ .

Comparison with Definition 6.2 shows that the sole difference between a neutrally stable strategy and an ESS concerns the inequality. Whereas an evolutionarily stable strategy must have strictly higher fitness than any mutant, in the resulting population mix, a neutrally stable strategy requires only that no mutant do better than the incumbent (possibly doing equally well) in the resulting population mix.

It can be proven (see Weibull, 1995) that every strategy in an evolutionarily stable set is neutrally stable. However, because not every game has a neutrally stable strategy, as the following example shows, it may be the case that no evolutionarily stable set exists.

**Example 6.4** Consider the game specified by the following payoff matrix:

	$S_1$	$S_2$	$S_3$
$S_1$	(1, 1)	(1, 0)	(0, 1)
$S_2$	(0, 1)	(1, 1)	(1, 0)
$S_3$	(1, 0)	(0, 1)	(1, 1)

One can easily verify that each pure strategy is a Nash equilibrium, as is the mixed strategy  $\sigma = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ . Notice that the pure strategy  $S_1$  is not neutrally stable, because

$$\pi(S_1|\epsilon S_3 + (1 - \epsilon)S_1) = \epsilon\pi(S_1|S_3) + (1 - \epsilon)\pi(S_1|S_1) = 1 - \epsilon$$

whereas

$$\pi(S_3|\epsilon S_3 + (1 - \epsilon)S_1) = \epsilon\pi(S_3|S_3) + (1 - \epsilon)\pi(S_3|S_1) = 1.$$

Similarly, one can show that neither  $S_2$  nor  $S_3$  are neutrally stable, as  $S_2$  can be invaded by  $S_1$  and  $S_3$  can be invaded by  $S_2$ . This leaves  $\sigma$  as the only remaining possibility for a neutrally stable strategy. However, one can easily show that  $\sigma$  can be invaded by any pure strategy, and so it is not neutrally stable, either.

The next theorem identifies some properties of evolutionarily stable sets.

### Theorem 6.6

- (a) If  $S \subset \Delta^{\text{ESS}}$ , then  $S$  is an evolutionarily stable set.
- (b) The union of evolutionarily stable sets is also an evolutionarily stable set.
- (c) If an evolutionarily stable set is the finite union of disjoint closed sets, then each such set is an evolutionarily stable set.

*Proof* (a) From Corollary 6.3, we know that there are only finitely many ESS, and so the set  $S$  must be finite. Thus, for each member  $s \in S$ , we may choose a neighborhood  $N_s$  such that  $N_s \cap S = \{s\}$ . Theorem 6.4 established that a strategy is an ESS if and only if it is locally superior, so for each  $s \in S$ , take  $N$  in the definition of an evolutionarily stable set to be the intersection of  $N_s$  and the set in which  $s$  is locally superior. This establishes that  $S$  is an evolutionarily stable set.

(b) Suppose that  $\{X_1, \dots, X_n\}$  is a collection of evolutionarily stable sets, and let  $Y = \bigcup_i X_i$ . If  $x \in Y$ , then  $x \in X_i$  for some  $i$ , and so there exists a neighborhood  $N$  such that  $\pi(x|y) \geq \pi(y|y)$  for all strategies  $y \in N$ , with the additional property that  $\pi(x|y) > \pi(y|y)$  for all strategies  $y \in N \cap Y \setminus X_i$ . Hence,  $Y$  is an evolutionarily stable set.

(c) Let  $Y$  be an evolutionarily stable set generated by a finite union of disjoint closed sets  $\{X_1, \dots, X_n\}$ , and let  $N$  denote a neighborhood whose existence is guaranteed by the definition of an evolutionarily stable set for  $Y$ . Now suppose that  $x \in X_i$ . Let  $N_i$  be a neighborhood of  $X_i$  which is disjoint from all of the other  $X_j$ , and consider  $U_i = N \cap N_i$ . By construction of  $U_i$ , it follows that  $X_i$  is an evolutionarily stable set.  $\square$

An alternative set-based concept of evolutionarily stability (originally due to Swinkels, 1992) is as follows:

**Definition 6.8** A set  $S \subset \Delta$  is said to be an *equilibrium evolutionarily stable set* if  $S$  is minimal with respect to the following property:  $S$  is a nonempty, closed subset of  $\Delta^{\text{NE}}$  and there exists a threshold  $\bar{\epsilon} \in (0, 1)$  such that, if  $s \in S$  and  $\sigma \in \Delta$ , with  $\epsilon \in (0, \bar{\epsilon})$  and  $y \in \beta((1 - \epsilon)s + \epsilon y)$ , then  $(1 - \epsilon)s + \epsilon y \in S$ .

**Theorem 6.7** Every equilibrium evolutionarily stable set  $S \subset \Delta^{\text{NE}}$  is a component of  $\Delta^{\text{NE}}$ .

*Proof* One can prove that  $\Delta^{\text{NE}}$  equals the union of a finite set of disjoint, closed, and connected sets (see Weibull, 1995, p. 16). The members of this set are known as the *components* of the set of Nash equilibria. We begin by showing that, for every equilibrium stable set  $S$ , there is an open set  $O$  such that  $\Delta^{\text{NE}} \cap O = S$ . (This establishes that there is a “buffer” surrounding  $S$  which is not in the set  $\Delta^{\text{NE}}$ .)

Suppose that this claim is false for some set  $S$ . Then we can find a strategy  $s \in S$  and a sequence of strategies  $\langle \mu_t \rangle_{t=1}^{\infty}$  with  $\mu_t \in \Delta$  such that

$$\sigma_t = \epsilon_t \mu_t + (1 - \epsilon_t)s \in \Delta^{\text{NE}} \cap \sim S$$

for all  $t$  sufficiently large, say  $t \geq T$ .

Because the support of  $\mu_t$  is contained in the support of  $\sigma_t$ , and each  $\sigma_t$  is a Nash equilibrium strategy, it follows that  $\mu_t \in \beta(\epsilon_t \mu_t + (1 - \epsilon_t)s)$ , for all  $t \geq T$ . But then it follows that  $\sigma_t \in S$  for all  $\epsilon_t < \bar{\epsilon}$ , where  $\bar{\epsilon}$  is the threshold mentioned in the definition of an equilibrium stable set. Thus we have a contradiction.

Because  $S$  is a closed set, it is either a component of  $\Delta^{\text{NE}}$  or the union of multiple components. But because of the minimality property of  $S$ , the latter possibility is ruled out.  $\square$

**Theorem 6.8** *Every evolutionarily stable set contains some equilibrium evolutionarily stable set. Any connected evolutionarily stable set is an equilibrium evolutionarily stable set.*

*Proof* Let  $S \subset \Delta^{\text{NE}}$  be an evolutionarily stable set, and let  $\sigma \in S$ . Pick a  $\mu \in \Delta$  and consider the strategy  $\tau = \epsilon\mu + (1 - \epsilon)\sigma$  for  $\epsilon \in (0, 1)$ . Now, if  $\mu$  is a best-response to  $\tau$ , then

$$\pi(\mu|\tau) \geq \pi(\sigma|\tau)$$

and so

$$\begin{aligned} \pi(\tau|\tau) &= \pi(\epsilon\mu + (1 - \epsilon)\sigma|\tau) \\ &= \epsilon\pi(\mu|\tau) + (1 - \epsilon)\pi(\sigma|\tau) \\ &\geq \pi(\sigma|\tau). \end{aligned}$$

Thus,  $\tau \in S$  for all  $\epsilon$  sufficiently small, by the definition of an evolutionarily stable set. Furthermore, note that  $S$  possesses the key property of an equilibrium evolutionarily stable set. The existence of a minimal subset  $T$  of  $S$  with the key property of an equilibrium evolutionarily stable set is guaranteed by Zorn's lemma. Finally, if  $S$  happens to be connected, then the only equilibrium evolutionarily stable set  $T$  of  $S$  must be  $S$  itself, from Theorem 6.7.  $\square$

## 6.4 Evolutionary dynamics

The primary shortcoming of the concept of an evolutionarily stable strategy is one shared by the underlying concept of a Nash equilibrium: it is *static*. Ideally, we would like to be able to say something about how populations behave in an out-of-equilibrium context, particularly regarding which evolutionarily stable strategies we might reasonably expect to find, in games having multiple ESS. And, as not all games have an ESS, it would still be good to be able to say something about the population's expected behavior in these cases, too.

### 6.4.1 The replicator dynamics

The first dynamical model studied in evolutionary game theory was the *replicator dynamics*, originally introduced by Taylor and Jonker (1978). As we will see, the replicator dynamics possesses a number of nice analytic properties; in addition, it has the virtue of admitting interpretation as either a model of (some types) of biological evolution or as a model of (some types) of cultural evolution.

Let  $S = \{s_1, \dots, s_n\}$  denote the set of pure strategies in the game, and let  $p_i(t) \in [0, 1]$  denote the proportion of pure strategy  $i$  in the population at time  $t$ . Unlike the previous section, here we assume players cannot use mixed strategies. If the only thing that matters is the frequency of a strategy in the population, then  $\vec{p}(t) = (p_1(t), \dots, p_n(t)) \in \mathbb{R}_+^n$  is a point representing the complete population state. In what follows, we shall typically suppress the explicit time parameter to simplify the notation.

In the most general case, the fitness of  $s_i$  may depend on the overall state of the population in a complicated manner. Let  $W_i(\vec{p})$  denote the fitness of  $s_i$  in the population. The average fitness of the population  $\phi(\vec{p})$  is just the weighted average of these individual fitnesses,  $\phi(\vec{p}) = \sum_i p_i \cdot W_i(\vec{p})$ .

**Definition 6.9** The *continuous replicator dynamics* for a game with the pure strategy set  $S = \{s_1, \dots, s_n\}$  is the dynamical system given by

$$\frac{dp_i}{dt} = p_i[W_i(\vec{p}) - \phi(\vec{p})]$$

for all  $i = 1, \dots, n$ .

It is often assumed that the fitness of each strategy depends only linearly on the underlying population distribution. (As we will see below, this occurs when players interact pairwise and at random, with  $p_i$  corresponding to the chance of interacting with  $s_i$ .) In this case,  $W_i(\vec{p})$  takes the particular form  $\sum_j p_j \pi(s_i|s_j)$  and  $\phi(\vec{p}) = \sum_i \sum_j p_i p_j \pi(s_i|s_j)$ . This system is the *continuous linear replicator dynamics*.

Let us first show how the replicator dynamics can represent biological evolution. Suppose each individual has a phenotype  $s_i$ , corresponding to a strategy in the underlying game. Let  $n_i$  denote the number of agents in the population with the phenotype  $s_i$ , with the total population size being  $N = \sum_{i=1}^m n_i$ . If the only thing that matters about each agent is their phenotype, and the only thing relevant for natural selection is the proportion of each phenotype in the population, all of the relevant information is thus contained in the state vector  $\vec{p} = (p_1, \dots, p_n)$ , where  $p_i = \frac{n_i}{N}$  for all  $i$ . (In Section 6.5 we consider what happens if the location of individuals, either spatially or within a network, has strategic importance.)

Let  $r_i$  be the growth rate of  $s_i$  and assume that the rate of change in the number of  $s_i$ -individuals is proportional to the number of  $s_i$ -individuals in the population. That is,

$$\frac{dn_i}{dt} = r_i n_i.$$

Given this expression for the rate of change of  $s_i$ , we calculate the rate of change for the total population as follows:

$$\frac{dN}{dt} = \frac{d}{dt} \left( \sum n_i \right) = \sum \frac{dn_i}{dt} = \sum r_i n_i = \sum r_i p_i N = \bar{r}N \quad (6.4)$$

where the constant  $\bar{r}$  is defined to be  $\sum r_i p_i$ .

Equation (6.4) provides an expression for the total change in the population over time, but what we are really interested in is how the relative frequencies of each phenotype change over time:

$$\begin{aligned} \frac{dp_i}{dt} &= \frac{d}{dt} \left( \frac{n_i}{N} \right) = \frac{N \frac{dn_i}{dt} - n_i \frac{dN}{dt}}{N^2} = \frac{r_i(p_i N)N - \bar{r}(p_i N)N}{N^2} \\ &= p_i s_i(r_i - \bar{r}). \end{aligned} \quad (6.5)$$

Denote the expected fitness of phenotype  $s_i$  in the population by  $W_i$ . Assuming that the current growth rate of  $n_i$  is approximately equal to the expected fitness of the underlying phenotype, Equation (6.5) may be re-expressed as

$$\frac{dp_i}{dt} = p_i (W_i - \phi),$$

because  $\bar{r} =_{\text{df}} \sum r_i p_i = \sum W_i p_i = \phi$ .

Let us now derive the replicator dynamics from a dissatisfaction-driven cultural learning model (see Weibull, 1995, and Björnerstedt and Weibull, 1999, for the original sources). It is worth noting that there are a number of other derivations of the replicator dynamics in cultural evolutionary contexts.

As before, suppose we have a population of agents who repeatedly interact and play a game with pure strategies  $S = \{s_1, \dots, s_n\}$ . In addition, let us suppose that each agent wants to maximize his or her expected payoff, and so they will occasionally change their strategy if they believe that they are not doing sufficiently well.

More precisely, assume that each agent periodically reviews her strategy. At the end of each review process, depending on the outcome, she may adopt a new strategy. Although each agent might have her own unique rate at which she reviews her strategy, let us assume that the review rate only depends on the strategy employed by an agent; hence, a person following  $s_i$  reviews her strategy at the rate  $r_i$ . Let  $x_i^j$  denote the probability that someone using  $s_i$  will exchange it for  $s_j$ .

Can we say anything more specific about the way in which an individual reviews her strategy? We can, if we make the following assumptions:

1. The number of times an agent reviews her strategy between time  $t$  and  $t + \Delta t$  does not affect the number of times that she reviews her strategy during another interval  $t'$  and  $t' + \Delta t'$ . That is, the number of times an agent reviewed her strategy in the past has no effect on the number of times she will review her strategy in the future.
2. The average rate that she reviews her strategy remains constant.

3. Each agent must finish reviewing her strategy before beginning to review it another time.

The first condition is the most suspect of the above assumptions, insofar as it requires that people do not become tired of reviewing their strategy, and are just as willing to commence a new review after a one-minute break as after a one-week break. However, assumptions 1–3 imply that the review times of each individual form a Poisson process. This is useful because the aggregate of a number of statistically independent Poisson processes is itself a Poisson process, and the rate of the aggregate process is the sum of the individual rates.

If each individual strategy review and strategy switch is independent of every other strategy review and strategy switch, and if the population is sufficiently large,<sup>5</sup> we can express in a compact form the proportion of the population that reviews and switches their strategy at any given time. In this limiting case, let us denote the proportion of the population following  $s_i$  that chooses to review it by  $p_i r_i$ . Similarly, let us denote the proportion of the population following  $s_i$  who switch to  $s_j$ , after review, by  $p_j r_j x_j^i$ . Hence the total rate at which people start using  $s_i$  is

$$\sum_{j \neq i} p_j r_j x_j^i$$

and the total rate at which people stop using  $s_i$  is

$$\sum_{j \neq i} p_i r_i x_i^j.$$

The overall rate of change for  $p_i$  is just the difference of these two:

$$\frac{dp_i}{dt} = \left( \sum_{j \neq i} p_j r_j x_j^i \right) - \left( \sum_{j \neq i} p_i r_i x_i^j \right).$$

A more useful form can be obtained by rearranging the expressions slightly.

$$\begin{aligned} \frac{dp_i}{dt} &= \left[ \left( \sum_j p_j r_j x_j^i \right) - p_i r_i x_i^i \right] - \left[ \left( \sum_j p_i r_i x_i^j \right) - p_i r_i x_i^i \right] \\ &= \left( \sum_j p_j r_j x_j^i \right) - p_i r_i x_i^i - \left( \sum_j p_i r_i x_i^j \right) + p_i r_i x_i^i \\ &= \left( \sum_j p_j r_j x_j^i \right) - p_i r_i \sum_j x_i^j. \end{aligned}$$

<sup>5</sup> A euphemism for saying that we are really looking at the infinite limit.

Because  $\sum_j x_i^j = 1$ , the rate of change of  $p_i$  is thus

$$\frac{dp_i}{dt} = \sum_{j=1}^n p_j r_j x_j^i - p_i r_i. \quad (6.6)$$

Provided that the review rates  $r_i$  and exchange probabilities  $x_i^j$  satisfy certain conditions,<sup>6</sup> Equation (6.6) gives a well-defined set of dynamical laws describing a process of cultural evolution (or strategic learning) for the population.

Equation (6.6) provides a set of dynamical laws, but it does not yet have the form of the replicator dynamics. This is because we have yet to specify the exact mechanism used to review and switch strategies. Gigerenzer *et al.* (1999) discuss a particular heuristic called “take the last.” According to this heuristic, a boundedly rational individual chooses the last option encountered. If people mix randomly, the probability that a person following  $s_i$  will adopt  $s_j$  is  $p_j$ . That is,

$$x_i^j = p_j. \quad (6.7)$$

Björnerstedt (1993) suggested a review rate with the following form:

$$r_i = a - bW_i$$

where  $a, b \in \mathbb{R}$  with  $b > 0$ , and  $\frac{a}{b} \geq W_i$ .

From this, it follows that

$$\begin{aligned} \frac{dp_i}{dt} &= \sum_j p_j r_j x_j^i - p_i r_i. \\ &= \sum_j p_j (a + bW_j) p_i - p_i (a - bW_i) \\ &= \sum_j (p_j p_i a - p_j p_i bW_j) - p_i (a - bW_i) \\ &= p_i a \sum_j p_j - p_i b \sum_j p_j W_j - p_i a + p_i bW_i \\ &= p_i a - p_i b\phi - p_i a + p_i bW_i \\ &= p_i b(W_i - \phi). \end{aligned}$$

The constant  $b$  is a free parameter, and if  $b = 1$  we obtain the continuous replicator dynamics.

Note that the simplex  $S^n$  is invariant under the replicator dynamics: if it is the case that the initial state of population lies on the simplex, and we denote the solution to the replicator equations by  $\tilde{p}(t)$ , then  $\tilde{p}(t) \in S^n$  for all  $t \in \mathbb{R}$ . This is easily verified by noting that the sum  $S = p_1 + \dots + p_n$  satisfies the equation  $\frac{dS}{dt} = (1 - S)\phi$ , for which  $S(t) = 1$  is a solution. In addition, if  $p_i(0) = 0$ , for any  $i$ , then  $p_i(t) = 0$  for all  $t$ . Because the replicator dynamics cannot introduce new strategies

<sup>6</sup> The  $r_i$  and  $p_i^j$  need to be Lipschitz continuous (see Weibull, 1995, pp. 153 and 232).

into a population if they are not initially present, variant dynamics, such as the replicator–mutator dynamics, or the Brown–von Neumann–Nash dynamics need to be used to model this behaviour.

One useful property of the continuous linear replicator dynamics is that its solutions are invariant under positive affine transformations, provided one suitably rescales time. If  $\pi$  is the fitness function, consider the transformed fitness function given by  $\bar{\pi} = a\pi + b$ , for some  $a, b \in \mathbb{R}$  with  $a > 0$ . If the current state of the population is  $\vec{p} = (p_1, \dots, p_n)$ , then

$$\begin{aligned}\frac{dp_i}{dt} &= p_i \left( \sum_{j=1}^n p_j \bar{\pi}(s_i|s_j) - \sum_{j=1}^n \sum_{k=1}^n p_j p_k \bar{\pi}(s_j|s_k) \right) \\ &= p_i \left( \sum_{j=1}^n p_j (a\pi(s_i|s_j) + b) - \sum_{j=1}^n \sum_{k=1}^n p_j p_k (a\pi(s_j|s_k) + b) \right) \\ &= ap_i \left( \sum_{j=1}^n p_j \pi(s_i|s_j) - \sum_{j=1}^n \sum_{k=1}^n p_j p_k \pi(s_j|s_k) \right).\end{aligned}$$

The effect of the constant  $a$  merely serves to speed up or slow down the rate that the population follows a trajectory, rather than change their shape. This enables us to assume, without loss of generality, that all payoffs are non-negative.

**Example 6.5** Weibull (1995, pp. 28–31) shows that there are only three fundamentally different types of symmetric two-player strategic form games, given that utility functions are only unique up to a positive affine transformation. Here are canonical examples of the three types and their behavior under the replicator dynamics.

*The Prisoners' dilemma.* Suppose two criminals were arrested by the police and put in separate cells. The prosecutor approaches each prisoner in turn and offers the following deal: if the suspect agrees to help the prosecutor, he will always get a better outcome than if he does not agree to help. However, if both suspects agree to help, each will go to prison for longer than if they did not agree to help. The payoff matrix to the left reflects one possible representation of this problem. The strategy labels  $C$  and  $D$  indicate whether the player “Defects” on his partner (i.e., agrees to help the prosecutor) or “Cooperates” with his partner (i.e., stays silent). The only Nash equilibrium of the game is  $(D, D)$ , which is not socially optimal. For this reason, the Prisoners’ dilemma is often used to model the strategic problem underlying cooperative behavior.

	$C$	$D$
$C$	(2, 2)	(0, 3)
$D$	(3, 0)	(1, 1)

The behavior of this game under the replicator dynamics is easily determined. Suppose that  $p_C = (1 - p_D) > 0$ . Because  $\pi(D|C) > \pi(C|C)$  and  $\pi(D|D) > \pi(C|D)$ , it follows that  $W_D > W_C$  and hence  $\frac{dp_D}{dt} > 0$  and  $\frac{dp_C}{dt} < 0$ . Any mixed population of Cooperators and Defects always converges to All Defect. A population which starts in the All Cooperate state will remain there, because  $\frac{dp_D}{dt} = p_D(W_D - \phi) = 0$ . (This is a general feature of the replicator dynamics: any strategy initially absent from the population will remain absent. The replicator–mutator dynamics, discussed in Section 6.4.2, provide an extension which lifts this restriction.)

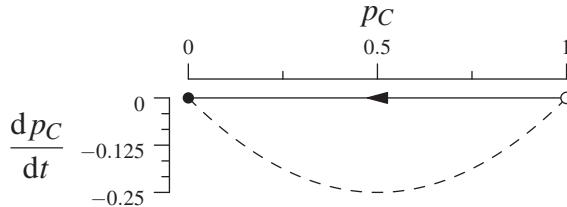
Specifically, the replicator equations for the Prisoners' Dilemma are:

$$\frac{dp_C}{dt} = p_C^2 - p_C \quad \frac{dp_D}{dt} = -\frac{dp_C}{dt} = p_C - p_C^2 \quad p_C(0) = p \quad p_D(0) = 1 - p$$

with solutions

$$p_C(t) = \frac{p}{-e^t - p + pe^t} \quad p_D(t) = \frac{e^t(-1 + p)}{-e^t - p + pe^t}.$$

The orbit of this system, along with its velocity, is shown below:



*A coordination game.* A pure coordination game awards players equal amounts of utility when all of them choose the same strategy; an impure coordination game awards players differing amounts of utility when they choose the same strategy.<sup>7</sup> The payoff matrix to the left represents a pure coordination game where the all-*A* equilibrium is preferred by both players. Note that all-*B* is also a Nash equilibrium and, furthermore, *B* is an evolutionarily stable strategy.

	<i>A</i>	<i>B</i>
<i>A</i>	(2, 2)	(0, 0)
<i>B</i>	(0, 0)	(1, 1)

<sup>7</sup> “Battle of the Sexes” is the canonical example of an impure coordination game. In that game, two individuals can attend either a boxing match or a ballet performance. One prefers boxing over ballet, the other ballet over boxing, but both prefer to do something together than alone. For example,

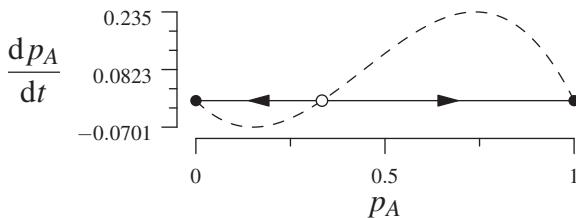
		Maggie	
		Boxing	Ballet
Billy		Boxing	(1, 2)
		Ballet	(0, 0)
			(2, 1)

Because this game is not symmetric, it falls outside the scope of games we are currently considering.

As above, the fact that  $p_A = (1 - p_B)$  simplifies the replicator equations to:

$$\begin{aligned}\frac{dp_A}{dt} &= 4p_A^2 - 3p_A^3 - p_A \quad p_A(0) = p \\ \frac{dp_B}{dt} &= 3p_A^3 + p_A - 4p_A^2 \quad p_B(0) = 1 - p.\end{aligned}$$

(The equations involving  $p_B$  are redundant, given that this is a two-strategy game.) One can check that  $p_A = \frac{1}{3}$  is a fixed point of the dynamics, but an unstable one. If the initial conditions have  $p_A > \frac{1}{3}$ , the system converges to an all-A state, and if  $p_A < \frac{1}{3}$ , it converges to all-B. The orbit and velocity of the system are as follows:



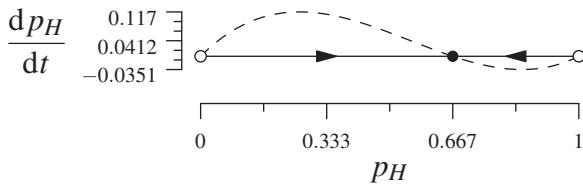
*The Hawk–Dove game.* This game was introduced as a model of territorial competition by Maynard Smith and Price (1973), and was one of the first games studied in evolutionary game theory. Suppose that members of a species compete for control of some resource. When a Hawk meets another individual, the Hawk engages in display behavior, potentially escalating to a conflict. If the Hawk's opponent is a Dove, the Dove immediately retreats upon the Hawk's escalation, ceding the resource to the Hawk. When two Hawks meet, a conflict occurs, ultimately resulting in one Hawk being hurt and the other obtaining the resource. Two Doves will share the resource.

	Hawk	Dove
Hawk	$\left(\frac{V-C}{2}, \frac{V-C}{2}\right)$	$(V, 0)$
Dove	$(0, V)$	$\left(\frac{V}{2}, \frac{V}{2}\right)$

Let  $V$  be the value of the resource, and  $C$  the cost incurred by the Hawk who is injured in a conflict. Assuming each Hawk has an equal chance of being hurt in a conflict, the payoffs are as in the matrix. If  $V > C$ , then the Hawk strategy is dominant and an evolutionarily stable strategy. The interesting case occurs when  $V < C$ , as then no pure strategy ESS exists. We can solve for the mixed strategy Nash equilibrium  $\sigma$  using the Bishop–Cannings theorem: if  $\sigma$  is a mixed-strategy Nash equilibrium, then  $\pi(\text{Hawk}|\sigma) = \pi(\text{Dove}|\sigma) = \pi(\sigma|\sigma)$ . Doing so reveals that the mixed-strategy Nash equilibrium, which is also an ESS, plays Hawk with probability  $\frac{V}{C}$ . The replicator equations for Hawk are:

$$\frac{dp_H}{dt} = \frac{1}{2}(1 - p_H)p_H(V - Cp_H) \quad p_H(0) = p.$$

If  $V = 2$  and  $C = 3$ , then the orbit and velocity is as follows:



□

As the number of strategies increases, so does the possible complexity of the underlying dynamics. For three strategy symmetric games using the replicator dynamics, Bomze (1983) provides a graphical classification of all 46 different kinds of games.

From the theory of differential equations, we know that solutions exist for the replicator dynamics. However, exact analytic solutions are typically hard to find, which means that one must either solve the equations using numerical methods, or else use other means to determine the long-term behaviour of the system. Because, in general, we are usually interested in the long-term behavior rather than the medium-run dynamics, let us now turn to some of these methods.

To begin, we need to distinguish several different stability concepts pertaining to dynamical systems. Stability concepts characterize the behavior of a dynamical system in the local region surrounding a *fixed point*,<sup>8</sup> a point  $\vec{p}$  where  $\frac{dp_i}{dt} = 0$  for all  $i = 1, \dots, n$ . In the following, when referring to “the replicator dynamics” I assume that the payoff matrix  $M$  has been fixed in advance.

**Definition 6.10** Let  $\vec{p} \in S^n$  be a fixed point of the replicator dynamics. Then  $\vec{p}$  is *stable* if, for every  $\epsilon > 0$  there is a  $\delta = \delta(\epsilon) > 0$  such that  $\|\vec{p}(0) - \vec{p}\| < \delta$  implies  $\|\vec{p}(t) - \vec{p}\| < \epsilon$ , for all  $t \geq 0$ . That is, every trajectory which passes sufficiently close to  $\vec{p}$  remains close to  $\vec{p}$  for all future times. This is known as “stability in the sense of Lyapunov,” after the Russian mathematician Aleksandr Lyapunov.

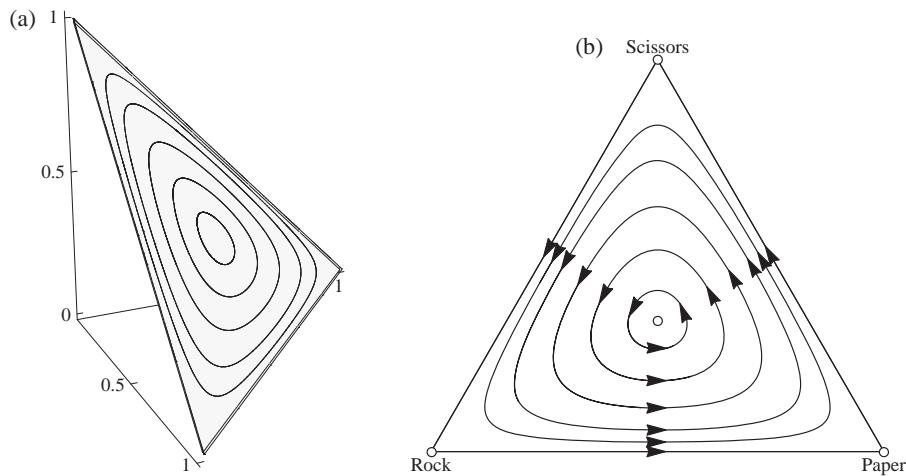
The intuition behind stability in the sense of Lyapunov is that no local “push” away from the fixed point exists. Another intuition concerning stability is that there is a local “pull” toward the fixed point.

**Definition 6.11** Let  $\vec{p}$  be a fixed point of the replicator dynamics. Then  $\vec{p}$  is *asymptotically stable* if it is stable and, in addition, there exists a  $\delta > 0$  such that if  $\|\vec{p}(0) - \vec{p}\| < \delta$  then  $\lim_{t \rightarrow \infty} \vec{p}(t) = \vec{p}$ .

**Definition 6.12** If the fixed point  $\vec{p}$  is stable but not asymptotically stable, then  $\vec{p}$  is said to be *neutrally stable*.

An example of a game with a fixed point that is neutrally stable, but not asymptotically stable, is the version of Rock–Paper–Scissors from Example 6.2. As

<sup>8</sup> Which is also known in the dynamical systems literature as an *equilibrium point*, so caution should be exercised to avoid conflating this sense of equilibrium with that of a Nash equilibrium.



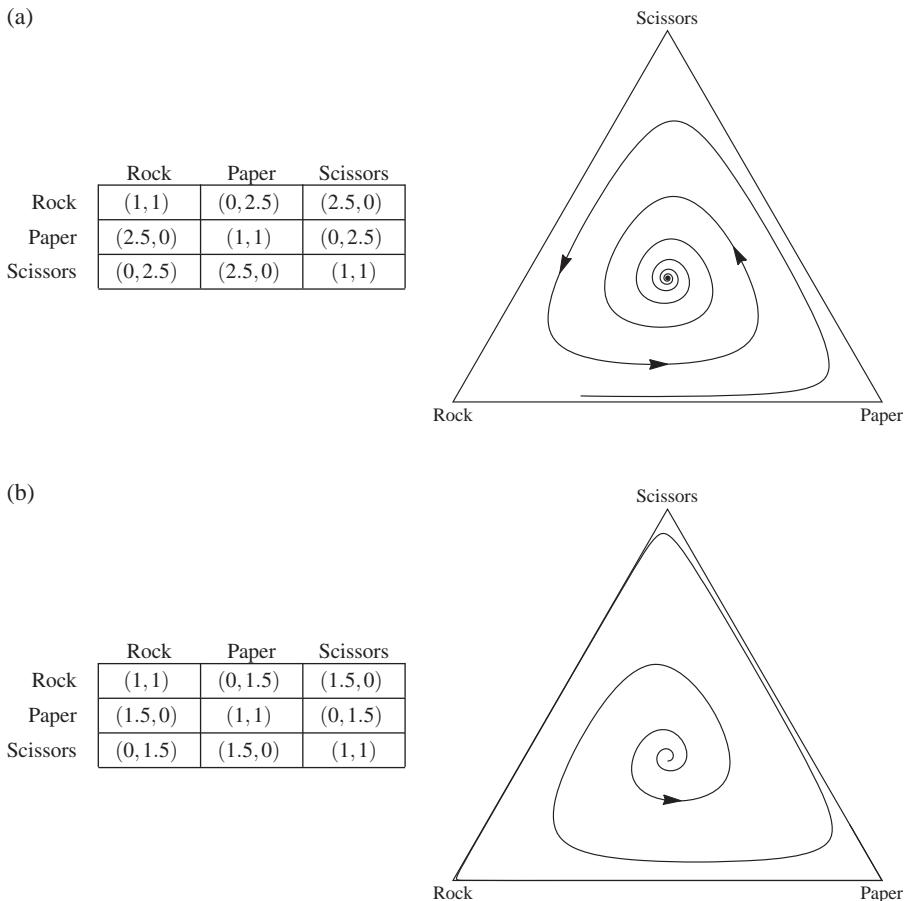
**Figure 6.2** Several trajectories for the game of Rock–Paper–Scissors, illustrating the neutral stability of the population state  $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ . (a) The simplex diagram, with flow trajectories, for the game of Rock–Paper–Scissors. (b) The simplex diagram for Rock–Paper–Scissors, rotated.

Figure 6.2 shows, with the exception of the fixed point at the population state  $\vec{p}_0 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ , all other points in the interior lie on orbits cycling around  $\vec{p}_0$ . It is obvious that  $\vec{p}_0$  is stable in the sense of Lyapunov, because, given any neighborhood  $N$  of  $\vec{p}_0$ , another neighborhood  $N' \subseteq N$  can be found such that all trajectories passing through  $N'$  stay within  $N$  (if  $N$  has radius  $\epsilon$ , simply let  $N'$  be the neighborhood with radius  $\frac{\epsilon}{2}$ ). Similarly, inspection clearly shows that  $\vec{p}_0$  is not asymptotically stable, so  $\vec{p}_0$  is neutrally stable.

If the payoffs to Rock–Paper–Scissors are slightly modified, though, the stability of the fixed point  $\vec{p}_0$  can change significantly. If the winning payoffs (e.g., that obtained by playing Rock against Scissors) are increased to  $2 + \epsilon$ , for any  $\epsilon > 0$ , all trajectories in the simplex interior converge to  $\vec{p}_0$ , as Figure 6.3a shows, and so  $\vec{p}_0$  is now asymptotically stable. However, if the winning payoffs are decreased to  $2 - \epsilon$ , for any  $\epsilon > 0$ , then  $\vec{p}_0$  becomes *unstable*, as shown in Figure 6.3b. This means that the fixed point in the original Rock–Paper–Scissors game is *structurally unstable*, as minor perturbations to the underlying vector field yield solutions which are not topologically equivalent.

An *attractor* of a dynamical system is a set of points invariant under the dynamics which also have the property that “nearby” states converge to the attracting set in the limit. (In addition, it is also required that no proper subset also counts as an attractor.) The set of points which converge to the attractor under the evolution of the dynamics is known as its *basin of attraction*. In Figure 6.3a, the basin of attraction for  $\vec{p}_0$  is the entire interior of the simplex.

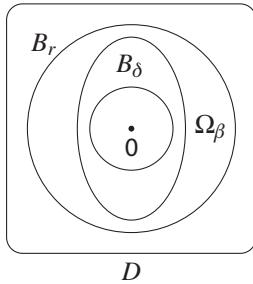
Visual inspection suffices for identifying states as neutrally or asymptotically stable in simple games, but as the number of strategies increases, eventually



**Figure 6.3** (a) Rock–Paper–Scissors, with asymptotically stable fixed point.  
(b) Rock–Paper–Scissors, with unstable fixed point.

visualization is no longer possible. Furthermore, this method requires obtaining a solution, either explicitly or numerically. We would like to be able to characterize the stability of fixed points without needing to refer to solutions to the equations; for this reason we turn to the topic known as Lyapunov stability.

By way of motivation, consider the dynamical system generated by a simple one-dimensional pendulum, where  $\theta \in (-\pi, \pi]$  denotes the displacement from the natural resting position in radians. In the absence of friction, the natural resting position occurs at  $\theta = 0$ . Conservation of energy implies that, if the system starts with some positive amount of energy  $E(0) = c$  (defined as the sum of potential and kinetic energy), the system will trace a closed orbit around the origin in the  $(\frac{d\theta}{dt}, \theta)$ -plane. In the presence of friction, energy will gradually be dissipated. This means that along every trajectory of the system, we have  $\frac{dE}{dt} \leq 0$ , and so the energy will eventually reach 0, showing that  $\theta = 0$  is asymptotically stable. The key insight of Lyapunov was the realization that a similar analysis can be performed using a broad family of such “generalized energy” functions.



**Figure 6.4** The nesting of sets about the fixed point  $\mathbf{x} = \mathbf{0}$  as in the proof of Theorem 6.9.

In the statement of the following theorem, it is assumed that the fixed point of the system occurs at the origin. This can be done without loss of generality because any fixed point may be translated to the origin through a change of variables. The following proof follows that given in Khalil (2002).

**Theorem 6.9** (Lyapunov stability theorem) *Let  $\dot{\mathbf{x}} = f(\mathbf{x})$  be an autonomous dynamical system with fixed point  $\mathbf{x} = \mathbf{0}$ , where  $f : D \rightarrow \mathbb{R}^n$  is a locally Lipschitz continuous map from the domain  $D \subseteq \mathbb{R}^n$  containing  $\mathbf{x} = \mathbf{0}$  into  $\mathbb{R}^n$ . Let  $V : D \rightarrow \mathbb{R}$  be a continuously differentiable function such that*

$$V(\mathbf{0}) = 0 \text{ and } V(\mathbf{x}) > 0 \text{ in } D \setminus \{\mathbf{0}\} \quad (6.8)$$

and

$$\dot{V}(\mathbf{x}) \leq 0 \text{ in } D. \quad (6.9)$$

Then  $\mathbf{x} = \mathbf{0}$  is stable. Furthermore, if  $\dot{V}(\mathbf{x}) < 0$  in  $D \setminus \{\mathbf{0}\}$  then  $\mathbf{x} = \mathbf{0}$  is asymptotically stable.

*Proof* Let  $\epsilon > 0$  be given. Fix  $r \in (0, \epsilon]$  such that the ball of radius  $r$  centered at the origin is contained entirely within the domain  $D$ , i.e.,

$$B_r = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq r\} \subset D.$$

Let  $\alpha = \min_{\|\mathbf{x}\|=r} V(\mathbf{x})$ , which means  $\alpha > 0$  from Equation (6.8). Choose  $\beta \in (0, \alpha)$  and let

$$\Omega_\beta = \{\mathbf{x} \in B_r : V(\mathbf{x}) \leq \beta\}.$$

It follows that  $\Omega_\beta$  lies within the interior of  $B_r$ . (If it did not, then some point  $\mathbf{p} \in \Omega_\beta$  would lie on the boundary of  $B_r$ ; however, this implies  $V(\mathbf{p}) \geq \alpha > \beta$ , contradicting the definition of  $\Omega_\beta$ .) In addition, any trajectory which begins in  $\Omega_\beta$  at  $t = 0$  stays inside  $\Omega_\beta$  for all times  $t > 0$ . This follows from Equation (6.9) because  $\dot{V}(\mathbf{x}(t)) \leq 0$  implies  $V(\mathbf{x}(t)) \leq V(\mathbf{x}(0)) \leq \beta$  for all  $t \geq 0$ .

Because  $\Omega_\beta$  is contained within  $B_r$ , it is closed and bounded and thus compact. From this, and the fact that  $f$  is Lipschitz-continuous, we know that whenever  $\mathbf{x}(0) \in \Omega_\beta$  that a unique solution exists. We will now show how the assumed

properties of  $V$  enable us to characterize the fixed point  $\mathbf{x} = \mathbf{0}$  as either stable or asymptotically stable.

The continuity of  $V(\mathbf{x})$ , and the fact that  $V(\mathbf{0}) = 0$ , means that there exists a  $\delta > 0$  such that  $\|\mathbf{x}\| \leq \delta$ , which implies  $V(\mathbf{x}) < \beta$ . Let  $B_\delta$  denote the ball centered at  $\mathbf{0}$ , and note that  $B_\delta \subset \Omega_\beta \subset B_r$ . Hence,  $\mathbf{x}(0) \in B_\delta$  implies  $\mathbf{x}(0) \in \Omega_\beta$ , which we have already shown implies  $\mathbf{x}(t) \in \Omega_\beta$ , and so  $\mathbf{x}(t) \in B_r$ . That is, we have shown for all  $t \geq 0$  that

$$\|\mathbf{x}(0)\| < \delta \implies \|\mathbf{x}(t)\| < r \leq \epsilon$$

i.e., the fixed point  $\mathbf{x} = \mathbf{0}$  is stable.

Now suppose that, in addition,  $\dot{V}(\mathbf{x}) < 0$  on the set  $D \setminus \{\mathbf{0}\}$ . In order to establish that the point  $\mathbf{x} = \mathbf{0}$  is asymptotically stable, we must show that  $\mathbf{x}(t) \rightarrow \mathbf{0}$  as  $t \rightarrow \infty$ . That is equivalent to showing for every  $a > 0$  that there is some time  $T > 0$  such that  $\|\mathbf{x}(t)\| < a$  whenever  $t > T$ . From the arguments given above, we know that for every  $a > 0$  we can choose a  $b > 0$  such that  $\Omega_b \subset B_a$ . Given this, it suffices to show that  $V(\mathbf{x}(t)) \rightarrow 0$  as  $t \rightarrow \infty$ . Now,  $V(\mathbf{x}(t))$  decreases monotonically with  $t$  and has a lower bound of 0, so we know  $V(\mathbf{x}(t))$  converges to some  $c \geq 0$  as  $t \rightarrow \infty$ . To establish that  $c = 0$ , let us assume that  $c > 0$  and show that this leads to a contradiction.

From the continuity of  $V(\mathbf{x})$  there is a  $d > 0$  such that  $B_d \subset \Omega_c$ . Because  $V(\mathbf{x})$  is monotonically decreasing, the limit  $V(\mathbf{x}(t)) \rightarrow c > d > 0$  implies that the trajectory  $\mathbf{x}(t)$  lies outside the ball  $B_d$  for all  $t$ . Let

$$-\gamma = \max_{d \leq \|\mathbf{x}\| \leq r} \dot{V}(\mathbf{x}),$$

which exists because  $\dot{V}(\mathbf{x})$  is a continuous function and the set  $\{d \leq \|\mathbf{x}\| \leq r\}$  is compact. Now  $-\gamma < 0$  because  $\dot{V}(\mathbf{x}) < 0$  on  $D \setminus \{\mathbf{0}\}$ . From this, we have

$$V(\mathbf{x}(t)) = V(\mathbf{x}(0)) + \int_0^t \dot{V}(\mathbf{x}(\tau)) d\tau \leq V(\mathbf{x}(0)) - \gamma t.$$

Notice that the right-hand side of the inequality eventually becomes negative for sufficiently large  $t$ , contradicting the assumption that  $V(\mathbf{x}) > 0$  on  $D \setminus \{\mathbf{0}\}$ , and so  $c = 0$ .  $\square$

One potential problem with applying the Lyapunov stability theorem is that it gives no guidance on how one is to find a suitable Lyapunov function  $V(\mathbf{x})$ . This matters because the theorem only provides a *sufficient* condition for stability, not a necessary condition. If one candidate function fails to satisfy conditions (6.8) or (6.9), that does not preclude their being others. However, as Hofbauer and Sigmund (2002) show, it is possible to define a local Lyapunov function for the continuous linear replicator dynamics.

**Definition 6.13** Let  $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_n) \in S^n$  be a population state for the continuous linear replicator dynamics with payoff matrix  $A$ . Then  $\hat{\mathbf{x}}$  is said to be an *evolutionarily stable state* if  $\hat{\mathbf{x}} \cdot A\mathbf{x} > \mathbf{x} \cdot A\mathbf{x}$  for all  $\mathbf{x} \neq \hat{\mathbf{x}}$  in a neighborhood of  $\hat{\mathbf{x}}$ .

**Theorem 6.10** *If  $\hat{\mathbf{x}} \in S^n$  is an ESS for the game with payoff matrix  $A$ , then  $\hat{\mathbf{x}}$  is an asymptotically stable rest point of the continuous linear replicator dynamics.*

*Proof* Let  $\hat{\mathbf{x}} \in S^n$  be an ESS. Following Hofbauer and Sigmund, consider the function defined for  $\mathbf{x} \in S^n$ :

$$P(\mathbf{x}) = \prod_{i=1}^n x_i^{\hat{x}_i}.$$

We now proceed to show that  $P(\mathbf{x})$  has a unique maximum at the point  $\hat{\mathbf{x}}$ , and that the function is a local Lyapunov function for the continuous replicator dynamics.

Consider the function  $f(x) = -\log(x)$ , defined on the nonnegative extended real line. (Following convention, we define  $0 \log(0) = 0 \log(\infty) = 0$ .) Jensen's inequality states that if  $f$  is a strictly convex function defined on some interval, then

$$f\left(\sum p_i x_i\right) \leq \sum p_i f(x_i)$$

for all  $x_1, \dots, x_n$  in the interval and every point  $(p_1, \dots, p_n) \in \text{int } S^n$ , where equality holds if and only if all the  $x_i$  are equal.

Now consider  $\mathbf{x}, \hat{\mathbf{x}} \in S^n$ . It follows that

$$\sum_{i=1}^n \hat{x}_i \log \frac{x_i}{\hat{x}_i} = \sum_{\hat{x}_i > 0} \hat{x}_i \log \frac{x_i}{\hat{x}_i} \leq \log \sum_{\hat{x}_i > 0} x_i \leq \log \sum_{i=1}^n x_i = \log 1 = 0.$$

And so

$$\sum_{i=1}^n \hat{x}_i \log x_i \leq \sum_{i=1}^n \hat{x}_i \log \hat{x}_i,$$

which shows that  $P(\mathbf{x}) \leq P(\hat{\mathbf{x}})$ , with  $P(\mathbf{x}) = P(\hat{\mathbf{x}})$  if and only if  $\mathbf{x} = \hat{\mathbf{x}}$ .

Now then, for all  $\mathbf{x} \in S^n$  where  $x_i > 0$  when  $\hat{x}_i > 0$ , we have

$$\begin{aligned} \frac{dP}{dt} &= \frac{d}{dt} (\log P) = \frac{d}{dt} \left( \sum_{i=1}^n \hat{x}_i \log x_i \right) = \sum_{\hat{x}_i > 0} \hat{x}_i \frac{\dot{x}_i}{x_i} \\ &= \sum_{i=1}^n \hat{x}_i ((A\mathbf{x})_i - \mathbf{x} \cdot A\mathbf{x}) = \hat{\mathbf{x}} \cdot A\mathbf{x} - \mathbf{x} \cdot A\mathbf{x}. \end{aligned}$$

The fact that  $\hat{\mathbf{x}}$  is evolutionarily stable means  $\hat{\mathbf{x}} \cdot A\mathbf{x} > \mathbf{x} \cdot A\mathbf{x}$  for all  $\mathbf{x} \neq \hat{\mathbf{x}}$  in a neighborhood of  $\hat{\mathbf{x}}$ , and so  $\frac{dP}{dt} > 0$  in that neighborhood. Thus,  $P$  is a local Lyapunov function and, from the Lyapunov stability theorem, we know that  $\hat{\mathbf{x}}$  is asymptotically stable.  $\square$

Although the last theorem identified a relationship between evolutionarily stable states and asymptotically stable rest points, we are often interested in knowing the relationship between the *Nash equilibria* of the underlying game and rest points of the replicator dynamics. Suppose that  $\vec{p}(t) = \langle p_1(t), \dots, p_n(t) \rangle$  is a solution to

the replicator dynamics for some payoff matrix  $M$  and initial condition  $\vec{p}(0)$ . The  $\omega$ -limit of  $\vec{p}(t)$ , denoted  $\omega(\vec{p}(t))$ , is the set of its accumulation points. That is,

$$\omega(\vec{p}(t)) = \{\vec{q} \in S^n \mid \vec{p}(t_k) \rightarrow \vec{q} \text{ for some sequence } t_k \rightarrow +\infty\}.$$

The  $\omega$ -limit of a path  $\vec{p}(t)$  is thus simply the set of points on the path all of whose neighborhoods get visited infinitely often. Because the simplex  $S^n$  is a compact set, it follows that  $\omega(\vec{p})$  is nonempty for all  $\vec{p}(t)$  which are solutions to the replicator equations. The two special cases where  $\omega(\vec{p}(t)) = \vec{p}(t)$  occur when  $\vec{p}(t)$  is a fixed point of the replicator dynamics, or a cyclical orbit.

### Theorem 6.11

1. If  $\sigma$  is a Nash equilibrium of a two-player symmetric game, then  $\sigma$  is a rest point of the continuous replicator dynamics.
2. If  $\sigma$  is the  $\omega$ -limit of an orbit in the interior of  $S^n$ , then  $\sigma$  is a Nash equilibrium.
3. If  $\sigma$  is Lyapunov stable, then  $\sigma$  is a Nash equilibrium.

*Proof* (1) Suppose that  $\sigma = (p_1, \dots, p_n)$  is a Nash equilibrium of a two-player symmetric game with strategy set  $S = \{s_1, \dots, s_n\}$ . Let  $c$  denote the value such that, according to the Bishop–Cannings theorem,  $\pi(s_i|\sigma) = c$  for all strategies  $s_i$  with positive support. According to the standard interpretation of mixed strategies, each  $p_i$  represents the probability with which the pure strategy  $s_i$  is played. Now, though, we shall interpret each  $p_i$  as the *proportion* of the population following  $s_i$ . Thus

$$W_i(\sigma) = \sum_{j=1}^n p_i \pi(s_i|s_j) = \pi(s_i|\sigma) = c$$

for all  $i$ , and so

$$\phi(\sigma) = \sum_{i=1}^n \sum_{j=1}^n p_i p_j \pi(s_i|s_j) = \sum_{i=1}^n p_i \pi(s_i|\sigma) = c.$$

Because  $W_i(\sigma) - \phi(\sigma) = 0$  for all  $i$ , so does  $\frac{d\phi}{dt}$ , which shows that the population state corresponding to the mixed strategy  $\sigma$  is a rest point of the continuous replicator dynamics.

(2) Assume that  $\sigma$  is the  $\omega$ -limit of some orbit  $\vec{p}(t)$  in the interior of  $S^n$ . Thus, for every sequence  $\langle t_k \rangle_{k=0}^\infty$  such that  $t_k \rightarrow +\infty$ , we know  $\vec{p}(t_k) \rightarrow \sigma$ . Suppose that  $\sigma$  is not a Nash equilibrium. Then there exists a pure strategy  $s_i$  such that  $\pi(s_i|\sigma) > \pi(\sigma|\sigma)$ . (Why? Because the payoff function for mixed strategies is just a linear combination of the payoff function for pure strategies, weighted by the respective probabilities, if  $\sigma$  is not a best-reply to itself, then there must be such a pure strategy.) In particular, this means that there is some fixed  $\epsilon > 0$  such that  $\pi(s_i|\sigma) - \pi(\sigma|\sigma) > \epsilon$ , which contradicts the claim that  $\sigma$  is the  $\omega$ -limit of the orbit  $\vec{p}(t)$ .

(3) Suppose that  $\sigma$  is Lyapunov stable but that  $\sigma$  is not a Nash equilibrium. From continuity of the payoff function, we know that there is some neighborhood  $N(\epsilon, \sigma)$  of  $\sigma$  such that  $\pi(\mu|\sigma) - \pi(\sigma|\sigma) > \epsilon$  for all  $\mu \in N(\epsilon, \sigma)$ . Because

$\frac{dp_i}{dt} > \epsilon$  at all points in this neighborhood,  $p_i$  increases exponentially, contradicting the assumption that  $\sigma$  was Lyapunov stable.  $\square$

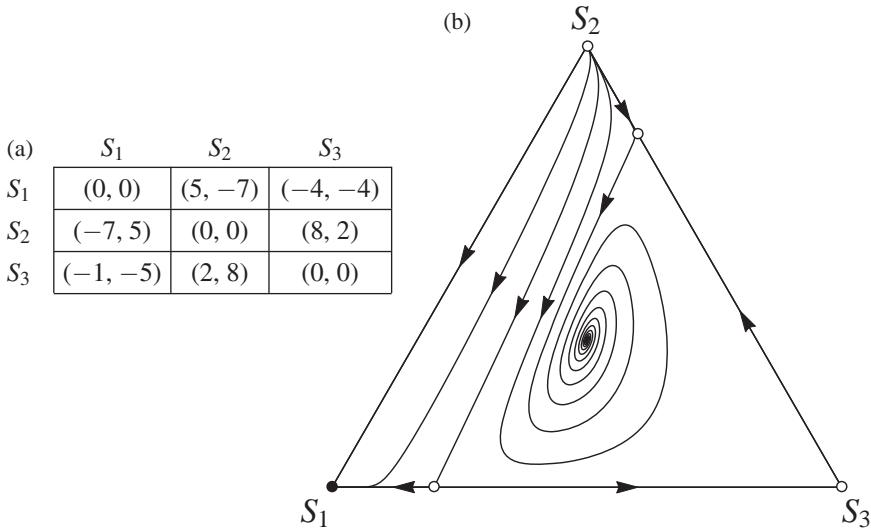
At this point, it is worth making a few remarks about subtleties which exist when relating rest points, stable points, and asymptotically stable points of the replicator dynamics, with Nash equilibria and evolutionarily stable strategies of the underlying game. To begin, notice that the version of Rock–Paper–Scissors described in Figure 6.3b shows that, although  $\sigma = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$  is a Nash equilibrium of the underlying game, the corresponding population state of the continuous replicator dynamics is an *unstable* fixed point; hence claim (1) of Theorem 6.11 cannot be strengthened without additional assumptions. This also serves to show that the converse of claim (3) is false, in general.

Claim (2) of Theorem 6.11 assumes  $\sigma$  to lie in the interior of  $S^n$ . This assumption is essential, for consider any version of Rock–Paper–Scissors and the initial population state  $\sigma_0 = (\frac{1}{10}, \frac{9}{10}, 0)$ . Because the replicator dynamics cannot introduce absent strategies, the  $\omega$ -limit of the trajectory passing through  $\sigma_0$  will be the state  $(1, 0, 0)$ , which – if interpreted as a mixed strategy – is clearly not a Nash equilibrium of the game. In addition, claim (2) only pertains to an  $\omega$ -limit consisting of a *single* state  $\sigma$ ; each orbit around the fixed point  $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$  in Figure 6.2 will be its own  $\omega$ -limit (consisting of a set of states), but none of the points on those orbits correspond to a Nash equilibrium of the game.

Furthermore, claim (2) only says that an asymptotically stable state in the interior of  $S^n$  is a Nash equilibrium. Because we know that every evolutionarily stable *strategy* corresponds to a population state which is, under the replicator dynamics, asymptotically stable, one might wonder whether claim (2) can be strengthened further. The game shown in Figure 6.5, taken from appendix D of Maynard Smith's *Evolution and the Theory of Games*, shows that this is not the case. First, observe that the state  $\mathbf{x} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$  in the interior of  $S^n$  is asymptotically stable (we shall see in a moment how to establish this analytically); however, so is the boundary state  $(1, 0, 0)$ . One can check that  $(1, 0, 0)$ , interpreted as a *strategy*, is an ESS. Because we know from Corollary 6.2 that a completely mixed ESS is unique, it follows that  $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$  cannot be an ESS. In particular, one can verify that the strategy  $\hat{\mathbf{x}} = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$  has the property that  $\hat{\mathbf{x}} \cdot A\mathbf{x} = \mathbf{x} \cdot A\mathbf{x}$  and  $\hat{\mathbf{x}} \cdot A\hat{\mathbf{x}} \geq \mathbf{x} \cdot A\hat{\mathbf{x}}$ .

Although one can show that a fixed point of a dynamical system is asymptotically stable by finding a Lyapunov function, often it is easier to establish this by studying its linearization. Suppose that  $\dot{\mathbf{x}} = f(\mathbf{x})$  is an autonomous nonlinear dynamical system, like the continuous replicator dynamics, with a fixed point at  $\mathbf{x}_0$ . We can write the system of equations more explicitly as the following:

$$\begin{aligned}\frac{dx_1}{dt} &= f_1(x_1, \dots, x_n) \\ \frac{dx_2}{dt} &= f_2(x_1, \dots, x_n) \\ &\vdots \\ \frac{dx_n}{dt} &= f_n(x_1, \dots, x_n).\end{aligned}$$



**Figure 6.5** Game from appendix D of Maynard Smith's Evolution and the Theory of Games. (a) Payoff matrix. (b) Flow trajectories for the replicator dynamics.

Now let  $\frac{\partial(f_1, \dots, f_n)}{\partial(x_1, \dots, x_n)}$  denote the *Jacobian* of the system, where

$$\frac{\partial(f_1, \dots, f_n)}{\partial(x_1, \dots, x_n)} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \dots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

is an  $n \times n$  matrix, with each entry a function of  $x_1, \dots, x_n$ , where this explicit dependency has been suppressed for brevity of notation. If we define  $\mathbf{A} = \left. \frac{\partial(f_1, \dots, f_n)}{\partial(x_1, \dots, x_n)} \right|_{\mathbf{x}=\mathbf{x}_0}$  to be the Jacobian evaluated at the point  $\mathbf{x}_0$ , then the *linearization of the system at  $\mathbf{x}_0$*  is simply

$$\dot{\mathbf{x}} = \mathbf{Ax}, \text{ for all } \mathbf{x} \in \mathbb{R}^n.$$

(In the case of the continuous replicator dynamics, note that the linearization should be restricted to  $\mathbf{x} \in S^n$ , rather than all of  $\mathbb{R}^n$ .)

The reason why we are interested in constructing the linearization of a dynamical system at a fixed point is because, in some cases, the flow trajectories of the linearized system are topologically equivalent to those of the original system in a neighborhood about the fixed point. Hence, if the fixed point of the original system is of the right kind, and if the fixed point of the linearized system is asymptotically stable, then so is the fixed point of the original system. The next few definitions and theorems will make this intuitive description precise.

**Definition 6.14** Let  $\mathbf{x}_0$  be a fixed point of a dynamical system. Then  $\mathbf{x}_0$  is said to be *hyperbolic* if all of the eigenvalues of the Jacobian, evaluated at  $\mathbf{x}_0$ , are nonzero.

The following statement (as well as a proof) of the Hartman–Grobman theorem can be found in Chicone (2000).

**Theorem 6.12** (Hartman–Grobman theorem) *Let  $\mathbf{x}_0$  be a fixed point of the autonomous dynamical system  $\dot{\mathbf{x}} = f(\mathbf{x})$  defined on  $\mathbb{R}^n$ , with trajectory  $\varphi$  and  $\psi$  the trajectory of the linearized system with  $\mathbf{x}_0$  shifted to the origin. If  $\mathbf{x}_0$  is a hyperbolic rest point, then there exists an open set  $U \subset \mathbb{R}^n$  such that  $\mathbf{x}_0 \in U$  and a homeomorphism  $G$  with domain  $U$  such that  $G(\varphi(\mathbf{x})) = \psi(G(\mathbf{x}))$  whenever  $\mathbf{x} \in U$  and both sides of the equation are defined.*

From this, one can prove the following, known as the “method of linearization” or, alternatively, as “Lyapunov’s indirect method.”

**Theorem 6.13** *Let  $\mathbf{x}_0$  be a fixed point of the autonomous dynamical system  $\dot{\mathbf{x}} = f(\mathbf{x})$ , where  $f : D \rightarrow \mathbb{R}^n$  is continuously differentiable and  $D$  is a neighborhood of  $\mathbf{x}_0$ . Let*

$$A = \left. \frac{\partial(f_1, \dots, f_n)}{\partial(x_1, \dots, x_n)} \right|_{\mathbf{x}=\mathbf{x}_0}.$$

Then

1. *The point  $\mathbf{x}_0$  is asymptotically stable if the real part of all the eigenvalues of  $A$  is negative.*
2. *the point  $\mathbf{x}_0$  is unstable if at least one of the eigenvalues of  $A$  is positive.*

*Proof* See Khalil (2002, pp. 139–142).  $\square$

**Example 6.6** As a concrete illustration of the method of linearization, consider the earlier claim that the state  $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$  for the game in Figure 6.5(a) was asymptotically stable. The replicator equations are as follows:

$$\begin{aligned} \frac{dx_1}{dt} &= f_1(x_1, x_2, x_3) = x_1(x_2(5 + 2x_1 - 10x_3) + (-4 + 5x_1)x_3) \\ \frac{dx_2}{dt} &= f_2(x_1, x_2, x_3) = x_2(2(4 - 5x_2)x_3 + x_1(-7 + 2x_2 + 5x_3)) \\ \frac{dx_3}{dt} &= f_3(x_1, x_2, x_3) = x_3(2x_2(1 - 5x_3) + x_1(-1 + 2x_2 + 5x_3)). \end{aligned}$$

And hence

$$\left. \frac{\partial(f_1, f_2, f_3)}{\partial(x_1, x_2, x_3)} \right|_{\mathbf{x}=(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})} = \begin{pmatrix} \frac{7}{9} & \frac{7}{9} & -\frac{17}{9} \\ -\frac{14}{9} & -\frac{8}{9} & \frac{19}{9} \\ \frac{4}{9} & -\frac{2}{9} & -\frac{5}{9} \end{pmatrix}.$$

Because the eigenvalues are  $-\frac{1}{3}$  and  $\frac{1}{6}(-1 \pm 3i\sqrt{7})$ , the real part of each is negative and the state is asymptotically stable.

Let us now turn attention to some issues similar in spirit to those addressed when we sought to characterise evolutionarily stable strategies. Recall that Theorem 6.1 established that no weakly dominated strategy could be an ESS. For the replicator dynamics, the analogous question would be whether strictly or weakly dominated strategies can avoid being eliminated from the population in the limit. The following result shows that, for a strictly dominated pure strategy, the answer is no.

**Theorem 6.14** *Suppose that the pure strategy  $s_i$  is strictly dominated. Then for any initial population state  $\mathbf{x}_0 \in \text{int } S^n$ , it is the case that  $x_i(t) \rightarrow 0$  as  $t \rightarrow \infty$  under the continuous replicator dynamics.*

*Proof* Let  $s_i$  be a pure strategy which is strictly dominated by some mixed strategy  $\sigma = (\sigma_1, \dots, \sigma_n)$ . Define

$$\epsilon = \min_{\mu \in \Delta} (\pi(\sigma|\mu) - \pi(s_i|\mu)).$$

Because the payoff function  $\pi$  is continuous, and the space of mixed strategies is closed and bounded, it follows that  $\epsilon > 0$ . Consider the function  $f_i : \text{int } S^n \rightarrow \mathbb{R}$  defined as follows:

$$f_i(\mathbf{x}) = \log(x_i) - \sum_{j=1}^n \sigma_j \log(x_j).$$

Now, because  $f_i$  is defined on the interior of the simplex  $S^n$ , if  $\mathbf{x}(t)$  denotes the trajectory beginning from some initial point  $\mathbf{x}_0 \in \text{int } S^n$ , we can consider the time derivative of  $f_i$  at any point  $\mathbf{x}(t)$ . This is

$$\begin{aligned} \dot{f}_i(\mathbf{x}) &= \sum_{j=1}^n \frac{\partial f_i(\mathbf{x})}{\partial x_j} \dot{x}_j \\ &= \frac{\dot{x}_i}{x_i} - \sum_{j=1}^n \sigma_j \frac{\dot{x}_j}{x_j} \\ &= \pi(s_i|\mathbf{x}) - \pi(\mathbf{x}|\mathbf{x}) - \sum_{j=1}^n \sigma_j (\pi(s_j|\mathbf{x}) - \pi(\mathbf{x}|\mathbf{x})) \\ &= \pi(s_i|\mathbf{x}) - \pi(\sigma|\mathbf{x}) \leq -\epsilon < 0. \end{aligned}$$

Because this holds for all  $t > 0$ , the value of  $f_i(\mathbf{x})$  converges to  $-\infty$ . The definition of  $f_i$ , and the fact that each  $x_k$  is between 0 and 1, means that  $x_i(t) \rightarrow 0$ .  $\square$

It turns out that this result can be extended to cover all pure strategies which are *iteratively* strictly dominated. The concept of a strategy which is iteratively strictly dominated is straightforward: given a strategy set  $S$  and a payoff matrix  $A$ , first construct the set  $S'$  by eliminating all the strictly dominated strategies from  $S$ . If  $S' = S$ , stop now. However, if  $S' \subset S$ , construct the payoff matrix  $A'$  by eliminating the appropriate rows and columns from  $A$ , and then construct the set  $S''$  by

eliminating from  $S'$  the strategies which are strictly dominated. (There may now be strategies which are strictly dominated which were not eliminated in the first pass.) Continue with this process until a set containing no strictly dominated strategies is obtained.

In a game theory course, one typically shows that iterated elimination of strictly dominated strategies will not throw away Nash equilibria (see, for example, Binmore *et al.*, 1985). Rational players, then, will never use a strictly dominated strategy. The following theorem, proven by Samuelson and Zhang (1992), establishes that the continuous replicator dynamics, likewise, will purge strictly dominated strategies in the limit.

**Theorem 6.15** *If a pure strategy  $s_i$  is iteratively strictly dominated, then for any initial population state  $\mathbf{x}(0) \in \text{int } S^n$ , it is the case that  $x_i(t) \rightarrow 0$  as  $t \rightarrow \infty$  under the continuous replicator dynamics.*

Although the above theorem specifically concerns the replicator dynamics, it is possible to generalise the result. (And, indeed, the Samuelson and Zhang (1992) paper referred to proves a more general result.) However, it is important to note that not all evolutionary dynamics lead to the elimination of strictly dominated strategies in all games. In particular, Sandholm (2010, p. 353) shows how one evolutionary dynamic (the *Smith dynamic*) allows the survival of a strictly dominated strategy in the game known as “Rock–Paper–Scissors with a feeble twin.”<sup>9</sup>

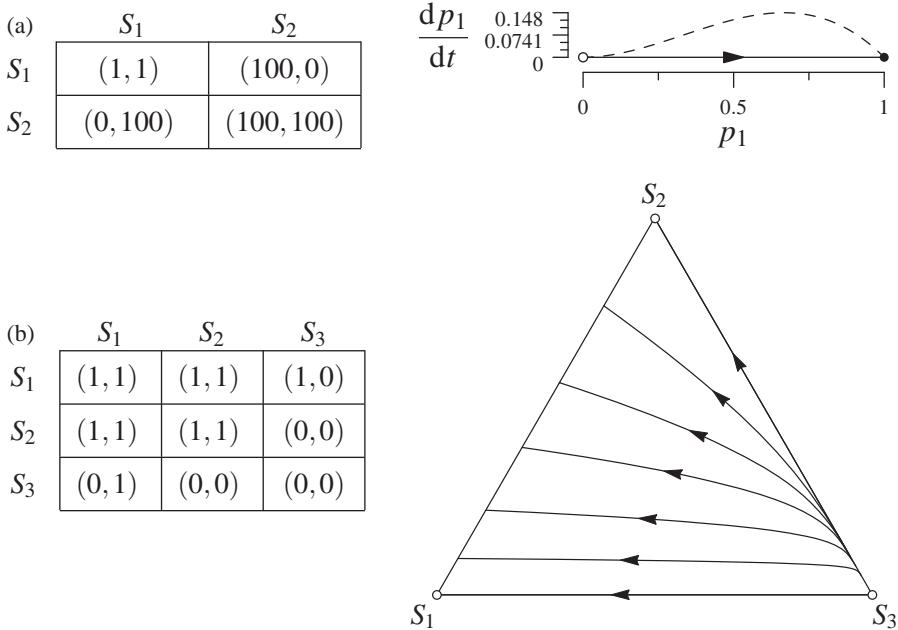
The case of weakly dominated strategies proves rather different. In the game shown in Figure 6.6(a),  $S_2$  is eliminated by the replicator dynamics: it is easy to calculate that  $\frac{dp_1}{dt} = p_1^2(1 - p_1)$ , which is positive for all  $p_1 \in (0, 1)$ . However, Figure 6.6(b) shows that weakly dominated strategies are not always eliminated. Strategy  $S_2$  is weakly dominated by  $S_1$ , but as  $S_2$  is indistinguishable from  $S_1$  in terms of its payoff, when played against either  $S_1$  or  $S_2$ , it remains in the population as  $S_3$  is driven out. In fact, any mix of  $S_1$  and  $S_2$  is a possible stable evolutionary outcome.

**Theorem 6.16** *Let  $s_i$  be a pure strategy which is weakly dominated by some strategy  $\sigma = (\sigma_1, \dots, \sigma_n) \in \Delta$ . If  $s_j$  is a pure strategy such that  $\pi(\sigma|s_j) > \pi(s_i|s_j)$ , then for any  $\mathbf{x}(0) \in \text{int } S^n$ , either  $x_i(t) \rightarrow 0$  or  $x_j(t) \rightarrow 0$  as  $t \rightarrow \infty$ , and possibly both.*

*Proof* Suppose that  $\sigma$  is a strategy weakly dominates the pure strategy  $s_i$ , and let  $s_j$  be another pure strategy with  $\pi(\sigma|s_j) > \pi(s_i|s_j)$ . As before, define  $f_i(\mathbf{x}(t)) : \text{int } S^n \rightarrow \mathbb{R}$  as

$$f_i(\mathbf{x}(t)) = \log(x_i(t)) - \sum_{k=1}^n \sigma_k \log(x_k(t)).$$

<sup>9</sup> This is a four-strategy variant of the normal Rock–Paper–Scissors with an additional strategy identical to Scissors except that all of its payoffs are reduced by  $\epsilon$ .



**Figure 6.6** Weakly dominated strategies are not always driven out by the replicator dynamics. (a) Elimination of a weakly dominated strategy. (b) Preservation of a weakly dominated strategy.

Hence

$$\begin{aligned} \frac{d}{dt} f_i(\mathbf{x}(t)) &= \frac{\dot{x}_i(t)}{x_i(t)} - \sum_{k=1}^n \sigma_k \frac{\dot{x}_k(t)}{x_k(t)} \\ &= \pi(s_i|\mathbf{x}(t)) - \pi(\mathbf{x}(t)|\mathbf{x}(t)) - (\pi(\sigma|\mathbf{x}(t)) - \pi(\mathbf{x}(t)|\mathbf{x}(t))) \\ &= \pi(s_i|\mathbf{x}(t)) - \pi(\sigma|\mathbf{x}(t)) \leq 0 \end{aligned}$$

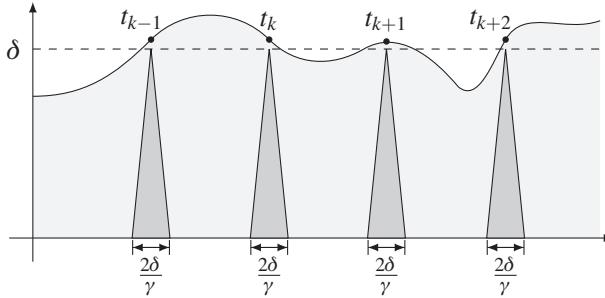
for all times  $t \geq 0$ .

Because  $\pi(\sigma|s_j) > \pi(s_i|s_j)$ , let  $\epsilon = \pi(\sigma|s_j) - \pi(s_i|s_j) > 0$ . Because  $\sigma$  weakly dominates  $s_i$ , it follows that for any population state  $\mathbf{x}(t) \in S^n$

$$\begin{aligned} \pi(\sigma|\mathbf{x}(t)) - \pi(s_i|\mathbf{x}(t)) &= \sum_{k=1}^n x_k(t)(\pi(\sigma|s_k) - \pi(s_i|s_k)) \\ &\geq x_j(t)(\pi(\sigma|s_j) - \pi(s_i|s_j)) \geq \epsilon x_j(t). \end{aligned}$$

(The above makes use of our earlier observation that solutions to the replicator dynamics are invariant under positive affine transformations, so that we may assume that the payoff function is nonnegative.) We have thereby shown that, for all  $t \geq 0$ ,

$$\frac{d}{dt} f_i(\mathbf{x}(t)) \leq -\epsilon x_j(t) \leq 0.$$



**Figure 6.7** The construction used to establish  $x_j(t) \rightarrow 0$  as  $t \rightarrow +\infty$  in the proof of Theorem 6.16.

Integrating  $\frac{d}{dt} f_i(\mathbf{x}(t))$  with respect to  $t$  yields

$$f_i(\mathbf{x}(t)) \leq f_i(\mathbf{x}(0)) - \epsilon \int_0^t x_j(\tau) d\tau.$$

Because  $x_j(t) \geq 0$  for all  $t$ , the integral  $\int_0^t x_j(\tau) d\tau$  either diverges to  $+\infty$  or converges to some real value  $r \in \mathbb{R}$  as  $t \rightarrow +\infty$ . In the first case,  $f_i(\mathbf{x}(t)) \rightarrow -\infty$  and so, from the definition of  $f_i$ , we know that  $x_i(t) \rightarrow 0$ .

In the second case, we will now show that  $x_j(t) \rightarrow 0$  follows from the uniform continuity of the replicator dynamics. First, note that we can place an upper-bound on the absolute value of  $\frac{d}{dt} f_i(\mathbf{x}(t))$ ; letting

$$\gamma = \max_{\mathbf{x} \in S^n} |\pi(s_j|\mathbf{x}) - \pi(\mathbf{x}|\mathbf{x})|,$$

which we know exists from the continuity of the payoff function  $\pi$  and the compactness of  $S^n$ , it then follows that

$$\left| \frac{d}{dt} f_i(\mathbf{x}(t)) \right| \leq \gamma, \text{ for all } t.$$

Now, if  $x_j(t)$  does *not* converge to 0 in the limit, then there must be a  $\delta > 0$  and an increasing, unbounded sequence of times  $\langle t_k \rangle_{k=1}^\infty$  such that  $x_j(t_k) > \delta$ , for all  $t_k$ . Note that we may assume that the distance between successive times in the sequence  $t_{k+1} - t_k$  exceeds  $\frac{2\delta}{\gamma}$ . As  $\gamma$  denotes the maximum rate of change of  $f_i(\mathbf{x}(t))$ , and  $x_j(t)$  is (by assumption) positive for all  $t$ , it must be the case that  $\int_0^t x_j(\tau) d\tau$  is greater than the sum of an infinite number of triangles, each of width  $\frac{2\delta}{\gamma}$  and height  $\delta$  (see Figure 6.7). This contradicts the assumption that  $\int_0^t x_j(\tau) d\tau \rightarrow r$ , and hence  $x_j(t) \rightarrow 0$  as  $t \rightarrow +\infty$ .  $\square$

Thus far we have identified a number of interesting properties of the replicator dynamics: an evolutionarily stable strategy corresponds to an asymptotically stable population state, strictly dominated strategies are eliminated, and we have identified a property which holds when weakly dominated strategies are preserved. In addition, it is possible to recapture a result, originally due to Fisher (1930),

known as the “fundamental theorem of natural selection.” This theorem says that, in some cases, selection pressure generates a monotonic increase in the average fitness of the population. The behaviour of the replicator dynamics in the case of the Prisoner’s Dilemma, which we saw in example 6.5, shows that the fundamental theorem is not true, in general. (A population starting with almost all Cooperators has a much higher fitness than the limiting population of all Defectors.)

**Definition 6.15** A game is said to be *doubly symmetric* if, for every pair of pure strategies  $s_i$  and  $s_j$ , it is the case that  $\pi(s_i|s_j) = \pi(s_j|s_i)$ .

**Theorem 6.17** Let  $\phi(\vec{p})$  denote the average fitness of the population for the continuous replicator dynamics. Then, for any doubly symmetric game,  $\frac{d\phi(\vec{p})}{dt} \geq 0$ .

*Proof* Let  $\vec{p} = (p_1, \dots, p_n)$  denote the current state of the population (with the time index suppressed), and assume that the underlying game is doubly symmetric. For ease of notation, we shall use  $w_{ij}$  to denote the payoff  $\pi(s_i|s_j)$ , and we shall suppress the dependency on  $\vec{p}$  when writing  $\phi$ , the average fitness of the population, and  $W_i$ , the fitness of strategy  $s_i$ .

For the continuous linear replicator dynamics,

$$\phi = \sum_{i=1}^n \sum_{j=1}^n p_i p_j w_{ij}$$

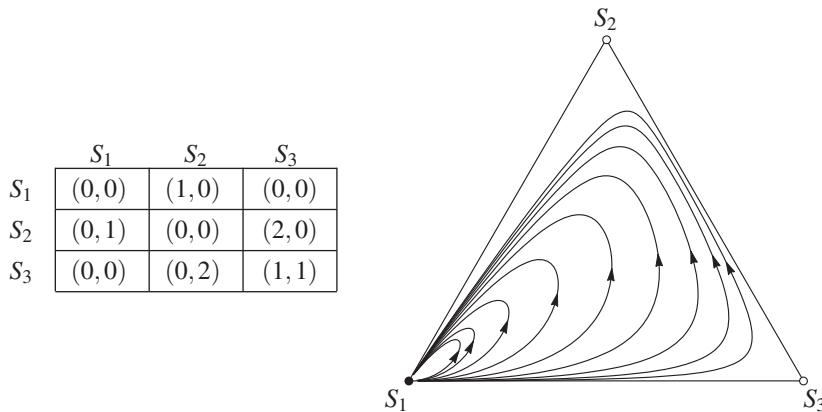
and so

$$\begin{aligned} \frac{d\phi}{dt} &= \sum_{i=1}^n \sum_{j=1}^n \dot{p}_i p_j w_{ij} + \sum_{i=1}^n \sum_{j=1}^n p_i \dot{p}_j w_{ij} \\ &= \left( \begin{array}{c} \dot{p}_1 p_1 w_{11} + \dot{p}_1 p_2 w_{12} + \cdots + \dot{p}_1 p_n w_{1n} \\ + \dot{p}_2 p_1 w_{21} + \dot{p}_2 p_2 w_{22} + \cdots + \dot{p}_2 p_n w_{2n} \\ \vdots \\ + \dot{p}_n p_1 w_{n1} + \dot{p}_n p_2 w_{n2} + \cdots + \dot{p}_n p_n w_{nn} \end{array} \right) + \left( \begin{array}{c} p_1 \dot{p}_1 w_{11} + p_1 \dot{p}_2 w_{12} + \cdots + p_1 \dot{p}_n w_{1n} \\ + p_2 \dot{p}_1 w_{21} + p_2 \dot{p}_2 w_{22} + \cdots + p_2 \dot{p}_n w_{2n} \\ \vdots \\ + p_n \dot{p}_1 w_{n1} + p_n \dot{p}_2 w_{n2} + \cdots + p_n \dot{p}_n w_{nn} \end{array} \right) \\ &= 2 \sum_{i=1}^n \sum_{j=1}^n \dot{p}_i p_j w_{ij} \end{aligned}$$

because  $w_{ij} = w_{ji}$ , as the game is doubly symmetric. Because  $\sum_{i=1}^n p_i(W_i - \phi)\phi = 0$  (because  $\phi = \sum_{i=1}^n p_i W_i$ ), it then follows that

$$\begin{aligned} \frac{d\phi}{dt} &= 2 \sum_{i=1}^n \sum_{j=1}^n p_i(W_i - \phi)p_j w_{ij} = 2 \sum_{i=1}^n p_i(W_i - \phi)W_i \\ &= 2 \sum_{i=1}^n p_i(W_i - \phi)W_i + \sum_{i=1}^n p_i(W_i - \phi)\phi \\ &= 2 \sum_{i=1}^n p_i(W_i - \phi)^2. \end{aligned}$$

□



**Figure 6.8** A population state can be a global attractor in the interior of the simplex without it being asymptotically stable.

One consequence of this result is that, for doubly symmetric games, the population state corresponding to an evolutionarily stable strategy is asymptotically stable under the replicator dynamics.

**Theorem 6.18** *For any doubly symmetric game, the following two statements are equivalent*

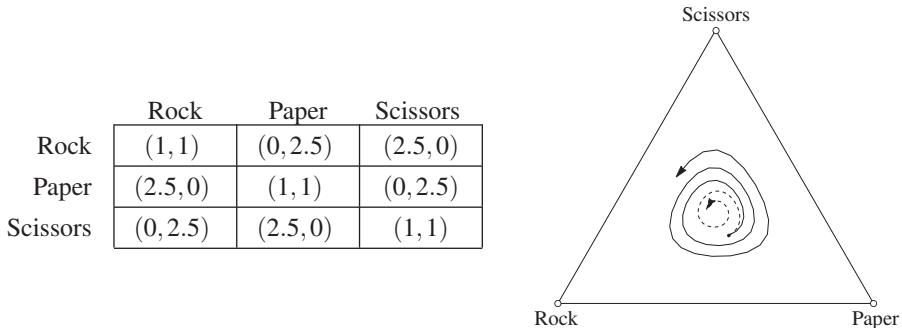
1.  $\sigma$  is an evolutionarily stable strategy.
2. The population state  $\sigma$  is asymptotically stable under the replicator dynamics.

*Proof* We have already shown that (1) implies (2) in Theorem 6.10. To see that (2) implies (1), note that if the state  $\sigma$  is asymptotically stable, then there is a neighborhood  $U$  around  $\sigma$  such that, for all initial states  $\mathbf{x}(0) \in U$ ,  $\mathbf{x}(t) \rightarrow \sigma$  as  $t \rightarrow +\infty$ . From the fundamental theorem, interpreting the population states as mixed strategies, it follows that for all  $\mu \in U$  that  $\pi(\mu|\mu) < \pi(\sigma|\sigma)$ , and hence  $\sigma$  is an evolutionarily stable strategy.  $\square$

Before we move on to consider other dynamical models, it is worth noting that, under the replicator dynamics, a population state may have the entire interior of the simplex as its basin of attraction without being asymptotically stable, much less being an evolutionarily stable state. To see this, consider the game shown in Figure 6.8. The state where everyone follows  $S_1$  has the interior of  $S^3$  as its basin of attraction, yet it is obviously not asymptotically stable: any interior  $\epsilon$ -displacement of the population away from the all- $S_1$  point will follow a trajectory away from all- $S_1$  before ultimately returning. This shows an attracting point of the replicator dynamics may not even be Lyapunov stable.

#### 6.4.1.1 Discrete replicator dynamics

Because real populations are, of course, discrete, let us now briefly consider a discretized version of the replicator dynamics. As before,  $p_i(t)$  denotes the proportion



**Figure 6.9** The discrete replicator dynamics, for two different values of the background fitness. Both trajectories were calculated using the Rock–Paper–Scissors payoff matrix to the left, starting from the common point  $(\frac{1}{3}, \frac{5}{12}, \frac{1}{4})$ ; however, the solid line set  $\alpha = 1$  and the dashed line set  $\alpha = 20$ .

of the population following  $s_i$  at time  $t$ , but where  $t$  is restricted to a nonnegative integer. The discrete replicator dynamics are given by the difference equation

$$p_i(t+1) = p_i(t) \cdot \frac{\alpha + \pi(s_i | \vec{p}(t))}{\alpha + \pi(\vec{p}(t) | \vec{p}(t))}$$

where  $\alpha \geq 0$  denotes the background fitness common to all individuals regardless of what strategy they follow. Increasing the value of  $\alpha$  reduces the influence of the payoff contribution, giving a discretization which more closely approximates the continuous version.

It is important to note that trajectories obtained from the discrete replicator dynamics can differ substantially from the continuous version. Figure 6.9 illustrates two cases generated from the same payoff matrix and initial condition, varying only the value of the background fitness. The payoff matrix is the version of the Rock–Paper–Scissors game from Figure 6.3(a), where we showed that the state  $\sigma = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$  was asymptotically stable under the continuous replicator dynamics. In the discrete case, with a background fitness of 1 the state  $\sigma$  is unstable, with the population evolving towards the boundary of the simplex; with a background fitness of 10, the population spirals in towards the state  $\sigma$ .

#### 6.4.1.2 Multipopulation models

Up to now, we have always assumed that the underlying game has been symmetric; that is, it made no difference to the payoff received whether you were assigned to be the Row or Column player. This was why we could use  $\pi(s_i | s_j)$  to denote the payoff of strategy  $s_i$  played against strategy  $s_j$ . With a game like Battle of the Sexes, this is no longer the case. In some cases, it might be reasonable to assume that players are assigned Row or Column at random, say via a coin toss. But if this were done with Battle of the Sexes, the game would collapse into a standard pure

coordination game, with the two pure strategy Nash equilibria each rewarding the players payoffs of 3/2.

		Column	
		$s_1$	$s_2$
Row	$s_1$	(1, 2)	(0, 0)
	$s_2$	(0, 0)	(2, 1)

Battle of the sexes

Alternatively, we could model this game as occurring between two interacting *populations* of individuals: one population corresponding to Row, the other population corresponding to Column. Every individual in the Row population interacts with one person from the Column population selected at random. Let  $\mathbf{x} = (x_1, x_2)$  represent the frequencies of strategies  $s_1$  and  $s_2$  in the Row population, and let  $\mathbf{y} = (y_1, y_2)$  be the corresponding frequencies for the Column population, where  $x_2 = 1 - x_1$  and  $y_2 = 1 - y_1$ . Then

$$\frac{dx_1}{dt} = x_1 (W_R^{s_1}(\mathbf{x}, \mathbf{y}) - \phi_R(\mathbf{x}, \mathbf{y})) \quad \frac{dy_1}{dt} = y_1 (W_C^{s_1}(\mathbf{x}, \mathbf{y}) - \phi_C(\mathbf{x}, \mathbf{y}))$$

where  $W_R^{s_1}(\mathbf{x}, \mathbf{y})$  is the fitness of  $s_1$  in the Row population, given the current population states  $\mathbf{x}$  and  $\mathbf{y}$ , with  $\phi_R(\mathbf{x}, \mathbf{y})$  being the average fitness of the Row population. The notation for the Column population is analogous.

Filling in some of the details, we have

$$\begin{aligned} W_R^{s_1}(\mathbf{x}, \mathbf{y}) &= y_1 \cdot \pi_R(s_1|s_1) + y_2 \cdot \pi_R(s_1|s_2) = y_1 \\ W_R^{s_2}(\mathbf{x}, \mathbf{y}) &= y_1 \cdot \pi_R(s_2|s_1) + y_2 \cdot \pi_R(s_2|s_2) = 2(1 - y_1) \\ \phi_R(\mathbf{x}, \mathbf{y}) &= x_1 y_1 + 2(1 - x_1)(1 - y_1) \end{aligned}$$

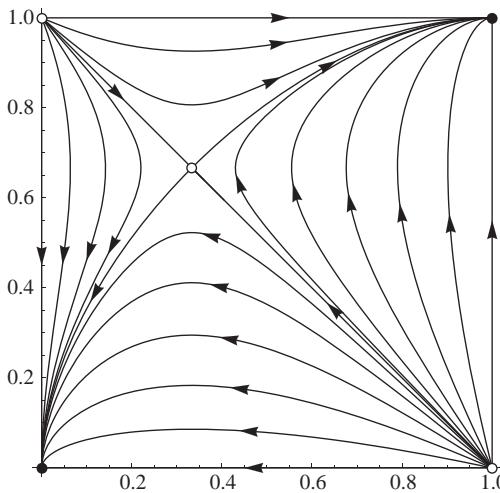
where the payoff function has a subscript to indicate whether it is the payoff for the Row or Column population. And so

$$\frac{dx_1}{dt} = (1 - x_1)x_1(3y_1 - 2) \quad \frac{dy_1}{dt} = (1 - 3x_1)(y_1 - 1)y_1.$$

As Figure 6.10 shows, the state space for the two-population Battle of the Sexes is the unit square, rather than the simplex. The original Battle of the Sexes game had three Nash equilibria: two in pure strategies, and one in mixed strategies. The population states corresponding to the mixed-strategy Nash equilibrium is an unstable saddle point, with the population states corresponding to the pure strategy Nash equilibria being asymptotically stable.

### 6.4.2 Replicator-mutator dynamics

**Definition 6.16** Assuming the general framework of the replicator dynamics, let  $Q = [q_{ij}]$  denote a *transition matrix* where  $q_{ij}$  is the probability of an individual



**Figure 6.10** The replicator dynamics for the two-population Battle of the Sexes.

following the pure strategy  $s_i$  mutating into one following  $s_j$ . (We interpret  $q_{ii}$  as the probability of an individual following  $s_i$  not mutating, and so  $q_{ii} = 1 - \sum_{j \neq i} q_{ij}$ .) The replicator–mutator dynamics states that the rate of change of  $p_i$  is given by

$$\frac{dp_i}{dt} = \sum_{j=1}^n p_j W_j(\vec{p}) q_{ji} - p_i \phi(\vec{p}).$$

Note that if  $q_{ij} = 0$  for  $j \neq i$ , then  $q_{ii} = 1$  for all  $i$ , and we obtain the continuous replicator dynamics as the limit outcome in the presence of no mutations.

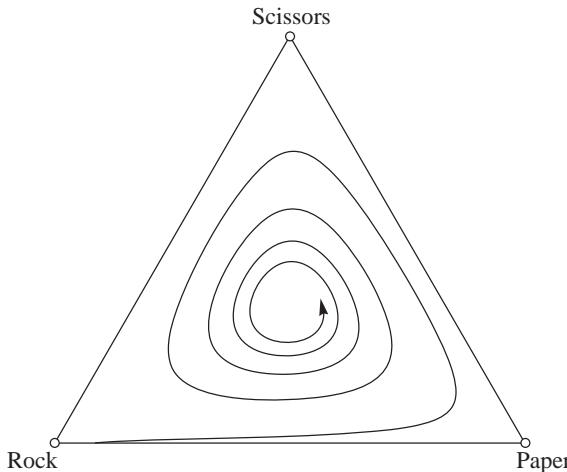
**Example 6.7** Recall the game of Rock–Paper–Scissors with the original payoff matrix from Example 6.2. We saw in Figure 6.2 that replicator dynamics gives rise to cyclical orbits around the state  $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ , and it is only when one modifies the payoffs as in Figure 6.3(a) that we get convergence to the population state corresponding to the mixed strategy Nash equilibrium.

Suppose that the transition matrix  $Q$  is defined as follows

$$Q = \begin{pmatrix} 1 - 2\epsilon & \epsilon & \epsilon \\ \epsilon & 1 - 2\epsilon & \epsilon \\ \epsilon & \epsilon & 1 - 2\epsilon \end{pmatrix}$$

for some small  $\epsilon$ . Then the replicator–mutator equation for the first strategy is

$$\begin{aligned} \frac{dp_1}{dt} = & p_1(p_1 + 2p_2)(1 - 2\epsilon) + p_3(2p_1 + p_3)\epsilon \\ & + p_2(p_2 + 2p_3)\epsilon - p_1(p_1 + p_2 + p_3)^2 \end{aligned}$$



**Figure 6.11** The game of Rock–Paper–Scissors evolving under the replicator–mutator dynamics.

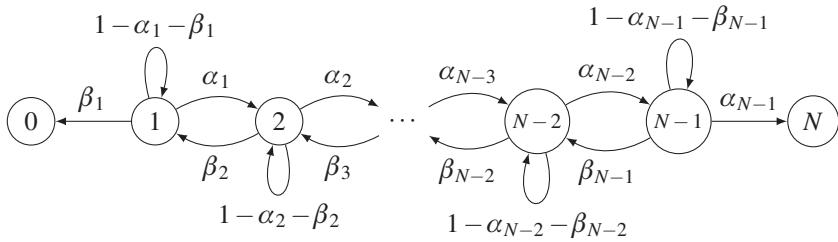
with corresponding equations for the other two strategies obtained in the natural way. Setting the mutation rate  $\epsilon = 0.01$  and selecting an initial population state  $\sigma = (0.914, .001, 0.085)$ , we obtain the solution shown in Figure 6.11. Note that, in contrast to the replicator dynamics, under the replicator–mutator dynamics the population evolves towards the state  $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ .

### 6.4.3 Finite population models

Both the replicator dynamics and the replicator–mutator dynamics assume a continuous population. When modeling a finite population, several questions need to be determined at the outset. First, is the population capable of increasing or decreasing in size over time? Second, how do interactions between individuals occur, and how are the fitness values calculated? If the population size is constant, a single player reproduces at a time (replacing another randomly selected player, who is eliminated) and the probability that a player of a fixed type reproduces is a function of the overall population state, the resulting model is that of a general birth–death process.

Figure 6.12 illustrates the Markov chain for a general birth–death process on a population of  $N$  individuals with two types,  $A$  and  $B$ . Let  $\alpha_i$  denote the probability of adding another  $A$  type to the population when  $i$  individuals of type  $A$  are present. Similarly, let  $\beta_i$  denote the probability of adding another  $B$  type to the population when  $N - i$  individuals of type  $B$  are present. We assume that once a type is no longer present in the population that the birth–death process has reached a fixed state, so there are two absorbing states to the Markov process.

Nowak (2006) provides an elegant analysis of the general birth–death process, which is reproduced below. To begin, let  $x_i$  denote the probability of reaching the



**Figure 6.12** The state diagram for the Markov chain underlying a general birth-death process.

state where all  $N$  individuals are of type  $A$  from the state containing  $i$  individuals of type  $A$ . Then

$$\begin{aligned}x_0 &= 0 \\x_i &= \beta_i x_{i-1} + (1 - \alpha_i - \beta_i) x_i + \alpha_i x_{i+1} \\x_n &= 1.\end{aligned}$$

Let  $y_i = x_i - x_{i-1}$ , for  $i = 1, \dots, N$ , and note that  $y_1 + \dots + y_N = x_N - x_0 = 1$ . In addition, let  $\gamma_i = \frac{\beta_i}{\alpha_i}$ , which allows us to derive the following expression:

$$\begin{aligned}x_i &= \beta_i x_{i-1} + (1 - \alpha_i - \beta_i) x_i + \alpha_i x_{i+1} \\0 &= -\beta_i(x_i - x_{i-1}) + \alpha_i(x_{i+1} - x_i) \\y_{i+1} &= \gamma_i y_i.\end{aligned}$$

Thus we see that  $y_1 = x_1$ ,  $y_2 = \gamma_1 x_1$ ,  $y_3 = \gamma_1 \gamma_2 x_2$ , and so on. We are then able to solve for  $x_1$  as follows:

$$\begin{aligned}1 &= y_1 + y_2 + y_3 + \dots + y_N \\1 &= x_1 + \gamma_1 x_1 + \gamma_1 \gamma_2 x_2 + \dots + \left( \prod_{k=1}^{N-1} \gamma_k \right) x_1 \\x_1 &= \frac{1}{1 + \sum_{j=1}^{N-1} \prod_{k=1}^j \gamma_k}.\end{aligned}\tag{6.10}$$

Finally, using the fact that  $y_i = x_i - x_{i-1}$ , we can readily solve for the general value of  $x_i$  as follows:

$$\begin{aligned}y_2 &= x_2 - x_1 = \gamma_1 x_1 \\x_2 &= \gamma_1 x_1 + x_1 = x_1(1 + \gamma_1)\end{aligned}$$

and

$$\begin{aligned}y_3 &= x_3 - x_2 = \gamma_1 \gamma_2 x_1 \\x_3 &= \gamma_1 \gamma_2 x_1 + x_2 \\&= x_1(1 + \gamma_1 + \gamma_1 \gamma_2)\end{aligned}$$

and so, in general,

$$x_i = \frac{1 + \sum_{j=1}^{i-1} \prod_{k=1}^j \gamma_k}{1 + \sum_{j=1}^{N-1} \prod_{k=1}^j \gamma_k}. \quad (6.11)$$

Now consider a generic symmetric two-strategy game, as below:

	A	B
A	$(a, a)$	$(b, c)$
B	$(c, b)$	$(d, d)$

Suppose that individuals play a game with a person selected from the population at random. If there are  $i$  players following strategy  $A$  and  $N - i$  players following strategy  $B$ , the expected payoff (fitness) of  $A$  is  $\alpha_i = \frac{a(i-1)+b(N-i)}{N-1}$  and the expected payoff (fitness) of  $B$  is  $\beta_i = \frac{c(i-1)+d(N-i-1)}{N-1}$ . If a player is selected to reproduce with probability proportional to his fitness, then the probability of a type  $A$  player being selected is  $\frac{i\alpha_i}{i\alpha_i + (N-i)\beta_i}$  and the probability of a type  $B$  player being selected is  $\frac{(N-i)\beta_i}{i\alpha_i + (N-i)\beta_i}$ . If the individual replaced is selected at random, then the transition probabilities for the birth-death process are

$$\alpha_i = \frac{i\alpha_i}{i\alpha_i + (N-i)\beta_i} \cdot \frac{N-i}{N}$$

and

$$\beta_i = \frac{(N-i)\beta_i}{i\alpha_i + (N-i)\beta_i} \cdot \frac{i}{N}.$$

In a pure population of type  $B$  individuals where a single  $A$ -mutant occurs, the probability that the  $A$ -mutant will spread to take over the population (also known as the fixation probability  $\rho_A$ ) is given by Equation (6.10). The ratio of the transition probabilities,  $\gamma_i = \beta_i/\alpha_i$ , simplifies to  $\beta_i/\alpha_i$ , and so

$$\rho_A = \frac{1}{1 + \sum_{j=1}^{N-1} \prod_{i=1}^j \frac{\beta_i}{\alpha_i}}.$$

Likewise, the fixation probability  $\rho_B$  for a single  $B$ -mutant in a pure population of type  $A$  individuals is

$$\begin{aligned} \rho_B &= 1 - x_{N-1} \\ &= \frac{\prod_{i=1}^{N-1} \frac{\alpha_i}{\beta_i}}{1 + \sum_{j=1}^{N-1} \prod_{i=1}^j \frac{\alpha_i}{\beta_i}}. \end{aligned}$$

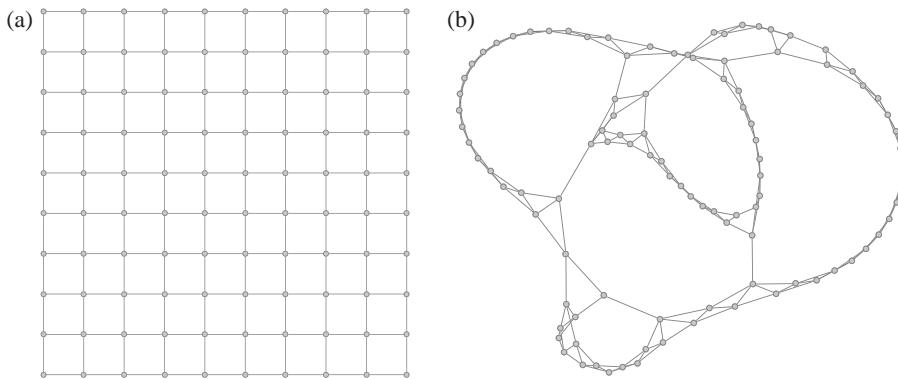
Because the ratio  $\frac{\beta_i}{\alpha_i} = \frac{c(i-1)+d(N-i-1)}{a(i-1)+b(N-i)}$ , one must be slightly careful when choosing payoffs in order to ensure that the denominator does not ever equal zero. For example, the version of the Prisoner's Dilemma discussed in Example 6.5 has  $b = 0$ , which means that  $\beta_i/\alpha_i$  is undefined when  $i = 1$ . If we simply boost all of the

payoffs by 1, so that  $a = 3, b = 1, c = 4$  and  $d = 2$ , we may then calculate the fixation probabilities for both Cooperate and Defect in populations of various sizes. When  $N = 10$ , the chance that a single cooperator can successfully invade a population of defectors is approximately 0.285%, whereas a single defector has about a 34% chance of successfully invading. When  $N = 100$ , a cooperating mutant has virtually no chance of taking over, whereas a defecting mutant has approximately a 25.9% chance. Finally, because one can show that the limit of the fixation probability for Defect as  $N \rightarrow \infty$  is 1/4, the continuous replicator dynamics cannot be thought of as the limit case of the birth–death process.

## 6.5 Local interaction models

None of the dynamics considered thus far are capable of modelling populations with any significant structure. (The best one can do is to allow for positive and negative correlation between strategies, or interaction across separate subpopulations.) As we will see, incorporating population structure into evolutionary games can significantly change the evolutionary outcomes. This is important for several reasons. First, human societies are structured, in that people are much more likely to interact with certain individuals than others; this can influence the outcomes of repeated play in games. Second, in the biological context, there are a number of cases where *spatial* structure affects the evolutionary outcome. For example, there exists one variety of *Escherichia coli* bacteria which produces a poison to which it is immune, but which will kill other strains of *E. coli*. Because producing the poison is costly, the poison-producing strain has a slightly lower fitness than the nonpoison-producing strain. In a large population consisting entirely of nonpoisoners, a small proportion of mutant poisoners would have lower than average fitness and, according to the replicator dynamics, would be predicted to be driven out. In laboratory experiments where a few poisoners are added to a well-stirred solution of nonpoisoners (thereby satisfying the replicator dynamic assumption that all pairwise interactions are equiprobable), this is exactly what one finds. However, when the experiment is performed on plates of agar, one finds that the poisoners are capable of invading and eventually taking over the population. (For further discussion, see Chao and Levin, 1981, and Durrett and Levin, 1997.)

One way to incorporate structure into evolutionary games is as follows. Let  $\Gamma$  be a two-player symmetric game where both players share a common strategy set of pure strategies  $S = \{s_1, \dots, s_n\}$ . Let  $G = (V; E_i, E_u)$  be a graph with vertex set  $V$  and two sets of undirected edges  $E_i$  and  $E_u$ . The set of vertices represents the players in the game and the edge sets represent the possible pairwise interactions between players; more precisely,  $E_i$  is the set of edges specifying whom individuals play the game with, and  $E_u$  is the set of edges specifying the individuals examined when determining whether to update one's strategy. It is not uncommon to assume that  $E_i = E_u$ , but as we shall see below, allowing these two sets to differ can have



**Figure 6.13** Two connected graphs representing different interaction structures. (a) A  $10 \times 10$  grid graph. (b) A random graph with 100 vertices generated using the Watts–Strogatz distribution with  $p = 0.025$ .

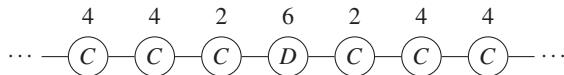
important consequences. In general, we assume that the graph  $G$  is connected, as otherwise we have several isolated populations within a single model. Figure 6.13 illustrates two possible local interaction models: one representing spatial location, and another one generated using the Watts–Strogatz model of “small world” networks (Watts and Strogatz, 1998).

There are a number of possible evolutionary dynamics one may consider. To begin, let us assume that the dynamics take place in discrete time as follows: each iteration, every player  $v \in V$  interacts with those players he is connected to by an edge in  $E_i$  (this is known as the *interaction neighborhood* of  $v$ ), receiving a payoff which is simply the sum of the payoffs from the individual games. After the interaction phase, each player  $v$  modifies his strategy according to the following rule:  $v$  compares his payoff to the payoff received by his highest-scoring neighbor in  $E_u$  (this is the *update neighborhood* of  $v$ ). If the payoff received by  $v$ 's highest-scoring neighbor is strictly greater than  $v$ 's payoff, then  $v$  adopts the strategy used by his highest-scoring neighbor. (If there is more than one such neighbor, then the tie is broken by a coin flip.) At this point, payoffs are reset to zero and a new interaction phase begins. This particular model is known as *imitate-the-best*.

**Example 6.8** Consider the Prisoners’ Dilemma as defined in Example 6.5, and let  $G$  be a simple cycle graph of  $N$  vertices, with  $E_i = E_u$ . If the population consists of an initial random mix of cooperators and defectors, what is the evolutionary outcome under imitate-the-best? In this case, it suffices to consider what happens when a single defector invades a population of cooperators, which looks like the following:



After each player has interacted with his neighbors, the payoffs will be:



And so, under imitate-the-best, the center defector's strategy spreads to his left and right neighbor:



The next round of interactions is what enables us to determine the evolutionary outcome. Note that the defectors on the left and right each receive a payoff of 4, because they interact with one cooperator and one defector. Their cooperating neighbors, on the boundary of the region of defectors and the region of cooperators, each receive a payoff of 2, an inferior score, and hence will modify their strategy. What strategy will these boundary cooperators adopt? Notice that each boundary cooperator is himself connected to another cooperator who received a payoff of 4, and so the tie-breaking rule of imitate-the-best is invoked. The boundary cooperators will thus randomly adopt the strategy of cooperator or defect with equal probability. However, this means that the region of defectors will never shrink, and will eventually spread across the entire cycle.  $\square$

**Example 6.9** One interesting property of local interaction models is that different evolutionary outcomes can occur even when the basic strategic properties of the underlying game remains the same. Consider the following three versions of the Prisoners' Dilemma, each of which has defect as the strongly dominant strategy (payoffs listed for row, as the game is symmetric):

	C	D
C	1	-0.1
D	1.1	0

(a)

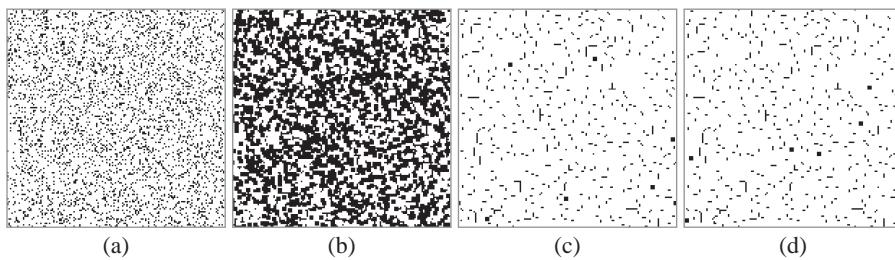
	C	D
C	1	-0.1
D	1.6	0

(b)

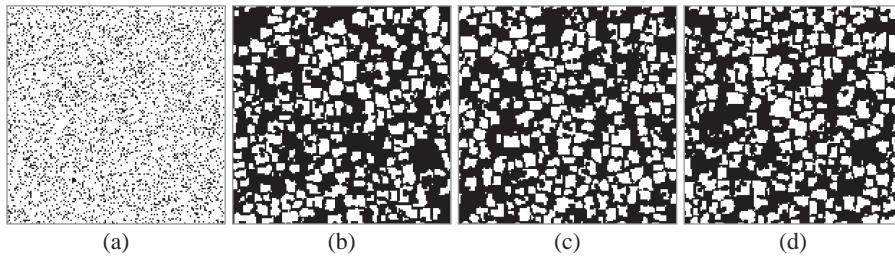
	C	D
C	1	-0.1
D	2.7	0

(c)

Now suppose that the Prisoners' Dilemma is played on a two-dimensional square lattice, where each player interacts with his eight nearest neighbors. Assume that the lattice wraps at the edges, making it topologically equivalent to a torus, which ensures that all players have an equal number of neighbors. Figure 6.14 illustrates a typical outcome, where cooperators are colored white and defectors are colored black. (The boundary of the lattice is drawn in gray so as to indicate the extent of the world.) Even though defect is the only Nash equilibrium of the game, and cooperate is a strongly dominated strategy, cooperators not only persist, but manage to drive out most of the defectors who appear in the second iteration.



**Figure 6.14** *The Prisoners' Dilemma played on a lattice, using payoff matrix (a). (a) Initial state. (b) Iteration 2. (c) Iteration 19. (d) Iteration 20.*



**Figure 6.15** *The Prisoners' Dilemma played on a lattice, using payoff matrix (b). (a) Initial state. (b) Iteration 100. (c) Iteration 200. (d) Iteration 300.*

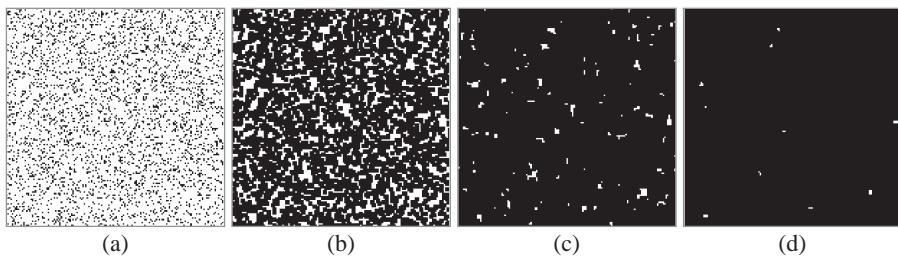
Comparing the outcome of Figure 6.14 with Figure 6.15, we see that the particular form of the payoff matrix makes an enormous difference. When the payoff earned by a defector when paired with a cooperator is increased from 1.1 to 1.6, the typical evolutionary outcome is a “chaotic” mix of defectors and cooperators. Depending on their relative spatial positioning, regions of cooperators can invade regions of defectors, and vice versa. This pattern of mutual invasion plays out across the lattice in a disorderly fashion that does not seem to settle down in any reasonable length of time. This fact was originally discovered by Nowak and May (1992) in the context of a slightly different game.<sup>10</sup>

Finally, Figure 6.16 shows that, when the payoff received by defectors is increased further, defectors manage to drive out cooperators after only a very few number of iterations. The point to note, by way of contrast, is that for all three payoff matrices the replicator dynamics converges to defect.  $\square$

10 In their original paper, Nowak and May consider the following game (payoffs listed only for row, as the game is symmetric):

	<i>C</i>	<i>D</i>
<i>C</i>	1	0
<i>D</i>	1.9	0

which they describe as the “prisoners’ dilemma”. That is not, strictly speaking, correct, because in that payoff matrix Defect is only weakly, not strongly, dominant. In addition, their implementation of the evolutionary dynamics has the peculiar feature that each person plays the game *with themselves*, which gives all cooperators a payoff boost of +1.



**Figure 6.16** *The Prisoners’ Dilemma played on a lattice, using payoff matrix (c). (a) Initial state. (b) Iteration 1. (c) Iteration 2. (d) Iteration 3.*

Much more could be said about how introducing structure affects the dynamics of evolutionary games. Those interested in further reading from a biological point of view should consult Nowak (2006), whereas those interested in an economic perspective would be well advised to read Jackson (2008). Applications of local interaction models to philosophical issues can be found in both Skyrms (2003) and Alexander (2007).

## 6.6 Further reading

Those interested in a short history of the origins of evolutionary game theory should consult Frank (1995), which discusses George Price’s influence on the work of John Maynard Smith in the early 1970s. (A lengthier discussion of the “evolutionary turn” in game theory can be found in Sugden, 2001.) The seminal paper of Maynard Smith and Price (1973) first introduced the concept of an evolutionarily stable strategy, and curiously features an analysis of the “Hawk–Mouse” game rather than the Hawk–Dove game.<sup>11</sup> As the literature on evolutionary games is vast, those interested in exploring particular topics are advised to consult the extensive references found in the books discussed below.

Maynard Smith (1982) provides an accessible introduction to evolutionary game theory, but is less mathematically sophisticated than more recent work. Weibull (1995) offers a more mathematically detailed discussion of evolutionary games, and some of the theorems and proofs presented in Sections 6.2 and 6.3 of this chapter can be found there. Samuelson (1997) approaches evolutionary game theory from the point of view of the equilibrium selection problem. Fudenberg and Levine (1998) and Young (1998) both consider the connection between evolutionary games and how less-than-perfectly rational players learn over repeated plays of a game. Hofbauer and Sigmund (2002) provides a treatment of evolutionary

<sup>11</sup> According to Frank (1995), George Price’s increasingly intense religious devotion led him to “[insist] that, in the Hawk–Dove game, the word ‘dove’ not be used because of its religious significance.”

game theory from a perspective appealing to both mathematical biologists and game theorists; that text offers a much more detailed analysis of the dynamics of evolutionary games than was possible here. Readers interested in what happens when the underlying game is given in the *extensive form*, rather than the strategic form (a topic which I entirely neglected), should consult Cressman (2003). Although Nowak (2006) addresses topics in evolutionary dynamics which lie outside the scope of evolutionary games, it features a number of chapters on evolutionary game theory covering finite population models, local interaction models, and spatial games. Those looking for a textbook featuring both traditional and evolutionary game theory would be well advised to consider Gintis (2009). Finally, Sandholm (2010) is a 540-page *tour de force* that gives a unified treatment of much of evolutionary game theory, with an extensive set of references.<sup>12</sup>

## References

- Alexander, J. M. (2007). *The Structural Evolution of Morality*. Cambridge: Cambridge University Press.
- Binmore, K., Shaked, A. and Sutton, J. (1985). Testing noncooperative bargaining theory: a preliminary study. *The American Economic Review*, **75**, 1178–1180.
- Binmore, K. (1992). *Fun and Games*. Lexington, MA: D. C. Heath and Company.
- Bishop, D. T. and Cannings, C. (1978). A generalised war of attrition. *Journal of Theoretical Biology*, **70**, 85–124.
- Björnerstedt, J. (1993). Experimentation, imitation, and evolutionary dynamics. Mimeo. Department of Economics, Stockholm University.
- Björnerstedt, J. and Weibull, J. (1999). Nash equilibrium and evolution by imitation. In K. J. Arrow, E. Colombatto, and M. Perlman (eds), *The Rational Foundations of Economic Behavior*. London: St. Martin's Press.
- Bomze, I. M. (1983). Lotka–Volterra equation and replicator dynamics: a two-dimensional classification. *Biological Cybernetics*, **48**, 201–211.
- Chao, L. and Levin, B. R. (1981). Structured habitats and the evolution of anticompetitor toxins in bacteria. *Proceedings of the National Academy of Sciences USA*, **78**, 6324–6328.
- Chicone, C. (2000). *Ordinary Differential Equations with Applications*. New York, NY: Springer.
- Cressman, R. (2003). *Evolutionary Dynamics and Extensive Form Games*. Cambridge, MA: The MIT Press.
- Durrett, R. and Levin, S. (1997). Allelopathy in spatially distributed populations. *Journal of Theoretical Biology*, **185**, 165–171.
- Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford: Oxford University Press.

<sup>12</sup> Be sure not to confuse Sandholm's *Population Games and Evolutionary Dynamics* with the similarly titled work by Hofbauer and Sigmund, *Evolutionary Games and Population Dynamics*.

- Frank, S. A. (1995). George price's contributions to evolutionary genetics. *Journal of Theoretical Biology*, **175**, 373–388.
- Fudenberg, D. and Levine, D. K. (1998). *The Theory of Learning in Games*. Cambridge, MA: The MIT Press.
- Gigerenzer, G., Todd, P. M. and the ABC Research Group. (1999). *Simple Heuristics That Make Us Smart*. Oxford: Oxford University Press.
- Gintis, H. (2000). *Game Theory Evolving*. Princeton, NJ: Princeton University Press.
- Gintis, H. (2009). *Game Theory Evolving*. Princeton, NJ: Princeton University Press, second edition.
- Hofbauer, J., Schuster, P. and Sigmund, K. (1979). A note on evolutionary stable strategies and game dynamics. *Journal of Theoretical Biology*, **81**, 609–12.
- Hofbauer, J. and Sigmund, K. (2002). *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press.
- Jackson, M. O. (2008). *Social and Economic Networks*. Princeton, NJ: Princeton University Press.
- Khalil, H. K. (2002). *Nonlinear Systems*. Upper Saddle Riner, NJ: Prentice Hall, third edition.
- Luce, R. D. and Raiffa, H. (1957). *Games and Decisions: Introduction and Critical Survey*. New York, NY: John Wiley and Sons, Inc.
- Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.
- Maynard Smith, J. and Price, G. (1973). The logic of animal conflict. *Nature*, **246**, 15–18.
- Nash, J. (1951). Non-cooperative games. *Annals of Mathematics*, **54**, 286–295.
- Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard, MA: Harvard University Press.
- Nowak, M. A. and May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, **359**, 826–829.
- Samuelson, L. (1997). *Evolutionary Games and Equilibrium Selection*. MIT Press series on economic learning and social evolution. Cambridge, Mas.: MIT Press.
- Samuelson, L. and Zhang, J. (1992). Evolutionary stability in asymmetric games. *Journal of Economic Theory*, **57**, 363–391.
- Sandholm, W. H. (2010). *Population Games and Evolutionary Dynamics*. Cambridge, MA: MIT Press.
- Skyrms, B. (2003). *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.
- Sugden, R. (2001). The evolutionary turn in game theory. *Journal of Economic Methodology*, **8**, 113–130.
- Swinkels, J. (1992). Evolutionary stability with equilibrium entrants. *Journal of Economic Theory*, **57**, 306–332.
- Taylor, P. D. and Jonker, L. B. (1978). Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, **40**, 145–156.
- Thomas, B. (1985). On evolutionarily stable sets. *Journal of Mathematical Biology*, **22**, 105–115.
- von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

- 
- Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, **393**, 440–442.
- Weibull, J. W. (1995). *Evolutionary Game Theory*. Cambridge, MA: The MIT Press.
- Young, H. P. (1998). *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton, NJ: Princeton University Press.

# 7 Choice, preference, and utility: probabilistic and deterministic representations

A. A. J. Marley and Michel Regenwetter

7.1	General remarks on the study of preference and choice	375
7.1.1	Comment on organization and assumed mathematical background knowledge	379
7.2	Choice probabilities: notation and definitions	381
7.3	Choice probabilities for binary preference relations	386
7.3.1	Binary preference relations	386
7.3.2	Choice probabilities induced by a single fixed binary preference relation	390
7.3.3	Choice probabilities for varying or uncertain preferences	394
7.3.3.1	Binary choice probabilities induced by rankings (strict linear orders)	397
7.3.3.2	Best choice probabilities induced by rankings (strict linear orders)	399
7.3.3.3	Best-worst choice probabilities induced by rankings (strict linear orders)	400
7.3.3.4	Ternary paired-comparison probabilities induced by strict partial orders, interval orders, semiorders, or strict weak orders	402
7.4	Algebraic theories and their real-valued representations	406
7.4.1	Real-valued representations of binary preference relations	406
7.4.2	Real-valued representations of weak orders over gambles	407
7.4.2.1	Notation	408
7.4.2.2	Ranked weighted utility representation	409
7.4.3	Axiomatizations of representations of gambles	409
7.4.3.1	Rank-dependent utility	410
7.4.3.2	Configural weighted utility	411
7.4.4	Joint receipts	412
7.4.5	Parametric forms for utility and weights	413
7.4.5.1	Parametric utility forms	413
7.4.5.2	Parametric weight forms	416
7.4.5.3	Multiplicative separability of utility and weight	418
7.4.5.4	Other parametric and weight forms	419
7.5	Choice probabilities for real-valued representations	419

7.5.1	Distribution-free random utility representations	420
7.5.1.1	Distribution-free random cumulative prospect theory: an example with Goldstein–Einhorn weighting and power utility	425
7.5.1.2	Axiomatizations of utility representations in probabilistic choice	426
7.5.2	Horse race models of choice and response time	427
7.5.2.1	Distribution-free horse race models	428
7.5.2.2	Luce's choice model derived from a horse race model	431
7.5.3	Context free linear accumulator models of choice and response time	432
7.5.3.1	Multiplicative drift rate variability	434
7.5.3.2	Relation of multiplicative LBA models with no start point variability and Fréchet drift rate variability to MNL models	435
7.5.3.3	Additive drift rate variability	438
7.5.4	Context-dependent models of choice and response time	439
7.5.4.1	Multiattribute linear ballistic accumulator (MLBA) model	441
7.5.4.2	The $2N$ -ary choice tree model for $N$ -alternative preferential choice	442
7.5.5	Context-dependent models of choice	443
7.6	Discussion, open problems, and future work	446
	Acknowledgments	447
	References	447

## 7.1 General remarks on the study of preference and choice

There is a huge theoretical literature on preference and choice, with a corresponding large literature on testing theories. In this chapter we focus almost entirely on theoretical work. In this section we review some empirical applications that motivate some of our theoretical primitives, such as the “sample spaces” for which probabilistic models are designed. In later sections, we provide references to relevant empirical work.

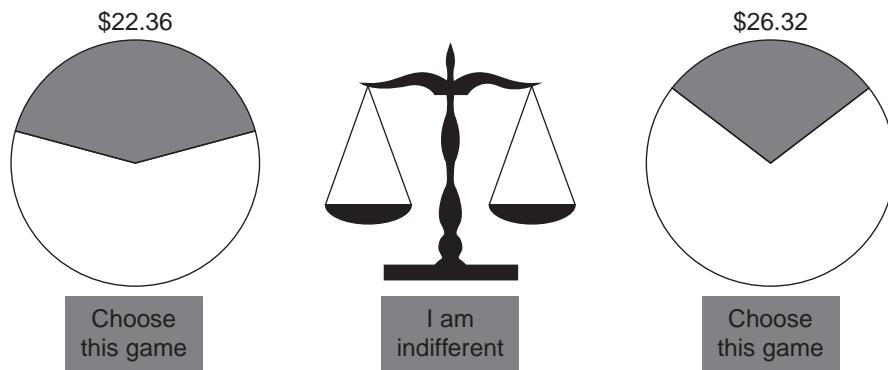
Recent important books, review articles, and handbooks, with a theoretical emphasis, include Barberá *et al.* (1999, 2004, 2013); Gilboa (2009); Luce (2000); and Wakker (2010). We focus on the past 15 years in this chapter and, for a significant part, restrict the content to general multiattribute options. While we provide some in-depth details for choice under uncertainty, we do not go into specialized models of intertemporal or sequential choice. Also, to keep the bibliography manageable, we often cite only one or two most relevant articles on a topic.

Historically, research on preference and choice has placed an emphasis on deterministic representations, most notably, algebraic models of preference, utility, and choice. A notorious challenge, for over 50 years, has been the question of adequate representation for data showing different choices in repeated presentations of the same task, both within a person and across decision makers.

This formal modeling challenge is sometimes associated with Duncan Luce, whose 1959 choice axiom was a major milestone in probabilistic characterizations of utility. Many other leading scholars have also studied and discussed ways to generalize algebraic models into probability models (see Blavatskyy and Pogrebna, 2010; Stott, 2006; Wilcox, 2008, and their citations). We will emphasize probabilistic representations of preference, utility and choice, and we will consider deterministic representations as the special case where all probability mass is concentrated on a single preference, a single utility function, or a single way of making a choice, and choice based on such a deterministic representation is error-free. The probabilistic generalizations we consider are based either on the premise that preferences or utilities are deterministic but responses are probabilistic (e.g., due to probabilistic error), or on the premise that preferences are probabilistic and responses are deterministic.

The focus is on *theory*, with data and model fitting introduced only when such is intimately connected to the theory. We take this approach mainly because of length restrictions, but also because there are numerous works on empirical evaluation and because there are ongoing developments and debates about the appropriate frequentist and/or Bayesian analyses of such data; we cite these empirical works throughout the chapter. Another challenge to much classic work on probabilistic choice is the assumption of context independence, which we now begin to address by summarizing several recent context-dependent choice models.

An important distinction in economics is that of *stated* versus *revealed* preference. Stated preference refers to the choices that a person makes in, say, a typical psychology experiment, where frequently the participants are college students who either receive course credit or an hourly payment for participation; sometimes an additional payment is made that is contingent on the person's choices in the experiment (see the example on ternary choice among gambles, below). Revealed preference refers to the actual choices made by a person in the real world, such as their grocery purchases or their participating, or not, in a weekly office lottery pool. As a result of scanner data, it is increasingly possible to study revealed preference for, say, food items, although much of those data are aggregated over consumers; recently, scanner data have become available for repeated choice occasions of a single individual. However, there remain areas where it is not possible to obtain revealed preferences, such as for new technology that is not yet generally available, or that is too expensive for general uptake (see the example on household renewable energy, below). Also, there are areas, such as medicine, where it is extremely difficult, if not impossible, to study revealed preferences in detail;



**Figure 7.1** A sample stimulus from a ternary paired-comparison study using binary lotteries. Figure 4 of the Online Supplement to Regenwetter and Davis-Stober (2012). Reproduced with permission from the American Psychological Association.

end-of-life decisions is one such area where important work is being carried out using stated preferences (Flynn *et al.*, 2015).

Lancsar and Swait (2014) summarize the literature on the external validity of stated preference with respect to revealed preference and other measures. We focus on models that are applied in experimental and consumer psychology to stated preference, although they are also frequently applied in economics to revealed preference.

We now present two examples of stated preference tasks, then give pointers to a more complex task involving both aspects of those tasks. The first task concerns stated ternary paired-comparisons between gambles (lotteries); the second concerns multiple choice between possible programs for microgeneration of electricity by households; and the more complex task concerns choices between risky or uncertain multiattribute health states. These examples are closely related to some of the formulations and models in Sections 7.3–7.5

Many descriptive, prescriptive, and normative theories of decision making share structural assumptions about the nature of individual preferences. Regenwetter and Davis-Stober (2012) investigated choices among pairs of lotteries where decision makers were permitted to express a lack of preference for either lottery (see the screen shot of an experimental trial in Figure 7.1). Their experimental study built on a seminal paper by Tversky (1969) and used lotteries that were contemporary dollar equivalents of lotteries in Tversky (1969), similar to Regenwetter *et al.* (2011a). The experiment was spread over three sessions of approximately one hour each. At the end of each session, the participant received a \$5 flat payment plus played one of his/her chosen cash lotteries from one trial (chosen uniformly at random) for real money. If the respondent had selected “no preference” on that trial, then one of the two lotteries of that trial was selected, with equal probability,

for real play. Regenwetter and Davis-Stober investigated the behavior of 30 participants separately and secured high statistical power by collecting 45 trials per gamble pair and per respondent and by exposing each person to three distinct sets, each containing 10 gamble pairs. They found that variability in choices between and within person could be explained very well by a parsimonious model in which ternary choice probabilities are marginal probabilities of an unknown probability distribution over strict weak orders. Strict weak orders will be introduced in Section 7.3.1 and the model that Regenwetter and Davis-Stober (2012) tested will be discussed, e.g., in Lemma 7.10 and in Theorem 7.9.

Marley and Islam (2012) report the results of an analysis of the data from a stated preference survey that was conducted to examine tradeoffs of features of solar panels for household-level electricity generation; this is a potentially very significant energy source as individual households account for one-third of all energy consumption in North America. Data for the survey were collected from 298 respondents by Pureprofile (a large online panel provider in Australia, North America, and other countries), with study participants screened based on owning a house in Ontario, Canada. The attributes and attribute levels studied were chosen based on extensive research, reviews of product/retailer ads/claims, websites, pilot study, etc. Each respondent was shown the list of attributes and their ranges before the choice task, and informed that the stated savings in energy cost and carbon emission were on an annual basis.

Figure 7.2 shows a sample screen shot of the first of the (20) choice sets in the survey; each choice set of four options was constructed using recently developed design theory that allows the researcher to efficiently “cover” the set of possible options. The response task was framed as a sequential choice process, with respondents instructed to choose the most preferred alternative out of four (Q1), then the least preferred out of the remaining three (Q2), and, finally, the most preferred out of the remaining two (Q3). Respondents were also asked to indicate (Q4, Figure 7.1) if they would choose none of the four options; we do not present models of the latter choice (see Orme, 2009, for models for both best/worst and such “anchor” questions). Each time a respondent selected a profile, that profile disappeared from the screen so as to restrict the respondent’s next choice to the remaining profiles. The (repeated) best/worst method of preference elicitation provides more information than, say, a best choice on each choice set. The models studied are closely related to those presented in Section 7.5, and gave parameter estimates that had reasonable properties – for instance, the estimates showed that a shorter payback time for a loan is preferred and that a monetary grant is preferred to a sales tax refund.

Our third example combines the structures of the first two examples. It has a temporal component (life expectancy), and, as noted above, we do not consider such structures in this chapter; however, the models we describe can be extended to such domains (see the references later in this paragraph). Here, each outcome in an option is a health state, which might be full health, death, or some other health

Features	Option A	Option B	Option C	Option D
Total initial investment including installation and connection to national grid (3KWh Capacity)	\$35,000	\$25,000	\$20,000	\$20,000
Energy cost saving	40%	10%	10%	20%
Carbon emission saving	0	1 tonne of CO <sub>2</sub>	0	1 tonne of CO <sub>2</sub>
Payback period	5 years	10 years	10 years	5 years
Tax Incentives & subsidy/rebates	Grant \$2,500	Refund of HST	Grant \$2,500	Grant \$2,500
Export reward as per micro-FIT program (pass all or excess capacity to national grid)	80 cents/KWh, Roof Mounted	80 cents/KWh, Roof Mounted	64 cents/KWh, Ground Mounted	80 cents/KWh, Roof Mounted
Yearly inflation on fossil fuel cost	3%	3%	3%	6%
Possibility of government policy changes about green energy technologies	No	No	Yes	No
% of local households already adopted one of these technologies	5%	5%	10%	10%

Q1. Which of the four options would you MOST likely choose?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Q2. Which of the remaining three options would you LEAST likely choose?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Q3. Which of the two remaining options would you MOST likely choose?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Q4. If you would choose none of the four options, check the box to the right:	<input type="checkbox"/>			

[Show All Options](#)

Scenario 1 of 20

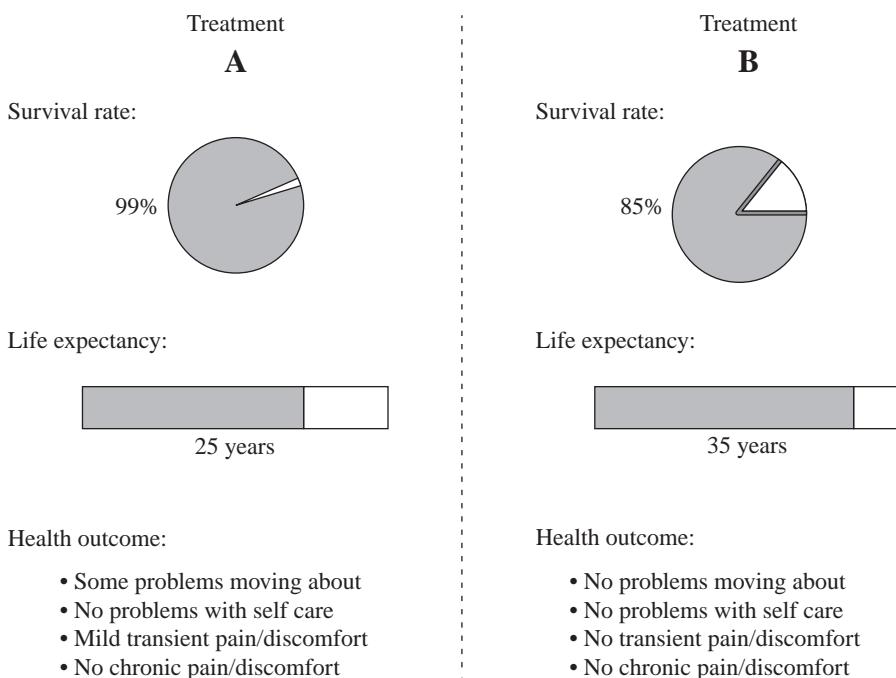
[Submit](#)

**Figure 7.2** A sample question from a study on preferences for microgeneration of electricity using solar panels. Figure 1 of Marley and Islam (2012). Reproduced with permission from Elsevier.

state defined in terms of various attributes; and an option consists of one or more such outcomes with an associated success (survival) rate and/or life expectancy. Figure 7.3 shows a pair of such options that might appear in a stated preference (choice) task, where the participant has to state which option is preferred. Matching tasks are also studied in this domain: for example, in the time-tradeoff (TTO) procedure, a respondent assigns a life expectancy to a reference health state (typically full health) such that she is indifferent between that state and a target health state with a specified life expectancy; and in the standard gamble (SG) procedure, a respondent assigns a survival rate to a reference health state (typically full health) with a specified life expectancy such that she is indifferent between that state and a target health state with the same life expectancy for sure (Bleichrodt and Pinto, 2006, review this literature). Most of the theories in this domain are algebraic, paralleling those in Section 7.4 but with multiattribute outcomes; however, recent work combines such representations with probabilistic models of the kind presented in Section 7.5 (Arons and Krabbe, 2013).

### 7.1.1 Comment on organization and assumed mathematical background knowledge

Section 7.2 discusses various classes of probability representations for response data using basic concepts of probability. These representations correspond to models of empirical data-generating processes, on which the other work builds. Section 7.3.1 reviews binary preference relations, and Section 7.3.2 discusses a



**Figure 7.3** *A hypothetical choice between two treatment options described by their success (survival) rate, life expectancy, and health outcome.*

range of probability models of choice behavior. In particular, Section 7.3.2 discusses models of probabilistic choice when individual preference is deterministic and variability is caused by, for instance, random errors in responding. Section 7.3.3 discusses models in which preferences themselves are probabilistic. Section 7.4 studies algebraic structures for choice options (mainly gambles) and presents conditions under which real-valued (utility) representations capture those structures. Section 7.5 proceeds to consider real-valued random variable representations to model both the choices made and the time to make them.

We expect that readers not familiar with all the relevant areas of mathematics may find some parts of the chapter harder to follow than others. In general, we assume familiarity with basic mathematical concepts in algebra, combinatorics, convex geometry, probability theory (including random variables), set theory, and real analysis: Sections 7.1–7.3 make extensive use of combinatorics and convex geometry, and Sections 7.4 and 7.5 of algebra and probability theory. Many sections are interdependent. We cross-reference relevant and related results across sections in an effort to allow readers to skip sections or subsections that they find too specialized and then go back and only read parts that are referenced in parts that they want to know about in depth. When we state a result without an associated citation (usually as a lemma), its proof is immediate.

In the early sections, we introduce a common notation for results, and open problems, for the several standard procedures in the study of (probabilistic) choice;

these include best choice; worst choice; best-worst choice; and ranking. As the notation is quite concise, we illustrate each new integrative concept with one or more examples.

## 7.2 Choice probabilities: notation and definitions

The primary purpose of this section is to provide formally concise definitions of various types of data-generating processes. We generally refer to these as “choice probabilities.” In many cases, these definitions characterize natural and very general parameter spaces of probabilistic models. We also provide characterizations of those parameter spaces using concepts from convex geometry. Our general concept of response probabilities aims to capture a variety of scenarios in which a decision maker is offered a subset  $X$  of choice alternatives from a finite master set  $\mathcal{A}$  with  $|\mathcal{A}| \geq 2$ , and is asked to provide information about the alternatives in that set  $X$ . The definition includes deterministic choice (and binary relations, as well as algebraic representations) as a special case. The information may range from choosing the better alternative in various pairs of alternatives, to indicating the decision maker’s favorite or least favorite alternative in offered sets of various sizes, to providing a complete rank ordering of the offered alternatives from most favorite to least favorite, etc. Different empirical paradigms differ in the types of choice sets  $X$  that they offer and in the responses they solicit.

The purpose of the first definition is to define response probabilities in a general fashion that can accommodate a variety of such paradigms. In this section, we do not yet model hypothetical constructs such as preferences and utility that presumably underlie response processes. Rather, we focus on formally characterizing the response processes themselves. Before one can develop a well-designed theory to account for data, one needs to understand one’s empirical space of observable quantities. Later sections discuss the (additional) constraints that decision models place on the data generating process.

We denote by  $P_{(X, R_X)}$  a probability distribution governing the response when a person is offered a set  $X$  of choice alternatives and is asked to indicate the choice alternatives ranked at the “rank positions” listed in an index set  $R_X$ . For instance, if  $R_X = \{1, 5\}$  and  $X = \{a, b, c, d, e\}$ , then the person is asked to report the option they prefer to all others (i.e., the “best”), as well as the option to which they prefer all others (i.e., the “worst”). The distribution  $P_{(\{a, b, c, d, e\}, \{1, 5\})}$  describes the probabilities of all 20 possible “best-worst” responses when asked to choose from  $\{a, b, c, d, e\}$ . When we later discuss models of hypothetical constructs like preferences or utilities, we will see how such models further constrain such response probabilities.

General examples of possible responses are: all possible ways of identifying a single best alternative (if  $R_X = \{1\}$ ); or all possible ways of identifying a single best and a single worst alternative (if  $R_X = \{1, |X|\}$ ); all possible preference

rankings (if  $R_X = \{1, 2, \dots, |X|\}$ ).<sup>1</sup> Space limitations prevent us from presenting detailed results on models of ranking behavior that use repeated best, then worst, or any other method. Similarly, with the exception of “ternary paired comparisons,” we will not consider cases in which a person states indifference or a lack of a strict preference among choice options.

Virtually all choice-based empirical decision making research currently picks and uses one single choice paradigm, most commonly best choice, followed by best-worst choice or ranking. However, there is no strong reason for limiting a study to one single such paradigm. Definition 7.1 explicitly sets the stage for the possibility that some trials in an experiment ask a person to pick the best out of a pair of options, some trials in an experiment ask the person to pick the worst out of a set of, say, four options, etc. The purpose of our general approach is to allow researchers more flexibility in their experimental design and to drop limiting yet unnecessary assumptions.

**Definition 7.1** Consider a finite master set  $\mathcal{A}$  of two or more choice alternatives. Let  $\Theta$  be a set containing elements  $\theta$  of the form  $\theta = (X, R_X)$  with each  $X \subseteq \mathcal{A}$  indicating an available choice set and each  $R_X \subseteq \{1, 2, \dots, |X|\}$  indicating the “rank positions” queried with respect to the set  $X$ . Let  $\Pi(R_X, X)$  denote the collection of all one-to-one mappings from  $R_X$  into  $X$ . A collection  $P = (P_\theta)_{\theta \in \Theta}$  is called a *collection of response probability distributions* if every  $P_\theta$  with  $\theta = (X, R_X)$  is a probability distribution over  $\Pi(R_X, X)$ , i.e.,

$$0 \leq P_{(X, R_X)}(\pi) \leq 1 \quad (\forall \pi \in \Pi(R_X, X)) \quad (7.1)$$

$$\text{and} \quad \sum_{\pi \in \Pi(R_X, X)} P_{(X, R_X)}(\pi) = 1. \quad (7.2)$$

Each  $\theta = (X, R_X)$  can be thought as a possible *query* to a decision maker: “Among the objects in  $X$ , indicate the objects at “rank positions”  $R_X$  within  $X$ ” and we can write  $\Pi(R_X, X)$  more generically as  $\Pi(\theta)$ . The collection  $\Pi(\theta)$  denotes all permissible responses that the decision maker may give to the query  $\theta$ .

For example, if  $R_X = \{1, |X|\}$  for each  $X$ , then the respondent is offered one or several sets  $X \subseteq \mathcal{A}$ , and when offered  $X$ , the respondent is asked to indicate the “best” and the “worst” option in  $X$ . Say, if  $\mathcal{A} = \{a, b, c, d, e\}$ , if  $X = \{a, b, c, d\}$ , then  $\Pi(R_X, X) = \Pi(\{1, 4\}, \{a, b, c, d\})$  is the collection of all possible responses in which one object among  $a, b, c, d$  is identified as “best” and one as “worst.” Mathematically, this is the collection of all one-to-one mappings from  $\{1, 4\}$  into  $\{a, b, c, d\}$ . If we were to ask a person to choose the best and the worst from  $\mathcal{A} = \{a, b, c, d, e\}$ , and we were also to ask her to choose the best from  $\{a, b, c\}$ , as well as the best among options  $b$  and  $c$ , then we would need to consider all one-to-one mappings from  $\{1, 4\}$  into  $\{a, b, c, d\}$ , all mappings from  $\{1\}$  into  $\{a, b, c\}$  and all mappings from  $\{1\}$  into  $\{b, c\}$  as possible responses. (We revisit this example after Definition 7.15.)

<sup>1</sup> We use the standard convention that  $\{1, 2, \dots, 2\}$  denotes the set  $\{1, 2\}$ .

The response probability distributions (7.1)–(7.2) can also be stated more generally as,

$$0 \leq P_\theta(\pi) \leq 1 \quad (\forall \pi \in \Pi(\theta))$$

and  $\sum_{\pi \in \Pi(\theta)} P_\theta(\pi) = 1.$

Notice, in particular, that, for any query  $\theta$ , any response  $\pi \in \Pi(\theta)$  satisfies  $\pi(i) = x$  if and only if the response  $\pi$  assigns  $x$  “rank position”  $i$ . Lemma 7.3 in Section 7.3 formally states this after Definition 7.9 formally introduces the concept of “rank.”

We now consider various special cases of this general framework, with  $\mathcal{A}$  fixed, and  $|\mathcal{A}| \geq 2$ . Our first special case of interest concentrates on situations where the decision maker is offered two distinct choice alternatives  $x$  and  $y$  at a time and we wish to model the probability that the decision maker indicates, say,  $x$  as the preferred alternative among the two.

**Definition 7.2** A collection of response probabilities  $(P_\theta)_{\theta \in \Theta}$  is a *collection of binary choice probabilities* if every  $\theta \in \Theta$  is of the form  $\theta = (X, \{1\})$ , with  $|X| = 2$ . A *complete collection of binary choice probabilities* on  $\mathcal{A}$  is a collection  $(P_\theta)_{\theta \in \Theta}$  of binary choice probabilities in which every two-element set of alternatives is an available choice set:  $(\{x, y\}, \{1\}) \in \Theta, \forall x, y \in \mathcal{A}$ , with  $x \neq y$ . We also write  $P_{xy}$  instead of  $P_{(\{x, y\}, \{1\})}(\pi)$ , where  $\pi(1) = x$ .  $P_{xy}$  is the probability that a decision maker chooses  $x$  when instructed to choose the preferred among two choice options  $x \neq y$  in  $\mathcal{A}$ . The notation  $P_{xy}$  is frequently used in the literature, as is  $p(x, y)$ .

Binary choice is usually viewed as meaning “binary best choice.” While uncommon, some scholars have also considered “binary worst choice,” that is,  $(P_\theta)_{\theta \in \Theta}$ , where every  $\theta \in \Theta$  is of the form  $\theta = (X, \{2\})$ , with  $|X| = 2$ . These scholars provided evidence that, contrary to the implicit assumption in most work, the probability of choosing  $x$  over  $y$  in binary choice need actually not match the probability of identifying  $y$  as the worst in the set  $\{x, y\}$  (Shafir, 1993, but Louviere *et al.*, 2015, critique this viewpoint).

We now move to more general situations, containing binary choice probabilities as special cases, where the decision maker is offered sets  $X$  of potentially varying sizes and asked to indicate either the best, or the worst, choice alternative among the available options in  $X$ , or both.

**Definition 7.3** Consider a collection of response probabilities  $(P_\theta)_{\theta \in \Theta}$ . The collection  $(P_\theta)_{\theta \in \Theta}$  is a *collection of best choice probabilities* if every  $\theta \in \Theta$  is of the form  $\theta = (X, \{1\})$ , with  $|X| \geq 2$ . The collection  $(P_\theta)_{\theta \in \Theta}$  is a *collection of worst choice probabilities* if every  $\theta \in \Theta$  is of the form  $\theta = (X, \{|X|\})$ , with  $|X| \geq 2$ . The collection  $(P_\theta)_{\theta \in \Theta}$  is a *collection of best-worst choice probabilities* if every  $\theta \in \Theta$  is of the form  $\theta = (X, \{1, |X|\})$ , with  $|X| \geq 2$ . A *complete collection of best choice probabilities* is a collection of best choice probabilities with  $(X, \{1\}) \in \Theta$ , for every subset  $X \subseteq \mathcal{A}$  with  $|X| \geq 2$ . Likewise, in a *complete collection of worst*

*choice probabilities*,  $(X, \{|X|\}) \in \Theta$ , for every subset  $X \subseteq \mathcal{A}$  with  $|X| \geq 2$  and, in a *complete collection of best-worst choice probabilities*,  $(X, \{1, |X|\}) \in \Theta$ , for every subset  $X \subseteq \mathcal{A}$  with  $|X| \geq 2$ .

For the special cases that we considered in Definition 7.3, we can simplify the notation to make it more mnemonic and to agree with other publications. We write  $B_X(x)$  to denote  $P_{(X, \{1\})}(\pi)$ , where  $\pi(1) = x$ , the probability that the decision maker identifies  $x \in X$  as the best alternative out of subset  $X \subseteq \mathcal{A}$  (that is, the alternative “ranked in position” 1 in  $X$ ). We write  $W_X(x)$  to denote  $P_{(X, \{|X|\})}(\pi)$ , where  $\pi(|X|) = x$ , the probability that the decision maker identifies  $x \in X$  as the worst alternative out of subset  $X \subseteq \mathcal{A}$  (that is, the alternative ranked in position  $|X|$  in  $X$ ). We write  $BW_X(x, y)$  to denote  $P_{(X, \{1, |X|\})}(\pi)$ , where  $\pi(1) = x, \pi(|X|) = y$ , the probability that the decision maker identifies  $x$  as the best alternative and  $y \neq x$  as the worst alternative out of subset  $X \subseteq \mathcal{A}$ .

We now consider “ordered  $n$ -tuples,” or “vectors of dimension  $n$ ” of values in the real-valued interval  $[0, 1]$ , or “points” in  $[0, 1]^n$ , with  $n \geq 2$ . Formally, these are mappings from  $\{1, 2, \dots, n\}$  into  $[0, 1]$  that map each “coordinate”  $i \in \{1, 2, \dots, n\}$  into a value in  $[0, 1]$ . For example,  $(0.5, .31, .723) \in [0, 1]^3$  is the mapping

$$1 \mapsto 0.5 \quad 2 \mapsto 0.31 \quad 3 \mapsto 0.723.$$

In probabilistic choice models, each permissible response is mapped into a probability. Hence, if there are  $n$  many possible responses, a response probability distribution is an ordered  $n$ -tuple in  $[0, 1]^n$  and that is precisely a mapping that maps each possible response into its probability. A special case of such a mapping will associate, with every permissible response, a degenerate probability, that is, a probability equalling one or zero. Such a mapping is a degenerate probability distribution that assigns probability one to exactly one elementary outcome. Hence, deterministic responses are a degenerate special case of a probabilistic choice model.

Much of what follows in later sections reviews extensive theoretical results on probabilistic choice models, as viewed through the lens of polyhedral combinatorics (combinatorial geometry). To prepare that discussion (especially Subsection 7.3.3), we cast the general parameter spaces in convex geometry terms.

**Definition 7.4** Let  $n \geq 2$  be an integer. An  *$n$ -dimensional unit hypercube* is the set  $[0, 1]^n$  of all mappings from  $\{1, 2, \dots, n\}$  into the real-valued interval  $[0, 1]$ . A *vertex* of an  $n$ -dimensional unit hypercube is a 0/1-valued vector with  $n$  coordinates, i.e., a mapping from  $\{1, 2, \dots, n\}$  into the set  $\{0, 1\}$ .

For example, a two-dimensional unit-hypercube  $[0, 1]^2$  is a square with sides of length 1, and its vertices are the vectors  $(0, 0), (0, 1), (1, 0)$ , and  $(1, 1)$ . A three-dimensional unit-hypercube  $[0, 1]^3$  is a cube with sides of length one. Its vertices are the eight “corners” of the cube. Geometrically in higher dimensions, a vertex of a hypercube is a “corner” of the hypercube.

**Lemma 7.1** Suppose that  $|\mathcal{A}| = n \geq 2$ . A complete collection of binary choice probabilities is contained in an  $n(n - 1)$ -dimensional unit hypercube. Noting that,

for all  $x \neq y$ ,  $P_{xy} = 1 - P_{yx}$ , a complete collection of binary choice probabilities can be reduced to a set of functionally independent binomial parameters by retaining, for each  $x \neq y$ , exactly one of  $P_{xy}$  or  $P_{yx}$ . These parameters form exactly an  $\binom{n}{2}$ -dimensional unit hypercube.<sup>2</sup> A complete collection of best (respectively, worst) choice probabilities is contained in a unit hypercube of dimension  $\sum_{k=2}^n \binom{n}{k} k$  and, after we drop redundant probabilities, the remaining multinomial parameters form a unit hypercube of dimension  $\sum_{k=2}^n \binom{n}{k} (k-1)$ . A complete collection of best-worst choice probabilities is contained in a unit hypercube of dimension  $\sum_{k=2}^n \binom{n}{k} k(k-1)$  and, after we drop redundant probabilities, the remaining multinomial parameters form a unit hypercube of dimension  $\sum_{k=2}^n \binom{n}{k} [k(k-1) - 1]$ .

The vertices of these hypercubes are the 0/1 vectors representing degenerate probability distributions, and these can be thought of as the error-free and uncertainty-free deterministic special case in which each response has a probability of either zero or one. In all these choice paradigms, probabilistic choice models can be characterized by the additional constraints they impose on the binomial or multinomial parameters. These constraints can be functional constraints on the permissible values of the parameters or, in a Bayesian framework, they can be distributional constraints that represent prior information.

We will later see how various algebraic models of preference or even utility can be embedded in the space of choice probabilities by representing deterministic preferences as certain vertices of these hypercubes, and how the constraints characterizing some probabilistic models take the form of convex polytopes formed by the convex hull of such vertices embedded in the hypercubes we just discussed (some of these terms will be defined later).

Before we proceed to models of choice probabilities, we also define “ternary paired-comparison probabilities,” the only paradigm in which we consider a decision maker who is allowed to indicate indifference or lack of preference.

**Definition 7.5** Consider a finite master set  $\mathcal{A}$  of two or more choice alternatives. A complete collection of ternary paired-comparison probabilities on  $\mathcal{A}$  is a collection  $(T_{x,y})_{\substack{x,y \in \mathcal{A} \\ x \neq y}}$  with

$$0 \leq T_{xy}, T_{yx} \leq 1 \quad \text{and} \quad T_{xy} + T_{yx} \leq 1 \quad (\forall x, y \in \mathcal{A}, x \neq y).$$

The ternary paired-comparison probability  $T_{xy}$  denotes the probability that the person indicates  $x$  as strictly preferable to  $y$ ,  $T_{yx}$  denotes the probability that the person indicates  $y$  as strictly preferable to  $x$ , and  $1 - T_{xy} - T_{yx}$  denotes the probability that the person indicates that neither alternative is strictly preferable to the other.

**Lemma 7.2** Suppose that  $|\mathcal{A}| = n \geq 2$ . A complete collection of ternary paired-comparison probabilities is contained in a  $n(n-1)$  dimensional unit hypercube. Note that, unlike binary choice probabilities, where for all  $x \neq y$ ,  $P_{xy} + P_{yx} = 1$ ,

<sup>2</sup> Note that this is different from a “probability simplex,” as the origin is a vertex of this hypercube, but is not a vertex of a probability simplex.

we now have  $T_{xy} + T_{yx} \leq 1$ . Hence, a complete collection of ternary paired-comparison probabilities cannot be reduced to a smaller set of nonredundant parameters, as all  $n(n - 1)$  parameters are nonredundant. Each pair  $T_{xy}, T_{yx}$  forms the two parameters of a trinomial. Furthermore, since each  $T_{xy} + T_{yx} \leq 1$ , the collection of ternary paired-comparison probabilities is a strict subset of the  $n(n - 1)$  dimensional unit hypercube (as the latter would allow any and all combinations of  $0 \leq T_{xy} \leq 1$  and  $0 \leq T_{yx} \leq 1$ ).

Various models of ternary paired-comparison probabilities are further characterized by (typically) convex polyhedra that are embedded within these spaces of trinomial parameters.

This section has characterized various types of response probabilities, independent of any theoretical model that might constrain which choice probabilities are and which are not compatible with a given theory of hypothetical constructs like preference and utility. Our next step is to formalize the hypothetical construct of “preference” and to discuss how such constructs constrain possible choice probabilities.

### 7.3 Choice probabilities for binary preference relations

#### 7.3.1 Binary preference relations

We start with the classical definitions of several classes of binary relations, as they form the basic building blocks for many formal models of decision and choice. Except where explicitly stated – for instance, in the definition of a “complete” binary relation – the elements  $x, y, z$  need not be distinct.

**Definition 7.6** A *binary relation*  $\succ$  on a set of choice alternatives  $\mathcal{A}$  is a subset of the Cartesian product of  $\mathcal{A}$  with itself, i.e.,  $\succ \subseteq \mathcal{A} \times \mathcal{A}$ . It is also standard to replace  $(x, y) \in \succ$  by the more compact and self-explanatory notation  $x \succ y$ . We write  $\neg[x \succ y]$  to denote that  $x \succ y$  does not hold. We use  $\wedge$  for logical AND and  $\vee$  for logical OR and  $\Rightarrow$  for logical implication. A binary relation  $\succ$  on  $\mathcal{A}$  is

- complete* if  $[x \succ y] \vee [y \succ x] \quad (\forall x, y \in \mathcal{A} \text{ with } x \neq y),$
- asymmetric* if  $[x \succ y] \Rightarrow \neg[y \succ x] \quad (\forall x, y \in \mathcal{A}),$
- transitive* if  $[x \succ y] \wedge [y \succ z] \Rightarrow [x \succ z] \quad (\forall x, y, z \in \mathcal{A}),$
- negatively transitive* if  $\neg[x \succ y] \wedge \neg[y \succ z] \Rightarrow \neg[x \succ z] \quad (\forall x, y, z \in \mathcal{A}).$

We will sometimes use the symbol  $\succsim$  for a binary relation that has a specific property defined next.

**Definition 7.7** A binary relation  $\succsim$  on a set of choice alternatives  $\mathcal{A}$  is

- strongly complete* if  $[x \succsim y] \vee [y \succsim x] \quad (\forall x, y \in \mathcal{A}).$

Whenever  $[x \succsim y] \wedge [y \succsim x]$ , we use the compact notation  $x \sim y$ .

There are a number of binary relations that satisfy some, but not others, of these axioms. We now review some of the most prominent relations.

**Definition 7.8** A *strict partial order* is an asymmetric and transitive binary relation. We denote the collection of all strict partial orders on  $\mathcal{A}$  by  $\mathcal{SPO}_{\mathcal{A}}$ . An *interval order* is a strict partial order  $\succ$  with the property

$$[w \succ x] \wedge [y \succ z] \Rightarrow [w \succ z] \vee [y \succ x] \quad (\forall w, x, y, z \in \mathcal{A}).$$

We denote the set of all interval orders on  $\mathcal{A}$  by  $\mathcal{IO}_{\mathcal{A}}$ . A *semiorder* is an interval order  $\succ$  with the property

$$[w \succ x] \wedge [x \succ y] \Rightarrow [w \succ z] \vee [z \succ y] \quad (\forall w, x, y, z \in \mathcal{A}).$$

The collection of all semiorders on  $\mathcal{A}$  is denoted by  $\mathcal{SO}_{\mathcal{A}}$ . A *strict weak order* is an asymmetric and negatively transitive binary relation. The collection of all strict weak orders on  $\mathcal{A}$  is denoted by  $\mathcal{SWO}_{\mathcal{A}}$ . A *strict linear order* is a transitive, asymmetric, and complete binary relation. The collection of all strict linear orders on  $\mathcal{A}$  is denoted by  $\mathcal{SLO}_{\mathcal{A}}$ . A *weak order* is a strongly complete and transitive binary relation. The collection of all weak orders on  $\mathcal{A}$  is denoted by  $\mathcal{WO}_{\mathcal{A}}$ .

Theorem 7.2 of Section 7.4.1 reviews the standard real-valued representations of the binary relations of Definition 7.8. Strict linear orders play a particularly important role in choice modeling. It is sometimes useful to view a strict linear order as a ranking and to consider the ranks of objects in the ranking. We define the latter concept next.

**Definition 7.9** Consider a finite master set  $\mathcal{A}$  of two or more choice alternatives. Let  $X \subseteq \mathcal{A}$  and let  $\succ \subseteq \mathcal{A} \times \mathcal{A}$  be a strict linear order on  $\mathcal{A}$ . The *rank* of  $x$  with respect to  $X$  and  $\succ$ ,  $\text{Rank}_{X, \succ}(x)$ , and  $\succ$  is given by

$$\text{Rank}_{X, \succ}(x) = |X| - |\{b \in X : x \succ b\}|.$$

When  $X = \mathcal{A}$  we also write  $\text{Rank}_{\succ}(x)$  instead of  $\text{Rank}_{\mathcal{A}, \succ}(x)$ .

We have informally used this concept in Definitions 7.1–7.3. The following lemma is useful for many of the models we discuss next. It formally states the tautology, used already informally in Section 7.2, that, for a strict linear order, the rank of the object at rank  $i$  equals  $i$ .

**Lemma 7.3** Let  $\mathcal{A}$  be a finite set,  $\succ$  a strict linear order on  $\mathcal{A}$ ,  $X \subseteq \mathcal{A}$ , and  $R_X \subseteq \{1, 2, \dots, |X|\}$ . For any  $\pi \in \Pi(R_X, X)$  we have, for any  $x \in X$ , and  $i \in R_X$ ,

$$\pi(i) = x \Leftrightarrow \text{Rank}_{X, \succ}(x) = i,$$

and thus:  $\text{Rank}_{X, \succ}(\pi(i)) = i$ .

Many decisions involve tradeoffs between two or more attributes, such as smaller, sooner rewards versus larger, later rewards or larger, less likely rewards versus smaller, more likely rewards, for example. Theories about how we choose among multiattribute options can be classified very broadly into two types:

according to *compensatory* theories, decision makers are able to tradeoff between attributes by “giving” a bit of one desirable attribute to “gain” a bit of another. Much of this chapter, starting with Section 7.4.2, discusses compensatory theories for preferences among monetary gambles. According to *heuristic* models, decision makers simplify the decision process through mental shortcuts, due, e.g., to cognitive limitations. The distinction between compensatory and heuristic decision making is central to the debate about rational versus irrational behavior. We now consider an example of a heuristic model.

When facing multiattribute options, a decision maker might have a separate preference relation for each attribute. These preference relations may combine somehow into an overall preference relation over the choice options. To this end, we now briefly consider ways in which two or more binary relations can be combined, e.g., to form a new binary relation. One way is through a lexicographic process:

**Definition 7.10** A *lexicographic* binary relation on  $\mathcal{A}$  is an ordered list of binary relations  $\succ_i$ ,  $1 \leq i \leq k$ , on  $\mathcal{A}$ .

For example, a lexicographic semiorder on  $\mathcal{A}$  is an ordered list of semiorders  $\succ_i$ ,  $1 \leq i \leq k$ , on  $\mathcal{A}$ .

**Lemma 7.4** A lexicographic binary relation can be reduced to a single binary relation  $\succ$ , say, by setting  $x \succ y$  if  $\exists i$  such that  $x \succ_i y$  and,  $\forall j < i : \neg[x \succ_j y]$  and  $\neg[y \succ_j x]$ . In other words,  $(x, y)$  belongs to  $\succ$  if and only if the first relation,  $\succ_i$ , in which either  $x \succ_i y$  or  $y \succ_i x$ , actually satisfies  $x \succ_i y$ .

In this fashion, we can, for instance, construct a binary relation

$$\succ = \{(a, c), (a, d), (b, d), (b, e), (c, e), (e, a)\} \quad (7.3)$$

on  $\mathcal{A} = \{a, b, c, d, e\}$ , that forms a lexicographic semiorder derived from the following two semiorders

$$\begin{aligned} \succ_1 &= \{(e, a)\}, \\ \succ_2 &= \{(a, c), (a, d), (a, e), (b, d), (b, e), (c, e)\}. \end{aligned}$$

Notice that the lexicographic semiorder  $\succ$  in Equation (7.3) violates transitivity because  $[a \succ c] \wedge [c \succ e] \wedge [e \succ a]$ , even though each of the semiorders  $\succ_1$  and  $\succ_2$  are transitive. Lexicographic semiorders are among the most prominent models of intransitive preferences (Tversky, 1969) and have attracted renewed attention in recent work (Birnbaum and Gutierrez, 2007; Brandstätter *et al.*, 2006; Regenwetter *et al.*, 2014). Note, however, that a lexicographic semiorder can very well be transitive. For example, the semiorder  $\succ_1$  above is a transitive lexicographic semiorder, as is  $\succ_2$ .

The next definition builds on Definition 7.10 to consider special lexicographic binary relations on  $\mathcal{A}$  that define certain kinds of lexicographic semiorders, including  $\succ$  in the example we just discussed. The basic idea is that there may be two

attributes that weakly order the objects in  $\mathcal{A}$  from best to worst in opposite directions (e.g., in Tversky, 1969, and in the gambles  $a-e$  of Section 7.5.1, the probability of winning is one attribute and the amount one can win is the other attribute. These two attributes trade off against each other). The model captures the idea that the decision maker, rather than trading off between attributes in a compensatory fashion, considers the attributes in a lexicographic fashion. Before we can state the definition of such a “simple lexicographic semiorder,” we need to define one more concept, namely that of a “trace.”

**Definition 7.11** Let  $\mathcal{A}$  be a set with  $|\mathcal{A}| = n \geq 2$ , and let  $\succ \in \mathcal{SO}_{\mathcal{A}}$ . The *trace* of  $\succ$ , denoted by  $\mathcal{T}_{\succ}$  is a binary relation given by

$$x \mathcal{T}_{\succ} y \Leftrightarrow \left\{ \begin{array}{l} [y \succ z] \Rightarrow [x \succ z] \quad (\forall z \in \mathcal{A}) \\ \quad \wedge \\ [w \succ x] \Rightarrow [w \succ y] \quad (\forall w \in \mathcal{A}). \end{array} \right.$$

The traces  $\mathcal{T}_{\succ}$  and  $\mathcal{T}_{\succ'}$ , for  $\succ, \succ' \in \mathcal{SO}_{\mathcal{A}}$ , are each other’s *reverses* if there exists a strict linear order  $\triangleleft \in \mathcal{SLO}_{\mathcal{A}}$  such that

$$\left. \begin{array}{l} [x \mathcal{T}_{\succ} y] \Rightarrow [x \triangleleft y] \\ \quad \wedge \\ [y \mathcal{T}_{\succ'} x] \Rightarrow [x \triangleleft y] \end{array} \right\} \quad (\forall x, y \in \mathcal{A}).$$

We call such a linear order  $\triangleleft$ , if it exists, *compatible* with  $\mathcal{T}_{\succ}$  and  $\mathcal{T}_{\succ'}$ .

**Lemma 7.5** *The trace of a semiorder is a strict weak order.*

We are now ready to define “simple lexicographic semiorders” as studied in Davis-Stober (2012).

**Definition 7.12** Let  $\mathcal{A}$  be a set with  $|\mathcal{A}| = n \geq 2$ , and let  $\succ_1, \succ_2 \in \mathcal{SO}_{\mathcal{A}}$  be two semiorders whose traces are each other’s reverses. Then the binary relation  $\succ$  resulting from  $\succ_1, \succ_2$  through the construction in Lemma 7.4 is called a *simple lexicographic semiorder*. For a fixed strict linear order  $\triangleright$  on  $\mathcal{A}$ , consider all  $\succ_1, \succ_2 \in \mathcal{SO}_{\mathcal{A}}$  with whose traces  $\triangleright$  is compatible with. We denote the collection of all resulting simple lexicographic semiorders  $\succ$  by  $\mathcal{SLSO}_{\mathcal{A}, \triangleright}$ .

The natural application of these concepts is in standard decision making paradigms in which decision makers face choice alternatives that are characterized by two attributes, such as time delay and reward, or probability and reward, that perfectly trade off with each other: in any pair of choice alternatives, the decision maker must figure out whether they prefer a little more of one attribute (e.g., probability of winning) in exchange for a little bit less of another attribute (e.g., magnitude of the reward), or vice versa. Simple lexicographic semiorders capture the idea that the decision makers employ a heuristic in which they lexically order the two attributes by their importance, that they visit the attributes in this order, and determine that they prefer one option to the other if the value of the attribute under consideration (e.g., chance of winning) in the former option “sufficiently” exceeds

the value of that attribute in the latter option. This type of process is a key theoretical alternative to “compensatory” models, according to which decisions makers genuinely trade off between competing attributes, e.g., by basing their pairwise preferences on weighted averages of attribute values. In particular, lexicographic semiorders need not be transitive.

This completes our review of binary preference relations, the leading mathematical models for the hypothetical construct of preference. We now proceed to discuss various kinds of probabilistic choice models that build on the concepts we have introduced, but aim to connect such hypothetical constructs to observable choice behavior. Different types of probabilistic choice models differ in how they expand from deterministic core hypothetical constructs to choice probabilities. Most importantly, they differ from each other regarding whether or not they model a fixed preference or varying preferences, and whether they model preferences as binary relations or, instead, are based on real-valued utilities and/or utility functions. We begin with the case of preferences that are binary relations.

### **7.3.2 Choice probabilities induced by a single fixed binary preference relation**

We start with a model that, to our knowledge, has so far only been considered for binary choice. The basic notion behind a “constant error”<sup>3</sup> model for binary choice is that a decision maker has a single “correct” response determined by a single fixed preference relation for each choice s/he needs to make, and that there is a constant probability  $\zeta$  of making an error in any given choice. Many authors also refer to this type of model as a *tremble* model or a *trembling hand* model. Whenever the decision maker is answering a question, there is a chance that s/he “trembles” and, say, presses the wrong key, or clicks the wrong icon on a screen.

**Definition 7.13** Let  $\mathcal{A}$  be a set with  $|\mathcal{A}| = n \geq 2$ , and let  $\succ \in \mathcal{SO}_{\mathcal{A}}$  be fixed. A collection  $P = (P_{\theta})_{\theta \in \Theta}$  of binary choice probabilities satisfies a *constant error model* (with regard to  $\succ$ ) with a constant error probability  $\zeta$  if the error probability  $\zeta < \frac{1}{2}$  and  $\forall x \neq y$ , with  $P_{xy}$  any one of the  $P_{\theta}$ ,

$$P_{xy} = \begin{cases} \zeta & \text{if } y \succ x, \\ 1 - \zeta & \text{if } x \succ y. \end{cases}$$

This model can be restated equivalently by labeling  $1 - \zeta$  the “probability of making a correct choice.” An interesting open problem for future work is to generalize the constant error model to other choice paradigms, such as the best-choice, worst-choice, or best-worst choice. In these more general settings one may need to differentiate between the probability of making an error and the probability of making a correct choice, and one may have to consider dependencies of such probabilities on the choice sets  $X$ , such as, say, the number of options,  $|X|$ , being

<sup>3</sup> To be precise, one should refer to a constant error *probability* or *rate*. We use the shorter terminology for convenience and to be consistent with the literature.

considered in a given choice trial. When there are multiple errors and/or correct “choices” that “combine” in a given trial, one may have to consider potentially complicated interdependencies among these component responses.

Historically, probabilistic choice models have been viewed as extensions of “core” algebraic models. We emphasize that this makes deterministic models special cases of probabilistic ones. We can embed deterministic preferences into choice probability models as special cases by considering the constant error and constant correct models where a person never trembles. Here, a deterministic, i.e., algebraic, model becomes a degenerate special case of a model of response probabilities.

We now consider more generally and formally how deterministic models are special cases of probabilistic ones. As in Definition 7.4, these concepts permit us to consider deterministic models and their probabilistic extensions through a convex geometry lens. Given a binary preference relation  $\succ$  on  $\mathcal{A}$ , a constant error model for  $\succ$  with  $\zeta = 0$ , is called a “vertex representation” of  $\succ$  (defined next) because it forms a “vertex” of a suitably defined “convex polytope” (see also Definition 7.18).

**Definition 7.14** Consider a finite set  $\mathcal{A}$  of two or more choice alternatives and a collection  $\Theta$  of elements of the form  $\theta = (X, R_X)$  with each  $X \subseteq \mathcal{A}$ , and each  $R_X \subseteq \{1, 2, \dots, |X|\}$ . Suppose that  $\succ$  is a strict linear order on  $\mathcal{A}$ . A *vertex representation*  $V_{\Theta}(\succ)$  of  $\succ$  with respect to  $\Theta$  is defined by the following collection of 0/1-coordinates:  $\forall(X, R_X) \in \Theta, \forall \pi \in \Pi(R_X, X)$ ,

$$V_{(X, \pi)}(\succ) = \begin{cases} 1 & \text{if } \text{Rank}_{X, \succ}(\pi(i)) = i, \forall i \in R_X, \\ 0 & \text{otherwise.} \end{cases} \quad (7.4)$$

The vertex representation in the definition uses a distinct coordinate for each  $(X, \pi)$ . We also call this a *full-dimensional representation*. In some cases, most notably the “mixture” models we discuss in the next section, it is natural to drop some of these coordinates because of redundancies (we give an example after Definition 7.15).

There are several cases of special interest, for which we can simplify the notation substantially (and illustrate the redundancy in coordinates). In the vertex representation for binary choice with a fixed strict linear order  $\succ \in \mathcal{SLO}_{\mathcal{A}}$ , one considers  $X = \{x, y\}, R_X = \{1\}$ , hence there are only two possible  $\pi \in \Pi(R_X, X)$ , namely

$$\pi_1 : 1 \mapsto x, \text{ and } \pi_2 : 1 \mapsto y.$$

Hence, with this choice of  $\pi$ , the vertex representation (7.4) for binary choice and a linear order preference  $\succ$  has the following 0/1-coordinates:

$$\begin{aligned} V_{(\{x, y\}, \pi)}(\succ) &= \begin{cases} 1 & \text{if } \text{Rank}_{\{x, y\}, \succ}(x) = 1, \text{Rank}_{\{x, y\}, \succ}(y) = 2, \\ 0 & \text{if } \text{Rank}_{\{x, y\}, \succ}(x) = 2, \text{Rank}_{\{x, y\}, \succ}(y) = 1, \end{cases} \\ &= \begin{cases} 1 & \text{if } x \succ y, \\ 0 & \text{if } y \succ x. \end{cases} \end{aligned}$$

Therefore, we rewrite the vertex representation of a preference  $\succ \in \mathcal{SLO}_{\mathcal{A}}$  for binary choice more mnemonically as,  $\forall x, y \in \mathcal{A}, x \neq y$  with  $(\{x, y\}, \{1\}) \in \Theta$ ,

$$V_{xy}(\succ) = \begin{cases} 1 & \text{if } x \succ y, \\ 0 & \text{if } y \succ x, \end{cases} \quad V_{yx}(\succ) = \begin{cases} 1 & \text{if } y \succ x, \\ 0 & \text{if } x \succ y. \end{cases}$$

Because  $V_{xy}(\succ) = 1 - V_{yx}(\succ)$ , it is standard in some applications to use only one coordinate per pair  $x, y$ .

A second case of special interest is the vertex representation for best choice with a fixed strict linear order  $\succ$ . Here,  $R_X = \{1\}$  and we can rewrite  $V_{(X, \pi)}(\succ)$  for  $\pi : 1 \mapsto x$  more mnemonically as a “vertex for best choice,”  $VB(\succ)$ , with coordinates

$$VB_{(X, x)}(\succ) = \begin{cases} 1 & \text{if } Rank_{X, \succ}(x) = 1, \\ 0 & \text{if } Rank_{X, \succ}(x) \neq 1. \end{cases}$$

In general, this gives  $|X|$  many 0/1-coordinates per set  $X$  under consideration. Because  $\sum_{x \in X} VB_{(X, x)}(\succ) = 1$ , for each each  $X$ , it is common in vertex representations of best choice to leave out one coordinate for each set  $X$ .

A third case is the vertex representation for worst choice with a fixed strict linear order  $\succ$ . We can also write this vertex representation more mnemonically as a “vertex for worst choice,”  $VW(\succ)$ , with coordinates

$$VW_{(X, x)}(\succ) = \begin{cases} 1 & \text{if } Rank_{X, \succ}(x) = |X|, \\ 0 & \text{if } Rank_{X, \succ}(x) \neq |X|. \end{cases}$$

Because  $\sum_{x \in X} VW_{(X, x)}(\succ) = 1$ , we can also leave out one of these 0/1-coordinates per set  $X$ .

Likewise, we write the vertex representation of best-worst choice with a fixed strict linear order  $\succ$ , as

$$VBW_{(X, x, y)}(\succ) = \begin{cases} 1 & \text{if } Rank_{X, \succ}(x) = 1 \wedge Rank_{X, \succ}(y) = |X|, \\ 0 & \text{if } Rank_{X, \succ}(x) \neq 1 \vee Rank_{X, \succ}(y) \neq |X|. \end{cases}$$

Because  $\sum_{\substack{x, y \in X \\ x \neq y}} VBW_{(X, x, y)}(\succ) = 1$ , we can also leave out one of these 0/1-coordinates per set  $X$ .

*Remark* The coordinates of a vertex representation form the coordinates of a vertex of a unit hypercube. In a full-dimensional representation, each such coordinate is a degenerate probability that equals zero or one, representing the deterministic strict linear order preference  $\succ$  as a special case of error-free choice (a constant error model with error rate  $\zeta = 0$  in the case of binary choice).

Using the vertex representations, we can generalize constant error probability by dropping the requirement of constant error rates  $\zeta$  and the limitation to binary choice. In these models, which are far more interesting than the constant error model, we now proceed to place only upper bounds on error rates. More abstractly, we permit all vectors of choice probabilities that are within a certain “distance” of the vector formed by the degenerate probability distribution that corresponds to deterministic choice. Here, we generally do not omit “redundant” coordinates.

**Definition 7.15** (Generalization of Regenwetter *et al.*, 2014) Consider a collection  $P = (P_\theta)_{\theta \in \Theta}$  of response probability distributions and a fixed strict linear order  $\succ$ . Let  $d = |\Theta|$  and  $\Delta$  be a (Minkowski) distance metric in  $\mathbb{R}^d$ . A *distance-based probabilistic specification* of a deterministic preference  $\succ$ , with distance  $\Delta$  and with an upper bound  $\tau > 0$ , states that the permissible response probabilities  $(P_\theta)_{\theta \in \Theta}$  that are allowable for fixed preference  $\succ$  must satisfy

$$\Delta((P_\theta)_{\theta \in \Theta}, \mathcal{V}_\Theta(\succ)) \leq \tau, \quad (7.5)$$

where  $\mathcal{V}_\Theta(\succ)$  is the (full-dimensional) vertex representation of Definition 7.14.

For example, if  $\Delta$  is the supremum distance, then Condition 7.5 specifies that the maximum deviation from deterministic choice cannot be more than  $\tau$  for any choice probability. This means that, according to the supremum distance specification,  $|P_\theta - V_\theta(\succ)| \leq \tau, \forall \theta \in \Theta$ . Suppose that  $\mathcal{A} = \{a, b, c, d\}$ , suppose that the decision maker's deterministic strict linear order preference is  $\succ = \{(a, b), (a, c), (a, d), (b, c), (b, d), (c, d)\}$  and suppose that the decision maker is asked to provide his best choice from  $\{a, b, c\}$ , his best-worst choice from  $\mathcal{A}$ , and make a binary choice among options  $b$  and  $c$ . (We have previously introduced this example after Definition 7.1.) Hence, the task is characterized by  $\Theta = \{(\{a, b, c\}, \{1\}), (\mathcal{A}, \{1, 5\}), (\{b, c\}, \{1\})\}$  and the vertex representation of  $\succ$  in  $\mathbb{R}^{11}$ ,  $\mathcal{V}_\Theta(\succ)$ , is given by the 11-dimensional vector whose coordinate values  $V_\theta(\succ)$  are

$$\begin{aligned} VB_{(\{a,b,c\},a)}(\succ) &= 1; & VB_{(\{a,b,c\},b)}(\succ) &= 0; & VB_{(\{a,b,c\},c)}(\succ) &= 0; \\ VBW_{(\mathcal{A},a,b)}(\succ) &= 0; & VBW_{(\mathcal{A},a,c)}(\succ) &= 0; & VBW_{(\mathcal{A},a,d)}(\succ) &= 1; \\ VBW_{(\mathcal{A},b,c)}(\succ) &= 0; & VBW_{(\mathcal{A},b,d)}(\succ) &= 0; & VBW_{(\mathcal{A},c,d)}(\succ) &= 0; \\ V_{bc}(\succ) &= 1; & V_{cb}(\succ) &= 0. \end{aligned}$$

Note in passing that, in this vertex representation, one best-choice, one best-worst choice, and one binary choice coordinate is redundant (derivable from the others, since each group sums to one). If the decision maker makes no errors, then s/he chooses deterministically in accordance with  $\succ$ , hence  $\tau = 0$  and therefore each  $P_\theta = V_\theta(\succ), \forall \theta \in \Theta$ , that is,

$$\begin{aligned} B_{(\{a,b,c\},a)}(\succ) &= 1; & B_{(\{a,b,c\},b)}(\succ) &= 0; & B_{(\{a,b,c\},c)}(\succ) &= 0; \\ BW_{(\mathcal{A},a,b)}(\succ) &= 0; & BW_{(\mathcal{A},a,c)}(\succ) &= 0; & BW_{(\mathcal{A},a,d)}(\succ) &= 1; \\ BW_{(\mathcal{A},b,c)}(\succ) &= 0; & BW_{(\mathcal{A},b,d)}(\succ) &= 0; & BW_{(\mathcal{A},c,d)}(\succ) &= 0; \\ P_{bc}(\succ) &= 1; & P_{cb}(\succ) &= 0. \end{aligned}$$

However, with  $\tau = .10$ , and  $\Delta$  the supremum distance, we obtain the constraints that each  $|P_\theta - V_\theta(\succ)| \leq 0.10, \forall \theta \in \Theta$ , that is,

$$\begin{aligned} B_{(\{a,b,c\},a)}(\succ) &\geq .9; & B_{(\{a,b,c\},b)}(\succ) &\leq .1; & B_{(\{a,b,c\},c)}(\succ) &\leq .1; \\ BW_{(\mathcal{A},a,b)}(\succ) &\leq .1; & BW_{(\mathcal{A},a,c)}(\succ) &\leq .1; & BW_{(\mathcal{A},a,d)}(\succ) &\geq .9; \\ BW_{(\mathcal{A},b,c)}(\succ) &\leq .1; & BW_{(\mathcal{A},b,d)}(\succ) &\leq .1; & BW_{(\mathcal{A},c,d)}(\succ) &\leq .1; \\ P_{bc}(\succ) &\geq .9; & P_{cb}(\succ) &\leq .1, \end{aligned}$$

in addition to the standard constraints for well-defined choice probabilities. When  $\Delta$  is the supremum distance, one of the two binary choice probabilities may be omitted because it is redundant, even in the distance-based specification. In general, however, as we consider other metrics, none of the coordinates may be redundant. (The coordinate redundancies become more important when we discuss “mixture models” in the next section.)

This completes our discussion of probabilistic choice models for a single fixed binary preference relation. The binary preference relation  $\succ$  under consideration could, in an experiment, be a stated hypothesis. It could also, however, be an unknown. In the latter case, the experimenter could specify several binary preference relations and take the union of their probabilistic specifications, such as a union of constant error models, or a union of distance specifications of those preference relations, then find the best-fitting model in that union of models. Even a union of such models with  $\succ$  an unknown “free parameter” of the model, still represents the preference relation  $\succ$  as unique and fixed. These models are subject to intense ongoing research (see, e.g., Guo and Regenwetter, 2014, for a recent example).

The best-known classical example of such a model is the *weak utility* model, also known as *weak stochastic transitivity*, for a complete collection of binary choice probabilities, according to which, writing  $u(x)$  for the utility of  $x$ ,

$$P_{xy} \geq \frac{1}{2} \quad \Leftrightarrow \quad \neg[y \succ x] \quad \Leftrightarrow \quad u(x) \geq u(y). \quad (7.6)$$

If the utility function  $u$  is one-to-one, but we do not otherwise fix  $u$ , then the preference relation  $\succ$  can be any strict linear order on  $\mathcal{A}$ , and this is a special case of the distance-based specification (7.5) for binary choice probabilities, where  $\Delta$  is the supremum distance and  $\tau = \frac{1}{2}$  (Regenwetter *et al.*, 2014).

We now move from a fixed, possibly unknown, preference to variable preferences. The following section heavily relies on the terminology and acronyms introduced in Definition 7.8.

### 7.3.3 Choice probabilities for varying or uncertain preferences

**Definition 7.16** Consider a finite master set  $\mathcal{A}$  of two or more choice alternatives. Consider a collection  $\succ_1, \succ_2, \dots, \succ_k$  of binary preference relations on  $\mathcal{A}$ . A collection of probabilities  $0 \leq P(\succ_j) \leq 1$  for  $(1 \leq j \leq k)$  with

$$\sum_{j=1}^k P(\succ_j) = 1,$$

is a *mixture* of preferences.

We can now define what it means for choices to be induced by a mixture of (i.e., a probability distribution over) preferences. When asking a person to provide options in certain rank positions in a given available set  $X$ , there are, in

general, many latent preference states with those options in those rank positions in  $X$ . The total probability of those preference states is the probability of that response.

**Definition 7.17** Consider a finite master set  $\mathcal{A}$  of two or more choice alternatives. Let  $(P(\succ_j))_{1 \leq j \leq k}$  be a mixture of strict linear order preferences on  $\mathcal{A}$ . Let  $\Theta$  be a set containing queries  $\theta$  of the form  $\theta = (X, R_X)$  with each  $X \subseteq \mathcal{A}$  and  $|X| \geq 2$ , and each  $R_X \subseteq \{1, 2, \dots, |X|\}$ . Let  $(P_\theta)_{\theta \in \Theta}$  be a collection of response probability distributions.  $(P_\theta)_{\theta \in \Theta}$  is a *mixture* (synonymously, a *random preference* or a *random relation*) model of  $(P(\succ_j))_{1 \leq j \leq k}$  if  $\forall \theta = (X, R_X) \in \Theta$  and  $\forall \pi \in \Pi(R_X, X)$ ,

$$P_\theta(\pi) = \sum_{\substack{j \in \{1, 2, \dots, k\} \text{ s.t.} \\ \text{Rank}_{X, \succ_j}(\pi(i)) = i, \forall i \in R_X}} P(\succ_j). \quad (7.7)$$

In words, in a mixture model, the probability of response  $\pi$  to query  $\theta$  is the total (marginal) probability of all those preference states (permissible strict linear orders) that rank each option at the same rank position in  $X$  as does the response  $\pi$ .

Recall that we have viewed  $[0, 1]^n$  formally as a collection of mappings, and alternatively, as a collection of “vectors,” “ $n$ -tuples,” or “points.” Thinking of the elements as points connects naturally to geometric analysis using Euclidean geometry. A *convex set* in  $[0, 1]^n$  is a set of points in  $[0, 1]^n$  such that, given any two points  $x, y$  in that set, any point in the line segment joining  $x$  with  $y$  also belongs to that set. The *convex hull* of a collection of points is the (unique) smallest convex set containing those points.

**Lemma 7.6** Consider a finite master set  $\mathcal{A}$  of two or more choice alternatives. Let  $(P(\succ_j))_{1 \leq j \leq k}$  be a mixture of strict linear order preferences on  $\mathcal{A}$ . For any single strict linear order  $\succ_j$ , with  $1 \leq j \leq k$ , we obtain the vertex representation of  $\succ_j$  by concentrating all probability mass on  $\succ_j$  in the mixture model (7.7), i.e., by setting  $P(\succ_j) = 1$  and  $P(\succ_i) = 0$ ,  $i \neq j$ ,  $1 \leq i \leq k$ . Hence, the mixture model (7.7) of  $(P(\succ_j))_{1 \leq j \leq k}$  forms the convex hull of the vertices  $\mathcal{V}_\Theta(\succ_j)_{1 \leq j \leq k}$ , where  $V_\theta(\succ_j)$ , with  $\theta = (X, \pi)$ ,  $|X| \geq 2$ , has the 0/1-coordinates given by

$$V_{(X, \pi)}(\succ_j) = \begin{cases} 1 & \text{if } \text{Rank}_{X, \succ_j}(\pi(i)) = i, \forall i \in R_X, \\ 0 & \text{otherwise.} \end{cases}$$

For example, suppose  $\mathcal{A} = \{a, b, c\}$ , let  $\succ_1 = \{(a, b), (a, c), (b, c)\}$ , i.e., the alphabetical ranking of  $a, b$  and  $c$ , set  $P(\succ_1) = 1$  in the mixture model (7.7), and consider best choice. Then

$$\begin{aligned} B_{\{a, b\}}(a) &= 1, B_{\{a, c\}}(a) = 1, B_{\{b, c\}}(b) = 1, B_{\{a, b, c\}}(a) = 1, \\ B_{\{a, b\}}(b) &= 0, B_{\{a, c\}}(c) = 0, B_{\{b, c\}}(c) = 0, B_{\{a, b, c\}}(b) = 0, B_{\{a, b, c\}}(c) = 0. \end{aligned}$$

If, instead,  $P(\succ_2) = 1$ , where  $\succ_2 = \{(a, b), (c, a), (c, b)\}$ , then

$$\begin{aligned} B_{\{a,b\}}(a) &= 1, B_{\{a,c\}}(c) = 1, B_{\{b,c\}}(c) = 1, B_{\{a,b,c\}}(a) = 1, \\ B_{\{a,b\}}(b) &= 0, B_{\{a,c\}}(a) = 0, B_{\{b,c\}}(b) = 0, B_{\{a,b,c\}}(b) = 0, B_{\{a,b,c\}}(c) = 0. \end{aligned}$$

If  $0 < P(\succ_1) = 1 - P(\succ_2) = p < 1$ , then

$$\begin{aligned} B_{\{a,b\}}(a) &= 1, B_{\{a,b\}}(b) = 0, B_{\{a,b,c\}}(c) = 0 \\ B_{\{a,c\}}(a) &= p, B_{\{b,c\}}(b) = p, B_{\{a,b,c\}}(a) = p, \\ B_{\{a,c\}}(c) &= 1 - p, B_{\{b,c\}}(c) = 1 - p, B_{\{a,b,c\}}(b) = 1 - p, \end{aligned}$$

so, as we vary  $0 \leq p \leq 1$ , the best-choice probabilities form the convex hull of the two vertices associated with  $\succ_1$  and  $\succ_2$ . Likewise, for general queries and general sets  $\mathcal{A}$  with two or more elements, the response probabilities in the mixture model (7.7) form the convex hull of the vertices associated with the permissible preference states  $\succ_j$ ,  $1 \leq j \leq k$ .

We can now utilize some helpful tools from convex geometry (Ziegler, 1995). In the definition below, and in all places where we use the concepts, we assume a finite-dimensional space.

**Definition 7.18** A *convex polytope* is the convex hull of a finite collection of points. Among all collections of points whose convex hull forms the given polytope, there is a (unique) minimal one; its elements are the *vertices* of the polytope. The set of vertices forms the *vertex description* of the polytope. A polytope is alternatively characterized as the intersection of finitely many (closed) half-spaces, where a *half-space* is formed by all the solutions to a given linear inequality. Hence, a polytope is the set of solutions of a finite system of linear inequalities. The set of points in the polytope that satisfy one or more fixed such inequalities with equality forms a *face* of the convex polytope. A polytope is *full-dimensional* in a given space, if no strict subspace also contains the polytope. Among all finite collections of linear inequalities whose set of solutions is the polytope, there is a (unique) minimal one; its elements are the *facet-defining inequalities*. The set of points in the polytope that satisfy a fixed facet-defining inequality with equality forms a *facet* of the polytope. The facets are also the faces of maximal dimension. The polytope is the solution set of its facet-defining inequalities; the latter forms its *minimal linear description*, also called its *facet description*.

For example, in a three-dimensional unit cube,  $[0, 1]^3$ , the vertices are the eight points with 0/1 coordinates in  $\mathbb{R}^3$ , namely  $(0, 0, 0), (1, 0, 0), \dots, (1, 1, 1)$ . The cube is a polytope formed by the convex hull of these eight vertices, but it is also the polytope formed by the intersection of six half-spaces. Writing  $p = (p_1, p_2, p_3) \in [0, 1]^3$  for a point in this cube, the polytope is characterized by six facet-defining inequalities

$$0 \leq p_1 \leq 1; 0 \leq p_2 \leq 1, 0 \leq p_3 \leq 1.$$

The facets are the six square-shaped, two-dimensional surfaces that we obtain by taking the points in the cube that satisfy one of these inequalities with equality, say, for instance,  $0 = p_1$ . The one-dimensional edges of the cube and the zero-dimensional vertices of the cube form faces of lower dimension that we obtain by taking points in the cube that simultaneously satisfy combinations of facet-defining inequalities with equality, say,  $[p_1 = 1] \wedge [p_2 = 1]$  (line segment, i.e., one-dimensional face) or  $[p_1 = 1] \wedge [p_2 = 0] \wedge [p_3 = 0]$  (vertex, i.e., zero-dimensional face).

Next we consider some examples of mixture models and discuss their minimal descriptions briefly.

### 7.3.3.1 Binary choice probabilities induced by rankings (strict linear orders)

**Definition 7.19** A complete system of binary choice probabilities is *induced by rankings (strict linear orders)* if there exists a probability distribution on  $\mathcal{SLO}_{\mathcal{A}}$  with  $P(\succ)$  denoting the probability of  $\succ \in \mathcal{SLO}_{\mathcal{A}}$ , such that,

$$P_{xy} = \sum_{\substack{\succ \in \mathcal{SLO} \\ a \succ b}} P(\succ) \quad (\forall x, y \in \mathcal{A}, x \neq y). \quad (7.8)$$

This model is also called the *linear ordering model for binary choice probabilities*.

Note that the left-hand side of the model (7.8) denotes the probabilities of observable binary choices, whereas the right-hand side refers to probabilities of latent (i.e., not directly observed) linear orders.

Models like the linear ordering model have a long tradition in several disciplines. We now consider what constraints one can derive from such models, e.g., in order to test them empirically. The linear ordering model (7.8) states that the binary choice probability of choosing  $x$  over  $y$  is the marginal probability of all linear orders in which  $x \succ y$ . If we concentrate on  $|\mathcal{A}| = 3$ , say,  $\mathcal{A} = \{a, b, c\}$ , there are six strict linear orders, which we can also write as rankings  $abc, bac, acb, cba, cab, bca$ . If we write  $P_{xyz}$  instead of  $P(\{(x, y), (x, z), (y, z)\})$  for the probability of latent ranking  $xyz$ , then  $P_{ab} = P_{abc} + P_{acb} + P_{cab}$ , according to (7.8). No matter what probability distribution we use in the right side of (7.8), as  $P_{abc} + P_{bac} + P_{acb} + P_{cba} + P_{cab} + P_{bca} = 1$ , it must be the case that

$$\begin{aligned} & P_{ab} + P_{bc} - P_{ac} \\ &= (P_{abc} + P_{acb} + P_{cab}) + (P_{abc} + P_{bac} + P_{bca}) - (P_{abc} + P_{bac} + P_{acb}) \\ &= P_{abc} + P_{cab} + P_{bca} \leq 1. \end{aligned}$$

Similarly,  $P_{ac} + P_{cb} - P_{ab} \leq 1$ . It is well known that the linear ordering model for binary choice probabilities for  $|\mathcal{A}| \leq 5$  holds if and only if the following *triangle inequalities* hold:

$$P_{xy} + P_{yz} - P_{xz} \leq 1, \quad (\forall \text{ distinct } x, y, z \in \mathcal{A}). \quad (7.9)$$

Now consider all linear orders on  $\mathcal{A}$  ( $|\mathcal{A}| \geq 2$ ), and for each linear order  $\succ \in \mathcal{SLO}_{\mathcal{A}}$ , consider the vertex representation for  $\succ$  for a complete system of

binary choice probabilities. The permissible binary choice probabilities under the linear ordering model (7.8) form the convex hull of the vertices  $\mathcal{V}(\succ)$  over all  $\succ \in \mathcal{SLO}$ , where, for each  $\succ$  and each pair of distinct  $x, y$ , we consider one of the coordinates

$$V_{xy}(\succ) = \begin{cases} 1 & \text{if } x \succ y, \\ 0 & \text{if } y \succ x \end{cases} \quad \text{or} \quad V_{yx}(\succ) = \begin{cases} 1 & \text{if } y \succ x, \\ 0 & \text{if } x \succ y, \end{cases}$$

and not the other. The linear ordering model of binary choice probabilities forms a convex polytope, because it can be restated as the convex hull of those vertices. A minimal description of the associated *linear ordering polytope of binary choice probabilities*,  $\mathcal{SLOP}_{\mathcal{A}}(\text{Binary})$ , is currently known only for small  $|\mathcal{A}|$  and for complete collections of binary choice probabilities. In particular, the triangle inequalities form a complete minimal description via facet-defining inequalities of  $\mathcal{SLOP}_{\mathcal{A}}(\text{Binary})$ , for  $|\mathcal{A}| \leq 5$  (see, e.g., Fiorini, 2001a; Koppen, 1995). Reinelt (1993) reviews complete descriptions for  $|\mathcal{A}| \leq 7$ . No such complete description is known for large  $|\mathcal{A}|$ , although many complicated necessary conditions are known (Charon and Hudry, 2010; Doignon *et al.*, 2007).

Regenwetter *et al.* (2011a) provided and discussed the first published quantitative statistical test of the triangle inequalities (7.9) on empirical data (using order-constrained inference methods). They used three sets of five choice alternatives that were intermixed with each other in a single experiment; one of these sets was a replication of a seminal experiment by Tversky (1969). The model fit extremely well on 17 of 18 participants, at a significance level of  $\alpha = 0.05$ . The one poor fit is consistent with the Type-I error rate of the test, if all participants were consistent with the model. Regenwetter *et al.* (2011a) concluded from their experiment that it is not so straightforward to demonstrate violations of transitivity when allowing for variability in preference, not just variability in choice (i.e., probabilistic preferences rather than probabilistic responses).

It is well known that either the weak utility model or the linear ordering model can hold without the other holding, but they can also both hold at the same time. Luce and Suppes (1965) discussed this and various other relationships among probabilistic choice models. One of the ongoing areas of related research, which we do not discuss here, is to further investigate the relationship between the models we have reviewed and “nonparametric Fechnerian” probabilistic choice models. By the latter, we mean models in which the binary choice probability can be any monotonic function of the strength of preference in favor of the preferred option.

Tversky’s (1969) original experiment aimed to demonstrate intransitive preferences by attempting to show violations of weak stochastic transitivity (7.6). Iverson and Falmagne (1985) had found Tverky’s (1969) data to be generally consistent with weak stochastic transitivity (7.6) when using a suitable order-constrained test. Regenwetter *et al.* (2010) found that weak stochastic transitivity (7.6) was also descriptive of their data.

### 7.3.3.2 Best choice probabilities induced by rankings (strict linear orders)

**Definition 7.20** A complete collection of best choice probabilities is *induced by rankings (strict linear orders)* if there exists a probability distribution on  $\mathcal{SLO}_{\mathcal{A}}$  with  $P(\succ)$  denoting the probability of  $\succ \in \mathcal{SLO}_{\mathcal{A}}$ , such that,

$$B_X(x) = \sum_{\substack{\succ \in \mathcal{SLO}_{\mathcal{A}} \\ Rank_{X,\succ}(x)=1}} P(\succ), \quad (\forall x \in X \subseteq \mathcal{A}). \quad (7.10)$$

In words, the probability that a decision maker indicates that  $x$  is best in  $X$  is the total probability of all those rankings over  $\mathcal{A}$  (strict linear orders on  $\mathcal{A}$ ) in which  $x$  is preferred to all other elements of  $X$ . The representation (7.10) is also called a *linear ordering model of best choice probabilities*. In Section 7.5, Theorem 7.10 presents a random utility representation of best choice probabilities induced by rankings, and Section 7.5.3 includes discussion of a stochastic process model for best choice, including response times.

Falmagne (1978) provided a set of necessary and sufficient conditions for the model of Definition 7.20, regardless of  $|\mathcal{A}|$ , and Barberá and Pattanaik (1986) rediscovered the result. Colonius (1984) and Fiorini (2004) developed shorter proofs of Falmagne's result using polyhedral combinatorics and related methods. The linear ordering model of best-choice probabilities forms a convex polytope, because it can be restated as the convex hull of the vertices  $VB(\succ)$ , for all  $\succ \in \mathcal{SLO}_{\mathcal{A}}$ . A minimal description of the associated *linear ordering polytope of best-choice probabilities*, which we denote as  $\mathcal{SLOP}_{\mathcal{A}}(\text{Best})$ , was given by Fiorini (2004). Falmagne's (1978) result used conditions in the form

$$\sum_{U : X \subseteq U \subseteq \mathcal{A}} (-1)^{|U \setminus X|} P_U(x) \geq 0, \quad (\forall X \subseteq \mathcal{A}, X \neq \emptyset, \forall x \in X),$$

where  $|U \setminus X|$  denotes the cardinality of the set  $U \setminus X = \{z \in U \text{ with } z \notin X\}$ . Provided that  $|\mathcal{A}| \geq 3$ , and for a given  $X$ , the corresponding inequality is facet-defining if and only if  $1 \leq |X| < |\mathcal{A}|$  (see Fiorini, 2004).

Exactly parallel results can be stated for the analogous *linear ordering polytope of worst-choice probabilities*, i.e., where

$$W_X(y) = \sum_{\substack{\succ \in \mathcal{SLO}_{\mathcal{A}} \\ Rank_{X,\succ}(y)=|X|}} Q(\succ), \quad (\forall x \in X \subseteq \mathcal{A}). \quad (7.11)$$

Note that now the sum is over final positions in the rank orders, i.e., over  $Rank_{X,\succ}(y) = |X|$ . de Palma *et al.* (2015) present results relating best and worst choice probabilities to each other when they are induced by the same rankings – that is, when  $P \equiv Q$  in (7.10) and (7.11).

Section 7.5, Theorem 7.10, discusses random utility formulations of the above material.

### 7.3.3.3 Best-worst choice probabilities induced by rankings (strict linear orders)

**Definition 7.21** A complete collection of best-worst choice probabilities is *induced by rankings (strict linear orders)* if there exists a probability distribution on  $\mathcal{SLO}_{\mathcal{A}}$  with  $P(\succ)$  the probability of  $\succ \in \mathcal{SLO}_{\mathcal{A}}$ , such that,

$$BW_X(x, y) = \sum_{\substack{\succ \in \mathcal{SLO}_{\mathcal{A}} \\ Rank_{X, \succ}(1)=x, Rank_{X, \succ}(|X|)=y}} P(\succ), \quad (\forall x, y \in X \subseteq \mathcal{A}, x \neq y). \quad (7.12)$$

In words, the probability that a decision maker indicates that  $x$  is best in  $X$  and  $y$  is worst in  $X$  is the total probability of all those rankings over  $\mathcal{A}$  (strict linear orders on  $\mathcal{A}$ ) in which  $x$  is preferred to all other elements of  $X$  and in which also all other elements of  $X$  are preferred to  $y$ . The representation (7.12) is also called a *linear ordering model of best-worst choice probabilities*. In Section 7.5, Theorem 7.11 presents a random utility representation of best-worst choice probabilities induced by rankings and, Section 7.5.3 includes discussion of a stochastic process model for best-worst choice, including response times.

For  $|\mathcal{A}| = 3$ , the model (7.12) yields, for distinct  $x, y, z$ ,

$$\begin{aligned} BW_{\{x,y,z\}}(x, y) &= P(\{(x, y), (x, z), (y, z)\}), \\ BW_{\{x,y\}}(x, y) &= BW_{\{x,y,z\}}(x, y) + BW_{\{x,y,z\}}(x, z) + BW_{\{x,y,z\}}(z, y). \end{aligned}$$

As we now show, by example and known results, the linear ordering model of best-worst choice probabilities raises interesting open theoretical questions which, in turn, lead to empirical question of relevance to best-worst scaling (Louviere *et al.*, 2015). We consider  $|\mathcal{A}| = 4$ , say,  $\mathcal{A} = \{a, b, c, d\}$  for the remainder of this subsection. Writing  $P_{wxyz}$  for the probability of the ranking  $wxyz$ , i.e.,

$$P_{wxyz} = P(\{(w, x), (w, y), (w, z), (x, y), (x, z), (y, z)\}),$$

the linear ordering model of best-worst choice probabilities states, e.g.,

$$\begin{aligned} BW_{\{a,b\}}(a, b) &= P_{abcd} + P_{abdc} + P_{adbc} + P_{acbd} + \cdots + P_{dcab}, \\ BW_{\{a,b,c\}}(a, c) &= P_{abcd} + P_{abdc} + P_{adbc} + P_{dabc}, \\ BW_{\{a,b,c,d\}}(a, d) &= P_{abcd} + P_{acbd}. \end{aligned}$$

It follows, e.g., that for all distinct  $x, y, z, w$  (relabelings of  $\mathcal{A}$ ),

$$\begin{aligned} BW_{\{x,y\}}(x, y) &= P_{wzxy} + P_{zwxy} + P_{zxwy} + P_{zywx} \\ &\quad + P_{wxzy} + P_{xwzy} + P_{xzwy} + P_{xzyw} \\ &\quad + P_{wxyz} + P_{xwyz} + P_{xywz} + P_{xyzw} \\ &= BW_{\{z,x,y\}}(z, y) + BW_{\{z,x,y\}}(x, y) + BW_{\{z,x,y\}}(x, z). \end{aligned} \quad (7.13)$$

Likewise, for all distinct  $x, y, z, w$  (relabelings of  $\mathcal{A}$ ),

$$\begin{aligned} & BW_{\{w,x,z\}}(x, z) + BW_{\{w,x,z\}}(x, w) + BW_{\{w,x,z\}}(z, w) \\ & = BW_{\{w,y,z\}}(y, z) + BW_{\{w,y,z\}}(y, w) + BW_{\{w,y,z\}}(z, w), \end{aligned} \quad (7.14)$$

which we state without proof.

Similarly, one can derive various inequality constraints. For example, the model directly implies that, for all distinct  $x, y, z, w$  (relabelings of  $\mathcal{A}$ ),

$$BW_{\{z,w,x,y\}}(w, y) \leq BW_{\{z,w,y\}}(w, y), \quad (7.15)$$

and also

$$\begin{aligned} BW_{\{x,y,w\}}(y, w) &= P_{zyxw} + P_{yzxw} + P_{yxzw} + P_{yxwz} \\ &\leq P_{zyxw} + (P_{zxyw}) + P_{yzxw} + P_{yxzw} \\ &\quad + P_{yxwz} + (P_{ywzx}) \\ &= BW_{\{x,y,z,w\}}(z, w) + BW_{\{x,y,z,w\}}(y, w) + BW_{\{x,y,z,w\}}(y, z). \end{aligned} \quad (7.16)$$

Consider all linear orders on  $\mathcal{A}$ , and for each linear order  $\succ \in \mathcal{SLO}_{\mathcal{A}}$ , and recall the vertex representation for  $\succ$  for a complete system of best-worst choice probabilities, with coordinates

$$VBW_{(X,x,y)}(\succ) = \begin{cases} 1 & \text{if } Rank_{X,\succ}(x) = 1 \wedge Rank_{X,\succ}(y) = |X|, \\ 0 & \text{if } Rank_{X,\succ}(x) \neq 1 \vee Rank_{X,\succ}(y) \neq |X|. \end{cases}$$

As we have seen in Remark 7.1, there are redundant response probabilities in best-worst choice because the choice probabilities for each fixed available set  $X$  sum to one:  $\sum_{\substack{x \neq y \\ x, y \in X}} BW_X(x, y) = 1, \forall X \subseteq \mathcal{A}$ . After dropping one redundant coordinate for each  $X$ , the collections of all possible best-worst choice probabilities on  $\mathcal{A} = \{a, b, c, d\}$  form a unit hypercube of dimension  $\sum_{k=2}^4 \binom{4}{k}[k(k-1)-1] = \binom{4}{2}[2(2-1)-1]\binom{4}{3}[3(3-1)-1]\binom{4}{4}[4(4-1)-1] = 37$ .

We can make various observations from the perspective of polyhedral combinatorics (see Definition 7.18). Consider the vertex representations of the linear orders on  $\mathcal{A} = \{a, b, c, d\}$  in this coordinate system. The convex hull of the 24 vertices  $VBW(\succ)$  (over all  $\succ \in \mathcal{SLO}_{\mathcal{A}}$ ) forms the linear ordering model of best-worst choice probabilities. Even after leaving out one coordinate per set  $X$ , the resulting polytope is not full dimensional. This follows directly, e.g., from Equations (7.13)–(7.14). We determined this description using public domain software, PORTA, the POlyhedron Representation Transformation Algorithm, of T. Christof and A. Löbel, 1997.<sup>4</sup> We found that there are 15 equations (including 5 nonredundant equations like Equation (7.13) and 4 nonredundant equations like Equation (7.14), yielding a polytope of dimension 22.

Because every convex polytope can be represented either as the convex hull of its vertices (here, 24 of the  $2^{22}$  vertices of a 22-dimensional unit hypercube), or as

<sup>4</sup> <http://www.iwr.uni-heidelberg.de/groups/comopt/software/PORTA/>. We thank Ying Guo for carrying out this analysis using PORTA.

the system of facet-defining inequalities, we can characterize all best-worst choice probabilities induced by rankings (on  $\mathcal{A} = \{a, b, c, d\}$ ) through such a minimal facet-defining system of affine inequalities. An analysis using PORTA yielded a system of 144 facet-defining inequalities, including 11 “trivial” inequalities of the form  $BW_X(x, y) \geq 0$ , 11 nonredundant inequalities of the form of Inequality 7.15 and 10 nonredundant inequalities of the form given in Inequality 7.16.

To our knowledge, a complete minimal description of the *linear ordering polytope of best-worst choice probabilities*, which we will denote  $\mathcal{SLOP}_{\mathcal{A}}(\text{Best-Worst})$ , is an open problem for  $\mathcal{A} > 4$ . Doignon *et al.* (2015) provide some preliminary, but general, results.

We have presented a number of different probabilistic choice models for choices induced by rankings, i.e., by strict linear orders. It is natural to consider alternative models for the latent binary preferences, such as strict partial orders, strict weak orders, semiorders, and interval orders. We consider one empirical paradigm for this, namely ternary paired comparison probabilities (Definition 7.5). We rely again on the terminology and acronyms introduced in Definition 7.8.

#### 7.3.3.4 Ternary paired-comparison probabilities induced by strict partial orders, interval orders, semiorders, or strict weak orders

**Definition 7.22** Consider a system of ternary paired-comparison probabilities on a finite set  $\mathcal{A}$ . The ternary paired-comparison probabilities are (1) *induced by strict partial orders*, or (2) *induced by interval orders*, or (3) *induced by semiorders*, or (4) *induced by strict weak orders*, if there exists a probability distribution  $P$  on (1)  $\mathcal{R} = \mathcal{SPO}_{\mathcal{A}}$ , or (2)  $\mathcal{R} = \mathcal{IO}_{\mathcal{A}}$ , or (3)  $\mathcal{R} = \mathcal{SO}_{\mathcal{A}}$ , or (4)  $\mathcal{R} = \mathcal{SWO}_{\mathcal{A}}$  such that,  $\forall x, y \in \mathcal{A}$ , with  $x \neq y$ , and writing  $P(\succ)$  for the probability of any  $\succ \in \mathcal{R}$ ,

$$T_{xy} = \sum_{\substack{\succ \in \mathcal{R} \\ x \succ y}} P(\succ). \quad (7.17)$$

As we review next, these models have been studied for small  $|\mathcal{A}|$ .

**Lemma 7.7** (Fiorini, 2001b) *Ternary paired-comparison probabilities on  $\mathcal{A} = \{a, b, c, d\}$  are induced by strict partial orders if and only if the following list of facet-defining inequalities for the partial order polytope on four objects are satisfied (for all suitable distinct relabelings  $w, x, y, z$  of  $a, b, c, d$ ):*

- $PO_1 : T_{xy} \geq 0,$
- $PO_2 : T_{xy} + T_{yx} \leq 1,$
- $PO_3 : T_{xy} + T_{yz} - T_{xz} \leq 1,$
- $PO_4 : T_{xy} + T_{yz} + T_{zx} - T_{yx} - T_{zy} - T_{xz} \leq 1,$
- $PO_5 : T_{xy} + T_{yz} + T_{zw} + T_{wx} - T_{yx} - T_{xz} - T_{wy} \leq 2,$
- $PO_6 : T_{xy} + T_{yz} + T_{zx} + T_{yw} + T_{wy} - T_{zy} - T_{yx} - T_{wz} - T_{wx} \leq 2,$
- $PO_7 : T_{xy} + T_{yz} + T_{xw} + T_{wz} - T_{yw} - T_{wx} - 2T_{xz} \leq 2,$
- $PO_8 : T_{xy} + T_{yz} + T_{zy} + T_{yw} - 2T_{yx} - 2T_{xz} - 2T_{zw} - 2T_{wy} \leq 3.$

The partial order polytope for four choice alternatives is a 12-dimensional polytope with 219 vertices and with 128 different facets. Taking into account all relabelings, it has 12 facets of type  $PO_1$ , 6 of type  $PO_2$ , 24 of type  $PO_3$ , 8 of type  $PO_4$ , 24 of type  $PO_5$ , 24 of type  $PO_6$ , 24 of type  $PO_7$ , as well as 6 of type  $PO_8$ .

**Lemma 7.8** (Regenwetter and Davis-Stober, 2011) Ternary paired-comparison probabilities on  $\mathcal{A} = \{a, b, c, d\}$  are induced by interval orders if and only if the following list of facet-defining inequalities for the interval order polytope on four objects are satisfied (for all suitable distinct relabelings  $x, y, z, v$  of  $a, b, c, d$ ):

$$\begin{aligned}
 IO_1 : & T_{xy} \geq 0, \\
 IO_2 : & T_{xy} + T_{yx} \leq 1, \\
 IO_3 : & T_{xy} + T_{yz} - T_{xz} \leq 1, \\
 IO_4 : & T_{xy} + T_{zv} - T_{xv} - T_{zy} \leq 1, \\
 IO_5 : & T_{xy} + T_{yz} + T_{zx} - T_{xz} - T_{zy} - T_{yx} \leq 1, \\
 IO_6 : & T_{xy} + T_{yz} + T_{zv} - T_{xz} - T_{xv} - T_{yv} - T_{zy} \leq 1, \\
 IO_7 : & T_{xy} + T_{yx} + T_{vz} - T_{xz} - T_{yz} - T_{vx} - T_{vy} \leq 1, \\
 IO_8 : & T_{xy} + T_{yz} + T_{zv} + T_{vx} - T_{xv} - T_{vz} - T_{zy} - T_{yx} \\
 & \quad - T_{xz} - T_{zx} - T_{yv} - T_{vy} \leq 1, \\
 IO_9 : & T_{xy} + T_{yx} + T_{zv} + T_{vz} - T_{xz} - T_{xv} - T_{yz} - T_{yv} \\
 & \quad - T_{zx} - T_{zy} - T_{vx} - T_{vy} \leq 1, \\
 IO_{10} : & T_{xy} + T_{yx} + T_{yz} + T_{zv} + T_{vy} - T_{xz} - T_{zy} - T_{yv} - T_{vx} \leq 2, \\
 IO_{11-12} : & -2 \leq 2T_{xy} + T_{yz} + T_{zv} + T_{vx} - T_{xv} - T_{xz} - T_{zy} - T_{yx} \\
 & \quad - T_{xv} - T_{vy} \leq 2, \\
 IO_{13-14} : & -3 \leq 2T_{xy} + 2T_{yz} + 2T_{zv} + 2T_{vx} - T_{yx} - T_{xv} - T_{vz} \\
 & \quad - T_{zy} - T_{xz} - T_{zx} - T_{yv} - T_{vy} \leq 3.
 \end{aligned}$$

This is a 12-dimensional polytope with 207 vertices and 191 facets. Taking into account all relabelings, it has 12 facets of type  $IO_1$ , 6 of type  $IO_2$ , 24 of type  $IO_3$ , 12 of type  $IO_4$ , 8 of type  $IO_5$ , 24 of type  $IO_6$ , 12 of type  $IO_7$ , 6 of type  $IO_8$ , 3 of type  $IO_9$ , 24 of type  $IO_{10}$ , 24 of type  $IO_{11}$ , 24 of type  $IO_{12}$ , 6 of type  $IO_{13}$ , and 6 of type  $IO_{14}$ .

**Lemma 7.9** (Regenwetter and Davis-Stober, 2011) Ternary paired-comparison probabilities on  $\mathcal{A} = \{a, b, c, d\}$  that are induced by semiorders form a 12-dimensional polytope with 183 vertices, characterized by 563 facets. These

include, e.g., (for all suitable distinct relabelings  $x, y, z, v$  of  $a, b, c, d$ ):

$$\begin{aligned}
 SO_1 : & T_{xy} \geq 0, \\
 SO_2 : & T_{xy} + T_{yx} \leq 1, \\
 SO_3 : & T_{xy} + T_{yz} - T_{xz} \leq 1, \\
 SO_4 : & T_{xy} + T_{zv} - T_{xv} - T_{zy} \leq 1, \\
 SO_5 : & T_{xy} + T_{yz} - T_{xv} - T_{vz} \leq 1, \\
 SO_6 : & T_{xy} + T_{yz} + T_{zx} - T_{xz} - T_{zy} - T_{yx} \leq 1, \\
 SO_7 : & T_{xy} + T_{yz} + T_{zv} - T_{xz} - T_{zy} - T_{xv} - T_{yx} \leq 1, \\
 SO_8 : & T_{xy} + T_{yz} + T_{zv} - T_{xz} - T_{zy} - T_{xv} - T_{yv} \leq 1, \\
 SO_9 : & T_{xy} + T_{yz} + T_{zv} - T_{vz} - T_{zy} - T_{xv} - T_{yv} \leq 1, \\
 SO_{10} : & T_{xy} + T_{yx} + T_{zv} - T_{xv} - T_{yv} - T_{zx} - T_{zy} \leq 1, \\
 & \vdots \\
 SO_{30} : & 2T_{xy} + 2T_{yz} + 2T_{zv} + T_{vx} - T_{xz} - T_{zx} - 2T_{xv} \\
 & \quad - T_{yx} - T_{yv} - T_{vy} - 2T_{zy} - T_{vz} \leq 2, \\
 SO_{31} : & 2T_{xy} + 2T_{yz} + 2T_{zv} + T_{vx} + T_{zx} + T_{vy} - 2T_{xz} - T_{xv} \\
 & \quad - T_{yx} - 2T_{yv} - 4T_{zy} - T_{vz} \leq 3.
 \end{aligned}$$

Taking into account all relabelings, there are 12 facets of type  $SO_1$ , 6 of type  $SO_2$ , 24 of type  $SO_3$ , 12 of type  $SO_4$ , 24 of type  $SO_5$ , 8 of type  $SO_6$ , 24 of type  $SO_7$ , 24 of type  $SO_8$ , 24 of type  $SO_9$ , 12 of type  $SO_{10}$ , ..., 24 of type  $SO_{30}$ , and 24 facets of type  $SO_{31}$ .

The most extensively studied model of ternary paired-comparison probabilities is the strict weak order polytope characterizing all ternary paired-comparison probabilities induced by strict weak orders. This model has been studied for  $|\mathcal{A}| \leq 5$ .

**Lemma 7.10** (Regenwetter and Davis-Stober, 2012) *The collection of all ternary paired comparison probabilities induced by strict weak orders on a five element set  $\mathcal{A} = \{a, b, c, d, e\}$  is a convex polytope characterized by 541 distinct vertices of the 20-dimensional unit hypercube. The facet description providing a minimal nonredundant collection of affine inequalities consists of 75,834 distinct inequalities. These include, e.g., (for all suitable distinct relabelings  $v, w, x, y, z$  of  $a, b, c, d, e$ ):*

$$\begin{aligned}
 T_{xy} &\geq 0, \\
 T_{xy} + T_{yx} &\leq 1, \\
 T_{xy} + T_{yz} - T_{xz} &\leq 1, \\
 3(T_{xy} - T_{yx} + T_{vy} - T_{yv} + T_{zw} + T_{wz}) + T_{xz} + T_{zx} + T_{yz} - T_{zy} - T_{vw} + T_{wv} \\
 &+ 3T_{yw} - T_{wy} - 3T_{xw} + T_{wx} - 3T_{vx} - T_{xv} - 3T_{vz} - T_{zv} \leq 4.
 \end{aligned}$$

Regenwetter and Davis-Stober (2012) provide all 75,834 facet-defining inequalities and they report a successful fit of this polytope to empirical ternary paired comparison data using order-constrained likelihood methods.

To our knowledge, a complete minimal description of the *strict weak order polytope of ternary paired-comparison probabilities*, which we will denote  $\text{SWOP}_{\mathcal{A}}(\text{Ternary})$ , is an open problem for  $\mathcal{A} > 5$ . For a recent discussion of the polytope's known facet structure, see Doignon and Fiorini (2002).

We now briefly consider an alternative type of ternary choice probabilities, namely we build on Definition 7.5 to define ternary-paired comparison probabilities that are consistent with compatible simple lexicographic semiorders and state a theorem about this model.

**Definition 7.23** Consider a finite set  $\mathcal{A}$  and a system of ternary paired comparison probabilities  $(T_{x,y})_{\substack{x,y \in \mathcal{A} \\ x \neq y}}$  on  $\mathcal{A}$ . Let  $\triangleright \in \mathcal{SLO}_{\mathcal{A}}$ . The ternary paired comparison probabilities are *induced by (compatible) simple lexicographic semiorders* if there exists a probability distribution  $P$  on  $\mathcal{SLSO}_{\mathcal{A}, \triangleright}$  such that,  $\forall x, y \in \mathcal{A}$ , with  $x \neq y$ , and writing  $P(\succ)$  for the probability of any compatible simple lexicographic semiorder  $\succ \in \mathcal{SLSO}_{\mathcal{A}, \triangleright}$ ,

$$T_{xy} = \sum_{\substack{\succ \in \mathcal{SLSO}_{\mathcal{A}, \triangleright} \\ x \succ y}} P(\succ). \quad (7.18)$$

Davis-Stober (2012, Theorems 1–3, Corollary 1) showed the following, using the famous “Catalan numbers”  $C_n$ .

**Theorem 7.1** Let  $\mathcal{A} = \{x_1, x_2, \dots, x_n\}$ , and let  $\triangleright \in \mathcal{SLO}_{\mathcal{A}}$  be a single fixed strict linear order on  $\mathcal{A}$ . Suppose that  $x_i \triangleright x_j \Leftrightarrow i < j$ . Let  $C_n = \frac{1}{n+1} \binom{2n}{n}$ . Then

$$|\mathcal{SLSO}_{\mathcal{A}, \triangleright}| = C_n C_{n+2} - C_{n+1}^2.$$

In particular, ternary paired-comparison probabilities induced by compatible simple lexicographic semiorders form a convex polytope of dimension  $n(n-1)$  with precisely  $C_n C_{n+2} - C_{n+1}^2$  many distinct vertices. This polytope has precisely  $\frac{5n^2-11n+8}{2}$  many facets. The facet-defining inequalities form seven distinct families and, for any finite  $n$ , are given by

$$\begin{aligned} T_{x_i x_j} - T_{x_i x_{j+1}} &\leq 0, & (\forall i, j \in \{1, 2, \dots, n\}, i < j < n), \\ T_{x_{i+1} x_j} - T_{x_i x_j} &\leq 0, & (\forall i, j \in \{1, 2, \dots, n\}, i + 1 < j), \\ T_{x_j x_i} + T_{x_i x_j} - T_{[j+1]i} - T_{x_i x_{j+1}} &\leq 0, & (\forall i, j \in \{1, 2, \dots, n\}, i < j < n), \\ T_{x_{i+1} x_j} + T_{x_j x_{i+1}} - T_{x_i x_j} - T_{x_j x_i} &\leq 0, & (\forall i, j \in \{1, 2, \dots, n\}, i + 1 < j), \\ -T_{x_i x_{i+1}} &\leq 0, & (\forall i \in \{1, 2, \dots, n-1\}), \\ -T_{x_i x_j} &\leq 0, & (\forall i, j \in \{1, 2, \dots, n\}, i > j), \\ T_{x_1 x_n} + T_{x_n x_1} &\leq 1. \end{aligned}$$

This completes our discussion of mixture models. There is an active ongoing research program to characterize mixture models of various kinds, in which only preferences consistent with a given theory of decision making (say, a certain theory

of risky or intertemporal choice) have positive probability (see Guo and Regenwetter, 2014; Regenwetter *et al.*, 2014, for recent examples).

## 7.4 Algebraic theories and their real-valued representations

### 7.4.1 Real-valued representations of binary preference relations

We now consider how some of the binary preference relations of Definition 7.8 are related to real-valued representations of utility.

**Theorem 7.2** *Let  $\mathcal{A}$  be a finite set of choice alternatives. Consider a binary preference relation  $\succ$ . Then the following hold:*

**Utility representations of strict weak orders:** *The relation  $\succ$  is a strict weak order on  $\mathcal{A}$  if and only if there exists a real-valued (utility) function  $u : \mathcal{A} \rightarrow \mathbb{R}$  which assigns values (utilities) to choice alternatives  $x \in \mathcal{A}$  via  $x \mapsto u(x)$  such that*

$$x \succ y \Leftrightarrow u(x) > u(y) \quad (\forall x, y \in \mathcal{A}, x \neq y). \quad (7.19)$$

**Utility representations of strict linear orders:** *The relation  $\succ$  is a strict linear order on  $\mathcal{A}$  if and only if there exists a one-to-one real-valued (utility) function  $u$  on  $\mathcal{A}$  satisfying Equivalence 7.19.*

**Utility representations of semiorders:** *The relation  $\succ$  is a semiorder on  $\mathcal{A}$  if and only if there exists a real-valued (utility) function  $u : \mathcal{A} \rightarrow \mathbb{R}$  which assigns real-valued utilities to choice alternatives  $x \in \mathcal{A}$  via  $x \mapsto u(x)$  and there exists a positive real number  $\varepsilon \in \mathbb{R}^{++}$ , such that*

$$x \succ y \Leftrightarrow u(x) > u(y) + \varepsilon \quad (\forall x, y \in \mathcal{A}, x \neq y).$$

**Utility representations of interval orders:** *The relation  $\succ$  is an interval order on  $\mathcal{A}$  if and only if there exists a real-valued (lower utility) function  $\ell : \mathcal{A} \rightarrow \mathbb{R}$  which assigns real-valued (lower bounds on) utilities to choice alternatives  $x \in \mathcal{A}$  via  $x \mapsto \ell(x)$  and a positive real-valued (threshold) function  $\tau : \mathcal{A} \rightarrow \mathbb{R}^{++}$  which assigns positive real values (thresholds of discrimination) to choice alternatives  $x \in \mathcal{A}$  via  $x \mapsto \tau(x) > 0$ , such that*

$$x \succ y \Leftrightarrow \ell(x) > \ell(y) + \tau(y) \quad (\forall x, y \in \mathcal{A}, x \neq y),$$

i.e.,

$$x \succ y \Leftrightarrow \ell(x) > u(y) \quad (\forall x, y \in \mathcal{A}, x \neq y),$$

with (upper utility) function  $u$  defined by:  $u(z) = \ell(z) + \tau(z)$ ,  $\forall z \in \mathcal{A}$  that assigns (upper bounds on) utilities to choice alternatives.

These relationships are well known (Roberts, 1979), and we leave the proof to the reader as an exercise.

### 7.4.2 Real-valued representations of weak orders over gambles

As mentioned in Section 7.1, Regenwetter and Davis-Stober (2012) investigated choices among pairs of gambles (lotteries) where decision makers were permitted to express a lack of preference for either lottery (see the screen shot of an experimental trial in Figure 7.1). This is an example from a major research area in the study of decision under uncertainty, that we now present in some detail.

We include the following highly selected results on axiomatization of deterministic representations of such gambles for two reasons. First, there is a vast related theoretical and empirical literature (see the citations in Section 7.1). Second, there are major challenges in extending this work to probabilistic representations and in empirically testing such extensions. Section 7.5.1 gives a selective summary of recent such extensions; important related work is presented in Stott (2006) and Regenwetter *et al.* (2014), with both testing probabilistic models of binary choice where the gambles satisfy *cumulative prospect theory* (Tversky and Kahneman, 1992, and below).

Luce and Marley (2005) develop<sup>5</sup> extensive theoretical relations among a variety of representations over gambles of gains (equally, over losses); some of the representations and results are fairly classical, others involve relatively new axiomatizations. Included are the class of ranked weighted utility; rank-dependent utility (which includes cumulative prospect theory as a special case); gains-decomposition utility; simple utility with weights that depend only on the relevant event, as used in the original prospect theory of Kahneman and Tversky (1979); and subjective expected utility. Marley and Luce (2005) discuss various data relevant to the theoretical properties presented in Luce and Marley (2005), and Marley *et al.* (2008) and Bleichrodt *et al.* (2008) each solve a different open problem in Luce and Marley (2005). In particular, Bleichrodt *et al.* (2008) complete the axiomatization of the *ranked additive representation* of ranked gambles, where a ranked gamble consists of a finite number of consequence–event pairs (called *branches*) ordered by the preferences among the consequences. The representation involves a sum of functions, one for each branch, that depends on two things: the utility of the consequence and the entire ordered vector of the event partition. The representation is developed using a form of ranked additive conjoint measurement that was axiomatized by Wakker (1991). Here, we illustrate results for this representation using the *rank weighted representation*, which is the special case of the ranked additive representation where the utility of each consequence–event branch is a product of the utility of the outcome for that branch and of a weight depending on the event associated with that branch and the entire ordered vector of the event partition.

We restrict the presentation to uncertain gambles (defined in the next section). For results on risky and/or ambiguous gambles, see Wakker (2010) and Abdellaoui *et al.* (2011). It is also important to remember that all the results in this section are

<sup>5</sup> A significant part of the material in Sections 7.4.2–7.4.4 is adapted, with newer results, from Luce and Marley (2005), with kind permission from Springer Science + Business Media B.V.

for gambles of gains, or, equivalently, for gambles of losses. Parallel results for mixed gambles, i.e., those with outcomes that can be gains and/or losses, can be derived in various ways. The representations for such mixed gambles depend on whether one considers the representations in both the gains and loss domains to be of the same form or of different forms; and, if the latter, how representations of mixed gambles are derived from those for gains/losses (Luce, 2000; Wakker, 2010). Section 7.4.5 presents some theoretical and empirical results for the mixed case.

#### 7.4.2.1 Notation

Let  $X$  denote the set of pure consequences<sup>6</sup> for which chance or uncertainty plays no role. A distinguished element  $e \in X$  is interpreted to mean *no change from the status quo*. We assume that there exists a *preference order*  $\succsim$  over  $X$  and that it is a weak order. Let  $\sim$  denote the corresponding indifference relation. A typical *first-order gamble of gains*  $g$  with  $n$  consequences is of the form

$$g = (x_1, C_1; \dots; x_i, C_i; \dots; x_n, C_n) \equiv (\dots; x_i, C_i; \dots),$$

where  $x_i$  are consequences and  $C_n = (C_1, \dots, C_i, \dots, C_n)$  is a (exhaustive and exclusive) partition of some “universal event”  $C(n) = \bigcup_{i=1}^n C_i$ . The underlying event  $C(n)$  is only “universal” for the purpose of this gamble. A (consequence, event) pair  $(x_i, C_i)$  is called a *branch*; thus, a gamble is a  $2n$ -tuple composed of  $n$  branches. We deal mostly with cases where each  $C_i \neq \emptyset$ . The results are confined to all gains, i.e., where each  $x_i \succsim e$  (or, equally, to all losses, i.e., where each  $x_i \precsim e$ ). Using our terminology, Figure 7.1 shows two first-order gambles (lotteries) of gains, each of which has two branches; and, using our notation, the gamble on the left is  $(\$22.36, C_1; \$0.00, C_2)$ , where  $C_1$  (respectively,  $C_2$ ) is the event associated with the black (respectively, white) area.

In stating the axioms and representations, it is convenient to assume that we have carried out a permutation of the indices such that the consequences are (rank) ordered from the most preferred to the least preferred, in which case we simply use the notation  $x_1 \succsim \dots \succsim x_i \succsim \dots \succsim x_n \succsim e$ . We then talk of rank ordered consequences.

We explore utility representations  $u$  onto real intervals of the form  $I = [0, \kappa[$ , where  $\kappa \in ]0, \infty]$ , that meet various, increasingly stronger, restrictions. Two conditions that are common to all representations we consider are:

$$g \succsim h \text{ iff } u(g) \geq u(h), \quad (7.20)$$

$$u(e) = 0. \quad (7.21)$$

We refer to these as *order-preserving representations*. Note that because  $I$  is open on the right, there is no maximal element in the structure.

<sup>6</sup> There is a slight inconsistency in the use here of  $X$  to denote the set of pure consequences and its use in the previous sections to denote a set of options. However, we think the retention of this notation is warranted because much of the literature in both areas uses  $X$  in this manner.

The framework above is for *uncertain gambles*, i.e., those where each outcome occurs if a particular event occurs, and probabilities are not given. We do discuss some representations for *risky gambles* – that is, those for which outcome probabilities are given – in Sections 7.4.5 and 7.5.1. We do not present any results for *ambiguous gambles* – those in which each outcome is associated with either an event or a probability (but not both) – Abdellaoui *et al.* (2011) present very nice empirical and (some) theoretical results for this domain.

#### 7.4.2.2 Ranked weighted utility representation

The following concept of a ranked weighted utility (RWU) representation, is defined in Marley and Luce (2001).

**Definition 7.24** An order-preserving representation  $u : \mathcal{D}_+ \xrightarrow{\text{onto}} I \subseteq \mathbb{R}_+$  is a *ranked weighted utility (RWU)* representation iff there exist weights  $S_i(\vec{\mathbf{C}}_n)$  assigned to each index  $i = 1, \dots, n$  and possibly dependent on the entire ordered partition  $\vec{\mathbf{C}}_n$ , where  $0 \leq S_i(\vec{\mathbf{C}}_n)$  and  $S_i(\vec{\mathbf{C}}_n) = 0$  iff  $C_i = \emptyset$ , such that, with rank-ordered consequences,

$$u(\dots; x_i, C_i; \dots) = \sum_{i=1}^n u(x_i)S_i(\vec{\mathbf{C}}_n). \quad (7.22)$$

Most theories of utility, including that for the RWU representation of Definition 7.24, have either explicitly or implicitly assumed *idempotence* (below). A major exception is the extensive series of papers on the *utility of gambling* that develops representations involving a relatively standard idempotent representation, plus a weighted entropy-type term; these results are summarized in Luce *et al.* (2009).

Consider a gamble with  $x_1 = \dots = x_i = \dots = x_n = x$  and consider the following property:

**Definition 7.25** *Idempotence* of gambles is satisfied iff, for every  $x \in X$  and every ordered event partition  $(C_1, \dots, C_i, \dots, C_n)$ ,

$$(x, C_1; \dots; x, C_i; \dots; x, C_n) \sim x. \quad (7.23)$$

If we assume idempotence along with RWU, (7.22), we obtain that

$$\sum_{i=1}^n S_i(\vec{\mathbf{C}}_n) = 1. \quad (7.24)$$

#### 7.4.3 Axiomatizations of representations of gambles

The RWU representation, (7.22), is of interest because it encompasses several representations in the literature including the standard rank-dependent representation. We now explore the additional conditions that are needed to reduce the ranked

weighted utility representation of Definition 7.24 to the rank-dependent utility representation. We then show that the rank-dependent representation includes cumulative prospect theory as a special case, and is closely related to Birnbaum's TAX representation.

#### 7.4.3.1 Rank-dependent utility

Define

$$C(i) := \bigcup_{j=1}^i C_j. \quad (7.25)$$

Consider a set of positive weights  $W_{C(n)}(C(i))$ ,  $i = 0, \dots, n$ , and define

$$W_i(\vec{C}_n) = \begin{cases} 0, & i = 0 \\ W_{C(n)}(C(i)), & 0 < i < n \\ 1, & i = n \end{cases}$$

**Definition 7.26** *Rank-dependent utility (RDU)* is the special case of RWU, (7.22), with weights of the form: for  $i = 1, \dots, n$ ,

$$S_i(\vec{C}_n) = W_i(\vec{C}_n) - W_{i-1}(\vec{C}_n).$$

Luce and Marley (2005, footnote 2) summarize the history of this representation and its various names. With certain additional assumptions it has also been called *cumulative prospect theory for gains (losses)* (Tversky and Kahneman, 1992).

The RDU representation exhibits the following property:

**Definition 7.27** *Coalescing* is satisfied iff for all rank-ordered consequences and corresponding ordered partitions with  $n > 2$  and with  $x_{k+1} = x_k$ ,  $k < n$ ,

$$\begin{aligned} & (x_1, C_1; \dots; x_k, C_k; x_k, C_{k+1}; \dots; x_n, C_n) \\ & \sim (x_1, C_1; \dots; x_k, C_k \cup C_{k+1}; \dots; x_n, C_n) \quad (k = 1, \dots, n-1). \end{aligned} \quad (7.26)$$

Note that the gamble on the left has  $n$  branches with at most  $n - 1$  consequences, whereas the one on the right has  $n - 1$  branches as well as at most  $n - 1$  consequences. The next result shows that coalescing is the key to RDU.

**Theorem 7.3** (Luce and Marley, 2005, Theorem 11) *For  $n > 2$ , the following statements are equivalent:*

1. RWU, Definition 7.24, idempotence, (7.23), and coalescing, (7.26), all hold.
2. RDU, Definition 7.26, holds.

There is a large literature on ways to arrive at idempotent RDU (see, e.g., Luce, 2000, for further discussion and references). None of those approaches seem simpler or more straightforward than the assumptions presented here.

### 7.4.3.2 Configural weighted utility

Birnbaum, over many years with various collaborators, has explored a class of representations that Marley and Luce (2005) called *configural weighted utility*; Birnbaum has typically called them *configural weight models*. In contrast to utility models that have been axiomatized in terms of behavioral properties, Birnbaum and his collaborators have shown that certain special configural weight models do or do not exhibit certain behavioral properties, and have compared how various of those properties fare relative to data. In contrast to this approach, Proposition 7.1 shows the close relation between the most frequently studied configural weighted utility model, called TAX, and the ranked weighted utility representation. According to Birnbaum and Navarrete (1998), TAX has the following form:<sup>7</sup>

**Definition 7.28** Let  $u$  be a utility function over ranked gambles and pure consequences,  $T$  a function from events into the nonnegative real numbers, and  $\omega_{i,j}(\vec{C}_n)$  mappings from ordered-event partitions to real numbers. Then, TAX is the following representation over gambles with rank-ordered consequences:

$$u(g_{\vec{C}_n}) = \frac{\sum_{i=1}^n u(x_i)T(C_i) + \sum_{i=1}^{n-1} \sum_{j=i+1}^n [u(x_i) - u(x_j)] \omega_{i,j}(\vec{C}_n)}{T(\vec{C}_n)}, \quad (7.27)$$

$$\text{where } T(\vec{C}_n) := \sum_{i=1}^n T(C_i).$$

The name TAX arises because Birnbaum describes the term on the right as imposing a tax from (respectively, to) lower-ranked consequences to (respectively, from) higher-ranked ones depending on whether the relevant weight is positive (respectively, negative); he usually assumes a particular form for the weights  $\omega_{i,j}(\vec{C}_n)$ . Marley and Luce (2005, proposition 6) proved the following result when the weights are unrestricted in form.<sup>8</sup>

### Proposition 7.1

(i) A TAX representation, (7.27), with  $S_i(\vec{C}_n)$ ,  $i = 1, \dots, n$ , defined by

$$S_i(\vec{C}_n) = \frac{T(C_i) + \sum_{j=i+1}^{n+1} \omega_{i,j}(\vec{C}_n) - \sum_{j=0}^{i-1} \omega_{j,i}(\vec{C}_n)}{T(\vec{C}_n)}, \quad (7.28)$$

<sup>7</sup> The notation is that of Marley and Luce (2005), and differs from that of Birnbaum and his collaborators in two, nonsubstantive, ways.

<sup>8</sup> Marley and Luce failed to include the conditions  $S_i(\vec{C}_n) \geq 0$  and  $S_i(\vec{C}_n) = 0$  iff  $C_i = \emptyset$  in their statement of part (i). There is also a typographic error in the final sum in the numerator of their (19) – each term  $\omega_{i,j}(\vec{C}_n)$  in that term should be  $\omega_{j,i}(\vec{C}_n)$ , as it is in our (7.28). The results in Marley and Luce that use the proposition remain valid as all those results assume an RWU representation, hence that  $S_i(\vec{C}_n) \geq 0$  and  $S_i(\vec{C}_n) = 0$  iff  $C_i = \emptyset$ .

where

$$\begin{aligned}\omega_{0,i}(\vec{\mathbf{C}}_n) &:= 0, \\ \omega_{i,n+1}(\vec{\mathbf{C}}_n) &:= 0,\end{aligned}$$

is an idempotent RWU representation, (7.22), provided each  $S_i(\vec{\mathbf{C}}_n) \geq 0$  and  $S_i(\vec{\mathbf{C}}_n) = 0$  iff  $C_i = \emptyset$ .

- (ii) Any idempotent RWU representation, (7.22), can be put (in many ways) in the form of a TAX representation, (7.27). One such has

$$T(C_i) > 0$$

and

$$\omega_{i,j}(\vec{\mathbf{C}}_n) = \begin{cases} T(\vec{\mathbf{C}}_n) \sum_{k=1}^i S_k(\vec{\mathbf{C}}_n) - \sum_{k=1}^j T(C_k), & i = 1, \dots, n-1, \\ & j = i+1 \\ 0, & i = 0 \text{ or } j = n+1, \\ & \text{or } j \neq i+1. \end{cases}$$

Thus, the general TAX model is equivalent to the idempotent ranked weighted utility representation (Section 7.4.2) when the TAX model satisfies the following two conditions:  $S_i(\vec{\mathbf{C}}_n)$  in (7.28) are nonnegative and  $S_i(\vec{\mathbf{C}}_n) = 0$  iff  $C_i = \emptyset$ . These two constraints correspond to requiring the TAX model to satisfy *co-monotonic consequence monotonicity*, i.e., the representation is strictly increasing in each consequence so long as the rank ordering is maintained; it is not known what constraints on the  $\omega_{i,j}(\vec{\mathbf{C}}_n)$  in (7.28) are sufficient for TAX to satisfy that condition. Marley and Luce (2005) explore various *independence properties* that are implied by the idempotent ranked weighted utility representation (Section 7.4.2), and hence by any TAX model that satisfies co-monotonic consequence monotonicity. Birnbaum (2008) re-summarizes various of the data included in Marley and Luce (2005), plus more recent data, and concludes that rank-dependent utility is not a satisfactory representation of his data, whereas TAX can handle the majority of the data. However, recent discussions of Birnbaum's analyses (which are often based either on descriptive modal choice or his *pattern-based true-and-error models*) suggest that further detailed analyses of data of individual decision makers are needed to place such a conclusion on a firm basis (see, e.g., Birnbaum, 2011; Regenwetter *et al.*, 2011b).

#### 7.4.4 Joint receipts

In this section we extend the domain  $\mathcal{D}_+$  of gambles to include the joint receipt of pure consequences and gambles. With  $X$  the set of pure consequences, for  $x, y \in X$ ,  $x \oplus y \in X$  represents receiving both  $x$  and  $y$ . When  $X$  denotes money, many authors assume that  $x \oplus y = x + y$ , but as discussed in Luce (2000), this is certainly not necessary and may well be false. When  $f$  and  $g$  are gambles,  $f \oplus g$

means having or receiving both gambles. Here, we assume  $\oplus$  to be a commutative operator.

The following concept of generalized additivity is familiar from the functional equation literature (see Ng, Chapter 3 of this Handbook):

**Definition 7.29** The operation  $\oplus$  has a *generalized additive representation*  $u : \mathcal{D}_+ \xrightarrow{\text{onto}} \mathbb{R}_+ := [0, \infty[$  iff (7.20), (7.21), and there exists a strictly increasing function  $\varphi$  such that

$$u(f \oplus g) = \varphi^{-1}(\varphi(u(f)) + \varphi(u(g))). \quad (7.29)$$

It is called *additive* if  $\varphi$  is the identity.

Note that  $V = \varphi(u)$  is additive and therefore one cannot distinguish between a generalized additive representation  $u$  and an additive representation  $V$  without considering additional structural assumptions about the properties of the utility representation; we consider such structural conditions in Section 7.4.5.

An important special case is when  $\varphi$  is the identity in which case  $u$  is additive over  $\oplus$ . This is, of course, a strong property. For example, if for money consequences  $x \oplus y = x + y$ , then additive  $u$  implies  $u(x) = \alpha x$  (with  $\alpha > 0$  for a monotonic increasing relation). For at least modest amounts of money – “pocket money” – this may not be unrealistic, as Birnbaum and collaborators have argued by fitting data (for example, Birnbaum and Beeghley, 1997, for judgment data, as well as Birnbaum and Navarrete, 1998, for choice).

Another important special case of generalized additivity is: for some  $\delta \neq 0$ ,

$$u(f \oplus g) = u(f) + u(g) + \delta u(f)u(g), \quad (7.30)$$

which form has been termed *p-additivity* (Ng *et al.*, 2009). This corresponds to the mapping  $\varphi(z) = \text{sgn}(\delta) \ln[1 + \delta u(z)]$ , where  $\text{sgn}(\delta)$  is the sign of  $\delta$ , in (7.29).

## 7.4.5 Parametric forms for utility and weights

### 7.4.5.1 Parametric utility forms

As pointed out after the definition of a generalized additive representation, Definition 7.29, one cannot distinguish between a generalized additive representation  $u$  and an additive representation  $V$  without considering additional structural assumptions about the properties of the utility representation; we now discuss such properties and the resulting representations.

**Definition 7.30** *Binary segregation* holds for gains iff for all gambles  $f, g$  of gains,

$$(f \oplus g, C_1, ; g, C_2) \sim (f, C_1, ; g, C_2) \oplus g. \quad (7.31)$$

Luce (2000, Theorems 4.4.4 and 4.4.6) shows the following:

**Theorem 7.4** *Let  $u$  be an order preserving representation of gambles over gains and  $W_{C(2)}$  a weighting function. Then any two of the following imply the third:*

1. *Binary segregation*, (7.31).
2. *Binary rank-dependent utility*:

$$u(f, C_1, ; g, C_2) = u(f)W_{C(2)}(C_1) + u(g)[1 - W_{C(2)}(C_1)].$$

3. *p-additivity*: for some real constant  $\delta$  that has the unit of  $1/u$ ,

$$u(f \oplus g) = u(f) + u(g) + \delta u(f)u(g),$$

and for some event  $K$  there is a weighting function  $W_K$  onto  $[0, 1]$  such that  $(u, W_K)$  forms a separable representation of the gambles  $(x, C, ; e, K \setminus C)$ , i.e.,

$$u(x, C, ; e, K \setminus C) = u(x)W_K(C).$$

This result thus gives a quite natural motivation for the assumption of a  $p$ -additive representation over gains, with a parallel result for losses. We are not aware of a parallel theorem over mixed gambles (gains and losses). Ng *et al.* (2009, Theorem 21) present an axiomatization of the rank-dependent form over both gains and losses in which the utility function is assumed to be  $p$ -additive over gains (over losses). They also assume segregation and derive binary RDU. A careful check would be needed to see whether their result can be derived by assuming binary RDU (which has been axiomatized by Marley and Luce, 2002), without assuming  $p$ -additivity, and then using the above result to derive  $p$ -additivity over gains (over losses).

Given the  $p$ -additive form, defining  $V(r) := \text{sgn}(\delta) \ln[1 + \delta u(r)]$ , we obtain the additive representation  $V(f \oplus g) = V(f) + V(g)$ . It is routine to show that the following representations are possible for mixed gambles (gains and losses) under the conditions of Ng *et al.* (2009, Theorem 21) and arguments paralleling those of Luce (2000, corollary, p. 152) show that they are the most general solutions:

- (i) If  $\delta = 0$ , then for some  $\alpha > 0$ ,

$$u(f) = \alpha V(f). \quad (7.32)$$

- (ii) If  $\delta > 0$ , then  $u$  is superadditive, i.e.,  $u(f \oplus g) > u(f) + u(g)$ , unbounded, and for some  $\kappa > 0$ ,

$$u(f) = \frac{1}{\delta}[e^{\kappa V(f)} - 1]. \quad (7.33)$$

- (iii) If  $\delta < 0$ , then  $u$  is subadditive, i.e.,  $u(f \oplus g) < u(f) + u(g)$ , bounded by  $1/|\delta|$ , and for some  $\kappa > 0$ ,

$$u(f) = \frac{1}{|\delta|}[1 - e^{-\kappa V(f)}]. \quad (7.34)$$

When the pure consequences are money, many economists assume that  $x \oplus y = x + y$ , in which case there is a constant  $A > 0$  such that  $V(x) = Ax$ . Note that then, for  $\delta > 0$ , where we have (7.33),  $u$  is strictly increasing and convex, which many identify as corresponding to risk seeking behavior; and for  $\delta < 0$ , where we have (7.33),  $u$  is strictly increasing and concave, which many identify as corresponding

to risk-averse behavior. Luce (2010a,b) includes further discussion of these representations, including their implications for interpersonal comparisons of utility.

The above representations have a common value for  $\delta$  for gains and losses, which, in the special case of money just discussed gives either risk-seeking behavior for both gains and losses, or risk-averse behavior for both gains and losses. However, it is more customary to assume that people are risk-seeking on gains and risk-averse on losses. Such behavior can be accommodated by having a  $p$ -additive function for pure consequences that has a  $\delta^+ > 0$  and  $\kappa^+ > 0$  for gains and a  $\delta^- < 0$  and  $\kappa^- > 0$  for losses – that is,

$$u(x) = \begin{cases} \frac{1}{\delta^+}(e^{\kappa^+ x} - 1), & x \geq 0, \\ \frac{1}{|\delta^-|}(1 - e^{-\kappa^- x}), & x < 0. \end{cases}$$

Another possible form has the utility of gains as well as the utility for losses bounded: there are positive constants  $a, b, A$ , and  $B$  such that

$$u(x) = \begin{cases} A(1 - e^{-ax}), & x \geq 0, \\ B(e^{bx} - 1), & x < 0. \end{cases}$$

Finally, the above representations assume that risk-seeking or risk-aversion is modelled by the shape of the utility function; in particular, risk-aversion is modelled by a concave utility function. However, rank-dependent utility (RDU) with  $u(x) = x^2$  and  $W(p) = p^2$  produces risk aversion, even though the utility function is strictly convex (Chateauneuf and Cohen, 1994, corollary 2).

Luce (2010a) proposed a condition that discriminates between the representations (7.32)–(7.34): consider the case where the pure outcomes are amounts of money, assume that  $r \oplus s = r + s$ , and let  $x, x', y, y'$  be such that  $x \succ x' \succ y \succ y'$ . Then Luce's condition is:

$$\delta = \begin{cases} 1 \\ 0 \\ -1 \end{cases} \text{ iff } (x + x', E; y + y', \Omega \setminus E) \begin{cases} \succ \\ \sim \\ \prec \end{cases} (x + y, E; x' + y', \Omega \setminus E),$$

where  $E$  is a chance event (with complement  $\Omega \setminus E$ , where  $\Omega$  is the “universal” event) such that for all amounts of money  $x, y$ ,  $(x, E; y, \Omega \setminus E) \sim (y, E; x, \Omega \setminus E)$ ; that is,  $E$  has (subjective) probability 1/2.

Davis-Stober and Brown (2013) extend Luce's original criterion to accommodate decision makers with risk preferences that vary as a function of the monetary environment (e.g., gambles of all gains, all losses, or mixed gains and losses); to a significant extent, their extension corresponds to testing both the  $\delta = -1, 0, +1$  cases covered by Luce's conditions and the mixed cases of  $\delta$  introduced above. Using a Bayesian modeling approach, they evaluated the repeated choices of 24 participants and found that one participant was best described as risk-seeking; 6 as risk-averse; 2 as risk-neutral; 6 as risk-seeking for losses and risk-averse for gains; 6 could not be described by a  $p$ -additive representation; and 1 of what they call *stakes sensitive* and 2 of what they call *mixed gamble*.

### 7.4.5.2 Parametric weight forms

A *risky gamble* is one where the events (of an uncertain gamble) are replaced by probabilities. For notational simplicity, let  $(x, p)$  denote a risky gamble in which the consequence  $x$  occurs with probability  $p$  and nothing otherwise. A consequence can be either a pure one (no risk), such as a sum of money, or a gamble such as  $(y, q)$ , where  $y$  is a pure consequence. In this section, the only gambles considered are of the (first-order) form  $(x, p)$  or the (second-order) form  $((y, q), p)$ , where  $x$  and  $y$  range over the pure consequences and  $p$  and  $q$  can be any probabilities. We assume that there is a weak order preference relation  $\succsim$  over first- and second-order gambles of the above form; and that  $\succsim$  is represented by an order-preserving representation  $u$  (see Equations 7.20–7.21).

**Definition 7.31** An order-preserving representation  $u$  is *separable* if there is a strictly increasing *weighting function*  $W : [0, 1] \rightarrow [0, 1]$  such that for each first-order gamble or pure consequence  $z$  and probability  $p$ ,

$$u(z, p) = u(z)W(p), \quad (7.35)$$

where  $u(x) =: u(x, 1)$ . Thus, given a pure consequence  $x$ , for a first-order gamble we have  $u(x, p) = u(x)W(p)$  and for a second-order gamble we have  $u((x, q), p) = u(x)W(q)W(p)$ .

We now present some results on parametric forms for the weights in first- and/or second-order binary risky gambles of the above forms. We focus on the case where there is a common representation over all the gambles, which can be either over gains, over losses, or over gains and losses (Prelec, 1998, extends the results discussed below to cases where the representations can differ for gains and losses).

Prelec (1998) was the first to axiomatize weighting functions assuming a separable representation (with the weights onto the interval  $[0, 1]$ ). He axiomatized three different forms; we summarize the results for one case.

The following definition is a variant of Prelec (1998, Definition 1) and Luce (2001, Definition 1):

**Definition 7.32** Let  $N \geq 1$  be any integer. Then *N-compound invariance* holds iff, for consequences  $x, y, u, v$ , and probabilities  $p, q, r, s \in ]0, 1[$  with  $p < q, r < s$ ,

$$(x, p) \sim (y, q), (x, r) \sim (y, s) \text{ and } (u, p^N) \sim (v, q^N)$$

imply

$$(u, r^N) \sim (v, s^N).$$

*Compound invariance* holds when N-compound invariance holds for all integers  $N \geq 1$ .

Prelec's (1998, proposition 1) then says, in essence, that in the presence of separability, (7.35), and a suitable density of consequences, compound invariance is

equivalent to the following *compound-invariance* family of weighting functions: there are constants  $\alpha > 0$ ,  $\beta > 0$  such that<sup>9</sup>

$$W(p) = \exp[-\beta(-\ln p)^\alpha]. \quad (7.36)$$

The necessity of compound invariance for this representation is seen by noting that the representation satisfies the property: for any real number  $\lambda \geq 1$ ,  $W(p^\lambda) = W(p)^{\lambda^\alpha}$ . Note that when  $\alpha = 1$  the representation reduces to a power function with positive exponent. Also, for  $\alpha < 1$ , the general representation has an inverse-S shape, which is the form found for many estimated weighting functions (Prelec, 1998; Wakker, 2010, section 7.7).

Compound invariance involves first-order gambles and a relatively complicated antecedent condition, plus the assertion that it holds for all integers  $N \geq 1$ . We now present a condition due to Luce (2001, Definition 2) that is equivalent to a weighting function belonging to the compound invariance family. This condition has the advantages that it has a simpler antecedent condition and only requires the assertion that the condition holds for the integers  $N = 1, 2$ , but has the (possible) disadvantage that it involves compound gambles. Note that the following definition and theorem include conditions stated for the doubly open unit interval, denoted  $]0, 1[$ .

**Definition 7.33** (Luce, 2001, Definition 1) Let  $N \geq 1$  be any integer. Then  $N$ -reduction invariance holds iff, for any consequence  $x$ , and probabilities  $p, q, r \in ]0, 1[$ ,

$$((x, p), q) \sim (x, r) \sim (y, s)$$

implies

$$((x, p^N), q^N) \sim (x, r^N).$$

Reduction invariance holds when N-reduction invariance holds for all integers  $N \geq 1$ .

**Theorem 7.5** (Luce, 2001, proposition 1) Suppose that a structure of binary gambles of the form  $(x, p)$  and  $((y, p), q)$  is weakly ordered in preference and has a separable representation (7.35) with  $W : [0, 1] \xrightarrow{\text{onto}} [0, 1]$ , where  $W$  is strictly increasing in  $p$ . Then the following two conditions are equivalent:

- (i)  $N$ -reduction invariance (Definition 7.33), holds for  $N = 2, 3$ .
- (ii)  $W$  satisfies (7.36).

Both Prelec (1998) and Luce (2001) generalize their condition to include other weight forms, including the exponential-power and hyperbolic-logarithm. However, all cases essentially assume a separable representation. The next section presents an extremely elegant joint derivation of a separable representation with a weighting function satisfying (7.36).

<sup>9</sup> Luce (2001, footnote 1) notes that Prelec focussed on the special case  $\beta = 1$ .

Diecidue *et al.* (2009) give additional preference foundations for parametric weighting functions under rank-dependent utility, namely, power functions  $p^a$  with  $a > 0$ ; exponential functions  $\frac{e^{cp}-1}{e^c-1}$  with  $c \neq 0$ ; and switchpower functions (separate power functions on a lower- and an upper-range); and Chechile and Barch (2013) review recent theoretical and empirical work on (risky) weighting functions and present new evidence in support of the Prelec function and their own *exponential odds function*.

#### 7.4.5.3 Multiplicative separability of utility and weight

We now summarize the very nice result by Bleichrodt *et al.* (2013) that, under weak regularity conditions, a weaker version of compound invariance implies that an order-preserving multiplicative representation exists for first-order gambles of the form  $(x, p)$  with the weight form satisfying (7.36).

**Definition 7.34** For real  $\lambda > 0$ ,  $\lambda$ -compound invariance holds iff, for nonzero consequences  $x, y, z$ , and nonzero probabilities  $p, q, r$ ,

$$(x, p) \sim (y, q), (x, q) \sim (y, r) \text{ and } (y, p^\lambda) \sim (z, q^\lambda)$$

imply

$$(y, q^\lambda) \sim (z, r^\lambda).$$

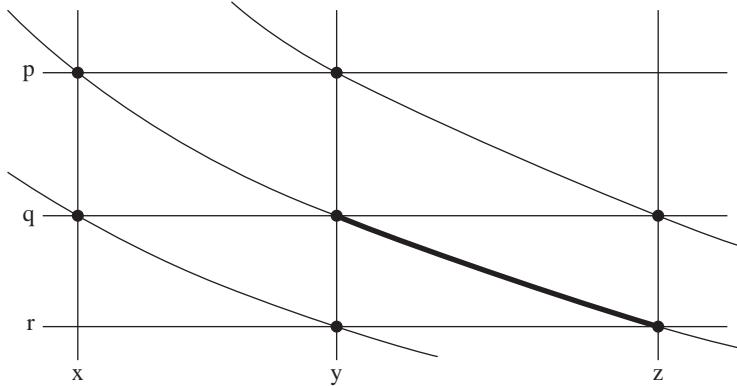
The main results of Bleichrodt *et al.* (2013) only use  $\lambda$ -compound invariance for the integer cases  $\lambda = 1, 2, 3$ , in which cases it agrees with the special case of N-compound invariance, Definition 7.32, with  $N = 1, 2, 3$ , and with  $x, y, u, v$  (respectively,  $p, q, r, s$ ) in N-compound invariance replaced by the special case  $x, y, y, z$  (respectively,  $p, q, q, r$ ) in  $\lambda$ -compound invariance.

Figure 7.4 illustrates 1-compound invariance.

We need various assumptions and terminology from Bleichrodt *et al.* (2013). We continue to use the representation  $(x, p)$  for a gamble where the pure consequence  $x$  is received with probability  $p$ , otherwise nothing. Assume that the set of such gambles is  $X \times [0, 1]$ , with  $X$  a nonpoint interval within  $\mathbb{R}^+$  containing 0, and that  $(0, p) \sim (x, 0) \sim (0, 0)$  for all  $p$  and  $x$ . The preference order  $\succsim$  on  $X \times [0, 1]$  is continuous if, for each gamble  $(x, p)$ ,  $\{(y, q) : (y, q) \succsim (x, p)\}$  and  $\{(y, q) : (y, q) \preceq (x, p)\}$  are closed subsets of the gamble space  $X \times [0, 1]$ . Strict stochastic dominance holds if  $(x, p) \succ (y, p)$  whenever  $x > y$  and  $p > 0$  and  $(x, p) \succ (x, q)$  whenever  $x > 0$  and  $p > q$ . Finally, the representation  $u$  on the set of gambles  $(x, p)$  is multiplicative if it is separable, Definition 7.31, with  $u(0) = 0$ ,  $W(0) = 0$ , and each of  $u$  and  $W$  is continuous and strictly increasing. Clearly, multiplicative representability implies strict stochastic dominance.

Bleichrodt *et al.*'s (2013) very nice main result shows that the multiplicative relation is a consequence when  $\lambda$ -compound invariance holds for  $\lambda = 1, 2, 3$ .

**Theorem 7.6** (Bleichrodt *et al.*, 2013, Theorem 3.4). *The following statements are equivalent for  $\succsim$  on  $X \times [0, 1]$ :*



**Figure 7.4** One-compound invariance. The indifference marked by a heavy line is implied by the other three indifferences. Figure 3.2 of Bleichrodt et al. (2013). Reproduced with permission from Elsevier.

- (i) A multiplicative representation exists.
- (ii)  $\sim$  is a continuous weak order that satisfies strict stochastic dominance,  $\lambda$ -compound invariance, Definition 7.34, for  $\lambda = 1, 2, 3$ , and  $(0, p) \sim (x, 0) \sim (0, 0)$  for all  $p$  and  $x$ .

#### 7.4.5.4 Other parametric and weight forms

The previous two sections presented axiomatizations for utility functions derived from  $p$ -additive representations, and weighting functions derived from compound invariance. A tremendous number of other utility and weighting functions have been suggested, although most have not been axiomatized in the way presented above. Stott (2006) summarizes many of the suggested utility functions for gambles over monetary gains and weighting functions over probabilities and then considers various combinations of utility and weighting function (in the cumulative prospect theory form) in conjunction with three standard functional forms for the (probabilistic) choice function. He fit these combinations to data on binary choices between gambles of the form discussed in the previous section, and concluded that the best fitting model had a power utility function, a Prelec weighting function, (7.36), and a multinomial logit (MNL) form, (7.53), for the (best) choice probabilities; Blavatskyy and Pogrebna (2010) discuss related work using a larger variety of econometric specifications and conclude that the particular probabilistic response function one uses can be pivotal for the relative performance of competing core deterministic theories. The next section presents recent axiomatizations of some of these (probabilistic) choice functions when expected utility holds.

## 7.5 Choice probabilities for real-valued representations

We begin this section with definitions and results on distribution-free random utility representations (i.e., those without a specific parametric form, such as

the Normal or Extreme Value). Later, in Section 7.5.3, we turn our attention to various of the parametric models that have been developed and applied in discrete choice surveys, mainly by economists and marketing scientists, and those developed and studied in controlled experiments, mainly by experimental economists and cognitive psychologists. We then attempt to integrate these in a common theoretical framework.

We begin with the classic models in psychology and economics, which were formulated mainly to account for choice probabilities. We then present extensions of those models to deal with more complicated phenomena, including the time to make a choice (response time). All of these models are *context-free* – that is, in a sense to be defined, the representation of a choice option is independent of the other available choice options. We then present *context-dependent* models – that is, where the representation of a choice option depends on the other available options.

### 7.5.1 Distribution-free random utility representations

**Definition 7.35** Consider a finite master set  $\mathcal{A}$  of two or more choice alternatives. A (distribution-free) *random utility model* for  $\mathcal{A}$  is a family of jointly distributed real random variables  $\mathbf{U} = (\mathbf{U}_{x,i})_{x \in \mathcal{A}, i \in \mathcal{I}}$  with  $\mathcal{I}$  some finite index set. If  $\mathbf{U} = (\mathbf{U}_{x,i})$  is a family of jointly independent random variables, the model is an *independently distributed random utility model*.

The realization of a random utility model at some sample point  $\omega$ , given by the real-valued vector  $(\mathbf{U}_{x,i}(\omega))_{x \in \mathcal{A}, i \in \mathcal{I}}$ , assigns to alternative  $x \in \mathcal{A}$  the utility vector  $(\mathbf{U}_{x,i}(\omega))_{i \in \mathcal{I}}$ . One possible interpretation of such a utility vector is that  $\mathcal{I}$  is a collection of attributes, and  $\mathbf{U}_{x,i}(\omega)$  is the utility of choice alternative  $x$  on attribute  $i$  at sample point  $\omega$ .

**Definition 7.36** Consider a finite master set  $\mathcal{A}$  of two or more choice alternatives. A (distribution-free) *unidimensional, noncoincident random utility model* on  $\mathcal{A}$  is a family of jointly distributed real random variables  $\mathbf{U} = (\mathbf{U}_x)_{x \in \mathcal{A}}$  with  $\Pr(\mathbf{U}_y = \mathbf{U}_z) = 0, \forall y \neq z \in \mathcal{A}$ .

The most common use of the term “random utility model” in the discrete choice literature (Train, 2003) refers to context-free, noncoincident, and unidimensional parametric models, where the random variables  $(\mathbf{U}_x)_{x \in \mathcal{A}}$  can be decomposed as follows:

$$(\mathbf{U}_x)_{x \in \mathcal{A}} = (u(x))_{x \in \mathcal{A}} + (\boldsymbol{\epsilon}_x)_{x \in \mathcal{A}},$$

and where  $u(x)$  is the deterministic real-valued utility of option  $x$  and  $(\boldsymbol{\epsilon}_x)_{x \in \mathcal{A}}$  is multivariate normal or multivariate extreme value. We present such models in some detail in Section 7.5.2.

We now present probabilistic generalizations of the deterministic utility representations, and results, of Theorem 7.2.

**Theorem 7.7** (Regenwetter and Marley, 2001) Let  $\mathcal{A}$  be a finite set with  $|\mathcal{A}| = N$ . Consider a family of  $k \times N$  many jointly distributed real-valued utility random variables  $\mathbf{U} = (\mathbf{U}_{x,i})_{x \in \mathcal{A}; i=1,\dots,k}$ . Then the collection  $(\mathbf{U})$  satisfies the following properties (with all  $w, x, y, z \in \mathcal{A}$ ):

**Random utility representations of strict linear orders:** a unidimensional (i.e.,  $k = 1$ ), noncoincident random utility model  $\mathbf{U}$  on  $\mathcal{A}$  induces a probability distribution  $\succ \mapsto P(\succ)$  on  $\mathcal{SLO}$  as follows. For any given  $\succ \in \mathcal{SLO}$ , writing  $x_i$  for the object at rank  $i$  in  $\succ$ , i.e.,  $\text{Rank}_{\mathcal{A}, \succ}(x_i) = i$ ,  $\forall i = 1, 2, \dots, N$ , let

$$P(\succ) = \Pr(\mathbf{U}_{x_1} > \mathbf{U}_{x_2} \cdots > \mathbf{U}_{x_N}).$$

**Random utility representations of strict weak orders:** a unidimensional (i.e.,  $k = 1$ ) random utility model on  $\mathcal{A}$  induces a probability distribution  $\succ \mapsto P(\succ)$  on  $\mathcal{SWO}$ , regardless of the joint distribution of  $\mathbf{U}$ , through

$$P(\succ) = \Pr\left(\left[\bigcap_{w \succ x} (\mathbf{U}_w > \mathbf{U}_x)\right] \cap \left[\bigcap_{\neg[y \succ z]} (\mathbf{U}_y \leq \mathbf{U}_z)\right]\right).$$

**Random utility representations of weak orders:** a unidimensional (i.e.,  $k = 1$ ) random utility model on  $\mathcal{A}$  induces a probability distribution  $\succsim \mapsto P(\succsim)$  on  $\mathcal{WO}$ , regardless of the joint distribution of  $\mathbf{U}$ , through

$$P(\succsim) = \Pr\left(\left[\bigcap_{w \succsim x} (\mathbf{U}_w \geq \mathbf{U}_x)\right] \cap \left[\bigcap_{\neg[y \succsim z]} (\mathbf{U}_y < \mathbf{U}_z)\right]\right).$$

**Random utility representations of semiorders:** a unidimensional (i.e.,  $k = 1$ ) random utility model on  $\mathcal{A}$  induces a probability distribution  $\succ \mapsto P(\succ)$  on  $\mathcal{SO}$ , regardless of the joint distribution of  $\mathbf{U}$  through, given a strictly positive real valued (constant) threshold  $\varepsilon \in \mathbb{R}^{++}$ ,

$$P(\succ) = \Pr\left(\left[\bigcap_{w \succ x} (\mathbf{U}_w > \mathbf{U}_x + \varepsilon)\right] \cap \left[\bigcap_{\neg[y \succ z]} (\mathbf{U}_y - \mathbf{U}_z \leq \varepsilon)\right]\right).$$

**Random utility representations of interval orders:** if  $k = 2$  and  $\Pr(\mathbf{U}_{x,1} \leq \mathbf{U}_{x,2}) = 1$ ,  $\forall x \in \mathcal{A}$ , then, writing  $\mathbf{L}_x$  for  $\mathbf{U}_{x,1}$  (lower utility) and  $\mathbf{U}_x$  for  $\mathbf{U}_{x,2}$  (upper utility),  $\mathbf{U}$  induces a probability distribution  $\succ \mapsto P(\succ)$  on  $\mathcal{IO}$ , through,

$$P(\succ) = \Pr\left(\left[\bigcap_{w \succ x} (\mathbf{L}_w > \mathbf{U}_x)\right] \cap \left[\bigcap_{\neg[y \succ z]} (\mathbf{L}_y \leq \mathbf{U}_z)\right]\right).$$

Conversely, each probability distribution on  $\mathcal{SLO}$ ,  $\mathcal{SWO}$ ,  $\mathcal{WO}$ ,  $\mathcal{SO}$ , or  $\mathcal{IO}$  can be represented in the above fashion (nonuniquely) by an appropriately chosen family of jointly distributed random variables.

We do not go into details for  $k > 2$ , but such cases are of interest, e.g., to multi-attribute preferences, “ $m$ -ary” relations, more general partial order preferences, and other cases (see, e.g., Regenwetter and Marley, 2001, for related discussions).

For the next results, we need various concepts from probability theory, which we describe in general terms. For formal definitions, we refer the reader to any textbook on probability theory. A probability space  $\langle \Omega, \sigma, \mathbb{P} \rangle$  is a triple consisting of a “sample space”  $\Omega$  of “elementary outcomes,” a “sigma-algebra”  $\sigma$  of “events,” and a probability measure  $\mathbb{P}$  that assigns a probability in  $[0, 1]$  to each event in the sigma-algebra. Each event is a suitable (“measurable”) set of elementary outcomes. For example, when rolling an unloaded six-sided die, each of the six sides is an elementary outcome, rolling an even number is an example of an event, and the probability of that event, for example, is  $\frac{1}{2}$ .

We now introduce function spaces abstractly followed by a small concrete illustration. By a function on  $\mathcal{A}$ , we mean a mapping from  $\mathcal{A}$  into the real numbers  $\mathbb{R}$ . The collection of all functions on  $\mathcal{A}$  is the space  $\mathbb{R}^{\mathcal{A}}$ . When  $\mathcal{A}$  contains  $N$  elements, this is  $\mathbb{R}^N$ , the  $N$ -dimensional reals.

For example,  $(2.5, 3.1, 7.23) \in \mathbb{R}^3$  is the mapping

$$1 \mapsto 2.5 \quad 2 \mapsto 3.1 \quad 3 \mapsto 7.23.$$

Let  $\mathcal{B}(\mathbb{R}^{\mathcal{A}})$  denote the “sigma-algebra of Borel sets” in  $\mathbb{R}^{\mathcal{A}}$ , that is, the smallest collection of subsets of  $\mathbb{R}^{\mathcal{A}}$  that is closed under countable union and which contains all all open sets and complements of open sets in  $\mathbb{R}^{\mathcal{A}}$ .

**Definition 7.37** Consider a finite master set  $\mathcal{A}$  of two or more choice alternatives. A *random function model* for  $\mathcal{A}$  is a probability space  $\langle \mathbb{R}^{\mathcal{A}}, \mathcal{B}(\mathbb{R}^{\mathcal{A}}), \mathbb{P} \rangle$ .

The idea behind a random function model is to define a probability measure  $\mathbb{P}$  on the space of (e.g., utility) functions on  $\mathcal{A}$ . This space, of course, contains all conceivable unidimensional, real-valued, utility functions on  $\mathcal{A}$ . We can now summarize key results about binary choice probabilities induced by linear orders, as given in Equation (7.8) of Definition 7.19.

**Theorem 7.8** Consider a finite set  $\mathcal{A}$  of choice alternatives and a complete collection  $(P_{xy})_{x,y \in \mathcal{A}, x \neq y}$  of binary choice probabilities. The binary choice probabilities are induced by strict linear orders (7.8) if and only if they are induced by a (distribution-free) unidimensional, noncoincident random utility model (Block and Marschak, 1960). Furthermore, this holds if and only if they are induced by a (distribution-free) random function model with one-to-one functions (Regenwetter and Marley, 2001). Formally, there exists a probability distribution on  $\mathcal{SLO}_{\mathcal{A}}$  with  $P(\succ)$  the probability of  $\succ \in \mathcal{SLO}_{\mathcal{A}}$ , such that Equation (7.8) holds, i.e.,

$$P_{xy} = \sum_{\substack{\succ \in \mathcal{SLO}_{\mathcal{A}} \\ x \succ y}} P(\succ), \quad (\forall x, y \in \mathcal{A}, x \neq y), \quad (\text{see Eq. 7.8})$$

if and only if there exists a family of jointly distributed random variables  $(\mathbf{U}_x)_{x \in \mathcal{A}}$  that are noncoincident, i.e.,  $\Pr(\mathbf{U}_x = \mathbf{U}_y) = 0, \forall x, y \in \mathcal{A}, x \neq y$ , such that

$$P_{xy} = \Pr(\mathbf{U}_x > \mathbf{U}_y), \quad (\forall x, y \in \mathcal{A}, x \neq y), \quad (7.37)$$

if and only if there exists a probability space  $\langle \mathbb{R}^{\mathcal{A}}, \mathcal{B}(\mathbb{R}^{\mathcal{A}}), \mathbb{P} \rangle$ , such that

$$\begin{aligned} P_{xy} &= \mathbb{P}(\{u \in \mathbb{R}^{\mathcal{A}} \mid u(x) > u(y), \text{ where } u \text{ is a one-to-one function}\}) \\ &\quad (\forall x, y \in \mathcal{A}, x \neq y). \end{aligned} \quad (7.38)$$

A complete collection  $(P_{xy})_{x,y \in \mathcal{A}, x \neq y}$  of binary choice probabilities satisfies any one of these three conditions if and only if

$$(P_{xy})_{x,y \in \mathcal{A}, x \neq y} \in \mathcal{SLOP}_{\mathcal{A}}(\text{Binary}).$$

Suck (2002) presents necessary conditions for the random variables in this theorem to be independent for a finite set  $\mathcal{A}$ , including necessary and sufficient conditions in the case when  $|\mathcal{A}| = 3$ . For  $|\mathcal{A}| > 3$ , we do not know which conditions are sufficient.

If we want to drop the requirement that the random utilities be noncoincident or that the utility functions be one-to-one functions, then we obtain the following theorem.

**Theorem 7.9** (Regenwetter and Marley, 2001; Regenwetter and Davis-Stober, 2012) Consider a finite set  $\mathcal{A}$  of choice alternatives and a complete collection  $(T_{xy})_{x,y \in \mathcal{A}, x \neq y}$  of ternary paired-comparison probabilities. The ternary paired-comparison probabilities are induced by strict weak orders (Equation (7.17) with  $\mathcal{R} = \mathcal{SWO}_{\mathcal{A}}$ ) if and only if they are induced by a (distribution-free) unidimensional random utility model. Furthermore, this holds if and only if they are induced by a (distribution-free) random function model. Formally, there exists a probability distribution on  $\mathcal{SWO}_{\mathcal{A}}$  with  $P(\succ)$  the probability of  $\succ \in \mathcal{SWO}_{\mathcal{A}}$ , such that Equation (7.17) holds, i.e.,

$$T_{xy} = \sum_{\substack{\succ \in \mathcal{SWO}_{\mathcal{A}} \\ x \succ y}} P(\succ) \quad (\forall x, y \in \mathcal{A}, x \neq y), \quad (\text{see Equation 7.17})$$

if and only if there exists a family of jointly distributed random variables  $(\mathbf{U}_x)_{x \in \mathcal{A}}$  such that

$$T_{xy} = \Pr(\mathbf{U}_x > \mathbf{U}_y), \quad (\forall x, y \in \mathcal{A}, x \neq y),$$

if and only if there exists a probability space  $\langle \mathbb{R}^{\mathcal{A}}, \mathcal{B}(\mathbb{R}^{\mathcal{A}}), \mathbb{P} \rangle$ , such that

$$T_{xy} = \mathbb{P}(\{u \in \mathbb{R}^{\mathcal{A}} \mid u(x) > u(y)\}), \quad (\forall x, y \in \mathcal{A}, x \neq y). \quad (7.39)$$

A complete collection  $(T_{xy})_{x,y \in \mathcal{A}, x \neq y}$  of ternary paired comparison probabilities satisfies any one of these three conditions if and only if

$$(T_{xy})_{x,y \in \mathcal{A}, x \neq y} \in \mathcal{SWOP}_{\mathcal{A}}(\text{Ternary}).$$

The equivalences in these two theorems state that whenever one of three mathematical representations for binary choice or ternary paired-comparison probabilities exists, the others exist as well. It is important to note, however, that each of the three representations can have different special cases of particular interest. For example, the linear ordering model specification can be specialized by considering

parametric ranking models of a particular type, such as “Mallows’  $\Phi$ -models,” in the right term of Equation (7.8) (Critchlow *et al.*, 1993). The random utility formulation can be specialized by considering parametric families of random utilities, such as logit or probit models, and many others in the right term of Equation (7.37) (Böckenholt, 2006; Train, 2003). Last but not least, the random function model can be specialized by considering only certain (utility) functions  $u$  that must satisfy a particular form of cumulative prospect theory (Stott, 2006) in the right-hand side of Equation 7.38 or Equation 7.39 (Regenwetter *et al.*, 2014). We will discuss these issues in more detail below. In particular, we give an example of a random function model containing cumulative prospect theory with Goldstein–Einhorn weighting functions, and with power utility functions, as the deterministic special case.

Similar results to those of Theorem 7.8 also hold for other nonbinary response probabilities induced by rankings (strict linear orders). We state these in the cases of best choice and best-worst choice probabilities.

**Theorem 7.10** *Consider a finite set  $\mathcal{A}$  of choice alternatives, and a complete collection  $(B_X(x))_{x \in X \subseteq \mathcal{A}}$  of best-choice probabilities on  $\mathcal{A}$ . The best choice probabilities are induced by strict linear orders if and only if they are induced by a (distribution-free) unidimensional, noncoincident random utility model (Block and Marschak, 1960). Furthermore, this holds if and only if they are induced by a (distribution-free) random function model with one-to-one functions (Falmagne, 1978). Formally, there exists a probability distribution on  $\mathcal{SLO}_{\mathcal{A}}$  with  $P(\succ)$  the probability of  $\succ \in \mathcal{SLO}_{\mathcal{A}}$ , such that*

$$B_X(x) = \sum_{\substack{\succ \in \mathcal{SLO}_{\mathcal{A}} \\ \text{Rank}_{X, \succ}(1)=x}} P(\succ), \quad (\forall x \in X \subseteq \mathcal{A}),$$

*if and only if there exists a family of jointly distributed random variables  $(\mathbf{U}_x)_{x \in \mathcal{A}}$  that are noncoincident (i.e.,  $\forall x, y \in \mathcal{A}, x \neq y, \Pr(\mathbf{U}_x = \mathbf{U}_y) = 0$ ) such that*

$$B_X(x) = \Pr(\mathbf{U}_x = \max_{y \in X} \mathbf{U}_y), \quad (\forall x \in X \subseteq \mathcal{A})$$

*if and only if*

$$B_X(x) = \mathbb{P}\left(\left\{u \in \mathbb{R}^{\mathcal{A}} \mid u(x) = \max_{y \in X} u(y), \text{ where } u \text{ is a 1-to-1 function}\right\}\right),$$

*with suitable choices of probabilities and random variables on the right-hand side. Any one these three conditions holds if and only if*

$$(BW_X(x))_{x \in X \subseteq \mathcal{A}} \in \mathcal{SLOP}_{\mathcal{A}}(\text{Best}).$$

**Theorem 7.11** *Consider a finite set  $\mathcal{A}$  of choice alternatives, and a complete collection  $(BW_X(x, y))_{x, y \in X, x \neq y, X \subseteq \mathcal{A}}$  of best-worst choice probabilities on  $\mathcal{A}$ . The best-worst choice probabilities are induced by strict linear orders if and only if they are induced by a (distribution-free) unidimensional, noncoincident random utility model (paralleling Block and Marschak, 1960). Furthermore, this holds if*

and only if they are induced by a (distribution-free) random function model with one-to-one functions. Mathematically,

$$BW_X(x, y) = \sum_{\substack{\succ \in \mathcal{SLO}^A \\ Rank_{X, \succ}(1)=x, Rank_{X, \succ}(|X|)=y}} P(\succ)$$

if and only if

$$BW_X(x, y) = P\left(\mathbf{U}_x = \max_{w \in X} \mathbf{U}_w, \mathbf{U}_y = \min_{z \in X} \mathbf{U}_z\right), \text{ with } \Pr(\mathbf{U}_x = \mathbf{U}_y) = 0$$

if and only if

$$BW_X(x, y) = \mathbb{P}\left(\left\{u \in \mathbb{R}^A \mid u(x) = \max_{w \in X} u(w), u(y) = \min_{z \in X} u(z)\right\} \text{ where } u \text{ is a one-to-one function}\right),$$

with suitable choices of probabilities and random variables on the right-hand side. Any one these three conditions holds if and only if

$$(BW_X(x, y))_{\substack{x, y \in X \subseteq A \\ x \neq y}} \in \mathcal{SLOP}_A(\text{Best-Worst}).$$

### 7.5.1.1 Distribution-free random cumulative prospect theory: an example with Goldstein–Einhorn weighting and power utility

We briefly discuss an example of a random function model for binary choice, as stated in Equation (7.38), in which we only consider certain functions as permissible (Regenwetter *et al.*, 2014). In fact, we will consider only finitely many functions and an unknown probability distribution over them.

For simplicity, consider the following five two-outcome gambles (Regenwetter *et al.*, 2011a), each of which offers some probability of winning a positive gain, and otherwise neither a gain nor loss. Gamble *a* offers a 7/24 probability of winning \$28, gamble *b* offers a 8/24 probability of winning \$26.60, gamble *c* offers a 9/24 probability of winning \$25.20, gamble *d* offers a 10/24 probability of winning \$23.80, gamble *e* offers a 11/24 probability of winning \$22.40. These lotteries were 2007 dollar equivalents of those used by Tversky (1969).

Consider cumulative prospect theory with a Goldstein–Einhorn weighting function (Stott, 2006) using weighting parameters  $\gamma \in [0, 1]$  and  $s \in [0, 10]$  and a power utility function using a parameter  $\alpha \in [0, 1]$ . Denote these assumptions by  $CPT - GE$ . According to this  $CPT - GE$ , a gamble  $(x, p)$  with probability  $p$  of winning  $x$  (and nothing otherwise) has a subjective numerical value of

$$u(x, p) = \frac{sp^\gamma}{sp^\gamma + (1-p)^\gamma} x^\alpha.$$

We consider only finitely many such functions  $u$ , all of which are 1-to-1 on the domain  $A = \{a, b, c, d, e\}$ . More specifically, we only consider those values  $\alpha, \gamma \in [0.001, 0.991]$  that are multiples of 0.01 and those values of  $s \in [0.01, 9.96]$  that are multiples of 0.05 (Regenwetter *et al.*, 2014). The resulting random function model turns out to be equivalent to a linear ordering model (7.8), in which only 11 of the 120 strict linear orders in  $\mathcal{SLOP}_{\{a,b,c,d,e\}}$  have positive probability. Hence,

the random function model for  $\mathcal{CPT} - \mathcal{GE}$  on  $\mathcal{A} = \{a, b, c, d, e\}$  forms a convex polytope with 11 vertices. Regenwetter *et al.* (2014) report that the facet-defining inequalities for this model are

$$0 \leq P_{ab} \leq P_{ac} \leq P_{ad} \leq P_{bc} \leq P_{ae} \leq P_{bd} \leq P_{be} \leq P_{cd} \leq P_{ce} \leq P_{de} \leq 1.$$

Regenwetter *et al.* (2014) also characterize three other random function models. For instance, the random function model for  $\mathcal{CPT} - \mathcal{GE}$  on a different set of two-outcome gambles has 487 distinct facet-defining inequalities. Regenwetter *et al.* (2014) find that these models perform poorly in quantitative order-constrained statistical tests on laboratory data, even though the full linear ordering model fits the same data extremely well (Regenwetter *et al.*, 2011a).

Random function models can be highly restrictive, hence make strong testable predictions. For example, suppose that there are two distinct elements  $x, y \in \mathcal{A}$  such that every function  $u$  under consideration satisfies  $u(x) > u(y)$ , i.e.,  $y$  is dominated by  $x$  no matter what utility function a decision maker uses. Then the random function model predicts that  $P_{xy} = 1$  and  $P_{yx} = 1 - P_{xy} = 0$ , i.e., a decision maker satisfying this random function model has probability zero of choosing the dominated option  $y$ . Guo and Regenwetter (2014) discuss a similar set of conditions for the perceived relative argument model (PRAM) of Loomes (2010). Section 7.5.3 presents distributional (parametric) random utility models that do not share such restrictive features. Before proceeding to that material, we briefly summarize recent literature on axiomatizations of utility representations in probabilistic choice.

### 7.5.1.2 Axiomatizations of utility representations in probabilistic choice

It is of theoretical interest to axiomatize random function models for (binary) choice between gambles where the functions representing the gambles all have a common representation among those summarized in Section 7.4.3 – for instance, each function has the rank-dependent form; or, as a possibly simpler first case, each function has the cumulative prospect theory representation presented above. Gul and Pesendorfer (2006) present such an axiomatization where each function has the expected utility form. That axiomatization involves an independence (linearity) axiom that can be written in our notation as: for any risky gambles  $g, h, k$  and probability  $p$ ,

$$p(g, h) = p[(g, p; k, 1 - p), (h, p; k, 1 - p)]. \quad (7.40)$$

A random function model with each function having the rank-dependent form will not, in general, satisfy this independence axiom – because the representation (value) of each of the gambles on the right-hand side then depends on the value of gamble  $k$  relative to the value of each of  $g$  and  $h$ ; the challenge is to find a parallel condition (or conditions) for rank-dependent and other cases.

We are aware of three other papers that axiomatize binary probabilistic choice between gambles with underlying expected utility representations; that work may lead to ideas for extensions to, say, rank-dependent representations. Dagsvik (2008,

Theorem 2) axiomatizes the representation: for gambles  $g = (x_1, p_1; \dots; x_n, p_n)$  and  $h = (y_1, q_1; \dots; y_n, q_n)$ ,

$$p(g, h) = \Psi [hV(g) - hV(h)],$$

where

$$V(g) = \sum_{i=1}^n u(x_i)p_i \text{ and } V(h) = \sum_{i=1}^n u(y_i)q_i,$$

with  $u$  real-valued;  $\Psi$  a continuous and strictly increasing cumulative distribution function defined on  $\mathbb{R}$  with  $\Psi(r) + \Psi(-r) = 1$ ; and  $h : \mathbb{R} \rightarrow \mathbb{R}$  is strictly increasing. Dagsvik and Hoff (2011) build on the extensive axiomatic work by Falmagne (1985, chapter 14) (and the many other researchers that they mention) in the use of the methods of dimensional analysis to justify specific functional forms in their representation; they illustrate their methods with the estimation of the utility of income using data from a Stated Preference (SP) survey.

Blavatskyy (2009, Theorem 1) axiomatizes the special case of the above representation when  $h$  is the identity function.

Dagvilk's and Blavatskyy's axiomatization each assume an independence (linearity) axiom similar to that of Gul and Pesendorfer stated in (7.40). That, and/or others of their axioms, may need to be revised in any axiomatization of probabilistic choice with, say, an underlying rank-dependent form.

Finally, Blavatskyy (2012) axiomatizes a representation for binary probabilistic choice between gambles that is related to that in Blavatskyy (2009); the revised representation includes natural endogenous reference points. This representation has several properties, some of empirical interest, including:  $p(g, h) \geq 1/2$  iff the expected utility of  $g$  is greater than the expected utility of  $h$ ;  $p(g, h) = 1$  if  $g$  dominates  $h$ ; and the closer lottery  $g$  is to the endogenous lower bound reference lottery  $g \wedge h$ , the closer is  $p(g, h)$  to 0. Blavatskyy (2012) discusses, but does not axiomatize, extensions where the axioms would lead to  $u$  being a non-expected utility (for instance, a ranked weighted utility form as in Section 7.4.2) which would extend the description of additional classic (deterministic) phenomena to the probabilistic case. A second application of this model would be to binary choice between multiattribute options (see Section 7.5.5 for work on choice between more than two such options for the above model and others). A further extension would involve choices between risky (or uncertain) gambles where the pure outcomes are multiattribute options. Bleichrodt *et al.* (2009) axiomatize a (deterministic) representation for uncertain gambles with multiattribute outcomes, obtaining a version of prospect theory (i.e., with referents on each attribute or option) with additive representation of the outcomes.

### 7.5.2 Horse race models of choice and response time

A *horse race model* of best choice is a model of the choice made, as well as, usually, the response time, where the choice probabilities, and the response times, are

induced by a shared unidimensional, noncoincident random utility model (see Definitions 7.20 and 7.36 and Theorem 7.10, and Equations (7.42) and (7.43)). Such models have been developed and applied in discrete choice surveys, mainly by economists and marketing scientists; and in experiments, mainly by experimental economists and cognitive psychologists.

We develop general results on horse race models of choice and response time for best choices, only, for three reasons: first, the developments for worst choices exactly parallel those for best choices; second, those for best-worst choices are much more complicated; third, although special cases yield numerous standard representations for choice probabilities, the models are unable to fit various context effects in choice and have restrictive relations between choice and response times. Thus, we later both restrict the models in certain ways, and extend them in others, to make them tractable and realistic as models of both choice and response time.

### 7.5.2.1 Distribution-free horse race models

Assume a master set  $\mathcal{A}$  of options, and let  $X \subseteq \mathcal{A}$  be any subset with two or more options. For some purposes, it is useful to enumerate  $\mathcal{A}$  as  $\mathcal{A} = \{c_1, c_2, \dots, c_n\}$  where  $n = |\mathcal{A}|$ , and a typical  $X \subseteq \mathcal{A}$  as  $X = \{x_1, x_2, \dots, x_m\}$  where  $n \geq m = |X|$ .

For each  $x \in X \subseteq \mathcal{A}$ ,  $T_X$  is a nonnegative (real) random variable denoting the time at which a choice is made,  $B_X$  is a random variable denoting the (single “best”) option chosen,<sup>10</sup> and  $B_X(x; t)$ ,  $t \geq 0$ , is the probability that option  $x$  is chosen as best in  $X$  after time  $t$ . Thus

$$B_X(x; t) = \Pr[B_X = x \text{ and } T_X > t]$$

may be thought of as a *survival function* for option  $x$  when presented in the set  $X$ . A collection of survival functions  $\{B_X(x; t) : x \in X \subseteq \mathcal{A}\}$  for a fixed  $X \subseteq \mathcal{A}$  is a *joint structure of choice probabilities and response times*. A joint structure is *complete* if the collection of survival functions ranges over all subsets of  $\mathcal{A}$ ; unless otherwise specified, we assume that the joint structure is complete.

We now investigate a class of models generating such a joint structure, all of which belong to the class of (possibly context-dependent) horse race random utility models (cf. Marley, 1989; Marley and Colonius, 1992; those papers focus on the context-free case – see below). We will see that this class of models is sufficiently broad to include many of the standard models of choice used in the discrete choice literature, although they are inadequate for context effects on choice and the details of response time (see Busemeyer and Rieskamp, 2014; Rieskamp *et al.*, 2006, for summaries of the limitations of the choice predictions of these models). Section 7.5.4 discusses context-dependent extensions of these models that overcome numerous of their limitations.

For each  $z \in X \subseteq \mathcal{A}$ , let  $T_X(z)$  be the time for a process (e.g., an “accumulator,” as detailed in Section 7.5.3) associated with  $z$  to produce a specified event (e.g.,

<sup>10</sup> The notation has to be generalized for worst and best-worst choice, and for other designs, such as where the person can select a subset of options to consider (a *consideration set*, as in Swait, 2001).

to reach a “threshold”) when  $X$  is set of available choice options. When we give details for specific models,  $\mathbf{T}_X(z)$  will depend on various parameters such as a threshold; drift rates; drift rate variability; start point variability; etc. There is no assumption that the  $\mathbf{T}_X(z)$  are independent of each other; however, we will see that the models become vacuous when their dependence on  $X$  is not restricted in some manner.

Consider a generic set  $X = \{x_1, \dots, x_m\} \subseteq \mathcal{A}$  with  $|X| = m$  and assume that there is a multivariate (survival) distribution  $S_X$  such that: for  $t_i \geq 0$ ,  $i = 1, \dots, m$ ,

$$S_X(t_1, \dots, t_n) = \Pr(\mathbf{T}_X(x_1) > t_1, \dots, \mathbf{T}_X(x_m) > t_m). \quad (7.41)$$

In particular, there can be (complicated) dependencies between the  $\mathbf{T}_X(x_i)$ ,  $i = 1, \dots, m$ . We use the survival form because various results are easier to state and prove in that form, plus that is the way the main results were stated in earlier papers (Marley, 1989; Marley and Colonius, 1992).

Given  $X \subseteq \mathcal{A}$  as the currently available choice set, let  $\mathbf{B}_X$  be the option chosen and  $\mathbf{T}_X$  the time it is chosen. Then we assume that, for  $x \in X$  and  $t > 0$ ,

$$B_X(x; t) = \Pr(\mathbf{B}_X = x \text{ and } \mathbf{T}_X > t) = \Pr(t < \mathbf{T}_X(x) = \min_{z \in X} \mathbf{T}_X(z)), \quad (7.42)$$

where the distributions are given by the  $S_X$  in (7.41). The above expressions give the joint choice and (tail) response time (RT) distribution. The choice probabilities are given by (7.42) with  $t = 0$ , i.e.,

$$B_X(x) = \Pr(\mathbf{B}_X = x) = \Pr(\mathbf{T}_X(x) = \min_{z \in X} \mathbf{T}_X(z)). \quad (7.43)$$

In the general form (7.42), the  $\mathbf{T}_X(z)$  can depend on  $X$ , i.e., they can be *context-dependent*, and/or they can be dependent on each other. When none of the  $\mathbf{T}_X(z)$  in (7.42) depend on  $X$ , we have a *context-free* horse race random utility model (cf. Marley and Colonius, 1992), and when they are independent of each other, we have an *independent horse race random utility model*; note that all four combinations, either context-dependent or context-free, combined with either independent or not independent, are possible. Parallel terms apply to the random variables in (7.43).

There are two major limitations of context-free horse race random utility models, namely that: for  $X \subseteq \mathcal{A}$ ,  $y \in \mathcal{A} - X$ ,

- i.  $\Pr(\min_{z \in X} \mathbf{T}_X(z) > t) \geq \Pr(\min_{z \in X \cup \{y\}} \mathbf{T}_X(z) > t)$ ,  
and
- ii.  $B_X(x; t) \geq B_{X \cup \{y\}}(x; t)$ .

The first relation means that response times do not increase with set size (Colonius and Vorberg, 1995, Lemma 2) and the second that the choice probabilities satisfy *regularity*: the probability of choosing an option cannot increase when an additional option is added to the choice set. Both of these properties have been reported to fail in various experiments (Busemeyer and Rieskamp, 2014; Teoderescu and

Usher, 2013). These and other limitations are addressed when we consider context-dependent horse race models (Section 7.5.4).

The following standard results hold for the class of horse race random utility models for best choice; parallel results can be developed for worst, and best-worst, choice. For simplicity, we assume that  $B_X(x; t) > 0$  for all  $t \geq 0$ , though the results are easily generalized to various weaker conditions (see Marley, 1992); and that all distributions are absolutely continuous, although, again, various of the results are valid when this is not the case (see the study by Marley, 1989, of Tversky's *elimination by aspects (EBA) model*).

**Theorem 7.12** (*Marley and Colonius, 1992, Theorem 1*) (a) Consider a joint structure of (best) choice probabilities and response times  $\{B_X(x; t) : x \in X\}$  on a fixed finite set  $X$ . If each  $B_X(x; t)$ ,  $x \in X$ , is absolutely continuous and positive for all  $t \geq 0$ , then there exist unique independent random variables  $t_X(x)$ ,  $x \in X$ , such that (7.42) holds with those random variables. (b) A (complete) joint structure of choice probabilities and response times  $\{B_X(x; t) : x \in X \subseteq \mathcal{A}\}$  can be uniquely represented by an independent horse race random utility model if the conditions of (a) hold and

$$\frac{(d/dt)B_X(x; t)}{\sum_{z \in X} B_X(z; t)}$$

is independent of  $X \subseteq \mathcal{A}$  for all  $t \geq 0$ .

Essentially, (a) says that any set of (best) choice probabilities and response times on a fixed finite set  $X$  can be represented by an independent horse race random utility model, and thus, indirectly, that any (complete) set of (best) choice probabilities and response times on (all) subsets of finite set  $\mathcal{A}$  can be fit by a *context-dependent* horse race random utility model with the random variables independent within each choice set; although the random variables may differ across subsets. Part (b) gives a set of conditions under which a (complete) joint structure of choice probabilities and response times can be written as a context-free horse race random utility model with a common set of independent random variables across all the choice sets. Part (b) can be viewed as a result for the joint representation of choice probabilities and response times that is somewhat analogous to Falmagne's (1978) result for the representation of choice probabilities.

These results can be interpreted as saying that the assumption that a horse race random utility model with independent random variables holds for a *single* choice set  $X$  is a descriptive theoretical language rather than an empirically falsifiable model. Dzhafarov (1993) and Jones and Dzhafarov (2014) discuss related issues when response time distributions are modeled by a deterministic stimulation-dependent process that terminates when it crosses a randomly preset criterion, a framework that is closely related to that of the linear ballistic accumulators (LBAs) presented in Section 7.5.3. Heathcote *et al.* (2014) respond to that discussion.

Given the above results, one can fit "any" choice and response time data using a context-dependent horse race random utility model with the random variables

independent within each choice set; in particular, one can fit the classic context effects in choice (see Section 7.5.4 for those effects). Thus, the goal of context-dependent models must be to motivate plausible constraints on the nature of the context dependencies – see Section 7.5.4 for recent models with such constraints.

### 7.5.2.2 Luce's choice model derived from a horse race model

The following results are stated for best choices; exactly parallel results hold for worst and best-worst.

Before proceeding, we need a definition:

**Definition 7.38** A complete set of choice probabilities on the subsets of a finite set  $\mathcal{A}$  satisfy *Luce's choice model* if all the choice probabilities are nonzero and there exists a ratio scale  $b$  such that for every  $x \in X \subseteq \mathcal{A}$  with  $|X| \geq 2$ ,

$$B_X(x) = \frac{b(x)}{\sum_{z \in X} b(z)}. \quad (7.44)$$

This model is equivalent to the *multinomial logit (MNL) model* (see Section 7.5.3), which is usually written with scale values  $u(z) = \log b(z)$ , i.e.,  $b(z) = \exp u(z)$ .

Luce's choice model, and therefore the MNL model, satisfies the following *independence of irrelevant alternatives (IIA)* condition: for any  $x, y \in X \subseteq T$  with  $B_{\{x,y\}}(x) \neq 0, 1$ ,

$$\frac{B_{\{x,y\}}(x)}{B_{\{x,y\}}(y)} = \frac{B_X(x)}{B_X(y)}.$$

This is a much stronger form of context-independence than the context-free condition introduced above. This fact is illustrated by the next theorem, which presents a strong condition under which a context-free horse race random utility model satisfies Luce's choice model.

For simplicity in the following theorem, we assume that all the choice probabilities are nonzero. The result can be generalized when this is not so by adding a connectivity and a transitivity condition (Luce, 1959, Theorem 4, p. 25).

**Theorem 7.13** (*Marley and Colonius, 1992, Theorem 2*). Consider an independent, context-free horse race random utility model where for each  $x \in X \subseteq \mathcal{A}$ ,  $B_X(x; t)$  is absolutely continuous and positive for all  $t \geq 0$  and  $B_X(x)$  is nonzero. If the option chosen,  $C_X$ , is independent of the time of choice,  $T_X$ , then the choice probabilities satisfy Luce's choice model, (7.44).

This derivation of Luce's choice model is interesting theoretically. However, as we now show, the resultant model of choice and response times is, in general, unsatisfactory. The assumption of Theorem 7.13 is that

$$\Pr(B_X = x \text{ and } T_X > t) = \Pr(B_X = x) \Pr(T_X > t). \quad (7.45)$$

From this, we easily derive

$$\Pr(\mathbf{T}_X > t | B(X = x) = \Pr(\mathbf{T}_X > t).$$

This result says that the distribution of response times is independent of the option chosen, which is generally incorrect – for example, the most preferred (“correct”) option is often chosen faster than less-preferred options (although experimental manipulations can give the opposite). Because there are considerable data showing that the Luce choice model does not provide a satisfactory description of various data (Busemeyer and Rieskamp, 2014; Rieskamp *et al.*, 2006), the above result should not concern us if it does not hold when the choice probabilities have a more general form than the Luce choice model. When the MNL model was found to be inadequate, it was generalized to a class of dependent random utility models that gave more general closed-form representations for choice probabilities – namely, the generalized extreme value (GEV) class (McFadden, 1978; Marley, 1989). Marley (1989) showed that there is a class of dependent horse race random utility models for choice and responses times that generates numerous members of that class of models (and related representations) for choice probabilities. However, regrettably, that class continues to have the property that the choices made and the time to make them are independent. Although models in the more general class have the major negative property of the probability of the choice made being independent of the time to make it – that is, (7.45) holds – Marley (1989) also shows that we can overcome that limitation by considering mixtures of models of the above form; to date, the properties of the latter class of models have not been studied in detail for response times. A further extension would be to add context-dependence, which is relatively easily done in the parametric representations of the class of models presented in Marley (1989). Rather than pursue that direction, we return to the most basic models in this class; reinterpret them in terms of accumulators with thresholds; summarize recent context-dependent models in that framework; and extend them to worst and best-worst choice. Most such accumulator models have the highly desirable property of the distribution of response times being dependent on the option chosen.

### **7.5.3 Context free linear accumulator models of choice and response time**

The class of models that we studied in Section 7.5.2 has the property that the option chosen is independent of the time of choice, (7.45). Note that if this condition holds, and we let  $B_X(x, t)$  be the probability that option  $x$  is chosen *at or before time t*, then it follows that

$$\begin{aligned} B_X(x, t) &= \Pr(B_X = x \text{ and } \mathbf{T}_X \leq t) \\ &= \Pr(B_X = x) \Pr(\mathbf{T}_X \leq t). \end{aligned}$$

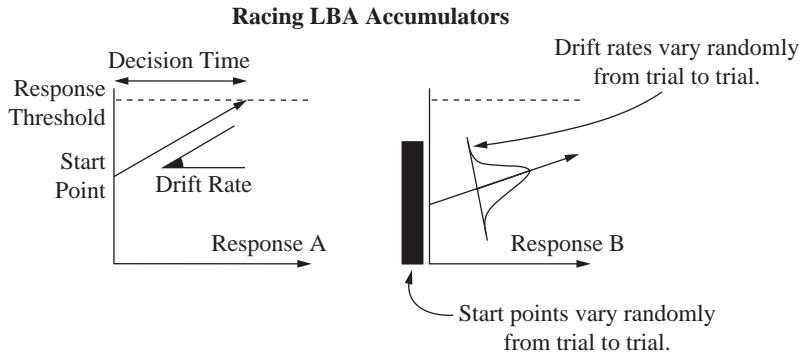
Thus we have the independence property in terms of the cumulative distribution, as well as for the survival distribution. Because most response time models are written in terms of cumulative distributions, we will use that form in the present section, even though we study models that do not satisfy the above independence property.

Context-free accumulator models are essentially horse race random utility models with additional structure imposed on the properties of the random variables; Section 7.5.4 presents context-dependent accumulator models. Here, we discuss the *linear ballistic accumulator (LBA) model* (Brown and Heathcote, 2008). Like other multiple accumulator models, the LBA is based on the idea that the decision maker accumulates evidence in favor of each (best and/or worst) choice, and makes a decision as soon as the evidence for any such choice reaches a threshold amount. The time to accumulate evidence to threshold is the predicted decision time, and the response time is the decision time plus a fixed offset ( $t_0$ ), the latter accounting for processes such as response production. The LBA shares this general evidence accumulation framework with many models (Busemeyer and Rieskamp, 2014), but has a practical advantage – namely, the computational tractability afforded by having an easily computed expression for the joint likelihood of a given response time and response choice among any number of options.

Figure 7.5 gives an example of an LBA decision between two options, A and B. The A and B response processes are represented by separate accumulators that race against each other. The vertical axes represent the amount of accumulated evidence, and the horizontal axes the passage of time. Response thresholds ( $b$ ) are shown as dashed lines in each accumulator, indicating the quantity of evidence required to make a choice. The amount of evidence in each accumulator at the beginning of a decision (the “start point”) varies independently between accumulators and randomly from choice to choice, usually assumed sampled from a uniform distribution:  $U(0, A)$ , with  $A \leq b$ . Evidence accumulation is linear, as illustrated by the arrows in each accumulator of Figure 7.5. The rate of accumulation is traditionally referred to as the “drift rate,” and this is assumed to vary randomly from accumulator to accumulator and decision to decision according to an independent distribution for each accumulator, reflecting choice-to-choice changes in factors such as attention and motivation. Rate means,  $d$ , and standard deviations,  $s$ , can differ between accumulators and options, although at least one value of  $s$  is usually fixed as a scaling constant.

In the following, we consider two classes of LBA – both assume subtractive start point variability, but one model assumes multiplicative drift rate variability, the second assumes additive drift rate variability; both classes assume the random variables  $\mathbf{T}_X(x)$  are independent. The first class relates in a natural way to the horse-race models in Section 7.5.2, the second class agrees with the assumptions in the original development of the LBA.

This section is confined to context-free models; Section 7.5.4 presents context-dependent models. Also, to agree with published papers, we use  $\mathbf{B}_x$  to denote the



**Figure 7.5** Illustrative example of the decision processes of the linear ballistic accumulator (LBA). Figure 1 of Hawkins et al. (2014a). Reproduced with permission from John Wiley & Sons.

distribution of time for option  $x$  to reach threshold, when the task is to choose the best of the available options.

### 7.5.3.1 Multiplicative drift rate variability

In the following summaries, we ignore the fixed offset ( $t_0$ ). Let  $\Delta$  be a random variable on the positive reals; this will be the distribution of the (multiplicative) drift rate variability. Let, for  $r > 0$ ,

$$\Pr(\Delta < r) = G(r), \quad (7.46)$$

where  $G$  is a cumulative distribution function (CDF). Let  $b$  be the threshold and let  $x$  be a typical choice option with drift rate  $d(x)$ , and let  $\Sigma$  be a second random variable on  $[0, A]$  with  $A \leq b$ ; this will be the distribution of the start point. Let, for  $h \in [0, A]$ ,

$$\Pr(\Sigma < h) = H(r).$$

For option  $x$ , the start point has distribution  $\Sigma$  and the drift rate has distribution  $\Delta d(x)$ ; also, assume the samples of  $\Sigma$  (respectively,  $\Delta$ ) for different  $x$ 's are independent, which we denote by  $\Sigma_x$  (respectively,  $\Delta_x$ ); thus, the subscripts do not imply dependence on  $x$  beyond the assumption of independent samples for different  $x$ 's. Then, with  $B_x = \frac{b - \Sigma_x}{\Delta_x d(x)}$  the random variable for the time taken for the LBA with (mean) drift rate  $d(x)$  to reach threshold, we have

$$\Pr(B_x < t) = \Pr\left(\frac{b - \Sigma_x}{\Delta_x d(x)} \leq t\right). \quad (7.47)$$

The major LBA applications to date have assumed that the *start point variability* is given by a random variable  $p$  with uniform distribution on  $[b - A, b]$ , in which case (7.47) can be rewritten as

$$\Pr(B_x < t) = \Pr\left(\frac{p_x}{\Delta_x d(x)} \leq t\right). \quad (7.48)$$

The above representation is a perfectly plausible multiplicative LBA model for any cumulative distributions  $G$  on the nonnegative reals and  $H$  on  $[0, A]$  with  $A \leq b$ . However, a major argument advanced for LBA models is their computational tractability in terms of probability density functions (PDF) and cumulative density functions (CDFs). In particular, it is usually assumed that the accumulators are independent, and therefore the main need is for the PDF and CDF of, say, (7.47) to be tractable. We summarize the form of the CDF corresponding to that expression, then summarize what is known about such forms.

With  $\mathbf{p}$  the uniform distribution on  $[b - A, b]$ , and the CDF of  $G$  given by (7.46), standard calculations show that  $\Pr(\mathbf{B}_x \leq t)$  of (7.48) can be written in terms of: the constants  $b$  and  $A$  and the term  $td(x)$ ; the mean of the distribution  $G$  when truncated to the interval  $[\frac{b}{td(x)}, \frac{b-A}{td(x)}]$ ; and the values of  $G\left(\frac{b-A}{td(x)}\right)$  and  $G\left(\frac{b}{td(x)}\right)$  (Terry *et al.*, 2015). Closed forms are known for the first quantity for various CDFs on the nonnegative reals, including for Gamma, inverted Beta, Fréchet, and Levy distributions (Nadarajah, 2009). However, one of the major motivations for the LBA framework was to achieve easily computable forms for the joint distribution of choice probabilities and choice times. In addition to independence, this requires the PDFs corresponding to the CDFs, above, to be easily computable; Terry *et al.* (2015) show that this is the case for various distributional assumptions (including the Fréchet, which we discuss in the next section), and present empirical tests of those assumptions.

Later, we describe the original LBA model, which has additive drift rate variability, in which case the PDFs are also easily computable and the model does not imply that the choice made is independent of the time of choice. However, before proceeding to that material, we take a step back to present a model with multiplicative drift rate variability that is equivalent (for choices) to the most basic, and frequently applied, random utility models of best, worst, and best-worst choice.

### 7.5.3.2 Relation of multiplicative LBA models with no start point variability and Fréchet drift rate variability to MNL models

We now present a multiplicative LBA with no start point variability and Fréchet drift scale variability that leads to a standard set of models for best, worst, and best-worst choice. Consider the special case of (7.48), where there is no start point variability, i.e.,  $A = 0$ , and so  $\mathbf{p}$  has a constant value  $b$ ; in this case, without loss of generality, we can set  $b = 1$ . Then (7.48) reduces to

$$\Pr(\mathbf{B}_x < t) = \Pr\left(\frac{1}{\Delta d(x)} \leq t\right).$$

We also assume that  $\Delta$  has a Fréchet distribution, i.e., there are constants  $\alpha, \beta > 0$  such that, for  $r \geq 0$ ,

$$\Pr(\Delta < r) = G(r) = e^{-(\alpha r)^{-\beta}}. \quad (7.49)$$

The following is then an obvious manner to obtain a set of models for best, worst, and best-worst choice, respectively, and the corresponding response times.<sup>11</sup> As above,  $B_X(x, t)$  denotes the probability of choosing  $x$  as the best option in  $X$  before time  $t$ . The corresponding notation for worst is  $W_X(y, t)$  and for best-worst is  $BW_X(x, t; y, t)$ .

We then assume:

- i. Best, with drift rates  $d(z)$ ,  $z \in X$ :

$$B_X(x, t) = \Pr \left( \frac{1}{\Delta_x d(x)} = \min_{z \in X} \frac{1}{\Delta_z d(z)} \leq t \right). \quad (7.50)$$

- ii. Worst, with drift rates  $1/d(z)$ ,  $z \in X$ :

$$W_X(y, t) = \Pr \left( \frac{1}{\Delta_y \frac{1}{d(y)}} = \min_{z \in X} \frac{1}{\Delta_z \frac{1}{d(z)}} \leq t \right). \quad (7.51)$$

- iii. Best-worst, with drift rates  $d(p)/d(q)$  and for all  $p, q \in X$ ,  $p \neq q$ ,

$$BW_X(x, t; y, t) = \Pr \left( \frac{1}{\Delta_{x,y} \frac{d(x)}{d(y)}} = \min_{\substack{p,q \in Y \\ p \neq q}} \frac{1}{\Delta_{p,q} \frac{d(p)}{d(q)}} < t \right) \quad (x \neq y). \quad (7.52)$$

Note that (7.52) implies that the best and the worst option are chosen at the same time; this is an unreasonably strong assumption, which we weaken later.

Then routine calculations applied to (7.50–7.52) (paralleling those in Marley, 1989) give: for  $X \subseteq \mathcal{A}$  and  $x, y \in X$ ,

$$B_X(x; t) = \frac{d(x)^\beta}{\sum_{z \in X} d(z)^\beta} \left[ 1 - \exp - \left\{ \left( \frac{b}{\alpha} \right)^\beta \sum_{z \in X} d(z)^\beta \right\} \right], \quad (7.53)$$

$$W_X(y, t) = \frac{1/d(y)^\beta}{\sum_{z \in X} 1/d(z)^\beta} \left[ 1 - \exp \left\{ \left( \frac{b}{\alpha} \right)^\beta \sum_{z \in X} \frac{1}{d(z)^\beta} \right\} \right], \quad (7.54)$$

and

$$\begin{aligned} BW_X(x, t; y, t) &= \frac{d(x)^\beta / d(y)^\beta}{\sum_{\substack{r,s \in X \\ r \neq s}} d(r)^\beta / d(s)^\beta} \left[ 1 - \exp \left\{ \left( \frac{b}{\alpha} \right)^\beta \sum_{\substack{r,s \in X \\ r \neq s}} \frac{d(r)^\beta}{d(s)^\beta} \right\} \right] \quad (x \neq y). \end{aligned} \quad (7.55)$$

<sup>11</sup> Remember that the subscript on  $\Delta_z$ ,  $\Delta_{x,y}$ , etc., indicates independent samples, not other dependence on  $z$  or  $x,y$ .

The corresponding choice probabilities  $B_X(x)$ ,  $W_X(y)$ ,  $BW_X(x, y)$  are given by these formula in the limit as  $t \rightarrow \infty$ , i.e., when the expressions in square brackets are set equal to 1. Now, for  $z, p, q \in \mathcal{A}$ ,  $p \neq q$ , let  $u(z) = \ln d(z)$ ,  $\epsilon_z = \ln \Delta_z$ ,  $\epsilon_{p,q} = \ln \Delta_{p,q}$ . Then the choice probabilities can equally well be written as: for all  $x, y \in X \in D(\mathcal{A})$ ,

$$B_X(x) = \Pr \left( u(x) + \epsilon_y = \max_{z \in X} [u(z) + \epsilon_z] \right), \quad (7.56)$$

$$W_X(y) = \Pr \left( -u(y) + \epsilon_y = \max_{z \in X} [-u(z) + \epsilon_z] \right), \quad (7.57)$$

and for all  $x, y \in X \in D(\mathcal{A})$ ,  $x \neq y$ ,

$$BW_X(x, y) = \Pr \left( u(x) - u(y) + \epsilon_{x,y} = \max_{\substack{p,q \in X \\ p \neq q}} [u(p) - u(q) + \epsilon_{p,q}] \right). \quad (7.58)$$

However, given that  $\Delta_z$  and  $\Delta_{p,q}$  are generated by Fréchet drift rate variability, i.e., (7.49), we have that  $\epsilon_z = -\ln \Delta_z$  and  $\epsilon_{p,q} = -\ln \Delta_{p,q}$  satisfy extreme value distributions.<sup>12</sup> When treated as a single model, the three models (7.56), (7.57), and (7.58) then satisfy an *inverse extreme value maximum<sup>13</sup> random utility model* (Marley and Louviere, 2005, Definition 11). Standard results (summarized by Marley and Louviere, 2005,<sup>14</sup> and the derivations above) show that the expression for the choice probabilities given by (7.56) (respectively, (7.57), (7.58)) agrees with a standard multinomial logit (MNL) form. Note that  $\beta$  is not identifiable from the choice probabilities, but it is, in general, identifiable when response times are also available (although the form predicted for the response time distribution is not suitable for data). In particular, when (7.53) and (7.54) both hold with  $F(t) \rightarrow 1$ , we have that for all  $x, y \in X$ ,  $x \neq y$ ,  $B_{\{x,y\}}(x) = W_{\{x,y\}}(y)$ ; empirically, this relation may not always hold (Shafir, 1993).

As already noted, each of the above models for choice and response times have the property that the option chosen is independent of the time of choice; this is not the case if start point variability is added (Terry *et al.*, 2015). Also, although the underlying utility maximization process may be cognitively plausible for best (or worst) choices, that for best-worst choice in (7.52) appears to require the participant to “simultaneously” compare all possible discrete pairs of options in the choice set, leading to a significant cognitive load; as already noted, it also implies that the best and the worst option are chosen at the same time. Fortunately, Marley and Louviere (2005) present the following plausible process

<sup>12</sup> With, in this case: for  $-\infty < t < \infty$   $\Pr(\epsilon_z \leq t) = \exp -(\alpha e^t)^{-\beta}$  and  $\Pr(\epsilon_{p,q} \leq t) = \exp -(\alpha e^t)^{-\beta}$ .

<sup>13</sup> We have added *maximum* to Marley and Louviere’s definition to emphasize that the random utility models of *choice* are written in terms of maxima, whereas the equivalent (“horse race,” accumulator) models of response time are written in terms of minima.

<sup>14</sup> Marley and Louviere used the case  $\alpha = \beta = 1$ .

involving separate best and worst choice processes that generates the same choice probabilities: the person chooses the best option in  $X$  and, independently, chooses the worst option in  $X$ ; if the resulting choices differ, then they are given as the responses; otherwise, the person repeats the process until the selected pair of options do differ. Then Marley and Louviere (2005, section 4.1.2, case 2, and above) show that if the best (respectively, worst) choices satisfy the MNL for best, (7.53) with  $F(t) \rightarrow 1$  (respectively, MNL for worst, (7.54) with  $F(t) \rightarrow 1$ ) choices, with  $v = -u$ , then the best-worst choices given by the above process satisfy the maxdiff model, (7.55) with  $F(t) \rightarrow 1$ ; in particular, we have a more cognitively plausible model than (7.58) for the best-worst choices. The next section presents the parallel (additive) LBA model which extends the above process to response times.

### 7.5.3.3 Additive drift rate variability

In contrast to the multiplicative drift rate model presented above, the linear ballistic accumulator (LBA) model assumes *additive drift rate variability* generated by independent normal random variates, truncated at zero – i.e., constrained to be nonnegative.<sup>15</sup> This model is a horse race random utility model, but does not have the undesirable property that the choices made and the time to make them are independent. Nonetheless, it has been shown to make extremely similar predictions for best, worst, and best-worst choice probabilities to those made by the multiplicative LBA model with no start point variability and Fréchet drift scale variability (Hawkins *et al.*, 2014a,b). The derivations for the additive model exactly parallel those described above with  $\Delta d(x)$  replaced by  $\text{trunc}(\mathbf{D}_z + d(z))$ , again with a major advantage of the additive model being that all formulae for CDFs and PDFs are computationally tractable when the start point distribution is uniform. We now briefly summarize recent applications of this model to some standard data on best, worst, and best-worst choice and in Section 7.5.4 summarize its extension to handle context effects in best choice and response times.

Hawkins *et al.* (2014a,b) fit four best-worst LBA models to each of three sets of data – two sets of best-worst choice data obtained in discrete choice experiments (often called “DCEs”), the first involving choice between aspects of dermatology appointments, the second between mobile phones; and one set of best-worst choice and response time data in a perceptual judgment task: choosing the rectangle with the largest (respectively, smallest) area in a set of four rectangles presented on a trial, with multiple presentations of various sets of rectangles. Each additive LBA model is based on processes similar to those already presented in the discussion of the multiplicative Fréchet model. In each fit of the models to only the best and/or worst choices, an essentially linear relationship was found between log drift rate and utility estimates for the corresponding MNL model; and goodness of fit, as determined by log-likelihood and root mean square error, was comparable.

<sup>15</sup> In their data analyses, Brown and Heathcote (2008) ignored the small percentage of times that the normal distributions could take on negative values; Heathcote and Love (2012) present the correct analysis in terms of truncated normals.

However, when both choice and response time data are fit, three of the models are inappropriate: one because it implies that best choices are always made before worst choices; the second because it implies the reverse; and the third because it implies that the best and worst choices are made simultaneously. The relevant one of the first two models might be appropriate in experiments where participants are forced, by the design, to respond in the order best, then worst (respectively, worst, then best); however, the third is implausible (for response times) under any reasonable circumstances. Participants in the perceptual judgment task were free to make the best and worst choice in whatever order they wished. The data show: large differences between participants in best-then-worst versus worst-then-best responding, but normally with both orders for each participant; changes in the proportion of response order as a function of choice difficulty; and changes in interresponse time due to choice difficulty. These data are incompatible with three of the models, and, as Hawkins *et al.* (2014b) show, they cannot be satisfactorily fit by mixtures of those models. Therefore, we restrict our (brief) presentation to the fourth model, called the *parallel (additive) LBA model*, which is an extension of Marley and Louviere's (2005, section 4.1.2, case 2, described in the previous section of this chapter) process model for choice to response time.

The parallel (additive) LBA model for best-worst choice assumes concurrent best and worst races, with, in general, different (drift rate) parameters in the two sets of races; Hawkins *et al.*'s (2014b) data analysis assumed that the worst drift rate is the reciprocal of the corresponding best drift rate. The option chosen as best (respectively, worst) is that associated with the first accumulator to reach threshold in the best (respectively, worst) race. Hawkins *et al.* (2014b) presented, and tested, an approximation to the full model, in that they allow the same option to be selected as both best and worst, which was not allowed in their experiment; their model also allows for vanishingly small times between a best choice and the corresponding worst choice, which are not physically possible. These predictions affect a sufficiently small proportion of their data that they are acceptable. The parallel model overcomes the drawbacks of the other three models described above by accounting for all general choice and response time trends observed in data – for instance, the model is able to capture inter- and intra-individual differences in response style, reflected by the data of those participants who prefer to respond first with the best option, or first with the worst option.

The fact that data on choice could be well-fit by each the four best-worst LBA models (Hawkins *et al.*, 2014a) but data on both choice and response time could only be well-fit that the parallel LBA model (Hawkins *et al.*, 2014b) gives support to the trend in consumer research to collect both the choices made and the time to make them.

#### 7.5.4 Context-dependent models of choice and response time

The general usage of the term *context effect* in the psychological study of (probabilistic) choice appears to be that the choice probabilities are not consistent with

either a random utility or a constant utility representation as defined in chapter 10 of Suppes *et al.* (1989). We do not attempt a formal definition here, but simply present some examples of standard (choice) context effects and of models that explain them.

We discuss three (context) effects, stated in terms of theoretical choice probabilities but which have also been demonstrated in data, that are incompatible with either a random utility model, or a constant utility model, or both. We then present the multiattribute LBA (MLBA) model (Trueblood *et al.*, 2014), for which the three context effects hold simultaneously for various sets of parameter values, and summarize the 2N-ary tree model for N-alternative preferential choice (Wollschläger and Diederich, 2012); both of these models also predict response times. Both models assume attribute-and-alternative wise comparisons; Noguchi and Stewart (2014) present eye-movement data in support of such processes. We then summarize three models of choice, only, that predict some, or all, of the context effects.

Busemeyer and Rieskamp (2014) and Rieskamp *et al.* (2006) summarize the following three context effects, and others, and the extant models that are, or are not, compatible with those effects holding; they do not discuss the MLBA or the 2N-ary tree model (which are described below).

### *Three standard context effects*

Following Trueblood *et al.* (2014), we present the context effects in terms of choice sets with three multiattribute options, although often some of them are stated in terms of a combination of two- and three-option choice sets. These effects have traditionally been observed in between subject designs; however, recent work demonstrates them within subjects in an inference task (Trueblood, 2012) and in a perceptual and a preference task (Trueblood *et al.*, 2013). We use slightly more general notation for psychological (subjective) representations than Trueblood *et al.* (2014) and describe their specific assumptions, as needed.

We restrict attention to options with two attributes; the theoretical formulation generalizes naturally to  $m \geq 2$  attributes.

Let  $\mathbf{r} = (r_1, r_2)$  denote a typical option, where  $r_i$  is an indicator for its objective value on attribute  $i$ ; this value can be quantitative (e.g., price, miles per gallon) or qualitative (e.g., camera/no camera on a mobile phone); most researchers, including Trueblood *et al.* (2014), have formulated and tested their models of context effects using values of the options on quantitative attributes. We say that  $\mathbf{r}$  strictly dominates  $\mathbf{s}$  iff  $r_i > s_i$  for  $i = 1, 2$ , and, equivalently say that  $\mathbf{s}$  is strictly inferior to  $\mathbf{r}$ .

We now state the general pattern of stimuli for each of three main context effects. The fixed options  $\mathbf{x}$  and  $\mathbf{y}$  are chosen so that they lie on a common indifference curve in the attribute space; that is, they satisfy  $x_1 + x_2 = y_1 + y_2$ .

*Attraction effect:* Consider  $\{\mathbf{x}, \mathbf{y}\}$  and a decoy  $\mathbf{a}_x$  (respectively,  $\mathbf{a}_y$ ) that is similar, but strictly inferior, to option  $\mathbf{x}$  (respectively,  $\mathbf{y}$ ). Then the attraction effect occurs when the probability of choosing  $\mathbf{x}$  is greater when the decoy is similar to  $\mathbf{x}$  than

when the decoy is similar to  $y$ , with a parallel result for  $y$ :

$$B_{\{x,y,a_x\}}(x) > B_{\{x,y,a_y\}}(x) \text{ and } B_{\{x,y,a_y\}}(y) > B_{\{x,y,a_x\}}(y).$$

Early (Huber *et al.*, 1982) and recent (Trueblood, 2012; Trueblood *et al.*, 2013) experiments demonstrate different magnitudes of the attraction effect for three different placements of decoys in the attribute space (so-called range, frequency, and range–frequency decoys).

*Similarity effect:* Consider  $\{x, y\}$  and a decoy  $s_x$  (respectively,  $s_y$ ) that is similar to  $x$  (respectively,  $y$ ), but not dominated by  $x$  (respectively,  $y$ ). The similarity effect occurs when the probability of choosing  $x$  is greater when the decoy is similar to  $y$  as compared to when it is similar to  $x$ , with a parallel result for  $y$ :

$$B_{\{x,y,s_y\}}(x) > B_{\{x,y,s_x\}}(x) \text{ and } B_{\{x,y,s_x\}}(y) > B_{\{x,y,s_y\}}(y).$$

*Compromise effect:* Consider options  $x$ ,  $y$ , and  $z$  where  $y$  is “between”  $x$  and  $z$ , and an option  $c_z$  that makes  $z$  “between”  $c_z$  and  $y$ . The compromise effect occurs when the probability of choosing  $y$  is greater when  $y$  is a compromise (“between”) alternative than when it is an extreme alternative, with a parallel result for  $z$ :

$$B_{\{x,y,z\}}(y) > B_{\{y,z,c_z\}}(y) \text{ and } B_{\{y,z,c_z\}}(z) > B_{\{x,y,z\}}(z).$$

In obtaining this effect,  $x$ ,  $y$ ,  $z$ , and are usually chosen so that they lie on a common indifference curve in the attribute space.

#### 7.5.4.1 Multiattribute linear ballistic accumulator (MLBA) model

The *multiattribute linear ballistic accumulator (MLBA) model* (Trueblood *et al.*, 2014) has four major components, which we summarize in turn: subjective value functions, context-dependent attribute weights, relative comparisons, and context-dependent drift rates. The choice probabilities satisfy a *random (binary) advantage model* (Marley, 1991b).

In the following,  $r$ ,  $s$ , denote arbitrary, but distinct, options. We present the model for options with two attributes.

##### *Subjective value functions*

Assume each option  $r = (r_1, r_2)$  has a psychological (subjective) representation  $(u_1(r), u_2(r))$ . As Trueblood *et al.* (2014) work with attributes that are measured by nonnegative real numbers, they consider the set of nonnegative valued vectors  $r = (r_1, r_2)$  in this attribute space, and map each “indifference” line  $r_1 + r_2 = c$ ,  $c$  a nonnegative constant, into a curve  $(u_1(r), u_2(r))$  in subjective space. For the cases that they study, they assume the functional forms: for  $i = 1, 2$ , there exists a nonnegative constant  $m$  such that

$$u_i(r) = r_i \frac{r_1 + r_2}{[(r_1)^m + (r_2)^m]^{\frac{1}{m}}}.$$

### *Context-dependent attribute weights*

For options  $\mathbf{r}$  and  $\mathbf{s}$ , there are constants  $\lambda_k > 0$ ,  $k = 1, 2$ , such that for attributes  $i = 1, 2$ ,

$$w_{r_i s_i} = \begin{cases} \exp(-\lambda_1 |u_i(\mathbf{r}) - u_i(\mathbf{s})|) & \text{if } u_i(\mathbf{r}) \geq u_i(\mathbf{s}) \\ \exp(-\lambda_2 |u_i(\mathbf{r}) - u_i(\mathbf{s})|) & \text{if } u_i(\mathbf{r}) < u_i(\mathbf{s}) \end{cases}.$$

These weight forms were selected as they are larger when differences in subjective attribute values are smaller, and the two values of  $\lambda$  allow different values for positive and negative subjective differences. Note that  $w_{r_1 s_1} + w_{r_2 s_2}$  is not equal to 1, whereas that would be so in numerous other “importance” or “attention” frameworks.

### *Relative comparisons*

For each pair of distinct options  $\mathbf{r}, \mathbf{s}$ , the value  $V_{rs}$  that enters into the context-dependent drift rate of the MLBA is

$$V_{rs} = w_{r_1 s_1} [u_1(\mathbf{r}) - u_1(\mathbf{s})] + w_{r_2 s_2} [u_2(\mathbf{r}) - u_2(\mathbf{s})].$$

### *Context-dependent drift rates*

The final stage gives the context-dependent drift rate for each  $\mathbf{r} \in X$ , where  $X$  is the currently available choice set:

$$d_X(\mathbf{r}) = I_0 + \sum_{\mathbf{s} \in X - \{\mathbf{r}\}} V_{rs}.$$

The constant  $I_0 \geq 0$  helps avoid non-termination in the LBA model.

Trueblood *et al.* (2014) show that this context-dependent LBA model can produce the three context effects described earlier, and all combinations of them, with numerous sets of parameter values. The context-dependent attribute weights lead to the similarity and/or attraction effect, and the compromise effect is produced by either, or both, the attention weights and the subjective value function. They also demonstrate that the model can handle various properties of the context effects when participants are under time pressure.

Note that the above positive results regarding the prediction of context effects by the MLBA model were obtained by adding a context-dependent “front end” to a context-free LBA model. This suggests taking a fresh look at various choice-only models that can handle some or all such context effects in choice, and extending them to response time by, say, adding a context-free LBA model that is driven by the “front-end” parameters of the choice model. We consider possible such (choice) models in Section 7.5.5.

#### **7.5.4.2 The 2N-ary choice tree model for N-alternative preferential choice**

The *2N-ary choice tree model* for N-alternative preferential choice (Wollschläger and Diederich, 2012) builds on earlier information sampling models, such as multi-alternative decision field theory (MADFT, Roe *et al.*, 2001) and the leaky competing accumulator (LCA) model (Usher and McClelland, 2001, 2004), in ways that resonate with the multiattribute linear ballistic accumulator (MLBA) model and

the parallel best-worst LBA model. In contrast to the MLBA model, the  $2N$ -ary choice tree model specifies details of the temporal accumulation of information, rather than having a front end that incorporates the overall effect of such accumulated information.

In contrast to MADFT and LCA, and in agreement with the general form of the parallel best-worst LBA model of Section 7.5.3, the  $2N$ -ary choice tree model assigns two counters to each of the  $N$  alternatives in the current choice set – one samples information in favor of the respective alternative, the other samples information against it. The difference of these two quantities for each alternative at any given time describes the preference state for that alternative at that time. Sampling time is discretized, with the length  $h$  of the sampling interval chosen arbitrarily with a shorter time interval leading to more precision in the calculation of expected choice probabilities and choice response times. At each time point, the model selects one (positive or negative) counter for one alternative, and increases its level by one; increasing only one counter state at a time with a fixed amount of evidence equal to one is equivalent to increasing all counter states at the same time with an amount of evidence equal to the probability with which these counters are chosen. Also, updating counters at discrete points in time creates a discrete structure of possible combinations of counter states which can be interpreted as a graph or, more precisely, as a ( $2N$ -ary) tree. There is a stopping rule that depends on the preference states associated with the alternatives; each preference state is the difference between the cumulative positive and negative count for the relevant alternative. The most general stopping rule considered by Wollschläger and Diederich (2012) has two separate thresholds for the difference count for each alternative, with one positive, which leads to choosing an option if its difference count reaches that threshold first, and the other negative, which leads to eliminating an option from further consideration if its difference count reaches that threshold first; in the latter case, later stages of the process are completed without counts for the eliminated option.

The  $2N$ -ary choice tree model has closed form solutions for choice probabilities and for response times when deadlines are imposed. As with the MLBA, it does not use inhibition or loss aversion for the effects. To date, it has not been shown whether the model can account for the similarity, attraction, and compromise effect with a common set of parameters. Nonetheless, both the MLBA and  $2N$ -ary choice tree model well-fit the data presented by its authors.

To date, the  $2N$ -ary choice tree model has been applied to best choice. However, as in the parallel best-worst LBA model of Section 7.5.3, the choice tree model has a mechanism for accepting (rejecting) options. Thus, it is of interest to extend the  $2N$ -ary choice tree model to best-worst choice and compare its fits to the data from such choice with those of the parallel best-worst LBA model.

### 7.5.5 Context-dependent models of choice

As noted in Section 7.5.4, the predictions of context effects by the MLBA model are obtained by adding a context-dependent “front end” to a context-free LBA

model, which lead us to suggests taking a fresh look at various choice-only models that can handle some or all such context effects in choice, and extending them to response time by, say, adding a context-free LBA model that is driven by the “front-end” parameters of the choice model.

Marley (1991a) describes various context-dependent choice models, the majority of them based on measures of binary advantage of one option relative to another. These include Rotondo’s (1986) binary advantage model, which is a generalization of Luce’s choice model that is not a random utility model; and a generalization of Rotondo’s model to a *random advantage model* – that is, where there exist (possibly dependent) random variables  $A_X(z)$ ,  $z \in X \subseteq \mathcal{A}$  of a particular form such that  $B_X(x) = \Pr[A_X(x) > A_X(y), y \in X - \{x\}]$ ; clearly, this representation corresponds to the choice probabilities given by a context-dependent horse race model (Section 7.5.2).

We now describe three other recent context-dependent choice models.

#### *Berkowitzsch, Scheibehenne, and Rieskamp*

Berkowitzsch *et al.* (2014) develop a choice-only version of decision field theory (DFT, Roe *et al.*, 2001) and compare its predictions to those of the multinomial logit and probit models; as with the original DFT (which predicts both choice and response time), this choice-only version of DFT can handle choice-based context effects. All the models under comparison accurately predict data from a relatively standard stated preference experiment (similar to those presented in Section 7.5.3), but the choice-only version of DFT is superior for designs where context effects are found; the latter is unsurprising, given that the tested multinomial logit and probit models are all random utility models (see Sections 7.5.3 and 7.5.4 for relevant commentary).

#### *Blavatskyy*

As discussed in Section 7.5.1, Blavatskyy (2012) developed a model for binary choice between risky gambles that has a natural extension to binary choice between multiattribute choice, and to choice between more than two such options. In this case, with

$$\mathbf{z} = (z_1, \dots, z_m)$$

where  $z_i$  is the level of option  $\mathbf{z}$  on attribute  $i$ , we would want to derive (axiomatize) or assume the existence of utility functions  $u_i(z_i)$ ,  $i = 1, \dots, m$ , such that, for multiattribute options  $\mathbf{r}, \mathbf{s}$  we have

$$\begin{aligned}\mathbf{r} \vee \mathbf{s} &= [\max(u_1(r_1), u_1(s_1)), \dots, \max(u_m(r_m), u_m(s_m))], \\ \mathbf{r} \wedge \mathbf{s} &= [\min(u_1(r_1), u_1(s_1)), \dots, \min(u_m(r_m), u_m(s_m))],\end{aligned}$$

and a function  $\varphi : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$  and a real-valued function  $u$  such that, say, for all  $\mathbf{r}, \mathbf{s}$ ,

$$p(\mathbf{r}, \mathbf{s}) = \frac{\varphi[u(\mathbf{r}) - u(\mathbf{r} \wedge \mathbf{s})]}{\varphi[u(\mathbf{r}) - u(\mathbf{r} \wedge \mathbf{s})] + \varphi[u(\mathbf{s}) - u(\mathbf{r} \wedge \mathbf{s})]} \quad (7.59)$$

or, alternatively, for all  $r, s$ ,

$$p(r, s) = \frac{\varphi[u(r \vee s) - u(s)]}{\varphi[u(r \vee s) - u(s)] + \varphi[u(r \vee s) - u(r)]}. \quad (7.60)$$

When  $u$  is additive, i.e.,

$$u(z) = \sum_{i=1}^m u_i(z_i),$$

the representations in (7.59) and (7.60) give the same choice probabilities.

Blavatskyy (2009, 2012) extends his model (and, consequently, the above model) of binary probabilistic choice to choices among two or more alternatives with the following algorithm: (i) Select an alternative at random from the choice set  $X$ ; call it  $x$ . (ii) Select another alternative at random from the choice set  $X - \{x\}$ ; call it  $y$ . (iii) Choose between  $x$  and  $y$  with  $x$  (respectively,  $y$ ) chosen with probability  $p(x, y)$  (respectively,  $p(y, x)$ ). (iv) Set the winner of step (iii) as the (new) alternative in step i. (v) Repeat these steps *ad infinitum*. Blavatskyy shows that the resultant set of objects involved in the representation of the choice probabilities for a set  $X$  are the *arborescences*<sup>16</sup> with the root of each arborescence a member of  $X$  and each directed edge  $(r, s)$  of the arborescence having weight  $p(r, s)$ ; and that the probability of choosing  $x$  from  $X$  is proportional to the sum of the products of these weights for each arborescence with root  $x$ . Blavatskyy (2012) summarizes several properties of this model, including the fact that, with appropriately selected binary choice probabilities, it predicts the asymmetric dominance effect and the attraction effect; in these predictions, he does not use a specific form for the choice probabilities, such as (7.59) or (7.60), above. Marley (1965) develops a related representation for discard (rejection, worst) probabilities motivated in a particular way from sets of preference (acceptance, best) choice probabilities where the latter satisfy regularity.

### Chorus

Using the notation above for the representation of multiattribute options, and constants  $\beta_i \geq 0$ , Chorus (2014) assumes that: for  $x \in X \subseteq \mathcal{A}$ ,

$$B_X(x) = \frac{e^{u_X(x)}}{\sum_{z \in X} u_X(z)},$$

where, for  $z \in X$

$$u_X(z) = - \sum_{s \in X - z} \sum_{i=1}^m \ln[1 + \exp(\beta_i[u_i(s_i) - u_i(z_i)])].$$

Chorus (2014) shows that this referent-dependent model explains various context effects. Using our earlier reasoning, the quantities  $e^{u_X(x)}$  could be entered as

<sup>16</sup> An arborescence is a directed graph in which, for a given vertex  $u$  (the root) and any other vertex  $v$ , there is exactly one directed path from  $u$  to  $v$ . Equivalently, an arborescence is a directed, rooted tree in which all edges point away from the root.

the drift rates in an LBA model. Other recent work includes the related approach of Huang (2012) based on ideal and reference points, and Rooderkerk *et al.* (2011), who present a choice model with a context-free partworth random utility component and a context-dependent random component. Bleichrodt (2007) develops an additive reference-dependent utility where the reference point varies across decisions and is one of the available options in each choice set, and Bleichrodt *et al.* (2009) axiomatize a representation for uncertain gambles with multiattribute outcomes, obtaining a version of prospect theory with referents on each attribute or option and with an additive representation of the outcomes.

## 7.6 Discussion, open problems, and future work

In the conclusion of their review chapter on preferential choice, Busemeyer and Rieskamp (2014) present an excellent summary of the similarities and differences between various approaches to the study of choice. The primary theme of that summary is that discrete choice models are primarily applied to the choices of large samples of people with relatively few data per person, with the primary goal of obtaining efficient estimates of economically relevant parameters, whereas process models account for choice, decision time, confidence, and brain activation, often with extensive data per person, with the primary goal of understanding the behavior of individuals. They agree that this is an oversimplification, and the results summarized or cited in our chapter illustrate ongoing multiple interactions between applied and basic research on choice, preference, and utility. A significant part of the reason why standard accrual-halting models (Townsend *et al.*, 2012) have not been used in applied areas is the “relatively high difficulty of estimating some sequential sampling models of decision making” (Berkowitzsch *et al.*, 2014). This difficulty is being alleviated by, for example, Berkowitzsch *et al.*’s (2014) derivation of a tractable closed-form for the choice probabilities given by an asymptotic form of decision field theory (DFT) and by the various linear ballistic accumulator (LBA) models for (best, worst, best-worst) choice and response times.

Nonetheless, there is currently no clear formulation of exactly what kind of (contextual) complexity is required in accrual-halting models, nor have such models been extensively used for prediction of, say, revealed choices (e.g., purchases) from stated choices (e.g., surveys) or of new phenomena. The (context-dependent) accrual-halting models that we have presented in this chapter involve numerous processes. Further empirical study along the lines of Teoderescu and Usher (2013), with follow-up theoretical work, is needed to clarify the conditions under which some or all these, and possibly other, structures are needed.

We chose not to present context-dependent versions of the material on choice probabilities in Section 7.2 and later related sections. We have developed the quite natural notation for such extensions, but substantive results are required for such notation to be useful – for instance, can one formulate a polytope for best choices

that satisfies one, or more, of the standard choice-based context effects? A second major limitation of that material on choice probabilities is just that – it is for choice probabilities, only, not response times. At this time, we have no concrete ideas of how to join polytopes for choice with representations of response time, but it sounds like a fascinating topic.

One open problem that we are somewhat optimistic about being able to solve (at least for small numbers of options) is the characterization of the linear ordering polytope of best-worst choice probabilities.

## Acknowledgments

We thank W. Batchelder (Editor) for his careful reading and detailed comments, which led us to correct and simplify the general framework of the chapter. Any remaining errors are our own. We thank E. Bokhari, C. Davis-Stober, J.-P. Doignon, A. Popova, and J. Trueblood for helpful comments on earlier partial drafts or portions of the text. Numerous other colleagues were extremely helpful with suggestions on the content and structure of the chapter. We thank Ying Guo for computing facet-defining inequalities for best-worst choice using PORTA. This research has been supported by Natural Science and Engineering Research Council Discovery Grant 8124-98 to the University of Victoria for Marley. The work was carried out, in part, whilst Marley was a Research Professor (part-time) at the Institute for Choice, University of South Australia Business School. National Science Foundation grants SES # 08-20009, SES # 10-62045, CCF # 1216016 and an Arnold O. Beckman Research Award from the University of Illinois at Urbana-Champaign each supported Regenwetter as Principal Investigator. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of colleagues, funding agencies, or employers.

## References

- Abdellaoui, M., Baillon, A., Placido, L. and Wakker, P. (2011). The rich domain of uncertainty: Source functions and their experimental implementation. *American Economic Review*, **101**, 695–723.
- Arons, M. and Krabbe, P. (2013). Probabilistic choice models in health-state valuation research: background, theories, assumptions and applications. *Expert Review of Pharmacoeconomics & Outcomes Research*, **13**, 93–108.
- Barberá, S. and Pattanaik, P. K. (1986). Falmagne and the rationalizability of stochastic choices in terms of random orderings. *Econometrica*, **54**, 707–715.
- Barberá, S., Hammond, P. J. and Seidl, C. (1999). *Handbook of Utility Theory: Volume 1: Principles*. New York, NY: Springer.
- Barberá, S., Hammond, P. J. and Seidl, C. (2004). *Handbook of Utility Theory: Volume 2: Extensions*. Boston, MA: Kluwer Academic Publishers.

- Barberá, S., Hammond, P. J. and Seidl, C. (2013). *Handbook of Utility Theory: Volume 3: Empirical Work and History*. New York, NY: Springer.
- Berkowitzsch, N. A. J., Scheibehenne, B. and Rieskamp, J. (2014). Testing multialternative field theory rigorously against random utility models. *Journal of Experimental Psychology: General*, **143**, 1331–1348.
- Birnbaum, M. (2008). New paradoxes of risky decision making. *Psychological Review*, **115**, 463–501.
- Birnbaum, M. (2011). Testing mixture models of transitive preference. Comment on Regenwetter, Dana, and Davis-Stober (2011). *Psychological Review*, **118**, 675–683.
- Birnbaum, M. and Beeghley, D. (1997). Violations of branch independence in judgments of the value of gambles. *Psychological Science*, **8**, 87–94.
- Birnbaum, M. and Gutierrez, R. (2007). Testing for intransitivity of preferences predicted by a lexicographic semiorder. *Organizational Behavior and Human Decision Processes*, **104**, 96–112.
- Birnbaum, M. and Navarrete, J. (1998). Testing descriptive utility theories: violations of stochastic dominance and cumulative independence. *Journal of Risk and Uncertainty*, **17**, 49–78.
- Blavatskyy, P. (2009). How to extend a model of probabilistic choice from binary choices to choices among more than two alternatives. *Economics Letters*, **105**, 330–332.
- Blavatskyy, P. (2012). Probabilistic choice and stochastic dominance. *Economic Theory*, **50**, 59–83.
- Blavatskyy, P. and Pogrebna, G. (2010). Models of stochastic choice and decision theories: why both are important for analyzing decisions. *Journal of Applied Econometrics*, **25**, 963–986.
- Bleichrodt, H. (2007). Reference-dependent utility with shifting reference points and incomplete preferences. *Journal of Mathematical Psychology*, **51**, 266–276.
- Bleichrodt, H. and Pinto, J. (2006). Conceptual foundations for health utility measurement. In Jones, A. (ed.), *The Elgar Companion to Health Economics*, pp. 347–358. Brackfield, VT: Edward Elgar.
- Bleichrodt, H., Rohde, K. and Wakker, P. (2008). Combining additive representations on subsets into an overall representation. *Journal of Mathematical Psychology*, **52**, 306–312.
- Bleichrodt, H., Schmidt, U. and Zank, H. (2009). Additive utility in prospect theory. *Management Science*, **55**, 863–873.
- Bleichrodt, H., Kothiyal, A., Prelec, D. and Wakker, P. (2013). Compound invariance implies prospect theory for simple gambles. *Journal of Mathematical Psychology*, **57**, 68–77.
- Block, H. D. and Marschak, J. (1960). Random orderings and stochastic theories of responses. In Olkin, I., Ghurye, S., Hoeffding, H., Madow, W. and Mann, H. (ed.), *Contributions to Probability and Statistics*. Stanford, CA: Stanford University Press, pp. 97–132.
- Böckenholt, U. (2006). Thurstonian-based analyses: past, present, and future utilities. *Psychometrika*, **71**, 615–629.
- Brandstätter, E., Gigerenzer, G. and Hertwig, R. (2006). The Priority Heuristic: making choices without trade-offs. *Psychological Review*, **113**, 409–432.
- Brown, S. D. and Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. *Cognitive Psychology*, **57**, 153–178.

- Busemeyer, J. R. and Rieskamp, J. (2014). Psychological research and theories on preferential choice. In Hess, S. and Dales, A. (eds), *Handbook of Choice Modelling* Cheltenham: Edward Elgar Publishing, pp. 49–72.
- Charon, I. and Hudry, O. (2010). An updated survey on the linear ordering problem for weighted or unweighted tournaments. *Annals of Operations Research*, **175**, 107–158.
- Chateauneuf, A. and Cohen, M. (1994). Risk seeking with diminishing marginal utility in a nonexpected utility model. *Journal of Risk and Uncertainty*, **9**, 77–91.
- Chechile, R. and Barch, D. (2013). Using logarithmic derivative functions for assessing the risky weighting function for binary gambles. *Journal of Mathematical Psychology*, **57**, 15–28.
- Chorus, C. (2014). Capturing alternative decision rules in travel choice models: a critical discussion. In Hess, S. and Dales, A. (eds), *Handbook of Choice Modelling* Cheltenham: Edward Elgar, pp. 290–310.
- Colonius, H. (1984). *Stochastische Theorien individuellen Wahlverhaltens (Stochastic theories of individual choice behavior)*. Berlin: Springer.
- Colonius, H. and Vorberg, D. (1995). Distribution inequalities for parallel models with unlimited capacity. *Journal of Mathematical Psychology*, **38**, 35–58.
- Critchlow, D. E., Fligner, M. A. and Verducci, J. S. (eds) (1993). *Probability Models and Statistical Analyses for Ranking Data*. New York, NY: Springer.
- Dagsvik, J. (2008). Axiomatization of stochastic models for choice under uncertainty. *Mathematical Social Sciences*, **55**, 341–370.
- Dagsvik, J. and Hoff, S. (2011). Justification of functional form assumptions of structural models: applications and testing qualitative measurement axioms. *Theory and Decision*, **70**, 215–254.
- Davis-Stober, C. (2012). A lexicographic semiorder polytope and probabilistic representations of choice. *Journal of Mathematical Psychology*, **56**, 86–94.
- Davis-Stober, C. and Brown, N. (2013). Evaluating decision making “type” under  $p$ -additive utility representations. *Journal of Mathematical Psychology*, **57**, 320–328.
- de Palma, A., Kilani, K. and Laffond, G. (2015). Relations between best, worst, and best-worst choices for probabilistic choice models. Manuscript, ENS Cachan.
- Diecidue, E., Schmidt, U. and Zank, H. (2009). Parametric weighting functions. *Journal of Economic Theory*, **144**, 1102–1118.
- Doignon, J., Fiorini, S. and Joret, G. (2007). Erratum to “facets of the linear ordering polytope: a unification for the fence family through weighted graphs (vol 50, pg 251, 2006)”. *Journal of Mathematical Psychology*, **51**, 341.
- Doignon, J.-P. and Fiorini, S. (2002). Facets of the weak order polytope derived from the induced partition projection. *SIAM Journal on Discrete Mathematics*, **15**, 112–121.
- Doignon, J.-P., Fiorini, S., Guo, Y., et al. (2015). On a probabilistic model of best-worst choice. Manuscript in progress.
- Dzhafarov, E. (1993). Grice-representability of response time distribution families. *Psychometrika*, **58**, 281–314.
- Falmagne, J.-C. (1978). A representation theorem for finite random scale systems. *Journal of Mathematical Psychology*, **18**, 52–72.
- Falmagne, J. C. (1985). *Elements of Psychophysical Theory*. New York, NY: Oxford University Press.

- Fiorini, S. (2001a). Determining the automorphism group of the linear ordering polytope. *Discrete Applied Mathematics*, **112**, 121–128.
- Fiorini, S. (2001b). *Polyhedral Combinatorics of Order Polytopes*. PhD thesis, Université Libre de Bruxelles.
- Fiorini, S. (2004). A short proof of a theorem of Falmagne. *Journal of Mathematical Psychology*, **48**, 80–82.
- Flynn, T., Huynh, E. and Corke, C. (2015). Advanced care planning: patients want less than doctors think. In *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge: Cambridge University Press, chapter 7.
- Gilboa, I. (2009). *Theory of Decision under Uncertainty*. Cambridge: Cambridge University Press.
- Gul, F. and Pesendorfer, W. (2006). Random expected utility. *Econometrica*, **74**, 121–146.
- Guo, Y. and Regenwetter, M. (2014). Quantitative tests of the Perceived Relative Argument Model: comment on Loomes (2010). *Psychological Review*, **121**, 696–705.
- Hawkins, G., Marley, A. A. J., Heathcote, A., et al. (2014a). Integrating cognitive process and descriptive models of attitude and preference. *Cognitive Science*, **38**, 701–735.
- Hawkins, G. E., Marley, A. A. J., Heathcote, A., et al. (2014b). The best of times and the worst of times are interchangeable. *Decision*, **1**, 192–214.
- Heathcote, A., Brown, S. and Wagenmakers, E.-J. (2014). The falsifiability of actual decision-making models. *Psychological Review*, **121**, 676–678.
- Heathcote, A. and Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in Psychology*, **3**, 1–19.
- Huang, J.-J. (2012). Further explanations for context effects: a perspective of ideal and reference points. *Quality and Quantity*, **46**, 281–290.
- Huber, J., Payne, J. and Puto, C. (1982). Adding asymmetrically dominated alternatives: violation of regularity and the similarity hypothesis. *Journal of Consumer Research*, **9**, 90–98.
- Iverson, G. J. and Falmagne, J.-C. (1985). Statistical issues in measurement. *Mathematical Social Sciences*, **10**, 131–153.
- Jones, M. and Dzhafarov, E. (2014). Unfalsifiability of major modeling schemes for choice reaction time. *Psychological Review*, **121**, 1–32.
- Kahneman, D. and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, **47**, 263–291.
- Koppen, M. (1995). Random utility representation of binary choice probabilities: critical graphs yielding critical necessary conditions. *Journal of Mathematical Psychology*, **39**, 21–39.
- Lancsar, E. and Swait, J. (2014). Reconceptualising the external validity of discrete choice experiments. *PharmacoEconomics*, **7**, 1–15.
- Loomes, G. (2010). Modeling choice and valuation in decision experiments. *Psychological Review*, **117**, 902–924.
- Louviere, J. J., Flynn, T. N. and Marley, A. A. J. (2015). *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge: Cambridge University Press.
- Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York, NY: John Wiley.
- Luce, R. D. (2000). *Utility of Gains and Losses Measurement – Theoretical and Experimental Approaches*. Mahwah, NJ: Lawrence Erlbaum Associates.

- Luce, R. D. (2001). Reduction invariance and Prelec's weighting functions. *Journal of Mathematical Psychology*, **45**, 167–179.
- Luce, R. D. (2010a). Behavioral assumptions for a class of utility theories: a program of experiments. *Journal of Risk and Uncertainty*, **41**, 19–37.
- Luce, R. D. (2010b). Interpersonal comparisons of utility for 2 or 3 types of people. *Theory and Decision*, **68**, 5–24.
- Luce, R. D. and Marley, A. A. J. (2005). Ranked additive utility representations of gambles: old and new axiomatizations. *Journal of Risk and Uncertainty*, **30**, 21–62.
- Luce, R. D. and Suppes, P. (1965). Preference, utility and subjective probability. In Luce, R. D., Bush, R. R. and Galanter, E. (eds), *Handbook of Mathematical Psychology*, volume III, pp. 249–410. New York, NY: Wiley.
- Luce, R. D., Marley, A. A. J. and Ng, C. (2009). Entropy-related measures of the utility of gambling. In: *Preference, Choice and Order: Essays in Honor of Peter C. Fishburn*. Springer-Verlag: Berlin, pp. 5–25.
- Marley, A. A. J. (1965). The relation between the discard and regularity conditions for choice probabilities. *Journal of Mathematical Psychology*, **2**, 242–253.
- Marley, A. A. J. (1989). A random utility family that includes many of the classical models and has closed form choice probabilities and choice reaction times. *British Journal of Mathematical and Statistical Psychology*, **42**, 13–36.
- Marley, A. A. J. (1991a). Aggregation theorems and multidimensional stochastic choice models. *Theory and Decision*, **30**, 245–272.
- Marley, A. A. J. (1991b). Context dependent probabilistic choice models based on measures of binary advantage. *Mathematical Social Sciences*, **21**, 201–218.
- Marley, A. A. J. (1992). A selective review of recent characterizations of stochastic choice models using distribution and functional equation techniques. *Mathematical Social Sciences*, **23**, 5–29.
- Marley, A. A. J. and Colonius, H. (1992). The “horse race” random utility model for choice probabilities and reaction times, and its competing risks interpretation. *Journal of Mathematical Psychology*, **36**, 1–20.
- Marley, A. A. J. and Islam, T. (2012). Conceptual relations between expanded rank data and models of the unexpanded rank data. *Journal of Choice Modelling*, **5**, 38–80.
- Marley, A. A. J. and Louviere, J. (2005). Some probabilistic models of best, worst, and best-worst choices. *Journal of Mathematical Psychology*, **49**, 464–480.
- Marley, A. A. J. and Luce, R. D. (2001). Ranked-weighted utilities and qualitative convolution. *Journal of Risk and Uncertainty*, **23**, 135–163.
- Marley, A. A. J. and Luce, R. D. (2002). A simple axiomatization of binary rank-dependent expected utility of gains (losses). *Journal of Mathematical Psychology*, **46**, 40–55.
- Marley, A. A. J. and Luce, R. D. (2005). Independence properties vis-à-vis several utility representations. *Theory and Decision*, **58**, 77–143.
- Marley, A. A. J., Luce, R. D. and Kocsis, I. (2008). A solution to a problem raised in Luce and Marley (2005). *Journal of Mathematical Psychology*, **52**, 64–68.
- McFadden, D. (1978). Modeling the choice of residential location. In Karlquist, A., Lundqvist, L., Snickars, F. and Weibull, J. (eds), *Spatial Interaction Theory and Planning Models*. Amsterdam: North-Holland, pp. 75–96.
- Nadarajah (2009). Some truncated distributions. *Acta Applied Mathematics*, **106**, 105–123.

- Ng, C., Luce, R. D. and Marley, A. A. J. (2009). Utility of gambling under  $p$ (olynomial)-additive joint receipt and segregation or duplex decomposition. *Journal of Mathematical Psychology*, **53**, 273–286.
- Noguchi, T. and Stewart, N. (2014). In the attraction, compromise, and similarity effects, alternatives are repeatedly compared in pairs on single dimensions. *Cognition*, **132**, 44–56.
- Orme, B. (2009). Anchored scaling in maxdiff using dual response. Sawtooth software research paper series, Sawtooth Software, Sequim, WA.
- Prelec, D. (1998). The probability weighting function. *Econometrica*, **66**, 497–527.
- Regenwetter, M., Dana, J. and Davis-Stober, C. P. (2010). Testing transitivity of preferences on two-alternative forced choice data. *Frontiers in Quantitative Psychology and Measurement*, **1**, 148.
- Regenwetter, M., Dana, J. and Davis-Stober, C. P. (2011a). Transitivity of preferences. *Psychological Review*, **118**, 42–56.
- Regenwetter, M., Dana, J., Davis-Stober, C. P. and Guo, Y. (2011b). Parsimonious testing of transitive or intransitive preferences: reply to Birnbaum (2011). *Psychological Review*, **118**, 684–688.
- Regenwetter, M. and Davis-Stober, C. (2011). Ternary paired comparisons induced by semi- or interval order preferences. In: Dzhafarov, E. and Perry, L. (eds), *Descriptive and Normative Approaches to Human Behavior*, volume 3 of *Advanced Series on Mathematical Psychology*. Singapore: World Scientific, pp. 225–248.
- Regenwetter, M. and Davis-Stober, C. P. (2012). Behavioral variability of choices versus structural inconsistency of preferences. *Psychological Review*, **119**, 408–416.
- Regenwetter, M., Davis-Stober, C. P., Lim, S. H., et al. (2014). QTTEST: quantitative testing of theories of binary choice. *Decision*, **1**, 2–34.
- Regenwetter, M. and Marley, A. A. J. (2001). Random relations, random utilities, and random functions. *Journal of Mathematical Psychology*, **45**, 864–912.
- Reinelt, G. (1993). A note on small linear-ordering polytopes. *Discrete & Computational Geometry*, **10**, 67–78.
- Rieskamp, J., Busemeyer, J. and Mellers, B. (2006). Extending the bounds of rationality: evidence and theories of preferential choice. *Journal of Economic Literature*, **44**, 631–661.
- Roberts, F. S. (1979). *Measurement Theory*. London: Addison-Wesley.
- Roe, R. M., Busemeyer, J. R. and Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, **108**, 370–392.
- Rooderkerk, R., Van Heerde, H. and Bijmolt, T. (2011). Incorporating context effects into a choice model. *Journal of Marketing Research*, **48**, 767–780.
- Rotondo, J. (1986). A generalization of Luce's choice axiom and a new class of choice models. In *Psychometric Society Conference Abstract*.
- Shafir, E. (1993). Choosing versus rejecting – why some options are both better and worse than others. *Memory & Cognition*, **21**, 546–556.
- Stott, H. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty*, **32**, 101–130.
- Suck, R. (2002). Independent random utility representations. *Mathematical Social Sciences*, **43**, 371–389.

- Suppes, P., Krantz, D. H., Luce, R. D. and Tversky, A. (1989). *Foundations of Measurement*, volume II. San Diego, CA: Academic Press.
- Swait, J. (2001). Choice set generation within the generalized extreme value family of discrete choice models. *Transportation Research Part B—Methodological*, **35**, 643–666.
- Teoderescu, A. and Usher, M. (2013). Disentangling decision models – from independence to competition. *Psychological Review*, **120**, 1–38.
- Terry, A., Marley, A. A., Wagenmakers, E. J., Heathcote, A. and Brown, S. D. (2015). Generalising the drift rate for linear ballistic accumulators. *Journal of Mathematical Psychology*, **68/69**, 49–58.
- Townsend, J., Houpt, J. and Silbert, N. (2012). General recognition theory extended to include response times: predictions for a class of parallel systems. *Journal of Mathematical Psychology*, **56**, 476–494.
- Train, K. E. (2003). *Discrete Choice Models with Simulations*. Cambridge: Cambridge University Press.
- Trueblood, J. (2012). Multialternative context effects obtained using an inference task. *Psychological Bulletin & Review*, **19**, 962–968.
- Trueblood, J., Brown, S., Heathcote, A. and Busemeyer, J. (2013). Not just for consumers: context effects are fundamental to decision making. *Psychological Science*, **24**, 901–908.
- Trueblood, J., Brown, S., Heathcote, A. and Busemeyer, J. (2014). The multiattribute linear ballistic accumulator model. *Psychological Review*, **121**, 179–205.
- Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, **76**, 31–48.
- Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, **5**, 297–323.
- Usher, M. and McClelland, J. (2001). On the time course of perceptual choice: the leaky competing accumulator model. *Psychological Review*, **108**, 550–592.
- Usher, M. and McClelland, J. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, **11**, 757–669.
- Wakker, P. (1991). Additive representations on rank-ordered sets. 1. The algebraic approach. *Journal of Mathematical Psychology*, **35**, 501–531.
- Wakker, P. (2010). *Prospect Theory For Risk and Ambiguity*. Cambridge: Cambridge University Press.
- Wilcox, N. (2008). Stochastic models for binary discrete choice under risk: a critical primer and econometric comparison. In Cox, J. and Harrison, G. (eds), *Risk Aversion in Experiments*, volume 12. Bingley: Emerald, Research in Experimental Economics, pp. 197–292.
- Wollschläger, L. and Diederich, A. (2012). The 2n-ary choice tree model for n-alternative preferential choice. *Frontiers in Psychology*, **3**, Article 189.
- Ziegler, G. M. (1995). *Lectures on Polytopes*. Berlin: Springer-Verlag.

# 8 Discrete state models of cognition

William H. Batchelder

8.1	Introduction	454
8.2	Discrete state hidden Markov chain models	458
8.2.1	Finite state Markov chains	458
8.2.2	Examples of hidden Markov chain models	460
8.3	Multinomial processing tree (MPT) models	466
8.3.1	Specification of MPT models	466
8.3.1.1	Parametric categorical models	466
8.3.1.2	Binary MPT models	469
8.3.1.3	Multi-link MPT models	472
8.3.2	Examples of MPT models	472
8.3.2.1	Model identifiability	477
8.3.2.2	Model validity	478
8.3.2.3	MPT models as measurement tools	479
8.3.3	Parametric constraints in MPT models	480
8.3.4	A string language for MPT models	486
8.3.4.1	A string language for BMPT models	486
8.3.4.2	Context-free and MPT string languages	489
8.3.4.3	Relating BMPT strings to BMPT parameterized full binary trees	493
8.4	Statistical inference for MPT models	494
8.4.1	The multinomial distribution	496
8.4.2	Parameterized multinomial models	497
8.4.3	Hierarchical Bayesian inference for MPT models	499
8.5	Relevant literature	500
	References	501

## 8.1 Introduction

This chapter is about a class of discrete state models that have been developed to understand data in a variety of experimental cognitive tasks. Discrete state models assume that observable responses to stimulus probes (items) are mediated by a finite set of unobservable (latent) states. Put simply, associated with each stimulus probe is a probability distribution over the latent states, and associated

with each latent state is a probability distribution over the observable responses. In this chapter, we will focus only on experimental tasks where the sets of possible stimulus probes and response alternatives are finite. Such experimental tasks involve paradigms where a participant must classify each of a set of  $M$  stimulus items,  $\mathbf{S} = \{S_1, \dots, S_M\}$ , into a set of  $K$  response categories,  $\mathcal{C} = \{C_1, \dots, C_K\}$ . A discrete state model specifies that there is a finite set of  $T$  latent (unobservable) cognitive states  $\mathbf{W} = \{w_1, \dots, w_T\}$ , along with two classes of probability distributions:

- (i)  $\forall m, \forall t, \Pr(w_t | S_m) \geq 0$ , where  $\sum_t \Pr(w_t | S_m) = 1$  and
- (ii)  $\forall t, \forall k; \Pr(C_k | w_t) \geq 0$ , where  $\sum_k \Pr(C_k | w_t) = 1$ .

Thus the probability that a stimulus item  $S_m$  is classified into a particular response category  $C_k$  is computed by summing response probabilities over the latent states, namely

$$\Pr(C_k | S_m) = \sum_t \Pr(C_k | w_t) \Pr(w_t | S_m). \quad (8.1)$$

Equation (8.1) reveals a common characteristic of all discrete state models, namely that an observable response can result from more than one latent state. Despite the simplicity of (8.1), the special character of a discrete state model depends on the formal structure of its latent states and their associated probabilistic processes. In the early history of cognitive modeling, most discrete state models were either adapted from Signal Detection Theory or Hidden Markov Chain (HMC) models. However, since the middle 1980s, most of the discrete state models have been in the large class of Multinomial Processing Tree (MPT) models.

Viewed generally, there are two main aspects to the study of discrete state models: (i) the formal nature of the models, and (ii) the nature of the experimental paradigms and statistical assumptions at play when a model is applied to data in a cognitive experiment. Most of this chapter will focus on the formal character of the models, with attention given to the experimental paradigms and statistical assumptions only when required to understand the models. Example 8.1 describes a discrete state model for signal detection, Section 8.2 describes some HMC models, Section 8.3 describes the properties of MPT models along with some examples, Section 8.4 describes particular experimental situations along with the statistical characteristics of the data and approaches to statistical inference for MPT models. Finally in Section 8.5, some relevant references for further study of the material presented in the chapter are suggested.

**Example 8.1** One of the earliest applications of discrete state models was in the area of signal detection. A frequent signal detection paradigm is the Yes/No task. In this task, experimental participants are exposed to a series of two types of items, signal items ( $S$ ), and noise items ( $N$ ). This example will focus on a simple

recognition memory task discussed in more detail later in Example 8.14. In a simple recognition memory experiment, participants first study a list of words, one at a time, and then they are given a test consisting of old studied words (signal items) and new, unstudied words (noise items, foils). The participant must classify each test word as Yes (old studied word) or No (new, not studied). Other examples include auditory signal detection, where the signal items are tones that include extra energy at a particular frequency, and medical decision-making situations, where the signal items are X-rays that exhibit a particular type of tumor. In such a task, the observables are the proportions of each response to each item type, and they can be viewed as a sample from the two-by-two probability matrix

$$\mathbf{P} = \begin{bmatrix} \Pr(\text{"Yes"}|S) & \Pr(\text{"No"}|S) \\ \Pr(\text{"Yes"}|N) & \Pr(\text{"No"}|N) \end{bmatrix}. \quad (8.2)$$

In standard signal detection terms, the top row of the matrix exhibits, respectively, the hit and miss rates, and the bottom row the false alarm and correct rejection rates.

Discrete state models for the Yes/No signal detection task are characterized by the assumption that there are detection thresholds for the two kinds of items, and when an item is not detected, the participant responds with a guessing bias,  $g \in [0, 1]$ , for a Yes response. Such models postulate that any test item falls into one of three latent states:  $w_1$ , detected as a signal item,  $w_2$ , detected as a noise item, and  $w_3$ , not detected. The most general version of the threshold model specifies parameters for each item type to fall into each discrete latent state. Let  $d_{it}$  be the probability that item type  $i$  leads to discrete state  $w_t$ , where  $i = 1$  for signal items and  $i = 2$  for noise items. The general threshold model specifies five nonnegative parameters  $\Theta = (d_{11}, d_{12}, d_{21}, d_{22}, g)$ , where  $d_{11} + d_{12} \leq 1$ ,  $d_{21} + d_{22} \leq 1$ ,  $0 \leq g \leq 1$ , where of course  $d_{i3} = 1 - (d_{i1} + d_{i2})$ .

Two matrices specify the model: a stimulus input matrix and a response output matrix. The stimulus matrix specifies the probabilities of getting into each of the three discrete states as a function of the item type, and the response matrix specifies the probability of a Yes or No response as a function of the discrete states. The stimulus matrix is given by

$$\mathbf{I}(\Theta) = \begin{bmatrix} d_{11} & d_{12} & (1 - d_{11} - d_{12}) \\ d_{21} & d_{22} & (1 - d_{21} - d_{22}) \end{bmatrix}$$

and the response matrix by

$$\mathbf{R}(\Theta) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ g & (1 - g) \end{bmatrix}$$

where the first column of the response matrix gives the probabilities of a Yes response given the discrete state and the second column the No probabilities. It

is easy to compute (8.2) from the stimulus and response matrices by matrix multiplication,

$$\begin{aligned}\mathbf{P}(\Theta) &= \mathbf{I}(\Theta) \cdot \mathbf{R}(\Theta) \\ &= \begin{bmatrix} d_{11} + (1 - d_{11} - d_{12})g & d_{12} + (1 - d_{11} - d_{12})(1 - g) \\ d_{21} + (1 - d_{21} - d_{22})g & d_{22} + (1 - d_{21} - d_{22})(1 - g) \end{bmatrix}\end{aligned}$$

and the resulting matrix is easily seen to include the four terms in (8.2), each as a function of the model parameters.

The general threshold model has too many parameters to uniquely interpret data in a single experimental condition involving a sample of signal and noise items. This is because its observable probability distributions are specified in terms of five model parameters, and the data structure is a product binomial structure with only 2 degrees of freedom because the row sums of (8.2) are both one. As a consequence, there are two special cases of the model that are usually employed when data come from a single experimental condition. The first is the single high-threshold (SHT) model, and the second is the double high-threshold (DHT) model. The SHT model sets  $d_{11} = D$ ,  $d_{12} = d_{21} = d_{22} = 0$ , and the DHT model sets  $d_{11} = d_{22} = D$ , and  $d_{12} = d_{21} = 0$ . In essence both of these models assume the detection thresholds are set sufficiently high so that detections are always accurate, and failures to detect lead to the guessing process parameterized by  $g$ . The difference between the two models is that the DHT model, but not the SHT model, allows new items to be detected as new. The hit ( $H$ ) and false alarm ( $F$ ) rates for the SHT model are,  $H = D + (1 - D)g$  and  $F = g$ , and for the DHT model, these rates are  $H = D + (1 - D)g$  and  $F = (1 - D)g$ . The DHT model is depicted as a MPT model in Figure 8.1 in Section 8.3 and will be discussed more there.

Other versions of the general threshold model either require data from multiple experimental conditions, where some of the parameters are equated across some of the conditions, or the experimental task is changed to accommodate more response options. One way to accomplish this is to have several groups of participants, where experimental efforts are made to keep the stimulus detection probabilities constant while varying the guessing probability. Varying the base rate of old items or varying the payoff for various correct responses sometimes accomplishes this. If successful, the result can be exhibited in a so-called receiver-operating characteristic (ROC), which plots the hit rate against the false alarm rate for each condition. For example, in the case of the DHT model, suppose several experimental groups,  $i = 1, \dots, I$  are run in a between-participants design, where  $D$  is constant but  $g_i$  varies. Then  $H = D + (1 - D)g_i$  and  $F_i = (1 - D)g_i$ , so the ROC curve is given by  $H_i = D + F_i$ . From this, it is easily seen that the ROC plot for the DHT is linear from  $(D, 0)$  to  $[1, (1 - D)]$ . Other signal detection models do not imply a linear ROC, and at the time of this writing there are a number of papers that have appeared that discuss whether or not a linear ROC plot is found in experimental data. Some references about ROC plots for between-participant designs will be given in Section 8.5.

## 8.2 Discrete state hidden Markov chain models

A major subclass of discrete state cognitive models is based on so-called Hidden Markov Chain (HMC) models. These models were especially popular in the 1960s and 1970s in the areas of concept identification, learning, memory, and game theory; however, they still have traction today. Unlike the discrete state signal detection models described in Example 8.1, these models concern dynamic changes in the response probabilities due to changes in the  $\Pr(w_t|S_m)$  of (8.1) over a series of discrete learning trials,  $n = 1, 2, \dots$ . These changes are often due to experimenter presented feedback (reinforcement) after each response to an item; however, they can also be due to uninstructed learning, namely changes that occur solely as a consequence of the response that is made. Before proceeding, a short formal review of Markov chains is useful.

### 8.2.1 Finite state Markov chains

Finite state Markov chains are a class of stochastic processes  $\{X_n|n = 1, 2, \dots\}$ , where each  $X_n$  is a random variable taking values in a finite state space  $W = \{w_1, \dots, w_T\}$ . If  $X_n = w_t$ , the chain is said to be in state  $w_t$  on trial  $n$ . For convenience we will use the state subscripts rather than the entire state descriptors whenever this will not cause confusion. A Markov chain specifies that whenever the process is in state  $u$  on any trial, there is a probability  $p_{uv}$  that the state on the next trial will be  $v$ , for  $1 \leq u, v \leq T$ . The central assumption is that these state-to-state transition probabilities are the same regardless of the trial number and the past history of state occupancies, that is

$$\forall w_u, w_v \in W, \forall n = 1, 2, \dots; \\ \Pr(X_{n+1} = v | X_n = u, X_{n-1} = t_{n-1}, \dots, X_1 = t_1) = p_{uv}.$$

In other words, the conditional distribution of a future state on trial  $n + 1$ , given the past history of state occupancies,  $X_n, \dots, X_1$ , depends only on the state occupied on trial  $n$ . Any such model generates a probability distribution over the set of state sequences  $\Omega_N = \prod_{n=1}^N W$ , for any  $N \geq 1$ .

In accord with the above description, a Markov chain model is specified by three entities: (1) the state space  $W$ , (2) a start vector  $S_1 = (p_t)_{1 \times T}$ , and (3) a state-to-state transition matrix  $\mathbf{T} = (p_{uv})_{T \times T}$ . The start vector gives the probability distribution over the state space on trial  $n = 1$ . The transition matrix  $\mathbf{T}$  gives the state-to-state transition probabilities, and by convention the rows are the current state and the columns are the next state, so  $\mathbf{T}$  is row stochastic, that is

$$\forall w_u \in W, \sum_t p_{ut} = 1.$$

There are a number of computational properties of a Markov chain that can be developed using the start vector and the transition matrix. For example, using induction it is easy to see that  $S_n$ , the probability distribution over the state space

on any trial  $n \geq 2$ , is given by  $\mathbf{S}_n = \mathbf{S}_1 \mathbf{T}^{n-1}$ . A somewhat deeper computational property of Markov chains is the extension of the one-step transition probabilities in  $\mathbf{T}$  to  $n$ -step transition probabilities given by

$$\forall n \geq 1, w_u, w_v \in W, p_{uv}^{(n)} = \Pr(X_{n+k} = v | X_k = u),$$

which of course are the same for any  $k \geq 1$ . The key to the extension is based on the so-called Chapman–Kolmogorov equations

$$\forall n, m \geq 1, \forall u, v; p_{uv}^{(n+m)} = \sum_{t=1}^T p_{ut}^{(n)} p_{tv}^{(m)}$$

where the equations reflect the fact that any transition from state  $u$  to  $v$  in  $n + m$  steps must pass through some state on trial  $n$ . If we display these  $n$ -step transition probabilities in a matrix  $\mathbf{T}^{(n)}$ , it is easy to show by induction and the Chapman–Kolmogorov equations that  $\mathbf{T}^{(n)} = \mathbf{T}^n$ , namely the  $n$ th power of the transition matrix.

Further computational properties of a Markov chain depend on the structure of the  $\mathbf{T}$ . One can define an accessibility relation  $A$  on  $W$  by

$$\forall w_u, w_v \in W, (u, v) \in A \Leftrightarrow \exists n \geq 1 \text{ with } p_{uv}^{(n)} > 0.$$

Thus,  $(u, v) \in A$  means that there is at least one sequence of state-to-state transitions with positive probability that starts with state  $u$  and ends in state  $v$ . Two states that are accessible from each other are said to communicate, and one can define a binary communication relation  $E$  by  $(u, v) \in E \Leftrightarrow [(u, v) \in A \wedge (v, u) \in A]$ . It is easy to show that the states of a Markov chain can be partitioned into mutually exclusive and exhaustive communicating classes because the communication relation  $E$  satisfies the properties of an equivalence relation, namely it is (i) reflexive,  $\forall w_t \in W, (t, t) \in E$ , (ii) symmetric,  $(u, v) \in E \Leftrightarrow (v, u) \in E$ , and (iii) transitive,  $[(u, t) \in E \wedge (t, v) \in E] \Rightarrow (u, v) \in E$ . A Markov chain is said to be irreducible if there is a single communicating class, and in our case where the number of states is finite, it follows that for an irreducible Markov chain, every state is recurrent in the sense that it is visited an infinite number of times. On the other hand, if the Markov chain is not irreducible, some states are not recurrent and these are called transient. Transient states will only be visited finitely many times (with probability one).

Of particular interest in applications of Markov chains is their limiting behavior, that is, what happens as  $n$  tends to infinity. In order to state the key result about limiting probabilities, it is necessary to define two more properties of the states. A state  $t$  is said to have period  $d$  if  $p_{tt}^n = 0$  whenever  $n$  is not divisible by  $d$ . For example, the two-state Markov chain with transition matrix

$$\mathbf{T} = \begin{bmatrix} a & 1-a \\ 1-b & b \end{bmatrix}$$

is easily seen to be irreducible if  $a, b < 1$ , have period  $d = 2$  if  $a = b = 0$ , and period  $d = 1$  if  $0 < a, b < 1$ . A state with period  $d = 1$  is said to be aperiodic. An

important limiting theorem for Markov chains says that for an irreducible Markov chain with aperiodic states,  $\lim_{n \rightarrow \infty} P_{uv}^{(n)} = \lambda_v$  exists and is independent of  $u$ . Thus,  $\Lambda = (\lambda_i)_{1 \times t}$ , independent of the start vector, constitutes a long-run probability distribution for the Markov chain. There are well-developed methods in linear algebra for determining the long-run probability distribution for an irreducible, aperiodic Markov chain; however, for this chapter we note only that if  $\Lambda$  is a probability distribution and satisfies  $\Lambda \cdot T = \Lambda$ , it will be the desired long-run distribution.

### 8.2.2 Examples of hidden Markov chain models

This section will describe three examples of models that employ Markov chains that were developed to study cognition. Section 8.5 will provide references and source books for further study of such models.

**Example 8.2** A frequent experimental paradigm in choice theory is a probability learning prediction experiment. In such an experiment, the participant predicts on each trial  $n \geq 1$ , which one of two lights will turn on, the left one  $L_0$  or the right one  $L_1$ . Let

$$X_n = \begin{cases} 1 & \text{if subject predicts } L_1 \text{ on } n \\ 0 & \text{if subject predicts } L_0 \text{ on } n \end{cases}.$$

In a probability learning experiment the experimenter selects a light on each trial with a Bernoulli process, with probability  $0 < \pi < 1$  that the right light will shine. Assume that the participant can occupy two states  $W = \{C_0, C_1\}$ , whose subscripts indicate the prediction on trial  $n$ . The learning dynamics for this example assumes that if a prediction is confirmed the participant will make the same prediction on the next trial; however, if the prediction is wrong, then the participant will shift to the other state on the next trial with probability  $0 \leq a \leq 1$ . Finally, assume the participant starts with equal likelihood of predicting the left or right light on trial 1.

From the above, it follows that the behavior of the stochastic process  $\{X_n | n \geq 1\}$  is given by a Markov chain with start vector  $S_1 = (1/2, 1/2)$  and transition matrix

$$\mathbf{T} = \begin{bmatrix} C_0 & C_1 \\ C_0 & C_1 \end{bmatrix} = \begin{bmatrix} (1 - \pi) + \pi(1 - a) & \pi a \\ (1 - \pi)a & \pi + (1 - \pi)(1 - a) \end{bmatrix}.$$

It is clear that the Markov chain is irreducible and aperiodic unless  $a = 0$ , in which event there are two communication classes of size one. Further, in case  $a \neq 0$ , it is easy to verify that  $\Lambda = [(1 - \pi), \pi]$  satisfies  $\Lambda T = \Lambda$ , so  $\Lambda$  is the long-run probability distribution for the Markov chain. The result is interesting for cognitive theory because it usually happens in probability learning prediction experiments that the relative frequency of predictions matches the Bernoulli probability used by the experimenter. This matching behavior is not optimal from the viewpoint of economic rationality, because predicting  $L_1$  if and only if  $\pi \geq 1/2$  will maximize

the number of correct predictions. While the model in the example predicts probability matching behavior, it is by no means the only learning model that has that property.

Thus far we have treated the state sequences of a Markov chain as leading to unique observable values of the random variables  $\{X_n | N \geq 1\}$ . However, it is HMC models that are typically employed as discrete state cognitive models. An HMC model is a Markov chain with the addition of state-to-response probability distributions. Suppose the observable responses of a cognitive experiment are a set of discrete categories  $\mathcal{C} = \{C_1, \dots, C_K\}$ , and that on each of a series of trials the random variables  $X_n$  take values in  $\mathcal{C}$ . Thus,  $\{X_n | n = 1, 2, \dots\}$  constitute the observables, and what is desired is to specify a stochastic process for them. Suppose the probability of observing a particular response category on a particular trial  $n$  depends on the state occupancy  $Y_n$  of a Markov chain. To handle this possibility, the model needs to specify a response matrix  $\mathbf{G} = (g_{tk})_{T \times K}$ , where

$$\forall w_t \in W, \forall C_k \in \mathcal{C}, g_{tk} = \Pr(C_k | w_t).$$

Of course,  $\mathbf{G}$  is row-stochastic because each row is a probability distribution over the response categories conditional on a particular row state of the chain. Given this addition, it is straightforward to compute the probability distribution over the responses on trial  $n$ . Let  $\mathbf{R}_n = [\Pr(X_n = C_k)]_{1 \times K}$ , then

$$\mathbf{R}_n = \mathbf{S}_1 \cdot \mathbf{T}^{n-1} \cdot \mathbf{G}. \quad (8.3)$$

The reason that the stochastic processes described above are called HMC models is that any observable response may occur from several states in the Markov chain, and this means that knowledge of the response on a trial does not necessarily reveal the latent state that gave rise to it. In fact, because of this property, the stochastic process  $\{X_n | n \geq 1\}$  generated by an HMC model is not, in general, a Markov chain.

To define an HMC model with  $T$  latent states and  $K$  response categories requires the specification of many quantities, specifically  $(T - 1)$  probabilities in  $\mathbf{S}_1$ ,  $T(T - 1)$  probabilities in  $\mathbf{T}$ , and  $T(K - 1)$  probabilities in  $\mathbf{G}$ . Consequently, when a particular HMC model is proposed, it is usually the case that many of the probabilities are specified to be zero or one, and the others are made functions of a parameter  $\Theta = (\theta_s)_{1 \times S}$ , with space  $\Omega_\Theta \subseteq \text{Re}^S$ . In this case, the result is a parameterized HMC model, with specifications of functions from  $\Omega_\Theta$  into the component probabilities of  $\mathbf{S}_1$ ,  $\mathbf{T}$ , and  $\mathbf{G}$ . Of course for any  $\Theta \in \Omega_\Theta$ , one has numerical values for these quantities, and one can proceed to calculate probabilities for properties of  $\{X_n | n \geq 1\}$  using (8.3).

**Example 8.3** One of the earliest successful HMC model is the so-called all-or-none model (AON) for a simple paired-associate learning experiment (Bower, 1961). In such an experiment there are several stimulus items and each is paired with exactly one of a small set of response alternatives such as numbered buttons. On each of a series of discrete trials, the stimulus items are presented one at a time for a test, and the participant attempts to guess the response for each item.

Immediately after a response is given to an item, the correct response appears for the participant to study. This procedure is repeated until the participant is able to correctly give the response to each stimulus for several consecutive test trials. The goal of the AON is to represent the processes whereby participants learn the stimulus response pairings.

The responses  $\{X_n | n \geq 1\}$  to a particular paired-associate item are coded as correct ( $C$ ) or error ( $E$ ). On any trial, the AON assumes that the stimulus item is in one of two latent, discrete states  $W = \{L, U\}$ , where  $L$  is called the learned state, and  $U$  the unlearned state. When a stimulus item is presented for test, the correct response is made if the item is in state  $L$ , and if the item is in state  $U$ , then there is a probability  $g \in (0, 1)$  that the correct response is guessed. This model assumes that all stimulus items start in state  $U$ , but on every study occasion where an item is in state  $U$ , there is a probability  $c \in (0, 1)$  of making a transition to state  $L$ . Further, once an item is in state  $L$  it remains there throughout the experiment. The name of the AON model derives from the fact that its specification implies that learning either completely occurs or it completely fails to occur.

The AON is specified by a parameter  $\Theta = (c, g) \in (0, 1)^2$ , a start vector  $\mathbf{S}_1 = (0, 1)$ , a transition matrix

$$\mathbf{T} = \begin{matrix} & L & U \\ \begin{matrix} L \\ U \end{matrix} & \begin{bmatrix} 1 & 0 \\ c & (1 - c) \end{bmatrix} \end{matrix},$$

and a response matrix

$$\mathbf{G} = \begin{matrix} & C & E \\ \begin{matrix} L \\ U \end{matrix} & \begin{bmatrix} 1 & 0 \\ g & (1 - g) \end{bmatrix} \end{matrix}.$$

From these quantities of the HMC model, it is easy to compute the state probabilities after study trial  $n \geq 2$ ,

$$\mathbf{S}_n = \mathbf{S}_1 \cdot \mathbf{T}^{n-1} = [1 - (1 - c)^{n-1}, (1 - c)^{n-1}],$$

and then the probability of the responses on test trial  $n \geq 2$  can be computed from (8.3) by

$$\mathbf{R}_n = \mathbf{S}_n \cdot \mathbf{G} = [1 - (1 - g)(1 - c)^{n-1}, (1 - g)(1 - c)^{n-1}].$$

Notice that a correct response on trial  $n$  could have been due to either a guess in state  $U$  or a correct response from state  $L$ , and as a consequence it is easy to show that the stochastic process,  $\{X_n | n \geq 1\}$ , for the all-or-none model is not a Markov chain.

From  $\mathbf{R}_n$ , the probability of an incorrect response over trials is given by

$$\Pr(X_n = E) = (1 - g)(1 - c)^{n-1},$$

and this shows that the error probability for the AON decreases geometrically from an initial value of  $(1 - g)$  to an asymptote of zero as  $n$  increases. Initially researchers

found this result quite striking, and it appeared to be paradoxical because the model implies, quite differently, that the error probability stays at a value of  $(1 - g)$  for one or more trials and then takes a step function to zero when learning has occurs. Further, the geometric error probability function results from quite different learning models, such as the linear operator model of Bush and Mosteller (1951) as described next.

The linear operator model (LOM) for simple paired-associate learning assumes that the error probability for an item undergoes a constant proportional decrease on every learning trial rather than a single step decrease to zero assumed by the AON. Specifically, the LOM is specified by two parameters, an initial error rate  $q_1 \in (0, 1)$  and a learning parameter  $\theta \in (0, 1)$ . The LOM assumes that the response random variables  $X_n$  are independent Bernoulli trials with marginal error probabilities given by the function  $P(X_n = E) = (1 - \theta)^{n-1} q_1$ , which exhibits a geometric drop in error probability over trials.

Clearly, by setting  $c = \theta$  and  $(1 - g) = q_1$ , the two models predict the same geometrically decreasing shape of the error curve. However, the result for the AON occurs because there is a geometric probability distribution over the trial on which learning occurs, and the geometric error curve is the result of averaging over the trial of learning. Despite the indistinguishable error curves for the two models, the models are quite different in other respects. In order to provide results that differentiate the AON from other learning models like the LOM, new statistics of the process need to be examined. For example, define a random variable  $L$  on  $\{X_n | n \geq 1\}$ , where  $L$  is the trial number of the last error (if no errors, set  $L = 0$ ). Then it is straightforward to derive for the AON that

$$\forall m \leq n, \Pr(E_m | L > n) = (1 - g),$$

and this result of a flat curve reveals the all-or-none character of the model. In the case of the LOM, the fact that the response random variables are independent results in

$$\forall m \leq n, \Pr(E_m | L > n) = (1 - \theta)^m q_1,$$

which exhibits a geometrically decreasing conditional error curve as opposed to the flat conditional error curve of the AON. Experimental work in simple paired-associate learning turned out to be more supportive of the AON than the LOM.

After the early success in understanding simple paired-associate learning data, more complicated HMC models were developed for a number of experimental learning paradigms. These models had more states than the AON, and new computational tools for the models were developed. Some of the relevant literature for these developments is presented in Section 8.5.

**Example 8.4** In this example, an HMC model for a paired comparison choice experiment is developed. In a paired comparison experiment there is a master set of choice objects  $C = \{c_1, \dots, c_M\}$ , such as items on a menu or items at a bookstore, and a participant is presented with pairs of choice objects and is required to pick the

one they prefer. It is presumed that for every two-element subset of the master set,  $\{c_i, c_j\}$ , there is a probability  $p_{i,j}$  that  $c_i$  is picked over  $c_j$ , and all these probabilities can be arrayed in a paired comparison matrix,  $\mathbf{P} = (p_{i,j})_{M \times M}$ , where  $p_{ii} \equiv 1/2$  and  $p_{ij} + p_{ji} = 1$ . First, an HMC model will be constructed for a particular pair, and then it will be extended to all pairs in  $\mathbf{P}$ .

Suppose a participant is confronted with a choice set  $\{c_i, c_j\}$  and required to select one of the objects. Define two parameters  $(\theta_i, \theta_j) \in (0, 1)^2$  that reflect the value of each object to the participant. Next, assume the participant decides that they “like”  $c_i$  with probability  $\theta_i$ , and independent of that decision, they decide that they like  $c_j$  with probability  $\theta_j$ . If they like one of the objects and fail to like the other, they choose it; however, if they like both or like neither, they go through the same process again until one of the objects is selected. We can represent this process by four states  $W = \{(c_i, \neg c_j), (c_i, c_j), (\neg c_i, \neg c_j), (\neg c_i, c_j)\}$ , where for example “ $\neg c_i$ ” means  $c_i$  not liked. From the description, it is easy to set up the start vector and transition matrix,

$$\mathbf{S}_1 = [\theta_i(1 - \theta_j), \theta_i\theta_j, (1 - \theta_i)(1 - \theta_j), (1 - \theta_i)\theta_j],$$

and

$$\mathbf{T} = \begin{matrix} & \begin{matrix} (i, \neg j) & (i, j) & (\neg i, \neg j) & (\neg i, j) \end{matrix} \\ \begin{matrix} (i, \neg j) \\ (i, j) \\ (\neg i, \neg j) \\ (\neg i, j) \end{matrix} & \begin{bmatrix} 1 & 0 & 0 & 0 \\ \theta_i(1 - \theta_j) & \theta_i\theta_j & (1 - \theta_i)(1 - \theta_j) & (1 - \theta_i)\theta_j \\ \theta_i(1 - \theta_j) & \theta_i\theta_j & (1 - \theta_i)(1 - \theta_j) & (1 - \theta_i)\theta_j \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{matrix}.$$

There are three observable response categories  $C_i, C_j, C_\emptyset$ , for pick  $c_i$ , pick  $c_j$ , and “No Choice,” respectively. The response matrix for the model is

$$\mathbf{G} = \begin{matrix} & \begin{matrix} C_i & C_\emptyset & C_j \end{matrix} \\ \begin{matrix} (i, \neg j) \\ (i, j) \\ (\neg i, \neg j) \\ (\neg i, j) \end{matrix} & \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{matrix}.$$

From these, the response distribution on trial  $n$  is given by applying (8.3)

$$\mathbf{R}_n = \mathbf{S}_1 \cdot \mathbf{T}^{n-1} \cdot \mathbf{G} = \left[ \theta_i(1 - \theta_j) \left( \frac{1 - A^n}{1 - A} \right), A^n, \theta_j(1 - \theta_i) \left( \frac{1 - A^n}{1 - A} \right) \right],$$

where  $A = \theta_i\theta_j + (1 - \theta_i)(1 - \theta_j)$ . The only difficulty in obtaining the result above is to compute the terms of  $\mathbf{T}^{n-1}$ . Their computations are straightforward by using a result from the theory of sums. Suppose  $S = \sum_{i=1}^n \alpha^{i-1}$  for some  $\alpha \neq 0$ , then  $S = (1 - \alpha^n)/(1 - \alpha)$ . This result is immediately obtained by noting that  $S - \alpha S = 1 - \alpha^n$ .

The HMC model defined above has several interesting properties that can be used to develop a discrete state model for a paired comparison experiment. First, the underlying four state Markov chain has two states that once either one of them

is entered, the chain remains in that particular state for all future trials. These states are called absorbing states of the Markov chain. Example 8.3 also has an absorbing state, namely the learned state  $L$ . It is obvious from the structure of  $\mathbf{T}$  that states  $(i, j)$  and  $(\neg i, \neg j)$  are transient; so with probability 1 the chain will eventually reach one of these absorbing states. Second, because the absorbing states correspond to choosing one of the two choice alternatives, we can define the choice in a paired comparison experiment as the object corresponding to the absorbing state entered by the chain. Therefore,

$$p_{i,j} = \frac{Pr(\exists n \geq 1 \text{ with } Y_n = C_i)}{Pr(\exists n \geq 1 \text{ with } Y_n = C_i) + Pr(\exists n \geq 1 \text{ with } Y_n = C_j)}.$$

From the response distribution,

$$\lim_{n \rightarrow \infty} \mathbf{R}_n = [\theta_i(1 - \theta_j), 0, \theta_j(1 - \theta_i)]/(1 - A),$$

so it is easy to see that

$$p_{i,j} = \frac{\theta_i(1 - \theta_j)}{\theta_i(1 - \theta_j) + \theta_j(1 - \theta_i)}. \quad (8.4)$$

Further, for any other choice set  $\{c_k, c_l\}$ , (8.4) would hold with the appropriate change of subscripts, so we have developed a discrete-state HMC paired comparison model for  $\mathbf{P}$  with parameter  $\Theta = (\theta_i)_{1 \times M} \in (0, 1)^M$  and entries given by (8.4).

The HMC paired comparison model reveals an interesting comparison to a well-known paired comparison model called the Bradley–Terry–Luce (BTL) model. The BTL model in one form, called the Luce ratio rule (Luce, 1959), has parameters  $\mathbf{V} = (v_i)_{1 \times M} \in (0, \infty)^M$ , and paired comparison probabilities are computed by  $p_{i,j} = v_i/(v_i + v_j)$ . Although the HMC model and BTL model may look quite different on the surface, they are in fact statistically equivalent (reparameterizations of each other) in the sense that each model can account for exactly the same set of paired comparison matrices  $\mathbf{P} = (p_{i,j})_{M \times M}$ . This can be seen by the one-to-one transformation  $\forall i, \theta_i = f(v_i) = v_i/(1 + v_i)$ . Plugging the new parameters into the HMC paired comparison model yields:

$$p_{i,j} = \frac{\theta_i(1 - \theta_j)}{\theta_i(1 - \theta_j) + \theta_j(1 - \theta_i)} = \frac{\frac{v_i}{(1 + v_i)(1 + v_j)}}{\frac{(v_i + v_j)}{(1 + v_i)(1 + v_j)}} = \frac{v_i}{(v_i + v_j)}.$$

Thus, for every  $\mathbf{V} \in (0, \infty)^M$  that generates a matrix of paired comparison probabilities from the BTL model, there is a  $\Theta \in (0, 1)^M$  that generates exactly the same matrix from the HMC paired comparison model and visa versa using  $v_i = f^{-1}(\theta_i) = \theta_i/(1 - \theta_i)$ . The BTL is usually thought of as a continuous model rather than a discrete state model, so the demonstration that it is statistically equivalent to a discrete state HMC model means that one should be careful in attributing the structure of a cognitive model to a continuum of states just because one form of the model has this property.

### 8.3 Multinomial processing tree (MPT) models

Multinomial processing tree (MPT) models constitute a large family of discrete state models for categorical data that have become quite popular in cognitive modeling. Informally, an MPT model assumes that the presentation of a stimulus item triggers one of a number of discrete state processing sequences, each of which in turn leads to an observable response category. The processing sequences consist of various successful or unsuccessful cognitive acts, each contingent on the outcome of the previous act. For example, suppose one is exposed to items in a memory experiment. The items might be attended to in several ways, and each way might influence how the items are stored in memory. Further, the way the items are stored in memory will influence the probabilities that they will be retrieved, which in turn will lead to different observable categorical responses. There may be multiple processing sequences that result in the same response category, so knowledge of the response category does not necessarily reveal the underlying processing sequence. In this way, MPT models share properties with HMC models.

The remainder of this section will have four major subsections. The first subsection will discuss the formal specification of the class of MPT models as a class of parameterized rooted tree structures. The second subsection will describe several specific MPT models for different experimental paradigms; and it will also discuss how one can validate an MPT model in a way that allows a researcher to interpret the parameters as actually tapping the cognitive processes for which they were developed. The third subsection will describe ways to reparameterize MPT models to capture testable constraints on the parameters. The fourth subsection will develop a context-free grammar that represents MPT models as special classes of symbol strings. This section will relate the string representation to the rooted tree specification. The statistical inference of MPT models will be discussed in Section 8.4, so Section 8.3 will only discuss statistical assumptions about the data when it is necessary to understand the formal underpinnings of the models.

#### 8.3.1 Specification of MPT models

Most discrete state models for cognition, including those discussed in earlier sections, can be parameterized to fit into a subclass of MPT models called binary MPT (BMPT) models. First, the class of BMPT models will be specified formally, and then its generalization to multi-link MPT (MMPT) models will be described. Models in either of these classes are parametric probability models for categorical data. Consequently, before detailing the formal properties of MPT models, it is useful to develop a general specification for parametric categorical models.

##### 8.3.1.1 Parametric categorical models

A categorical probability model specifies a collection of probability distributions over a finite set of manifest (observable) categories  $\mathcal{C} = \{C_1, \dots, C_K\}$ , for some

positive integer  $K$ . Such models attempt to understand categorical data, which consist of observations, each of which falls into one and only one member of  $\mathcal{C}$ . For example, when a die is rolled, the result falls into one of six categories, or when a person is born, the month of their birth can be classified into one of 12 categories. The collection of all possible probability distributions on  $\mathcal{C} = \{C_1, \dots, C_K\}$  is known as the probability simplex in  $\text{Re}^K$ , and it is given by

$$\Lambda_K = \left\{ \mathbf{p} \in \text{Re}^K \mid p_k \geq 0, \sum_k p_k = 1 \right\}.$$

A parametric model for categorical data specifies probability distributions over a category system such as  $\mathcal{C} = \{C_1, \dots, C_K\}$  in terms of a parameter  $\Theta = (\theta_1, \dots, \theta_S)$ , for some positive integer  $S$ . Parameters are drawn from a parameter space  $\Omega_\Theta \subseteq \text{Re}^S$ . In particular, the model provides a function  $p : \Omega_\Theta \rightarrow \Lambda_K$ , such that for each  $\Theta \in \Omega_\Theta$ ,  $p(\Theta)$  selects one of the probability distributions specified by the model. In the following, let  $p(\Omega_\Theta) \subseteq \Lambda_K$  denote the range of the function  $p$ , namely all the probability distributions specified by the model. In the case of scientifically motivated categorical models for cognition, one designs the parameter space to consist of component parameters,  $\theta_s$ , that are hypothesized to tap underlying latent cognitive acts that combine to produce observed categorical data. This is the validity assumption of a parametric model, and it will be discussed later in Section 8.3.2 in the context of MPT models; however, for now it is important to be clear about several mathematical aspects concerning the parameter space of a categorical model that are sometimes misunderstood by users.

First, from a mathematical perspective the parameter space of a categorical model is not at all unique, and one purpose of the parameter space is to serve as an index set for the model's probability distributions in the sense that each  $\Theta \in \Omega_\Theta$  indexes a particular probability distribution  $p(\Theta) \in \Lambda_K$ . There are an unlimited number of alternative ways to index a model's probability distributions. For example, suppose  $\Phi = (\phi_1, \dots, \phi_V) \in \Omega_\Phi \subseteq \text{Re}^V$ , for some positive integer  $V$ , is another parameter with associated parameter space  $\Omega_\Phi$ . Further, suppose that there is a bijection (one-to-one, onto) function  $h : \Omega_\Theta \rightarrow \Omega_\Phi$ . Then,  $\Omega_\Phi$  can serve as an alternate parameter space for the model with parameter space  $\Omega_\Theta$  by defining  $q : \Omega_\Phi \rightarrow \Lambda_K$  by  $\forall \Phi \in \Omega_\Phi, q(\Phi) = p[h^{-1}(\Phi)]$ . Clearly, the range of  $q$  is identical to the range of  $p$  because  $h$  is a bijection. As a result, both models are said to be statistically equivalent because they entail exactly the same subset of the probability simplex  $\Lambda_K$ . Obviously, by observing how the alternative form of the model was constructed with parameter space  $\Omega_\Phi$ , it is easy to see that there are an unlimited number of alternative statistically equivalent models for any given parametric model for categorical data.

A second aspect of the parameter space for a model is that it is usually selected so that it satisfies additional mathematical properties over just serving as an index set for its probability distributions. One such property is that the function  $p$  may satisfy the property that it is one-to-one. This property is useful if categorical data

are used to estimate a model's parameter because a particular probability distribution over the categories in  $p(\Omega_\Theta)$  determines a unique parameter. In such a case, the model is called *globally identified*, and this simplifies the statistical inference for the model as described in Section 8.4. A weaker and nevertheless useful property of the parameter space for a model is a property called *locally identified*. Local identification holds if the function  $p$  has the property that for any  $\Theta \in \Omega_\Theta$  there exists an open neighborhood  $N(\Theta)$  including  $\Theta$  such that  $\forall \Theta' \in N(\Theta)$  with  $\Theta' \neq \Theta$ ,  $p(\Theta') \neq p(\Theta)$ . Finally, another useful mathematical property of the parameter space is that there may be a metric on a model's parameter space such that when two parameters  $\Theta, \Theta'$  are close, so too are their probability distributions  $p(\Theta)$  and  $p(\Theta')$ . Such a property may allow one to use calculus to compute things like

$$\frac{\partial p(\Theta)}{\partial (\theta_s)}$$

to explore properties of the model. It is not necessary for our purposes to discuss this matter in mathematical detail here, because when specific models are discussed later, these properties will become clear.

**Example 8.5** Suppose we flip a possibly biased coin twice and count the number of heads. If the two flips are independent and identically distributed (i.i.d.), then we can illustrate all the aspects of the previous discussion of parametric categorial models. First, there are three categories  $\mathcal{C} = \{C_0, C_1, C_2\}$ , where the subscript refers to the number of heads. The probability simplex is  $\Lambda_3 = \{\mathbf{p} \in \text{Re}^3 | p_k \geq 0, \sum p_k = 1\}$ . One parameter space for the model is  $\Omega_\theta = \{\theta | 0 \leq \theta \leq 1\}$ , where  $\theta$  is interpreted as the probability of a head on any given flip. Then  $p : \Omega_\theta \rightarrow \Lambda_3$  is defined by

$$\forall \theta \in \Omega_\theta, p(\theta) = [(1 - \theta)^2, 2\theta(1 - \theta), \theta^2].$$

A second parameter space could be  $\Omega_\phi = \{\phi | 0 \leq \phi \leq 1/2\}$ . Clearly,  $h : \Omega_\theta \rightarrow \Omega_\phi$  defined by  $h(\theta) = \theta/2$  is a bijection. From this, it is easy to see that  $q(\phi) = p[h^{-1}(\phi)]$  establishes a mapping  $q : \Omega_\phi \rightarrow \Lambda_3$  defined by  $q(\phi) = [(1 - 2\phi)^2, 4\phi(1 - 2\phi), 4\phi^2]$  with exactly the same range as  $p$ .

It is easily seen that both models are globally identified; for example,  $\forall \theta, \theta' \in \Omega_\Theta, p(\theta) = p(\theta') \iff \theta = \theta'$ , and of course the models are locally identified because they are globally identified. Finally,  $p$  is differentiable,

$$\frac{\partial p(\theta)}{\partial (\theta)} = [-2(1 - \theta), 2(1 - 2\theta), 2\theta].$$

A slight modification of the situation provides an example of a model that is locally but not globally identified. Let the two flips be scored as same,  $C_0$  (HH or TT) or different,  $C_1$ . Then  $\forall \theta \in \Omega_\theta, p(\theta) = [\theta^2 + (1 - \theta)^2, 2\theta(1 - \theta)]$ . Because, for example,  $p(\frac{1}{4}) = p(\frac{3}{4})$ , the model is not globally identified; however, as long as  $\theta \in (\frac{1}{8}, \frac{3}{8})$ ,  $p(\theta) = p(\frac{1}{4})$  implies that  $\theta = \frac{1}{4}$ . A similar argument covers the entire parameter space  $\Omega_\theta$ , so the model is locally identified.

### 8.3.1.2 Binary MPT models

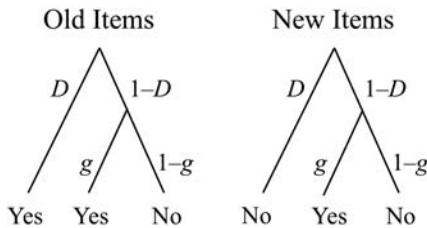
In this subsection, the definition of BMPT models is developed. There are four aspects to a BMPT model for a single category system: (i) observable categories, (ii) latent parameters, (iii) tree architecture, and (iv) computational rules. The first aspect consists of  $K$  response categories,  $\mathcal{C} = \{C_1, \dots, C_K\}$ . The assumption is that participants are presented with an experimental task in which each of their responses to a stimulus item may be scored into exactly one of the categories in  $\mathcal{C}$ . In fact, as we will see later, a usual situation in BMPT modeling is that the participant is exposed to items from several classes of items, each with its own set of response categories; however, first the case of a single category system is defined.

The second aspect of a BMPT model consists of a set of  $S$  functionally independent parameters,  $\{\theta_s | s = 1, \dots, S\}$ , each with full range,  $0 \leq \theta_s \leq 1$ . By functionally independent is meant that each of the parameters can take any value in  $[0, 1]$  irrespective of the values taken by the other parameters. It is useful to array the parameters of a BMPT model in a vector  $\Theta = (\theta_1, \dots, \theta_S)$ ; thus, the parameter space of a BMPT model with  $S$  parameters is the  $S$ -dimensional unit interval  $\Omega_\Theta = [0, 1]^S$ . Each parameter is interpreted as the probability of the occurrence of some latent (unobservable) event, e.g., storing an item in memory, retrieving an item from storage, matching two items, clustering items, making an inference, detecting a signal, discriminating between items, or guessing a response. Some of these possibilities will become clear in the examples in Subsection 8.3.2.

The third aspect of a BMPT model is its tree architecture consisting of cognitive processing branches, each leading to a particular response category in  $\mathcal{C}$ . The tree architecture consists of a particular type of digraph called a *full binary tree* (FBT). To proceed, it is useful to define some terms from graph theory. A digraph,  $\mathbf{D} = (V, A)$ , consists of a finite set of nodes,  $V$ , and a set of directed edges (arcs, links)  $A \subseteq V \times V - \{(v, v) | v \in V\}$ . If  $(u, v) \in A$ , we refer to  $u$  as a parent of  $v$ , and correspondingly  $v$  is called a child of  $u$ . An FBT consists of a single initial node (the root), intermediate nodes, and terminal nodes (the leaves). Each node in the tree except the root has exactly one parent and each nonterminal node in the tree has exactly two children. Such a tree is usually depicted with the root at the top and the leaves at the bottom.

Figure 8.1 presents the DHT model described earlier in Example 8.1, in the form of a BMPT model with two full binary trees, one for the class of old items and the other for the class of new items. Both trees have two internal nodes, three leaves, and four links. Each branch (sequence of links) from the initial node to a leaf corresponds to a possible information processing sequence leading to a response category as described next.

The specification of a BMPT model involves assigning a category to each leaf in the tree, and a parameter,  $\theta_s \in \{\theta_s | s = 1, \dots, S\}$ , or a number,  $x \in (0, 1)$ , to each nonterminal node. If a particular parameter  $\theta_s$  is assigned to a nonterminal node, then the left link from that node is associated with the success, with probability  $\theta_s$ , and the right link the failure, with probability  $(1 - \theta_s)$ , of the corresponding latent cognitive act represented by the parameter. The nodes with number assignments



**Figure 8.1** A BMPT representation of the DHT model in terms of a tree for Old items and a tree for New items. The parameter  $\Theta = (D, g)$  consists of  $D$ , the detection parameter; and  $g$ , the guessing parameter.

work similarly, where  $x$  is the probability of taking the left link and  $(1 - x)$  is the probability of taking the right link. For example, at the root at the top of the left tree in Figure 8.1, the parameter  $D$  is assigned, and the effect of this is to associate probability  $D$  to the left link and  $(1 - D)$  to the right link, as depicted in the figure. Finally, each leaf corresponds to one of the two observable categories. Note that a particular category or parameter may appear several times in a BMPT tree, for example in the left tree of Figure 8.1, the “Yes” category appears twice because for the DHT model there are two different processing sequences that can lead to a “Yes” response.

An interesting feature of BMPT models is that the tree continuation at any internal node  $v$  is itself a BMPT model with  $v$  serving as the root. For example, the tree continuation at the right child of the tree for Old items in Figure 8.1 is itself a BMPT model with one internal node serving as the root and two leaves. This result suggests that any BMPT model can be constructed recursively from a series of BMPT models, one for every internal node. This property will play an important role in Subsection 8.3.4.

The computational component of any BMPT model provides the probabilities of each observable category in terms of the parameters and numbers assigned to the internal nodes. First, each branch in the tree is a sequence of links from the root node into one of the categories. Branch probabilities are computed as products of the parameters or numbers associated with the links on the branch. From this specification, link probabilities have a special form. Let  $B_{jk}$  be the  $j$ th branch in a tree leading to category  $C_k$ , then

$$\Pr(B_{jk}|\Theta) = c_{jk} \prod_{s=1}^S \theta_s^{a_{jk,s}} (1 - \theta_s)^{b_{jk,s}} \quad (8.5)$$

where  $c_{jk}$  is the product of numbers associated with the links of  $B_{jk}$ , or set to one if there are no numerical links associated with that branch, and  $a_{jk,s}$  and  $b_{jk,s}$  are, respectively, the number of links on  $B_{jk}$  associated with parameters  $\theta_s$  and  $1 - \theta_s$ . Because each branch ends in a particular category, the category probabilities,  $p_k(\Theta)$  are sums of branch probabilities,

$$p_k(\Theta) = \sum_{j=1}^{I_k} \Pr(B_{jk}|\Theta) = \sum_{j=1}^{I_k} c_{jk} \prod_{s=1}^S \theta_s^{a_{jk,s}} (1 - \theta_s)^{b_{jk,s}} \quad (8.6)$$

where  $I_k$  is the number of branches leading to category  $C_k$ . Given the structure of (8.6), it is easy to show that

$$\forall \Theta \in \Omega_\Theta = [0, 1]^S, p_k(\Theta) \in [0, 1], \sum_{k=1}^K p_k(\Theta) = 1.$$

This result shows that any FBT with parameters, numbers, and categories assigned to the nodes according to the specification rules gives rise to a collection of probability distributions over the observable response categories.

Note that in Figure 8.1, there are two branches in the left tree for Old items that lead to a “Yes” response. From (8.5), the left-most one can be written as

$$\Pr[B_{1,Y}|(D, g)] = 1 \cdot D^1(1 - D)^0 g^0 (1 - g)^0,$$

and the second one can be written as

$$\Pr[B_{2,Y}|(D, g)] = 1 \cdot D^0(1 - D)^1 g^1 (1 - g)^0.$$

Putting these together using (8.6) yields  $\Pr(\text{“Yes”}|Old) = D + (1 - D)g$ , and this is the same probability as the upper left term in the matrix  $\mathbf{P}(\Theta)$  in Example 8.1 for the DHT model with  $d_{11} = d_{22} = D, d_{12} = d_{21} = 0$ .

As mentioned earlier, most BMPT models are developed for experimental paradigms involving several classes of items, where each participant responds to items in each class. In this case, a parameterized tree for each class of items describes the model. The only modifications to the above description is that there must be a set of response categories and a tree specified for each class of items; however, any given parameter can appear in several trees. In addition, the model generates probability distributions in a product simplex, where each tree is associated with a probability simplex for its category system. For example, the DHT model in Figure 8.1 has two trees, one for the class of Old items and the other for the class of New items. Each item class has two response categories. However, it is important to see that a “Yes” response to an Old item is not the same category as a “Yes” response to a New item. The model maps the parameters into the product simplex given by  $\forall \Theta \in \Omega_\Theta, p(\Theta) \in \Lambda_2 \times \Lambda_2$ .

It is worth observing that in the case where more than one branch leads to the same category, BMPT models are not necessarily log-linear models. Log-linear models are a large class of statistical measurement models for categorical data structures that are well understood. These models have the property that the log of a category probability is linear in the parameters (or their logs). From (8.6) it is clear that this property may not hold for BMPT models since the log of a sum is not the sum of the logs; however, if branches are one-to-one with categories, (8.5) shows that this subclass of BMPT models is in the log-linear family, for example

$$\log[\Pr(B_{jk}|\Theta)] = \log(c_{jk}) + \sum_s [a_{jk,s} \log \theta_s + b_{jk,s} \log(1 - \theta_s)].$$

### 8.3.1.3 Multi-link MPT models

The class of MPT models also includes models where nonterminal nodes of the tree can have two or more than two children, each corresponding to a different processing outcome at that node. These MPT models are called multi-link MPT (MMPT) models, and Figures 8.2 and 8.5 of the next section are examples of MMPT models. The only important difference from the specification for BMPT models is that for MMPT models, a parameter vector or a numerical vector representing the probability distribution over the children of that node is assigned to each nonterminal node of the tree. For example, if an internal node has three children, then one needs to assign a parameter  $\Phi^{(3)} = (\phi_1, \phi_2, \phi_3)$  with space  $\Lambda_3$  or a numerical vector in  $\Lambda_3$  to that node.

More generally, in the specification of the parameters for an MMPT model, one needs to treat parameters assigned to nodes with two children as vectors, for example  $\Phi_s^{(2)} = [\theta_s, (1 - \theta_s)]$ , rather than the one-dimensional form for a BMPT parameter, for example  $\theta_s$ . Then, the parameter set for an MMPT model consists of  $S$  parameter vectors of various dimensionalities,

$$\Phi = \{\Phi_s^{(d_s)} = (\phi_{1s}, \dots, \phi_{ds}) | d_s \geq 2, 1 \leq s \leq S\},$$

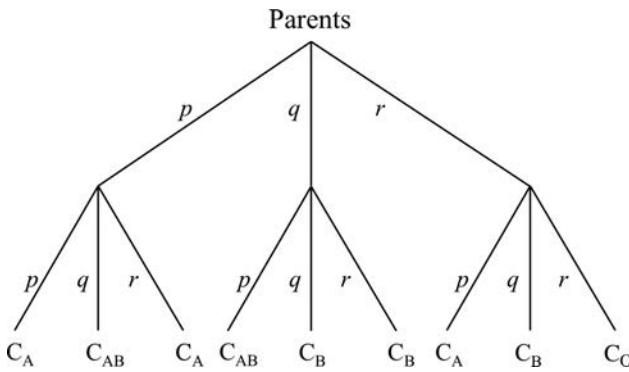
with parameter space  $\Omega_\Phi = \prod_{s=1}^S \Lambda_{d_s}$ .

There are experimental paradigms where it is natural to specify an MMPT rather than a BMPT model, for example in Figure 8.4 there are internal nodes with three possible guessing alternatives. However, it is not difficult to show that every multi-link MPT model is statistically equivalent to a BMPT model where, as mentioned earlier, statistical equivalence means that both models entail exactly the same set of probability distributions over the response category systems. The key idea for generating statistically equivalent BMPT models is shown in Example 8.6. The reparameterization of MMPT models as BMPT models is useful for conducting statistical inference for the MMPT model; however, it is important to note that the parameters in a statistically equivalent BMPT model may fail to retain the substantive meaning associated with the MMPT model parameters.

### 8.3.2 Examples of MPT models

In this subsection, three examples of MPT models will be presented. The first is a famous model from statistical genetics, which is an area that developed a number of multigenerational discrete state models for gene counting that turned out to satisfy the properties of MPT models. However, the field of statistical genetics did not develop a general theory for their class of models such as the MPT class described here. The other examples in this section are two of the first and most-used MPT models for cognitive tasks. Following the examples, the section will have subsections on model identifiability and on model validity.

**Example 8.6** Figure 8.2 presents a MMPT model for the ABO blood group model (Bernstein, 1925). In the model, there are four phenotypic categories,  $C_A$ ,  $C_B$ ,  $C_{AB}$ ,  $C_O$ , corresponding to the four blood types that can be detected by



**Figure 8.2** The ABO blood group model presented as an MMPT. The categories are observable blood group types.

medical procedures. The tree has nine two-link branches corresponding to the nine combinations of mother's gene, A, B, or O, with father's gene. In the model, O is a recessive gene and A and B are dominant. The model has a single parameter  $\Phi^{(3)} = (p, q, r) \in \Lambda_3$ , which represents the probability distribution for the hypothetical equilibrium proportions of A, B, and O genes in the population, which are assumed to hold independently for both members of a couple that has an offspring. One can see that there are three branches that lead each to blood types A and B, including two that have an O gene that is not expressed in the phenotype. In fact the only way to get a detectable O blood type is for both the mother and father to donate an O gene, and this has probability  $r^2$ . The equations for the ABO model are

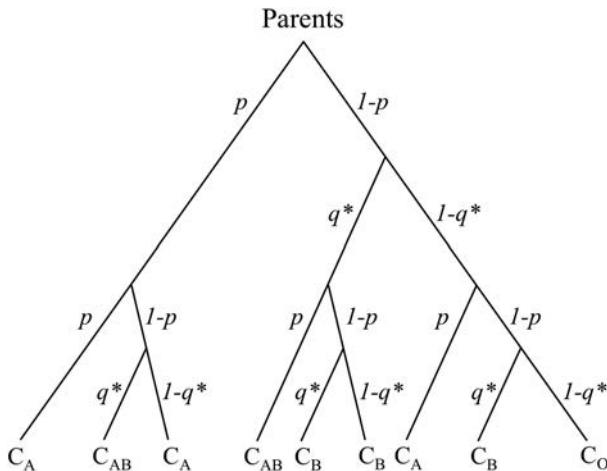
$$p_A = p^2 + 2pr, p_B = q^2 + 2qr, p_{AB} = 2pq, p_O = r^2.$$

The ABO model exhibits a property shared by most MPT models, namely it makes approximate substantive assumptions in service of simplifying the statistical structure of the model. In this case, the simplifying assumptions are that the gene pool has reached equilibrium, that both mother's and father's genes are sampled from the equilibrium distribution, and that there is no mate selection on the basis of blood type. Despite these simplifications, the model has served well in the field to measure gene frequencies in different populations.

It is possible to reparameterize the ABO model into a BMPT model with parameter space  $\Theta = (p^*, q^*) \in \Omega_\Theta = [0, 1]^2$ . The result is presented in Figure 8.3. Given the original MMPT parameter  $\Phi^{(3)} \in \Lambda_3$ , the new parameter  $p^* = p$  has the same meaning as before, and has space  $p^* \in [0, 1]$ ; however, the new parameter  $q^*$  can be related to the original MMPT parameters by

$$q^* = \begin{cases} q/(1-p) & \text{if } 0 \leq p < 1 \\ 0 & \text{if } p = 1 \end{cases}.$$

It is easy to see that the category probabilities for the new BMPT model match those of the original MMPT model with the parameter replacements. For example, from the new tree  $p_A = p^{*2} + 2p^*(1-p^*)(1-q^*)$ , and with replacements

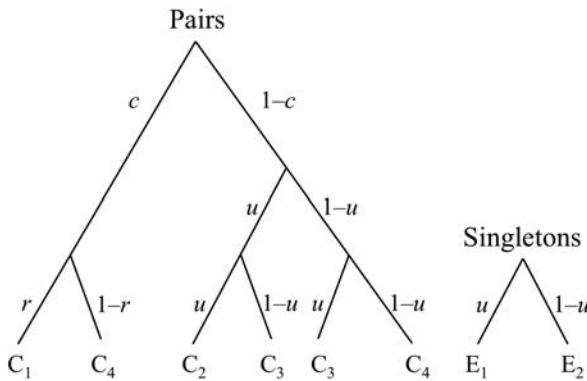


**Figure 8.3** A BMPT reparameterization of the ABO blood group model. The parameters are  $(p, q^*) \in [0, 1]^2$ .

$p_A = p^2 + 2p(1-p)[1 - q/(1-p)] = p^2 + 2pr$ , the same probability distribution occurs as for the MMPT model in Figure 8.2.

From this demonstration, a little work shows that the MMPT model in Figure 8.2 and the reparameterized BMPT model in Figure 8.3 are statistically equivalent. Further, using similar techniques, it is not difficult to show that for any MMPT model, one can construct a statistically equivalent BMPT model. The idea is that at any selected node in the MMPT model that is assigned a parameter vector  $\Phi^{(d)} = (\phi_1, \dots, \phi_d)$ , one constructs  $d - 1$  BMPT parameters as follows. First  $\theta_1 = \phi_1$ , and, following the pattern exhibited by the parameterization in Figure 8.3, one sets for  $1 < i < d$ ,  $\theta_i = \phi_i / (1 - \sum_{j=1}^{i-1} \phi_j)$ . The new parameters are functionally independent each with space  $[0, 1]$  as defined like  $q^*$  above. Once these new parameters are constructed, one replaces the selected node with  $d$  children in the MMPT model with a right branching binary tree with  $d - 1$  left-end links and one right-end link). At each of these end links, in sequence, one continues the tree as it continued in the original MMPT tree for each child of the selected node. When all the internal nodes in the MMPT model are replaced, in any order, by suitable right-branching binary trees, the resulting tree is a BMPT tree that is statistically equivalent to the original MMPT model. For example, the BMPT tree in Figure 8.3 is obtained in this way by four replacements, one for each of the four internal nodes of the MMPT tree in Figure 8.2.

**Example 8.7** Figure 8.4 depicts the two BMPT trees for the pair-clustering model by Batchelder and Riefer (1986). The model was designed to separately measure memory storage capacity and memory retrieval capacity in human memory. The



**Figure 8.4** A BMPT representation of the pair-clustering model in terms of a tree for clusterable pairs and a tree for singleton items.

data for the model come from a specially designed free recall task, where participants first study a list of words, one at a time, and then on a later test they try to recall as many list words as they can in any order. The list of words contains two classes of items: semantically clusterable word pairs, e.g., car and train, daisy and rose, or oxygen and hydrogen, and singleton words that have no clusterable partner. Responses to pairs are scored into four response categories:  $C_1$ , both words recalled successively;  $C_2$  both words recalled, but not successively;  $C_3$  only one of the two words recalled;  $C_4$  neither word recalled. The singleton words are classified into two response categories:  $E_1$  recalled; and  $E_2$  not recalled. The model is presented in Figure 8.4.

The model specifies three component parameters, each designed to measure a different cognitive capacity: a storage parameter  $c$  represents the probability that a word pair is clustered and stored in memory during study, a retrieval parameter  $r$  represents the conditional probability that a clustered word pair is recalled at test, and a parameter  $u$  represents the probability that an unclustered word is both stored in memory and retrieved. Thus the parameter for the model is  $\Theta = (c, r, u) \in \Omega_\Theta = [0, 1]^3$ . Just as in Figure 8.1, the pair-clustering model has two trees for the two classes of stimulus items. Note that there are two branches in the left tree that lead to category  $C_3$  and two that lead to  $C_4$ . In addition, the parameter  $u$  appears three times in the tree for pairs and once in the tree for singletons.

The category probabilities for the pairs are given by

$$\begin{aligned} p_{11} &= cr, \\ p_{12} &= (1 - c)u^2, \\ p_{13} &= 2(1 - c)u(1 - u), \\ p_{14} &= c(1 - r) + (1 - c)(1 - u)^2, \end{aligned}$$

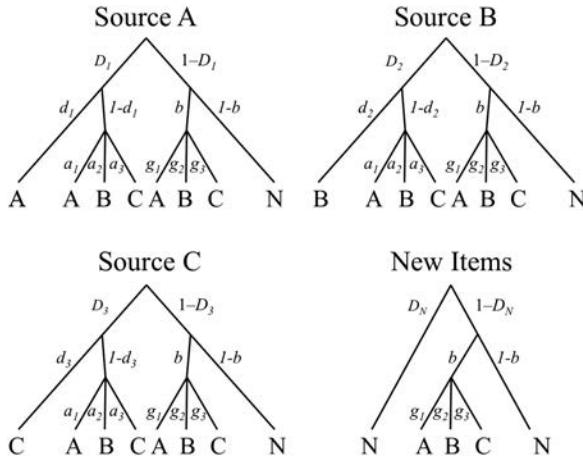
and the singleton probabilities are given by  $p_{21} = u$ ,  $p_{22} = (1 - u)$ . Thus, the probability distributions for the model fall into a product simplex, namely  $p(\Theta) \in \Lambda_4 \times \Lambda_2$ .

The pair-clustering model exhibits several simplifying assumptions. For example, it is assumed that if a pair is clustered and retrieved, the two words are recalled successively ( $C_1$ ), and if not clustered, then the two members of the pair cannot be recalled successively. There is, of course, a small probability that non-clustered items could be recalled successively by chance; however, to include that possibility in the model would make computation much more complicated. A second simplifying assumption is that the recall probability of singletons is equated to the recall of unclustered items in a pair. This assumption is testable as one can compare this version of the model with one parameter  $u$  with a model that allows a different parameter, say  $a$ , that appears instead of  $u$  in the singleton tree.

A final simplifying assumption is that the response categories for the pairs do not take the presentation order during list study into consideration. For example, it is possible to define two categories for sequential recall of both members of a pair depending on whether or not the recall order matched the order of the words in a pair during the study trial. The advantage of adding these categories is that the data structure would have more degrees of freedom, and this would allow more parameters to be worked into the model; however, the disadvantage is that the observed frequency of responses in each category would be decreased, as they would be spread out over the additional categories. This tradeoff between more categories but sparser category counts occurs in developing many MPT models, and it is the theorist's judgment as to which set of response categories is the most productive one to model in any given situation.

**Example 8.8** Another popular area for MPT modeling is a source monitoring task. Such a task is an extension of a simple recognition memory task discussed earlier in Example 8.1, where the items on the study list come from different classes defined by their presentation source, e.g., different speakers, different sensory modalities, or from different spatial locations. In the test phase, participants must decide if an item is Old or New, but also they must identify the presentation source of items they designate as Old items.

A general source monitoring model for three sources is presented in Figure 8.5. It is an MMPT model because the internal nodes near the bottom of the trees have three children corresponding to guessing which source an item may have come from. The model has a tree for each of the three classes of presentation sources for the old items (A, B, or C) as well as a tree for new items, and in total there are 29 processing branches. The model has 10 parameters: four “detection parameters,”  $D^{(2)} \in \Lambda_2$ , three “discrimination parameters,”  $d^{(2)} \in \Lambda_2$ , and three other parameters for guessing processes,  $b^{(2)} \in \Lambda_2$ ,  $a^{(3)} \in \Lambda_3$ ,  $g^{(3)} \in \Lambda_3$ . These 10 parameters have a total of 22 component parameters, but the actual number of free parameters is 12 because the binary parameters each contribute one free parameter and  $a_3 = (1 - a_1 - a_2)$ ,  $g_3 = (1 - g_1 - g_2)$ . Each of the four trees has a parameter  $D^{(2)}$  associated with its root that gives the probability of detecting or failing to detect if an item is old or new. On the three trees for old items, on the detection  $D$  link is a parameter  $d^{(2)}$  for the probability that the source of a detected



**Figure 8.5** An MMPT model representation of the source monitoring model for three sources, A, B, or C. The categories are the three sources, A, B, or C, and N for a new, unstudied item.

old item is or is not correctly discriminated. Successful discriminations lead to correct responses to items; however, failures to detect, with probability  $(1 - D)$ , and failures to discriminate, with probability  $(1 - d)$  lead to guessing processes. In particular,  $b^{(2)}$  refers to the probability that a nondetected item is or is not biased into one of the old source categories, and the  $a_i$  and  $g_i$  refer to the probability of guessing a particular source, respectively, for nondiscriminated old items and nondetected items.

There are four categories for each tree, and while they have the same labels (A, B, C, N) in each tree, they have different meanings in each tree because they refer to different classes of studied items. Letting 1, 2, 3, 4 stand, respectively, for A, B, C, N, the parameter for the model along with its space is

$$\Theta = \left[ \left( D_i^{(2)} \right)_{i=1}^4, \left( d_i^{(2)} \right)_{i=1}^3, b^{(2)}, a^{(3)}, g^{(3)} \right] \in \Omega_\Theta = (\Lambda_2)^8 \times (\Lambda_3)^2,$$

and

$$\forall \Theta \in \Omega_\Theta, p(\Theta) \in (\Lambda_4)^4,$$

namely the product simplex of four copies of  $\Lambda_4$  having  $4 \times 3 = 12$  degrees of freedom.

### 8.3.2.1 Model identifiability

In Section 8.3.1.1, it was pointed out that for measurement purposes it is useful for an MPT model to be globally identified. In such a case, the function from the model's parameter space to the proper simplex for the category system (or product simplex if there are several category systems, each associated with a BMPT tree) is one-to-one, and this means that knowledge of the probability distribution over

the category system leads to a single possible model parameter. As Section 8.4 shows, under certain data sampling assumptions, one can gain knowledge of this probability distribution, so in these cases one can measure the probabilities of the various cognitive acts associated with the model parameters.

One fact about identifiability is obvious, namely, if an MPT model has more free parameters than degrees of freedom in the probability simplex (or product simplex), the model cannot be globally identified. However, if there are no more free parameters than degrees of freedom in the simplex, a model may or may not be globally identified. One way to check to see if a model is globally identified is to assume two values of the parameters,  $\Theta, \Theta'$ , and then work with the system of equations  $\forall k, p_k(\Theta) = p_k(\Theta')$ . If one can show that necessarily  $\theta_s = \theta'_s$ , then the parameter  $\theta_s$  is globally identified.

For example, consider the BMPT model for clusterable pairs in Figure 8.4. It has three parameters and the pair simplex is  $\Lambda_4$  with three degrees of freedom. Assume two parameters for the pairs  $\Theta = (c, r, u)$ ,  $\Theta' = (c', r', u')$ . Next, equate the category probabilities,  $\forall k, p_k(\Theta) = p_k(\Theta')$ . From this, we have

$$\frac{p_3}{p_4} = \frac{2u(1-u)}{(1-u)^2} = \frac{2u'(1-u')}{(1-u')^2},$$

and this implies  $u = u'$ , so  $u$  is globally identified. Next, using this fact and equating expressions for  $p_2$  establishes that  $c$  is globally identified, and from equating expressions for  $p_1$  establishes that  $r$  is globally identified, thus the BMPT model for clusterable pairs is globally identified.

More generally, determining if an MPT model is globally identified by this method is not in general straightforward, as the system of nonlinear equations obtained by the restriction  $p(\Theta) = p(\Theta')$  can become quite unwieldy. For example, the reader is invited to try and check to see if the general source monitoring model for three sources is globally identified. Section 8.5 will provide some references for further study.

### 8.3.2.2 Model validity

In the examples of the previous subsection, the MPT parameters were interpreted as tapping various cognitive acts such as clustering a pair of related items, retrieving a clustered pair, detecting an item as old, discriminating the source of a detected item, and guessing. However, just because a theorist claims that the parameters are tied to these hypothetical cognitive acts does not mean that they are so tied. In fact, the ability of an MPT model to fit data in a cognitive paradigm has little to do with whether or not the parameters are connected to the underlying cognitive acts that they were postulated to tap. The reason for this is discussed in Section 8.3.1, where it is shown that given any particular parameterization of a categorical model, there are an unlimited number of other parameterizations that generate statistically equivalent models. Statistically equivalent models have exactly the same capacity to fit data, so the ability of a model to fit data does not say anything about the validity (substantive interpretation) of a particular parameterization.

As a consequence, a key question in the development of an MPT model is whether the model's parameters are in fact valid measures of their respective associated cognitive processes. Ideally, a parameter designed to measure, say, memory retrieval should be influenced by experimental factors that are known to affect memory retrieval, and not by factors that are designed to affect other, unrelated cognitive processes such as memory storage or guessing. Thus, determining validity typically involves fitting the model to data from a series of experiments, where one tries to show that a particular parameter is selectively influenced by experimental manipulations that are known theoretically to affect the cognitive act that the parameter is designed to measure.

Almost all the many papers designing new MPT models include validity studies of this sort. For example, published studies with the pair-clustering model have shown that providing the category names as retrieval cues during retrieval of the clusters increases estimates of the cluster retrieval parameter  $r$  without significantly changing the storage parameter  $c$ . On the other hand, increasing the study time or the proximity of the clusterable pairs during the presentation of the study list has been shown to selectively influence the cluster storage parameter  $c$ . For another example, varying the proportion of Old and New items in a simple recognition memory experiment discussed in Example 8.1 has been shown to affect the guessing parameter  $g$  but not the detection parameter  $D$ . On the other hand, using low-frequency words (words that do not occur often in everyday printed material) rather than high-frequency words in a recognition experiment increases the estimated value of  $D$ . This result is consistent with the theory that in recognition memory one must discriminate between words on the study list and words encountered in everyday experience. Such discrimination should be easier for words that are rarely used in daily life as opposed to more frequently used words.

Of course, showing that the interpretation of the parameters of an MPT model are selectively influenced by experimental manipulations as they ought to be does not establish the model as the scientifically correct model for its experimental paradigm. For example, there are several quite different recognition memory models that have parameters for detecting (or discriminating) Old items from New items, as well as parameters for biasing responses like the parameter  $g$  in the DHT model. These models also tend to satisfy the selective influence studies just described for the DHT model. This is not surprising, as many MPT models like the DHT model are based on approximations to more elaborate cognitive processing assumptions. The bottom line is that establishing model validity through selective influence studies is a necessary but not a sufficient condition to establish the model as a useful tool to measure latent cognitive processes. Section 8.5 has more literature on selective influence studies for MPT models as well as for other cognitive models.

### 8.3.2.3 MPT models as measurement tools

To use a validated MPT model like the pair-clustering model or the ABO blood group model as a measurement tool, it is important to consider the experimental

conditions under which it is used. In experimental cognitive psychology, a frequent strategy is to treat models as though they are scientific theories and seek experiments to falsify them. The objective is to seek the correct theory as one that survives successive falsification efforts. This may be a good scientific strategy but not a good measurement strategy. MPT models are not intended as final theory, but as useful approximations to more exact theories. So when a researcher finds a variation on an experiment situation like the pair-clustering task where the model fails to account for the data, it is important to consider what one should do. If the researcher is seeking a correct scientific theory, one should either reject the pair-clustering model or else expand it to handle the new experimental situation. On the other hand, if an MPT model has already passed validation tests and has proven useful in certain experimental situations, one should not reject the model but instead restrict the experimental conditions where it is used. A term that characterizes this strategy is *cognitive psychometrics*. Unfortunately, in the view of the author, cognitive psychologists have too frequently followed the scientific strategy rather than the measurement strategy, and as a consequence a large number of models in areas like learning, memory, classification, and choice response time have been dropped from consideration due to their inability to fit certain types of data after an initial positive reception. This scientific strategy applied to behavioral experiments in cognition has left behind numerous, useful models in search of the correct scientific model, and worse, at the time of this writing there are no universally accepted models in these areas. In this, one is reminded of the first half of the last century where many behaviorally based models of the learning process were invented and became popular, and yet today they have almost no play in learning or cognitive psychology.

### 8.3.3 Parametric constraints in MPT models

It is often the case that a researcher wants to test a hypothesis for an MPT model that involves a restriction on its parameter space. Such restrictions result in nested models in the sense that the set of probability distributions that can be generated by the parameters of the restricted model is a subset of the probability distributions that can be generated by parameters of the unrestricted model. For example, consider a BMPT with parameter  $\Theta = (\theta_1, \dots, \theta_S) \in \Omega_\Theta = [0, 1]^S$ . Two simple parameter restrictions are equating a particular parameter to a number, e.g.,  $\theta_s = 0.50$ , or equating two parameters, e.g.,  $\theta_s = \theta_t$ . Both of these restrictions are easily accommodated in the BMPT model. In the first case, one replaces each assignment of  $\theta_s$  to an internal node of the FBT with the number  $x = 0.50$ , and in the second case, one can drop  $\theta_t$  (or alternatively  $\theta_s$ ) from  $\Theta$ , and everywhere that  $\theta_t$  is assigned to an internal node one assigns  $\theta_s$  instead.

Unlike the simple parameter restrictions above, there are others that cannot be accommodated by simple changes in their associated BMPT tree. These parametric restrictions lead to violations of the BMPT requirement that all parameters are free to vary independently in their full space  $[0,1]$ . Examples of such restrictions are

$\theta_s = \theta_u\theta_v$  or  $\theta_s \leq \theta_t$ . In the first case, parameter  $\theta_s$  has a full space [0,1]; however, its value is restricted by the values of  $\theta_u$  and  $\theta_v$ . In the second case, the restriction is called a parametric order constraint, and the parameter space for the pair  $(\theta_s, \theta_t)$  is  $\Omega_{(\theta_s, \theta_t)} = [(x, y) | 0 \leq x \leq y \leq 1]$ , which has exactly half the area of [0,1]<sup>2</sup>. Thus, because of the violation of the requirement that all parameters of a BMPT model are functionally independent and free to vary in [0,1], the original BMPT model subject to either of these parameter constraints is not a member of the class of BMPT models. It is, of course, possible to conduct statistical inference for a BMPT model with such parameter constraints; however, it is inconvenient because one has to write appropriate code that accommodates the restrictions.

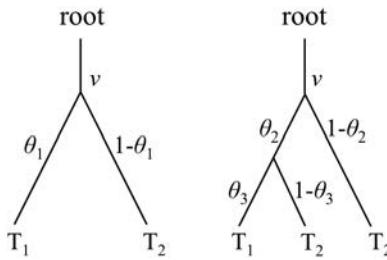
As will be seen in Section 8.4, there is a general approach and associated software that can conduct statistical inference for any model in the class of BMPT models, and this software becomes unavailable when parametric constraints remove a model from the BMPT class. However, it turns out that for the constraints mentioned above, as well as for a number of others, the constrained BMPT model can be reparameterized into a statistically equivalent (unconstrained) BMPT model, and one can use the general BMPT software on the statistically equivalent BMPT to handle inference for the constrained model. The remainder of this section will discuss two types of parameter constraints, and establish some reparameterization theorems in both cases. The first type concerns the case where some of the parameters are defined as functions of the others, and the second type will concern parametric order constraints. Theorem 8.1 concerns the first case.

**Theorem 8.1** *Let  $\Theta = (\Theta_1, \Theta_2)$  be the parameter for a BMPT model where  $\Theta_1 = (\theta_1, \dots, \theta_t)$ ,  $\Theta_2 = (\theta_{t+1}, \dots, \theta_S)$ , and suppose the first  $t$  parameters are specified by the remaining  $S - t$  parameters by a function  $\Theta_1 = [f_1(\Theta_2), \dots, f_t(\Theta_2)]$ . Then the restricted model with parameter  $\Theta_2$  is statistically equivalent to a BMPT model if each  $f_s(\Theta_2)$  for  $s = 1, \dots, t$ , is of one of the following forms:*

- (i)  $f_s(\Theta_2) = \alpha, \alpha \in (0, 1)$
- (ii)  $f_s(\Theta_2) = \prod_{k=t+1}^S \theta_k^{n_k}, n_k \in \{0, 1\}, k = t + 1, \dots, S$
- (iii)  $f_s(\Theta_2) = \sum_{k=t+1}^S \mu_k \theta_k^{\rho_k}$

where the  $\rho_k$  are positive integers and the  $\mu_k$  are nonnegative reals, with  $\sum_k \mu_k = 1$ .

Theorem 8.1 is adapted from Observation 5 in Hu and Batchelder (1994), and the complete proof can be found in the appendix of that article. However, it is easy to illustrate the ideas behind the proof. Clearly, form (i) is handled immediately because it involves replacing the assignment of a BMPT parameter with a number. To illustrate form (ii), suppose a BMPT model with parameter  $\Theta = (\theta_1, \theta_2, \theta_3)$ , and suppose the parameter  $\theta_1$  is restricted by equating it to a product of the other



**Figure 8.6** Portion of a BMPT model before and after reparameterization with  $\theta_1 = \theta_2\theta_3$ .

two parameters,  $\theta_1 = \theta_2\theta_3$ , as described earlier. This is a version of form (ii) with  $\Theta_2 = (\theta_2, \theta_3)$  and  $\eta_2 = \eta_3 = 1$ . Suppose  $\theta_1$  is assigned to an internal node  $v$  and its two children have left and right tree continuations, respectively,  $T_1$  and  $T_2$ . As pointed out in subsection 8.3.1.2, these tree continuations, along with  $v$  serving as the root, constitute a BMPT model. Next, rewrite the structure of this BMPT model as shown in Figure 8.6.

From Figure 8.6 it is possible to compute the probabilities of the tree continuations at  $v$  for the new structure and compare it to the former structure with the parameter constraint. The result is

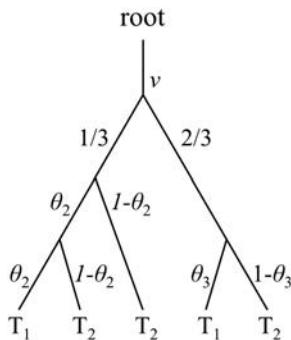
$$\begin{aligned}\Pr(T_1|v) &= \theta_2 \cdot \theta_3 = \theta_1, \\ \Pr(T_2|v) &= \theta_2(1 - \theta_3) + (1 - \theta_2) = (1 - \theta_1).\end{aligned}$$

Next, one needs to perform the same tree modification at all other internal nodes that are assigned  $\theta_1$ . The result is a BMPT model with parameters  $\Theta = (\theta_2, \theta_3) \in [0, 1]^2$  that is statistically equivalent to the original model subject to the parameter restrictions. More complex cases of form (ii) are easily handled in the same way, namely, more levels of the tree are introduced starting at a node assigned to the restricted parameter.

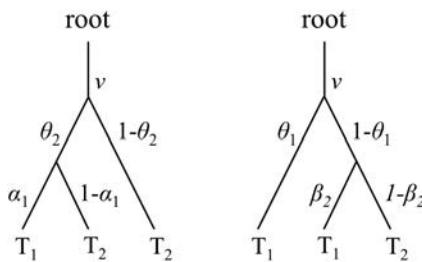
Form (iii) works similarly except the replacement rules are different. Consider the same original BMPT model with parameter  $\Theta = (\theta_1, \theta_2, \theta_3)$ , except that  $\theta_1$  is restricted by  $\theta_1 = \theta_2^2/3 + 2\theta_3/3$ . This is a case of form (iii) with  $\mu_2 = 1/3$ ,  $\mu_3 = 2/3$ ,  $\rho_2 = 2$ ,  $\rho_3 = 1$ . Figure 8.7 shows the tree substitution that reparameterizes the original BMPT model with parameter  $\Theta = (\theta_1, \theta_2, \theta_3)$  subject to the constraint that  $\theta_1 = \theta_2^2/3 + 2\theta_3/3$ . The resulting model is a BMPT model with parameter  $\Theta = (\theta_2, \theta_3) \in [0, 1]^2$ , and further it is statistically equivalent to the original model with the parameter constraint as can be seen by

$$\begin{aligned}\Pr(T_1|v) &= \theta_2^2/3 + 2\theta_3/3 = \theta_1, \\ \Pr(T_2|v) &= (1 - \theta_2)/3 + \theta_2(1 - \theta_2)/3 + 2(1 - \theta_3)/3 = (1 - \theta_1).\end{aligned}$$

Other forms that fit (iii) are handled in the same way. This means that if one wanted to test a hypothesis where each of several BMPT parameters was restricted by one of the three forms in Theorem 8.1, successive substitutions at each internal node



**Figure 8.7** Portion of a BMPT model after reparameterization with  $\theta_1 = \theta_2^2/3 + 2\theta_3$ .



**Figure 8.8** The replacement structures for two methods of reparameterization to handle simple order constraints such as  $0 \leq \theta_1 \leq \theta_2 \leq 1$ . The method on the left retains parameter  $\theta_2$  and the method on the right retains parameter  $\theta_1$ .

assigned one of the restricted parameters would lead to a BMPT model with the unrestricted parameters that is statistically equivalent to the original model with the parameter restrictions. This new model would be nested in the original BMPT model without the restrictions, so statistical hypotheses tests of the restrictions as described in Section 8.4 could be conducted.

The other kind of parametric constraint on a BMPT model described earlier is an order constraint such as  $0 \leq \theta_1 \leq \theta_2 \leq 1$ . Such a constraint violates the property of BMPT models that the parameters are free to vary independently because the value of  $\theta_2$  restricts the possible values of  $\theta_1$ . It is possible to replace  $\theta_1$  with a new parameter  $\alpha_1 \in [0, 1]$ , and construct a BMPT model (with no parameter constraints) with the same number of parameters that is statistically equivalent to the original BMPT model with the order constraint. The idea comes from noting that if  $\theta_1 = \alpha_1\theta_2$ , then the order restriction is guaranteed to hold. The new tree with parameters  $\Phi(\alpha_1, \theta_2) \in \Omega_\Phi = [0, 1]^2$  is obtained by making the substitution in the left panel of Figure 8.8 for every node assigned to  $\theta_1$ . Note that with this substitution

$$\Pr(T_1|v) = \alpha_1 \cdot \theta_2 = \theta_1,$$

$$\Pr(T_2|v) = \theta_2(1 - \alpha_1) + (1 - \theta_2) = (1 - \theta_1).$$

There is another method to accommodate the order constraint above which allows the researcher to retain the parameter  $\theta_1$  rather than  $\theta_2$ . The idea comes from the fact that  $\theta_1 \leq \theta_2 \iff (1 - \theta_2) \leq (1 - \theta_1)$ . Then one can add a parameter  $\beta_2 \in [0, 1]$  in place of  $\theta_2$ , and note that the order constraint is satisfied by  $(1 - \theta_2) = (1 - \beta_2)(1 - \theta_1)$ . As before, a BMPT tree can be constructed with parameter  $\Phi = (\theta_1, \beta_2) \in \Omega_\Phi = [0, 1]^2$  that is statistically equivalent to the original model with the order constraint. The approach is to substitute the structure in the right panel of Figure 8.8 at every internal node in the original model that is assigned  $\theta_2$ . Note that with this substitution

$$\begin{aligned}\Pr(T_1|v) &= \theta_1 + (1 - \theta_1)\beta_2 = \theta_2, \\ \Pr(T_2|v) &= (1 - \theta_1)(1 - \beta_2) = (1 - \theta_2).\end{aligned}$$

A more complex set of order constraints would be  $\theta_1 \leq \theta_2 \leq \dots \leq \theta_n$ . Such an order might occur naturally in a multi-trial learning or memory experiment. For example, in a multi-trial version of the pair-clustering model in Example 8.6, it would make sense that the storage parameter  $c$  and the retrieval parameter  $r$  would be monotonically nondecreasing over successive learning trials. It is easy to expand the two approaches to order constraints in Figure 8.8 to handle this situation by retaining  $\theta_1$  and introducing new unconstrained parameters  $\{\beta_i|i = 2, \dots, n\}$  or retaining  $\theta_n$  and introducing new unconstrained parameters  $\{\alpha_i|i = 1, \dots, n - 1\}$ . For example, the relationship

$$(1 - \theta_i) = (1 - \theta_1) \prod_{j=2}^i (1 - \beta_j), \quad 1 < i \leq n$$

is easily built into a BMPT model with parameter

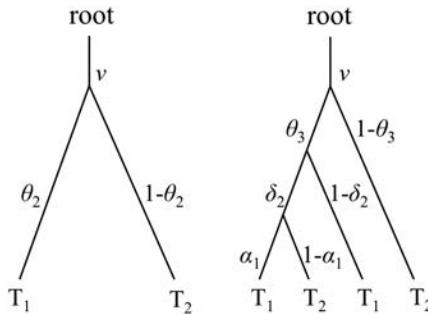
$$\Phi = (\beta_2, \dots, \beta_n, \theta_1) \in \Omega_\Phi = [0, 1]^n.$$

There is an approach to reparameterizing a linear order on BMPT parameters that is not based on extending the methods illustrated in Figure 8.8, and it is sometimes useful in reparameterizing parametric order constraints more complicated than a linear order. The approach involves expressing a parameter as a convex combination of two parameters, one below it and the other above it in the linear order. To simplify, suppose  $\theta_1 \leq \theta_2 \leq \theta_3$ . This approach first introduces  $\alpha_1 \in [0, 1]$  and handles  $\theta_1$  by  $\theta_1 = \alpha_1\theta_3$ . Then a new parameter  $\delta_2 \in [0, 1]$  is introduced, and  $\theta_2$  is expressed by the convex combination

$$\theta_2 = \delta_2\theta_1 + (1 - \delta_2)\theta_3 = \theta_3[1 - \delta_2(1 - \alpha_1)],$$

which assures that its value is constrained to lie between  $\theta_3$  and  $\theta_1$ . Figure 8.9 shows how this mathematical expression can lead to a BMPT model with parameter  $\Phi = (\alpha_1, \delta_2, \theta_3) \in \Omega_\Phi = [0, 1]^3$  that is statistically equivalent to the original BMPT model with the linear order restriction.

**Example 8.9** In general, if there are  $S$  parameters in a BMPT model, one can express any set of order constraints by a binary relation  $A$  on  $V = \{\theta_s|s =$



**Figure 8.9** The replacement structure for a node assigned to  $\theta_2$  in the linear order  $\theta_1 \leq \theta_2 \leq \theta_3$  based on constraining it to be between two parameters in the linear order.

$1, \dots, S\}$ , where  $(\theta_s, \theta_t) \in A \iff \theta_s \leq \theta_t$ . It is obvious that the digraph relation in  $\mathbf{D} = (V, A)$  is transitive

$$[(\theta_s, \theta_t) \in A, (\theta_t, \theta_u) \in A] \implies (\theta_s, \theta_u) \in A$$

and asymmetric,  $(\theta_s \leq \theta_t) \Rightarrow \neg(\theta_t \leq \theta_s)$ , relation. An open problem at the time of this writing is to characterize exactly which digraphs of order constraints on the parameters admit to a statistically equivalent BMPT model with no order constraints and exactly the same number of parameters as the original order-constrained BMPT model. It can be shown that not all order constraint digraphs can be so reparameterized; for example, a BMPT model with parameter  $V = \{\theta_1, \theta_2, \theta_3, \theta_4\}$  and subject to order constraints

$$A = \{(\theta_3, \theta_1), (\theta_3, \theta_2), (\theta_4, \theta_1), (\theta_4, \theta_2)\}$$

cannot be so reparameterized. In seeking a solution to this problem, it is also of interest to see if it is necessary to discover other methods of reparameterization not revealed in the methods illustrated in Figures 8.8 and 8.9.

**Example 8.10** Thus far, only order constraints in BMPT models have been discussed. There are also issues of accommodating order constraints in MMPT models. Suppose  $\Phi_s^{(3)} = (\phi_{1s}, \phi_{2s}, \phi_{3s})$  and  $\Phi_t^{(3)} = (\phi_{1t}, \phi_{2t}, \phi_{3t})$  are two parameters in an MMPT model. There are two kinds of order constraints that one might want to impose. The first would be within parameter constraints, for example  $\phi_{1s} \leq \phi_{2s} \leq \phi_{3s}$ , and the second would be constraints between parameters, e.g.,  $\phi_{1s} \leq \phi_{1t}$ . It is known that some of these constraints can be handled by designing a statistically equivalent MMPT model with the same number of parameters and no parameter constraints; however, so far it is an open question to completely describe the situations that cannot be handled by a statistically equivalent MMPT model with the same number of parameters as the original MMPT model with the order constraints.

### 8.3.4 A string language for MPT models

In Section 8.3.1.2 it was pointed out that BMPT models have a recursive property in the sense that any internal node can serve as a root node for a BMPT model that consists of the remainder of the tree that falls under that node. Based on this property, it is possible to capture the entire class of BMPT models in the form of a string language. In other words, as will be shown, a BMPT model can be viewed as a left to right string of symbols, and given such a string one can derive its FBT representation. This means, for example, that BMPT models can be transmitted easily between computers and entered as strings into inference software. Further, this string language will turn out to derive from a context-free grammar, and as such a number of useful properties of the models can be computed using methods developed for context free languages.

This section will be divided into three subsections. First, in Section 8.3.4.1 the BMPT model string language will be axiomatized along with an extension to MMPT models. Second, in Section 8.3.4.2 the structure of context-free languages will be discussed, and properties of those languages will be used to analyze BMPT model strings. Finally, Section 8.3.4.3 will show the connection between the BMPT string language and BMPT models viewed as parameterized FBTs in Section 8.3.1.

#### 8.3.4.1 A string language for BMPT models

In the following definition, BMPT models will be defined recursively as strings of category and parameter symbols. The definition has structural axioms for the construction of a BMPT string and response axioms for how the string leads to parameterized probability distributions over its categories. To facilitate the definition, one addition is needed to the class of BMPT models defined in terms of FBTs in Section 8.3.2. The addition allows a single category to be a BMPT model. In this case, the root is also a leaf, and the interpretation is that category occurs with probability one. First, the string language for BMPT models will be presented, and then its extension to MMPT models will be described.

**Definition 8.1** Let  $\mathbf{C}$  be a set of categories,  $\mathbf{X} = (0,1)$ , and  $\Theta$ , disjoint from  $\mathbf{C}$  and  $\mathbf{X}$ , be a set of functionally independent parameters each with space  $[0,1]$ . Then the class  $\mathbf{B}$  of BMPT models is defined by the following five axioms:

#### BMPT structural axioms

**Axiom 1** (*Category membership*). If  $C \in \mathbf{C}$ , then  $C \in \mathbf{B}$ .

**Axiom 2** (*String construction*). If  $M_1, M_2 \in \mathbf{B}$  and  $\theta \in \Theta$  or  $x \in \mathbf{X}$  then  $\theta M_1 M_2 \in \mathbf{B}$  and  $x M_1 M_2 \in \mathbf{B}$ .

**Axiom 3** (*String deconstruction*). If  $M \in \mathbf{B} - \mathbf{C}$ ,  $\exists M_1, M_2 \in \mathbf{B}$  and  $\theta \in \Theta$  or  $x \in \mathbf{X}$  such that  $M = \theta M_1 M_2$  or  $M = x M_1 M_2$ .

## BMPT response axioms

### Axiom 4 (Atomic responses)

$$\forall M \in \mathbf{C}, \forall C \in \mathbf{C}; \Pr(C|M) = \begin{cases} 1 & \text{if } M = C \\ 0 & \text{if } M \neq C \end{cases}$$

### Axiom 5 (Conditional responses)

$$\forall M \in \mathbf{B-C}, \text{ with } M = \theta M_1 M_2 \text{ or } M = x M_1 M_2, \forall C \in \mathbf{C}$$

$$\Pr(C|M) = \theta \Pr(C|M_1) + (1 - \theta) \Pr(C|M_2)$$

or

$$\Pr(C|M) = x \Pr(C|M_1) + (1 - x) \Pr(C|M_2).$$

Axioms 1, 2, and 3 serve to construct the class of BMPT models as strings. Axiom 1 asserts that single categories are “atomic” BMPT models in  $\mathbf{B}$ . Axiom 2 shows how to recursively construct new members of  $\mathbf{B}$  from other members of  $\mathbf{B}$ . Note that in terms of the definition, the new model is a string of symbols to be read from left to right, not a parameterized FBT as in the construction in Section 8.3.1. In Section 8.3.4.3, the connection between the definitions of BMPT models in terms FBTs and strings will be explicated. Axiom 3 says that any nonatomic BMPT model can be deconstructed into a string consisting of a parameter or a number followed by two models in  $\mathbf{B}$ . It is designed to delimit the class  $\mathbf{B}$  to only those strings that can be recursively constructed from Axioms 1 and 2. There are other axioms that would serve as this closure property; however, Axiom 3 is designed to facilitate the statement of the response axioms. Note that Axiom 3 does not say the decomposition of  $M \in \mathbf{B} - \mathbf{C}$  is unique, and this property is crucial for applying Axiom 5. In Section 8.3.4.2 this uniqueness property will be shown in Theorem 8.3, which is made possible by introducing context-free grammars. Axioms 4 and 5 are the response axioms. Axiom 4 is obvious, and Axiom 5 shows how to start with a nonatomic BMPT string and compute response probabilities recursively as the string is deconstructed into unique constituent BMPT strings.

**Example 8.11** Consider the BMPT model for old items in Figure 8.1. Coding the categories by  $Y$  and  $N$ , here is a derivation of that model as a string from Definition 8.1. First,  $Y, N \in \mathbf{B}$  by Axiom 1. Then, as  $g \in \Theta$ ,  $gYN \in \mathbf{B}$  by Axiom 2. Finally, as  $D \in \Theta$ , Axiom 2 with  $M_1 = Y, M_2 = gYN$  yields the string  $DYgYN$ . To see that this string yields the same category probabilities as the tree for old items in Figure 8.1, Axioms 4 and 5 can be used. The first line below comes from applying Axiom 5, and the second line involves Axiom 5 again and then Axiom 4.

$$\begin{aligned} \Pr(Y|DYgYN) &= D\Pr(Y|Y) + (1 - D)\Pr(Y|gYN) \\ &= D + (1 - D)[g\Pr(Y|Y) + (1 - g)\Pr(Y|N)] \\ &= D + (1 - D)g. \end{aligned}$$

Thus, the derivation from the BMPT string yields exactly the same category probability distribution as the FBT definition of the Old item tree. The next example involves a more complex derivation.

**Example 8.12** Here is a derivation of the model for the pairs in Figure 8.4 in Example 8.4 from the BMPT string axioms. By Axiom 1,  $C_1, C_2, C_3, C_4 \in \mathbf{B}$ . Then using Axiom 2,

$$rC_1C_4 \in \mathbf{B}, uC_2C_3 \in \mathbf{B}, uC_3C_4 \in \mathbf{B}.$$

Again by Axiom 2,  $uuC_2C_3uC_3C_4 \in \mathbf{B}$ , and again using Axiom 2 we have

$$crC_1C_4uuC_2C_3uC_3C_4 \in \mathbf{B}.$$

This final string represents the BMPT model for the pairs. Next is the computation of the parameterized category probabilities from the string. This is illustrated only for  $\Pr(C_4|crC_1C_4uuC_2C_3uC_3C_4)$ . Using Axiom 3 yields  $\theta = c$ ,  $M_1 = rC_1C_2$ ,  $M_2 = uuC_2C_3uC_3C_4$ , and by Axiom 5,

$$\begin{aligned} \Pr(C_4|crC_1C_4uuC_2C_3uC_3C_4) \\ = c\Pr(C_4|rC_1C_4) + (1 - c)\Pr(C_4|uuC_2C_3uC_3C_4). \end{aligned}$$

Next, taking the left term, and using Axioms 3 and 4,

$$\begin{aligned} c\Pr(C_4|rC_1C_4) &= c[r\Pr(C_4|C_1) + (1 - r)\Pr(C_4|C_4)] \\ &= c[r \cdot 0 + (1 - r) \cdot 1] \\ &= c(1 - r), \end{aligned}$$

and in several similar steps the right term becomes

$$(1 - c)\Pr(C_4|uuC_2C_3uC_3C_4) = (1 - c)(1 - u)^2.$$

Finally, putting things together yields

$$\Pr(C_4|crC_1C_4uuC_2C_3uC_3C_4) = c(1 - r) + (1 - c)(1 - u)^2,$$

and this is the same expression for the category probability  $\Pr(C_4|\Theta = (c, r, u))$  as obtained for the FBT formulation in Example 8.7. It is left as an exercise to show that the other three parameterized category probabilities from the string representation of the pair model match the expressions obtained in Example 8.7.

By using the steps illustrated in Example 8.12, it can be shown that one can construct a string for any parameterized FBT discussed in Section 8.3.2 and that the parameterized category probability distributions from the FBT and string representations are identical. However, the derivations from Definition 8.1 can be quite tedious, and it turns out that there is an easier way to see the connection between the strings and the FBT representations covered next in Section 8.3.4.2. However, before discussing this connection, it is important to exhibit the modifications in the BMPT model string axioms to cover MMPT models.

Two important modifications are needed to define the class  $\mathbf{M}$  of MMPT model strings. First, the class  $\Phi$ , as a substitute for  $\Theta$ , must be modified to include parameter vectors of the form  $\Phi_s^{(d_s)} = (\phi_{1s}, \dots, \phi_{ds})$ , for  $d_s \geq 2$ , each with spaces

$\Omega_{\Phi_s^{(d_s)}} = \Lambda_{d_s}$ . In particular, for an internal node that has two children in an MMPT model, one represents the parameter as  $\Phi_s^{(2)} = (\phi_{1s}, \phi_{2s})$  rather than by  $\theta_s$  as in the BMPT model representation. In addition,  $\mathbf{X}$  must include numerical vectors  $\mathbf{x} \in \Lambda_d$ , for  $d \geq 2$ , in  $\mathbf{X}$ . Second, when a parameter or numerical vector is inserted in the string, it must be followed by a number of models in  $\mathbf{M}$  equal to the dimensionality of the vector. **Axiom 1** remains essentially the same, and the modification of Axiom 2 for MMPT models  $\mathbf{M}$  is as follows.

**Axiom 2 (MMPT string construction).** If  $\Phi^{(d)} = (\phi_1, \dots, \phi_d) \in \Phi$  or  $\mathbf{x}^{(d)} = (x_1, \dots, x_d) \in \mathbf{X}$ , for some  $d \geq 2$ , and  $M_1, \dots, M_d \in \mathbf{M}$ , then  $\Phi^{(d)}M_1, \dots, M_d \in \mathbf{M}$  and  $\mathbf{x}^{(d)}M_1, \dots, M_d \in \mathbf{M}$ .

The deconstruction axiom is a straightforward modification of Axiom 3 for the BMPT strings.

**Axiom 3 (String deconstruction).** If  $M \in \mathbf{M} - \mathbf{C}$ ,  $\exists \Phi^{(d)} \in \Phi$  or  $\mathbf{x} \in \Lambda_d$  for  $d \geq 2$  and  $M_1, \dots, M_d \in \mathbf{M}$  such that  $M = \Phi^{(d)}M_1, \dots, M_d$  or  $M = \mathbf{x}M_1, \dots, M_d$ .

Next, the two response axioms are as follows. First, **Axiom 4** for the MMPT strings is the same as its counter part for BMPT strings, and finally

**Axiom 5 (Conditional responses).** If  $M = \Phi^{(d)}M_1, \dots, M_d \in \mathbf{M} - \mathbf{C}$ , then

$$\forall C \in \mathbf{C}, \Pr(C|M) = \sum_{i=1}^d \Phi_i^{(d)} \Pr(C|M_i),$$

with a similar breakdown for  $\mathbf{x} \in \Lambda_d$ .

Of course, as with the BMPT string axioms, it is important to establish that the deconstruction in Axiom 4 is unique, and as before the way to do this is explained in Section 8.3.4.2.

**Example 8.13** Consider the ABO blood group model in Example 8.6. The string in  $\mathbf{M}$  for that model is easy to construct. Let  $\Omega^{(3)} = (p, q, r) \in \Phi$ . Now,  $A', B', AB, O \in \mathbf{C}$  (using prime to distinguish  $A$  and  $B$  from  $AB$ ) so by Axiom 1  $A', B', AB, O \in \mathbf{M}$ . Next, using Axiom 2,  $\Omega^{(3)}A'ABA' \in \mathbf{M}$ ,  $\Omega^{(3)}ABB'B' \in \mathbf{M}$ ,  $\Omega^{(3)}A'B'O \in \mathbf{M}$ . Finally, again using Axiom 2

$$\Omega^{(3)}\Omega^{(3)}A'ABA'\Omega^{(3)}ABB'B'\Omega^{(3)}A'B'O \in \mathbf{M},$$

and this is the ABO blood model as a string in  $\mathbf{M}$ . It is left as an exercise to show that the response rules applied to the string give the same parameterized category probability distributions as the MMPT model in Example 8.6.

### 8.3.4.2 Context-free and MPT string languages

In order to relate the two definitions of BMPT models, one as a parameterized FBT in Section 8.3.2.2 and the other as strings that satisfies the BMPT axioms in Section 8.3.4.1, it is helpful to use some results in the theory of formal grammars. A formal language in the mathematical theory of grammars is defined as set of

strings of symbols, read from left to right, constructed in accord with some system of formation rules. Most interesting formal languages have infinitely many strings that are generated recursively from production rules. Much of the study of formal languages centers on the so-called Chomsky hierarchy, e.g. Chomsky (1963), where grammars are ordered by their generative power. Each class of grammars, e.g., regular, context-free, context-sensitive, and recursively enumerable, is associated with a type of automaton, e.g., finite automaton, pushdown automaton, linear-bounded automaton, and Turing machine. The grammar generates strings from its axioms and the associated automaton processes strings to see if they are among the strings that could be generated from the grammar. In our case, we need to use languages generated from context-free grammars (CFGs) and their associated pushdown automata to understand the connection between the two definitions of BMPT models.

To facilitate the definition of a CFG, the so-called Kleene star operation applied to a set of strings is needed. First, we use the symbol  $\epsilon$  for the empty string consisting of no symbols. Let  $A$  be a set of finite length strings of symbols, then the Kleene star operation applied to  $A$  is defined as

$$A^* = \{s_1, s_2, \dots, s_k | \forall 0 < k < \infty, \forall 1 \leq i \leq k, s_i \in A\} \cup \{\epsilon\}.$$

**Definition 8.2** A context-free grammar is a quadruple  $(V, \Sigma, R, S)$  where

1.  $V$  is a finite set called the variables;
2.  $\Sigma$  is a finite set disjoint from  $V$  called the terminals;
3.  $R$  is a finite set of productions (or replacement rules), where each production is composed of a variable, an arrow, and a string, and written in the form  $\alpha \rightarrow A$ , where  $\alpha \in V$  and  $A \in (V \cup \Sigma)^*$ ;
4.  $S \in V$  is called the start symbol.

It turns out that a CFG generates a set of strings called words in a context-free language, CFL. The alphabet of a CFL consists of the terminals of a CFG and therefore the words in the CFL are a subset of  $\Sigma^*$ . The words of a CFL are exactly those strings that can be generated by (derived from) the CFG. The way a word is derived is by starting with the start symbol  $S$  and then applying a series of productions until the resulting string consists only of terminals. Obviously the initial production must be of the form  $S \rightarrow A$ , and then  $S$  is replaced by the string  $A$ . Next, some variable  $\alpha$  in  $A$  is selected anywhere in  $A$ , and a production of the form  $\alpha \rightarrow B$  is applied, and  $B$  replaces  $\alpha$  in  $A$ . This continues until only terminals remain in the resultant string, and the result is a word in the CFL generated by the CFG. With this background, it is now possible to define a CFG that links the FBT definition and the string definition of BMPT models together.

**Definition 8.3** The BMPT CFG is  $G_{BMPT} = (\{M\}, \{\theta, C\}, R, M)$  where  $M$  is the only variable, the terminals are  $\{\theta, C\}$ ,  $M$ , is the start symbol, and  $R$  consists of two productions:  $P_1 : M \rightarrow C$  and  $P_2 : M \rightarrow \theta MM$ .

The choice of notation in Definition 8.3 is obviously suggestive of  $M$  for model,  $C$  for category, and  $\theta$  for parameter; however, it is important to remember they are without meaning at this point. When Definition 8.3 is compared to Definition 8.1, certain similarities are evident. For example, the first production is used to change the variable into a terminal  $C$ , and this is analogous to Axiom 1 in Definition 8.1. Also, the second production has a clear similarity to Axiom 2 of Definition 8.1. The language generated by the CFG in Definition 8.3 can be called the BMPT CFL, denoted by  $L_{BMPT}$ .

To illustrate, the following is a word that can be generated from  $G_{BMPT}$ :

$$M \rightarrow \theta \overline{MM} \rightarrow \theta C \overline{M} \rightarrow \theta C \theta \overline{MM} \rightarrow \theta C \theta C \overline{M} \rightarrow \theta C \theta CC,$$

where a line is placed over the variable that is selected for the next production. This word in  $L_{BMPT}$  is related to the string language construction of the BMPT model for old items in Example 8.11. The key is to substitute actual categories and actual parameters (and numbers if any) in place of the generic parameter and category symbols of Definition 8.3. The result is  $DYgYN$ , the same string as in Example 8.11.

In the case of the more complex pair-clustering model string for the pairs, an appropriate word from  $L_{BMPT}$  is derived as follows:

$$\begin{aligned} M &\rightarrow \theta \overline{MM} \rightarrow \theta \theta \overline{MMM} \rightarrow \theta \theta C \overline{MM} \rightarrow \theta \theta CC \overline{M} \rightarrow \theta \theta CC \theta \overline{MM} \\ &\rightarrow \theta \theta CC \theta \theta \overline{MM} \rightarrow \theta \theta CC \theta \theta C \overline{M} \rightarrow \theta \theta CC \theta \theta CC \overline{M} \\ &\rightarrow \theta \theta CC \theta \theta CC \theta \overline{M} \rightarrow \theta \theta CC \theta \theta CC \theta C \overline{M} \\ &\rightarrow \theta \theta CC \theta \theta CC \theta CC. \end{aligned} \tag{8.7}$$

The steps in the derivation involved using either  $P_1$  or  $P_2$  on a selected variable (always  $M$  of course). The series of productions is continued until only terminal symbols are left in the string so it becomes a word. It is interesting to note that the word above derived from  $G_{BMPT}$  is identical to the string for the pairs in the pair-clustering model derived from Definition 8.1 in Example 8.12, when appropriate replacements are made for the parameter and category symbols. The result is  $crC_1C_4uuC_2C_3uC_3C_4 \in \mathbf{B}$ .

It turns out that any string that can be generated by the three BMPT structural axioms in Definition 8.1 is a string that can be generated from Definition 8.3 if categories in  $\mathbf{C}$  and parameters in  $\Theta$  or numbers in  $\mathbf{X}$ , respectively, replace single the symbols  $C$  or  $\theta$  in the appropriate word in  $L_{BMPT}$ . We denote the class of such strings by  $L_{BMPT}^*$ . With this result, the structural axioms in Definition 8.1 have been tied to the theory of CFGs.

Now that we have established a connection between words in  $L_{BMPT}$  and strings from the BMPT structural axioms, it remains to provide a way to determine if a string in  $\{C, \theta\}^*$  is a word in  $L_{BMPT}$ . We have seen from Example 8.12 how complex it would be to examine a string and try to derive it from the axioms; however, it turns out that it is possible to develop a simple, straightforward algorithm for

deciding if a symbol string is in  $L_{BMPT}$ . The result comes from several propositions about the words in  $L_{BMPT}$ . First it is necessary to describe some properties of words in  $L_{BMPT}$ .

**Definition 8.4** Let  $W = x_1x_2 \dots x_N$  for  $N \geq 1$ , be a word in  $L_{BMPT}$ . Define

$$\forall i, T_W(x_i) = \begin{cases} 1 & \text{if } x_i = \theta \\ -1 & \text{if } x_i = C \end{cases}$$

and for  $1 \leq i \leq N$ ,  $S_W(i) = \sum_{j=1}^i T_W(x_j)$ .

Now we are in a position to completely characterize all the words in  $L_{BMPT}$ .

**Theorem 8.2** Let  $W = x_1x_2 \dots x_N \in \{\theta, C\}^*$ , for some  $N \geq 1$ . Then

$$W \in L_{BMPT} \Leftrightarrow (\forall 1 \leq i < N, S_W(i) \geq 0) \wedge (S_W(x_N) = -1).$$

*Proof* Suppose  $W \in L_{BMPT}$  and let  $|W|$  denote the number of symbols in  $W$ . Suppose  $|W| = 1$ . Then  $W$  consists of a single symbol  $C$  and the condition is satisfied. It is obvious from the two production rules that any  $W \in L_{BMPT}$  will have an odd number of symbols. We proceed by induction. Clearly, if  $|W| = 3$ ,  $W = \theta CC$ , and it is easy to see that the condition holds. Now assume  $W \in L_{BMPT}$  with an odd number of symbols  $N \geq 5$  and that the condition holds for all  $V \in L_{BMPT}$  with  $|V| \leq N$ . It is clear that the first production in generating  $W$  was  $M \rightarrow \theta M_1 M_2$ , where the subscripts allow us to separate the resulting productions into those that occur starting with  $M_1$  and those that occur starting with  $M_2$ . Clearly,  $W = \theta W_1 W_2$ , where  $W_i \in L_{BMPT}$  is the word generated by productions that begin with the start symbol  $M_i$ ,  $i = 1, 2$ . Let  $|W_i| = N_i$ , now  $N_1, N_2 < N$  because  $W$  has  $N$  symbols, so by the inductive hypothesis the condition holds for both  $W_1$  and  $W_2$ . From this it is easy to verify that the condition holds for  $W$ . First note that  $S_W(1) = 1$ , further because  $W_1 \in L_{BMPT}$ , for  $1 < i \leq N_1$ ,  $S_W(i) \geq 0$  and  $S_W(|W_1| + 1) = 0$ . Because the running count is zero at the start of the first symbol in  $W_2 \in L_{BMPT}$  and because  $N_2 < N$ , it is clear that for  $N_1 + 1 < i < N_1 + N_2$ ,  $S_W(i) \geq 0$  and finally  $S_W(1 + N_1 + N_2) = -1$ . This establishes that all  $W \in L_{BMPT}$  satisfies the condition.

For the reverse, only the key ideas will be given. Results from other sources show that any string in  $\{\theta, C\}^*$  that satisfies the condition in Theorem 8.2 is in  $L_{BMPT}$ . This result was proved for an isomorphic grammar to the one in Definition 8.3 (Mäkinen, 1998). This grammar will be the appropriate one to describe FBTs, and it is discussed in Section 8.3.4.3. The basic idea is to establish that  $G_{BMPT}$  is a so-called left Szilard language. This means that all words that can be derived from the grammar can be derived unambiguously by applying productions only to the left-most nonterminal symbol in a string that evolves from the start symbol to a string of terminals. The derivation of the word in  $L_{BMPT}$  for the pair-clustering model illustrates a left derivation. Given the fact that all words in  $L_{BMPT}$  are exactly those that can be obtained by left derivations, the sequence of  $S_W(i)$  for a string  $W \in \{\theta, C\}^*$  that satisfies the condition can be shown to track exactly a permissible

sequence of productions of the grammar from the start symbol  $M$  that lead to  $W$  unambiguously.  $\square$

Recall that the string deconstruction axiom (Axiom 3, Definition 8.1) for BMPT models asserted that if a BMPT string  $M$  is not a single category then there are BMPT strings  $M_1, M_2$  and a parameter  $\theta$  (or number  $x$ ) such that  $M = \theta M_1 M_2$ . It was important for the response axioms that this decomposition be unique. Now that we have Theorem 8.2, it is easy to establish the uniqueness.

**Theorem 8.3** *Consider Definition 8.1 and suppose  $M = \theta M_1 M_2 = \theta' M_3 M_4$  where  $M, M_1, M_2, M_3, M_4 \in \mathcal{B}$ ,  $\theta, \theta' \in \Theta$ . Then  $\theta = \theta'$ ,  $M_1 = M_3$ ,  $M_2 = M_4$ .*

*Proof* Clearly  $\theta = \theta'$  because it is the first symbol in  $M$ . Clearly if  $M_1 = M_3$ , then  $M_2 = M_4$  because of the identity of the two strings. Toward a contradiction suppose  $M_1 \neq M_3$ , and without loss of generality suppose  $|M_1| = n_1 < |M_3| = n_3$ . Note that the string of symbols in  $M_1$  is exactly the same as the first  $n_1$  symbols in  $M_3$ . Now from Theorem 8.2, it is clear that  $S_{M_1}(n_1) = -1$  and  $S_{M_3}(n_1) \geq 0$ , as these two sums should be equal we have a contradiction.  $\square$

#### 8.3.4.3 Relating BMPT strings to BMPT parameterized full binary trees

Next, it is desirable to tie the strings in  $L_{BMPT}^*$  to BMPT models described as parameterized FBTs in Section 8.3.2. The key to relating the definition of BMPT models in terms of FBTs to  $L_{BMPT}^*$  comes from approaches in computer science to provide succinct codings of binary trees. The approach replaces a FBT by a sequence of symbols, where each symbol is a node in the tree. The idea is to visit each node once and only once in a sequence in such a way that the FBT can be reconstructed from its associated string. In graph theory, moving through a tree in a prescribed order is called a *tree traversal*. There are many codings of binary trees in the literature and some of them are developed as CFGs. To connect the traversal of a FBT to strings in  $L_{BMPT}$  a recursive data structure algorithm called a *preorder traversal* is needed. A preorder traversal visits all the nodes in the FBT recursively by following an order defined by:

1. Visit the root.
2. Traverse the left subtree, if it exists, with a preorder recursive call.
3. Traverse the right subtree, if it exists, with a preorder recursive call.

To illustrate a preorder traversal of a FBT, let's examine several FBTs that we have presented. Because we have not numbered the nodes, we will use in their place the parameters and categories that are associated with them. For example, the traversal of the left tree in Figure 8.1 would start with the root node. Then the node for the left child would be visited next. That node is a leaf, so the search shifts to the right subtree. The next unvisited node is the right child of the root, and it is visited as a root, and finally the two leaves under that node are visited, first the left child and then the right child. The result is  $DYgYN$ , and this is exactly the same as the BMPT string for that model shown in Section 8.3.4.1. A second example comes

from the pair tree in Figure 8.4. The preorder traversal (now coding internal nodes by  $\theta$  and leaves by  $C$ ) of the pair tree is  $\theta\theta CC\theta\theta CC\theta C$ . This string is identical to the one generated from  $G_{BMPT}$  in Section 8.3.4.1.

It turns out that Mäkinen (1998) provided a CFG for the preorder traversal of the class of FBTs.

**Definition 8.5** The FBT preorder traversal CFG is given by  $G_{PFBT} = (\{S\}, \{a, b\}, R, S)$ , where  $S$  is the only variable, the terminals are  $a, b$ , and  $R$  consists of two productions:  $P_1 : S \rightarrow b$  and  $P_2 : S \rightarrow aSS$ , and  $S$  is the start symbol.

It is obvious that the two grammars in Definitions 8.3 and 8.4 are isomorphs of each other in the sense that if  $S, a, b$  are replaced, respectively, by  $M, \theta$ , and  $C$ , the resulting quadruples are identical. Let  $L_{PFBT}$  refer to the words generated by  $G_{PFBT}$ , by the same symbol replacement, we have  $L_{PFBT} = L_{BMPT}$ . Of course, the intended interpretation, or semantics, for the two grammars are completely different, because for  $G_{PFBT}$  the symbol  $a$  is interpreted as an internal node and  $b$  as a leaf. Based on the fact that the two grammars are isomorphs, a theorem connecting the two definitions of BMPT models can be stated. The key is to substitute parameter or number assignments to each internal node  $a$  and a category for every leaf  $b$  in  $L_{PFBT}$ . Let the resulting class of strings be denoted by  $L_{PFBT}^*$ .

**Theorem 8.4**  $L_{PFBT}^* = L_{BMPT}^*$ .

*Proof* Because  $L_{PFBT} = L_{BMPT}$  and the rules for substituting parameters and category symbols are identical, it is clear that  $L_{PFBT}^* = L_{BMPT}^*$ .  $\square$

The result of Theorem 8.4 is that any BMPT model represented as a parameterized FBT can also be represented by a string in  $L_{BMPT}^*$  and vice versa. Both equivalent representations have uses in various contexts. For example, when describing the psychological rationale for a BMPT model, the FBT representation is by far the most natural to use. On the other hand, when transmitting BMPT models between users or as input to special statistical software packages, the string language is to be preferred because of its succinctness.

## 8.4 Statistical inference for MPT models

In this section some mathematical properties of MPT models relevant to conducting statistical inference will be discussed. The goal will be to examine a typical data structure in experiments relevant to MPT modeling, and then to capture mathematical properties of MPT models that are useful for inference with such data. In particular, it will be assumed that the reader has some knowledge of mathematical statistics and statistical inference.

Many cognitive experiments share a data structure that is assumed by most discrete state models of cognition. This structure arises from a standard experimental design where there are one or more experimental groups of participants (subjects,

observers), and within each experimental group, each participant is required to make an observable (manifest) categorical response to each of a set of stimulus probes (hereafter items). Examples of items are old and new words in a recognition memory experiment described in Example 8.1, choice of an object in a paired comparison choice experiment in Example 8.4, or pairs and singletons in a pair-clustering task in Example 8.7. Corresponding responses, respectively, would be “Yes” or “No” as to whether a word is recognized from an earlier study list, selecting the most preferred object from a pair, or how the recall performance on clusterable pairs is classified. For the discussion of statistical inference in this section, it is useful to formalize the above experimental situation. First, suppose there is a single class of items with  $M$  members, and the responses of  $N$  participants are classified in the category set  $\mathcal{C} = \{C_1, \dots, C_K\}$ . This data structure can be represented by a matrix of random variables  $\mathbf{X} = (X_{ij})_{N \times M}$ , where each  $X_{ij}$  is the categorical response of participant  $i$  to item  $j$ . Each of these random variables has space (set of possible values) in  $\mathcal{C}$ , and for convenience in some formula only the category subscript will be used to describe the response.

In the case of  $L$  classes of items, with  $M_l$  members in class  $l$ , and  $K_l$  possible response categories for items in class  $l$ , the relevant category system is denoted by  $\mathcal{C}_l = \{C_{1l}, \dots, C_{K_l l}\}$ . In this case, the data structure would be a three-way structure consisting of random variables  $X_{ij,l} \in \mathcal{C}_l$  indicating the response of participant  $i$  to item  $j$  in class  $l$ . Then the entire data structure for such an experiment consists of a collection of  $L$  random matrices  $\mathbf{X} = \{\mathbf{X}_1 = (X_{ij,1})_{N \times M_1} | 1 \leq l \leq L\}$ .

**Example 8.14** In a simple recognition memory experiment discussed in Example 8.1, a participant is exposed to a study list of word items, one at a time. Then they are presented with a test list of items consisting of  $M_1$  old studied words and  $M_2$  unstudied new items (foils), and for each item they are required to respond with a “Yes” or “No” response, indicating whether or not they think the item was in the studied list. Usually, the study list items and the test list items are interspersed and presented in a different random order for each participant. In this example we will assume that every participant receives the same set of study items and foils, although possibly in different orders.

To record data in this situation, a particular order of the studied items and of the foil items is selected. There are  $L = 2$  item classes, corresponding to old studied words and new foils, and the response categories for old and new words are the same, namely  $\mathcal{C}_l = \{1, 0\}$ , where 1 codes a “Yes” response and 0 a “No” response. Then the data structure for such an experiment consists of two random matrices  $\mathbf{X} = \{\mathbf{X}_1 = (X_{ij,1})_{N \times M_1}, \mathbf{X}_2 = (X_{ik,2})_{N \times M_2}\}$ . Of course, the actual data collected in the experiment are a realization  $\mathbf{x}$  of  $\mathbf{X}$ .

Statistical inference for a MPT model depends on the sampling assumptions that one makes about the data collected in a particular cognitive paradigm. It was established in Section 8.3 that the probability distribution for a particular participant responding to a particular item in the  $l$ th class of items is a member of the probability simplex  $\Lambda_{K_l}$ , where  $K_l$  is the number of observable categories for that

item. However, with  $N$  participants responding to the same  $M_l$  items, it is inevitably the case that there will be several observations that will fall into the same category. This is where it is important to characterize the sampling assumptions used to carry out statistical inference for the model. First, it is important to realize that the sampling assumptions of the data are not part of the formal specification of a MPT model. The same model can be analyzed in different ways depending on the sampling structure of the data as well as the selected type of inference.

The simplest sampling assumption about the  $NM_l$  participant-by-item random variables for the  $l$ th item class is that they constitute a random sample from a particular distribution in  $\Lambda_{K_l}$ . Recall that a random sample is a collection of independent and identically distributed (i.i.d.) random variables. In this section, the assumption that the data are the observations of a random sample will be explored first in Subsections 8.4.1 and 8.4.2, and then in Subsection 8.4.3 Bayesian methods to deal with more complex assumptions about the sampling distribution behind the data will be considered. In order to present results for MPT models under the i.i.d. assumption, it is helpful to first describe the most general MPT model known as the multinomial distribution.

### 8.4.1 The multinomial distribution

The multinomial distribution can be viewed as a generalization of the familiar binomial distribution, except each observation falls into one of a set of  $K \geq 2$  categories,  $\mathcal{C} = \{C_1, \dots, C_K\}$ , instead of just  $K = 2$  categories for the binomial distribution. The multinomial distribution is for the case where observations of  $N$  independent and identically distributed random variables are sorted into the  $K$  categories with some common distribution  $\mathbf{p} = (p_1, \dots, p_K) \in \Lambda_K$ . In such a case, one observes a category count vector,  $\mathbf{d} = (n_k)_{k=1}^K$ , where each  $n_k$  is a nonnegative integer and  $\sum_k n_k = N$ . Such a count vector is a realization of a random vector  $\mathbf{D} = (D_k)_{k=1}^K$ , with a sample space (set of possible values)

$$S_{\mathbf{D}} = \left\{ (n_k)_{k=1}^K \mid n_k \in \{0, 1, \dots\}, \sum_k n_k = N \right\}.$$

The multinomial distribution on  $K$  categories has two parameters,  $N \geq 1$  and  $\mathbf{p} \in \Lambda_K$ , written  $\text{Multi}(N, \mathbf{p})$ . then the probability distribution for  $\mathbf{D}$  is given by

$$\Pr[\mathbf{D} = (n_k) | N, \mathbf{p}] = \begin{cases} N! \prod_k \frac{p_k^{n_k}}{n_k!} & \text{if } (n_k) \in S_{\mathbf{D}} \\ 0 & \text{otherwise} \end{cases}. \quad (8.8)$$

$\text{Multi}(N, \mathbf{p})$  is easily derived by noting that each sequence of observations consistent with the count vector  $(n_k)$  has probability  $\prod_{k=1}^K p_k^{n_k}$ , and combinatorially there are

$$\binom{N}{n_1 \dots n_K} = \frac{N!}{\prod_k n_k!}$$

such sequences. It is straightforward to calculate the first few marginal moments of  $\text{Multi}(N, \mathbf{p})$ :

$$E(D_k) = Np_k, \text{Var}(D_k) = Np_k(1 - p_k), \text{Cov}(D_j, D_k) = -Np_jp_k.$$

The multinomial distribution can be viewed as an MMPT model with a single internal node (the root),  $K$  leaves, and where the internal node is assigned the parameter  $\mathbf{p}^{(K)}$ , and the leaves are assigned to the categories.

### 8.4.2 Parameterized multinomial models

It is a consequence of the computational rules of a BMPT model that if the category counts in a particular class of stimulus items are obtained as i.i.d. observations, then the MPT model becomes a specially parameterized multinomial model with probability distribution

$$Pr(\mathbf{D} = (n_1, \dots, n_k) | \Theta) = N! \prod_{k=1}^K \frac{p_k(\Theta)^{n_k}}{n_k!}, \quad (8.9)$$

where  $n_k$  is the number of counts in category  $C_k$ ,  $N = \sum n_k$ , where the  $p_k(\Theta)$  are given by (8.6) in the case of BMPT models.

In the case of multiple stimulus item classes, if responses to all items are independent, and within class responses are i.i.d., (8.9) just becomes a product of parameterized multinomial distributions, one for each item class. Such a data structure is called a product multinomial. For example, for the DHT model of Figure 8.1, the probability distribution for the parameterized product binomial model is given by

$$\begin{aligned} & Pr[< o_Y, o_N >, < Y, n_N > | (D, g)] \\ &= N! \frac{[D + (1 - D)g]^{n_1}}{n_1!} \frac{[(1 - D)(1 - g)]^{n_2}}{n_2!} \\ &\bullet M! \frac{[(1 - D)g]^{m_1}}{m_1!} \frac{[D + (1 - D)(1 - g)]^{m_2}}{m_2!}, \end{aligned}$$

where 1 and 2 code Old and New responses,  $N$  and  $M$  are the numbers of old and new items, and  $N = n_1 + n_2$ ,  $M = m_1 + m_2$ .

So far, (8.9) and the special case above for the model in Figure 8.1 can be viewed as providing a probability distribution over the possible count patterns given a particular fixed parameter  $\Theta$ . Statistical inference typically involves a fixed, observed data pattern, and what is needed is the so-called likelihood function that gives the probability of the fixed data pattern as a function of the parameter. The likelihood function  $L(\Theta | \mathbf{D})$  has exactly the same form as (8.9) except  $\mathbf{D}$  is fixed at some observation and  $\Theta$  varies in its domain, namely its space  $\Omega_\Theta$ . The likelihood function for a model is used in statistical inference in many different ways including classical likelihood based inference as well as Bayesian inference discussed in Section 8.4.3.

Under the sampling assumptions leading to (8.9), Hu and Batchelder (1994) provide classical methods for statistical inference based on the likelihood function for the entire class of globally identifiable BMPT models. Recall that global identifiability means that at most one parameter vector generates any particular probability distribution over the categories. Thus, in principle, knowledge of the category probabilities uniquely determines the (“best fitting”) underlying parameter vector  $\Theta$ . The first step in this approach is to derive a general likelihood function for a BMPT model under the i.i.d. assumption. Equation (8.9) gives the probability of various observations in terms of a fixed BMPT model parameter  $\Theta$ ; however, if the count vector has a fixed value and the model parameter varies, the same functional form provides the likelihood function. Using the form for  $p_k(\Theta)$  in (8.6), the general likelihood function for a BMPT model with a single tree becomes

$$L(\Theta|\mathbf{D}) = N! \prod_{k=1}^K \frac{\left( \sum_{j=1}^{I_k} c_{jk} \prod_{s=1}^S \theta_s^{a_{jk,s}} (1 - \theta_s)^{b_{jk,s}} \right)^{n_k}}{n_k!}. \quad (8.10)$$

An essential aspect of statistical inference is to use the data to learn something about the parameter. One main approach is to obtain a so-called maximum likelihood estimate (MLE) of the parameter denoted by  $\hat{\Theta} = (\hat{\theta}_1, \dots, \hat{\theta}_S)$ . The MLE gives the values of the component parameters that maximize the likelihood function. In the case that the model is globally identified, it is easy to show that  $\hat{\Theta}$  is unique. However, for complex BMPT models, it may be computationally difficult to globally maximize  $L(\Theta|\mathbf{D})$ . However, it turns out that BMPT models have a special structure that makes it easy to find the MLE. In particular, if one knew all the branch frequencies  $n_{jk}$ ,  $1 \leq j \leq I_k$ , then the MLE of each  $\theta_s$  can be obtained in closed form by

$$\hat{\theta}_s = \frac{\sum_{k=1}^K \sum_{j=1}^{I_k} a_{jk,s} n_{jk}}{\sum_{k=1}^K \sum_{j=1}^{I_k} (a_{jk,s} + b_{jk,s}) n_{jk}}. \quad (8.11)$$

Informally, this result occurs because the formula represents the number of times a node assigned to  $\theta_s$  is reached divided into the number of times the left link, associated with  $\theta_s$ , is taken rather than the right link, associated with  $(1 - \theta_s)$ . In other words, it is an extension of the fact that the MLE of the category probabilities in  $\text{Multi}(N, \mathbf{p})$  is easily shown to be  $\hat{p}_k = n_k/N$ . Of course, one does not know the branch frequencies for a BMPT model unless  $\forall k, I_k = 1$ ; however, they are subject to  $n_k = \sum_{j=1}^{I_k} n_{jk}$ , because the category counts  $n_k$  are observable. While (8.11) only pertains to a BMPT model with a single tree, the approach is easily generalized to the case with several classes of items, each with its own category system and associated MMPT.

The key to the approach in Hu and Batchelder (1994) is to regard the branch frequencies as missing data and employ the so-called *Expectation-Maximization*

*algorithm* (EM). The EM algorithm is an iterative approach to obtaining the MLE that is used when it is easy to obtain the MLE if certain missing data are known. The M-step of the algorithm calculates the MLE conditional on conjectured values of the missing data, and the E-step recalculates values of the missing data given the conditional MLE. There is considerable theory about this approach, e.g., Dempster *et al.* (1977), showing that given a particular starting value (a conjectured MLE) successive iterations of the E-step and M-step will under most conditions yield a local maximum of the likelihood function. There is a software package in Hu and Phillips (1999) based on the EM algorithm for BMPT models that enables a researcher to design an MPT model graphically, load in the data, and obtain MLEs, confidence intervals, goodness-of-fit measures, do hypothesis testing within and between participant groups, and run Monte Carlo simulations.

### 8.4.3 Hierarchical Bayesian inference for MPT models

In the previous two subsections the sampling assumption that observations within a class of items are i.i.d. was made. The i.i.d. assumption is very strong because it assumes not only independence (or exchangeability in the Bayesian sense), but it entails the assumptions that there are no individual differences (homogeneity) in participants and in the items, i.e., that all participant by item responses are generated by the model with the same parameter. Typically, participants are drawn from college classes where grades from A to F are given for different performance due to heterogeneous ability levels, so it seems questionable to assume that the performance of these same participants in a cognitive experiment would derive from homogeneous cognitive abilities. The assumption of homogeneity in items is arguably more acceptable because in a typical cognitive experiment, the items are equated on known sources of heterogeneity. For example, items in a simple recognition memory experiment discussed in Example 8.1 typically are equated on such factors as concreteness, imagery, word length, and frequency of occurrence in written text, etc. In any event, in the seven or so decades since cognitive models of any type have been used to analyze experimental data, the assumption of i.i.d. within a class of items has been the norm.

It is true that cognitive modeling papers have often acknowledged concerns about the assumption of i.i.d.; however, only recently, in the last decade or so, have modelers learned how to relax the assumptions of homogeneity in participants and/or items. The key to this development is to specify the model's hierarchically. Detailed references to this approach will be given in Section 8.5; however, the basic idea of a hierarchical BMPT model is as follows. Suppose  $\Theta \in \Omega_\Theta$  is the parameter for a BMPT model. For the i.i.d. approach discussed earlier, one assumes that each participant's data are generated from the model with the same value of the parameter. The hierarchical assumption is that each participant's value of  $\Theta$  is sampled i.i.d. from some hierarchical distribution  $g(\Theta|\Delta)$ , where  $\Delta = (\delta_1, \dots, \delta_T) \in \Omega_\Delta$ . A frequent simplification of  $g(\Theta|\Delta)$  is to assume that each component parameter in  $\Theta$  has its own hierarchical distribution, for example  $g(\Theta|\Delta) = \prod_{s=1}^S g(\theta_s|\Delta)$ . Of

course, to specify a hierarchical BMPT model, one has to specify the base model in the usual way and then add a specification of the hierarchical distribution. This has been facilitated by Bayesian hierarchical inference, where the model parameters for participants or items are assumed to be a sample from a hierarchical distribution whose parameters themselves have known numerical distributions. Chapter 9 of this volume is an entire chapter on hierarchical Bayesian inference for cognitive models, and in addition there is a whole issue of the *Journal of Mathematical Psychology* on this subject (Lee, 2011).

## 8.5 Relevant literature

The organization of this section will be to suggest references to further knowledge in the material discussed in this chapter. The suggested references will be organized in sequence to correspond to the sections in this chapter.

Concerning Section 8.1, Macmillan and Creelman (2004) is a good reference for different signal detection models and experimental designs; however, it does not focus very much on threshold models. Luce (1963) is an example of a successful discrete state model of signal detection. Much recent theoretical controversy centers on comparing the discrete state and continuous state signal detection model as they are applied to recognition memory studies. Batchelder and Alexander (2013), Bröder *et al.* (2013), and Klauer and Kellen (2011) are among a large number of articles in this vein, and they will have good references to some of the others.

Concerning Section 8.2, for further background in Markov chain theory a good source is Ross (2014), and special treatment of the mathematics of HMC models is in Wickens (1982). Hidden Markov chain models were popular in concept identification and human memory experimental paradigms. For substantive models in the area of concept identification, Millward and Wickens (1974) is a good reference, and for memory paradigms, HMC models are reviewed in Greeno and Bjork (1973) and Restle and Greeno (1970). A comparison between the linear operator model and the all-or-none model of simple learning is in Batchelder (1975). HMC models ceased to be of major theoretical importance in learning and memory sometime in the late 1970s, when a variety of more elaborate and detailed cognitive models of learning, memory, and choice were introduced, such as global memory models and later connectionist models. Nevertheless, HMC models have continued to serve as psychometric models for psychological measurement. For example, they have been applied to early development and to aging, where standardized paradigms provide learning sequences for different age groups that can be used to estimate age-related trends in latent storage and retrieval processes (e.g. Alexander *et al.*, 2016; Batchelder *et al.*, 1997). The development of an HMC model for paired comparison scaling is in Batchelder *et al.* (2009).

Concerning Section 8.3, discrete state models in the area of statistical genetics are discussed in Elandt-Johnson (1971) and Weir (1990). The MPT

family for cognitive modeling was developed and formalized, including its classical statistical inference, in Batchelder and Riefer (1986), Riefer and Batchelder (1988), Hu and Batchelder (1994), Knapp and Batchelder (2004), and Wu *et al.* (2010). The treatment of MPT models as strings in a context-free language is in Purdy and Batchelder (2009). More on the mathematical structure of languages and their relationship to automata is covered in Partee, Meulen, and Wall (1990).

Smith and Batchelder (2008) provide nonparametric methods to detect individual differences in participants and/or items for categorical data. Several papers have provided Bayesian hierarchical inference for BMPT models that can handle heterogeneity in participants or items, e.g., Klauer (2010), Smith and Batchelder (2010), and Matzke *et al.* (2015). It is expected that future applications of MPT models will increasingly utilize Bayesian inference to handle heterogeneity.

Batchelder and Riefer (1999) describe 30 MPT models for different experimental paradigms, and a decade later, Erdfelder *et al.* (2009), in the lead article to a special issue of *Zeitschrift für Psychologie* on MPT modeling, describe 70 MPT models and variants in over 20 research areas as well as references to available software to analyze MPT models. The pair-clustering model is presented in Batchelder and Riefer (1986), models of source monitoring are discussed in Batchelder and Batchelder (2008), and various special cases of the source monitoring model in Figure 8.4 have been used in a variety of research settings, including illusory correlations, face recognition, social categorization, and a variety of clinical settings involving special populations of participants; see Erdfelder *et al.* (2009) for a complete list. The use of MPT models as measurement tools, especially in clinical settings, is discussed in Batchelder (1998, 2009), Batchelder and Riefer (2007), and Riefer *et al.* (2002).

## References

- Alexander, G. E., Satalich, T. A., Shankle, W. R. and Batchelder, W. H. (2016). A cognitive psychometric model for the psychodiagnostic assessment of memory related deficits. *Psychological Assessment*, **29**, 279–293.
- Batchelder, W. H. (1975). Individual differences and the all-or-none vs incremental learning controversy. *Journal of Mathematical Psychology*, **12**, 53–74.
- Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, **10**, 331.
- Batchelder, W. H. (2009). Cognitive psychometrics: using multinomial processing tree models as measurement tools. In: Embretson, S. E. (ed.), *Measuring Psychological Constructs: Advances in Model Based Measurement*. Washington, DC: American Psychological Association Books.
- Batchelder, W. H. and Alexander, G. E. (2013). Discrete-state models: comment on Pazzaglia, Dube, and Rotello Psychological Bulletin, **139**, 1204–1212.
- Batchelder, W. H. and Batchelder, E. (2008). Metacognitive guessing strategies in source monitoring. In: Dunlosky, J. and Bjork, R. A. (eds), *Handbook of Metamemory and Memory*, New York, NY: psychology press, pp. 211–244.

- Batchelder, W. H. and Riefer, D. M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical and Statistical Psychology*, **39**, 129–149.
- Batchelder, W. H. and Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, **6**, 57–86.
- Batchelder, W. H. and Riefer, D. M. (2007). Using multinomial processing tree models to measure cognitive deficits in clinical populations. In: Neufeld R.W.J. (ed.), *Advances in Clinical Cognitive Science: Formal Modeling of Processes and Symptoms*. Washington, DC: APA, pp. 19–50.
- Batchelder, W. H., Chosak-Reiter, J., Shankle, W. R. and Dick, M. B. (1997). A multinomial modeling analysis of memory deficits in Alzheimer's disease and vascular dementia. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, **52**, 206–215.
- Batchelder, W. H., Hu-X. and Smith, J. B. (2009). Multinomial processing tree models for discrete choice. *Zeitschrift für Psychologie/Journal of Psychology*, **217**, 149–158.
- Bernstein, F. (1925). Zusammenfassende Betrachtungen über die erblichen Blutstrukturen des Menschen. *Molecular and General Genetics MGG*, **37**, 237–270.
- Bower, G. H. (1961). Application of a model to paired-associate learning. *Psychometrika*, **26**, 255–280.
- Bröder, A., Kellen D.-Schütz J. and Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*, **21**, 916–944.
- Bush, R. R. and Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, **58**, 413.
- Chechile, R. A. (1998). A new method for estimating model parameters for multinomial data. *Journal of Mathematical Psychology*, **42**, 432–471.
- Chomsky, N. (1963). Formal properties of grammars. In: Luce, R. D., Bush-R. R. and Galanter, E. (eds), *Handbook of Mathematical Psychology: Vol. II*. New York, NY: Wiley.
- Dempster, A. P., Laird-N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.
- Elandt-Johnson, R. C. (1971). *Probability Models and Statistical Methods in Genetics*. New York, NY: John Wiley & Sons, Inc.
- Erdfelder, E., Auer, T.-S., Hilbig, B. E., et al. (2009). Multinomial processing tree models. *Zeitschrift für Psychologie/Journal of Psychology*, **217**, 108–124.
- Greeno, J. G. and Bjork, R. A. (1973). Mathematical learning theory and the New “Mental Forestry”. *Annual Review of Psychology*, **24**, 81–116.
- Hu, X. and Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, **59**, 21–47.
- Hu, X. and Phillips, G. A. (1999). GPT. EXE: A powerful tool for the visualization and analysis of general processing tree models. *Behavior Research Methods, Instruments, & Computers*, **31**, 220–234.
- Klauer, K. C. (2010). Hierarchical multinomial processing tree models: a latent-trait approach. *Psychometrika*, **75**, 70–98.
- Klauer, K. C. and Kellen, D. (2011). The flexibility of models of recognition memory: an analysis by the minimum-description length principle. *Journal of Mathematical Psychology*, **55**, 430–450.

- Knapp, B. R. and Batchelder, W. H. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *Journal of Mathematical Psychology*, **48**, 215–229.
- Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, **55**, 1–7.
- Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York, NY: Wiley.
- Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, **70**, 61.
- Macmillan, N. A. and Creelman, C. D. (2004). *Detection Theory: A User's Guide*. New York, NY: Psychology Press.
- Mäkinen, E. (1998). Binary tree code words as context-free languages. *The Computer Journal*, **41**, 422–424.
- Matzke, D., Dolan C. V., Batchelder, W. H. and Wagenmakers, E-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, **80**, 205–235.
- Millward, R. B. and Wickens, T. D. (1974). Concept identification models. In: Krantz, D. H., Atkinson, R. C., Luce, R. D. and Suppes, P. (eds), *Developments in Mathematical Psychology (Vol. 1): Learning, Memory, Thinking*. San Francisco, CA: Freeman.
- Partee, B. H., Meulen, A. T. and Wall, R. (1990). *Mathematical Models in Linguistics*. Boston, MA: Springer.
- Purdy, B. P. and Batchelder, W. H. (2009). A context-free language for binary multinomial processing tree models. *Journal of Mathematical Psychology*, **53**, 547–561.
- Restle, F. and Greeno, J. G. (1970). *Introduction to Mathematical Psychology*. Reading, MA: Addison-Wesley.
- Riefer, D. M. and Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, **95**, 318.
- Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D. and Manifold, V. (2002). Cognitive psychometrics: assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, **14**, 184.
- Ross, S. M. (2014). *Introduction to Probability Models*. New York, NY: Academic Press.
- Smith, J. B. and Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review*, **15**, 713–731.
- Smith, J. B. and Batchelder, W. H. (2010). Beta-MPT: multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, **54**, 167–183.
- Weir, B. S. (1990). *Genetic Data Analysis. Methods for Discrete Population Genetic data*. Sunderland, MA: Sinauer Associates, Inc. Publishers.
- Wickens, T. D. (1982). *Models for Behavior: Stochastic Processes in Psychology*. San Francisco, CA: WH Freeman.
- Wu, H., Myung J. I. and Batchelder, W. H. (2010). On the minimum description length complexity of multinomial processing tree models. *Journal of Mathematical Psychology*, **54**, 291–303.

# 9 Bayesian hierarchical models of cognition

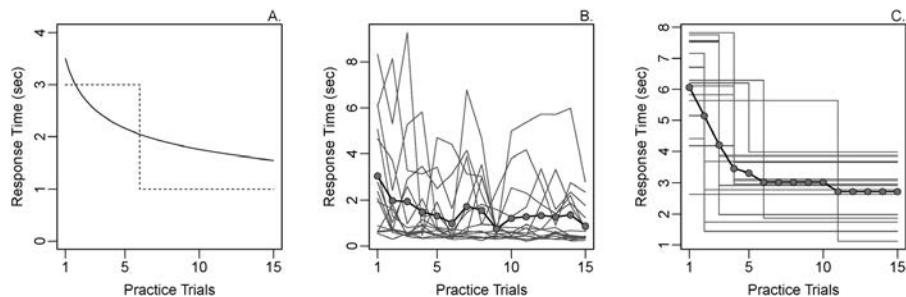
Jeffrey N. Rouder, Richard D. Morey, and Michael S. Pratte

9.1 Introduction: the need for hierarchical models	504
9.2 Bayesian basics	508
9.2.1 Frequentist and Bayesian conceptualizations of probability	508
9.2.2 Bayes rule	510
9.2.3 Sampling: an approach to Bayesian analysis with more than one parameter	514
9.3 Bayesian hierarchical models are simple and natural	516
9.4 Comparing hierarchical models	524
9.5 Hierarchical models for assessing subliminality	527
9.5.1 A hierarchical model	529
9.5.2 Extending the hierarchical model	533
9.6 Hierarchical models for signal-detection experiments	537
9.6.1 Consequences of aggregation in memory experiments	538
9.6.2 A hierarchical dual-process model of recognition memory	540
9.6.3 Applications of the hierarchical memory model	542
9.7 Concluding remarks	544
References	546

## 9.1 Introduction: the need for hierarchical models

Those of us who study human cognition have no easy task. We try to understand how people functionally represent and process information in performing cognitive activities such as vision, perception, memory, language, and decision making. Fortunately, experimental psychology has a rich theoretical tradition, and there is no shortage of insightful theoretical proposals. Also, it has a rich experimental tradition, with a multitude of experimental techniques for isolating purported processes. What it lacks, however, is a rich statistical tradition to link theory to data. At the heart of the field is the difficult task of trying to use data from experiments to inform theory, that is, to understand accurately the relationships within the data and how they provide evidence for or against various theoretical positions.

The difficulty in linking data to theory can be seen in a classic example from Estes (1956). Estes considered two different theories of learning: one in which



**Figure 9.1** Estes' (1956) example of the difficulty of linking learning-curve data to learning theories. (A). Predictions: the solid and dashed lines show predictions from the gradual-decrease and all-at-once models of learning, respectively . (B). Data from Reder and Ritter (1992). The gray lines show the times for 15 individuals as a function of practice; the circles are means across individuals, and these means decrease gradually with practice. (C). Hypothetical noise-free data from the all-at-once learning model. Individuals' data are shown as thin gray lines. The mean, shown with points, nonetheless decreases gradually. This panel shows that the mean over individuals does not reflect the structure of any of the individuals.

learning was gradual, and another where learning happened all at once. These two accounts are shown in Figure 9.1A. Because these accounts are so different, adjudicating between them should be trivial: one simply examines the data for either a step function or a gradual change. Yet, in many cases, this task is surprisingly difficult. To see this difficulty, consider the data of Reder and Ritter (1992), who studied the speed up in response times from repeated practice of a mathematics tasks. The data are shown in Figure 9.1B, and the gray lines show the data from individuals. These individual data are highly variable, making it impossible to spot trends. A first-order approach is to simply take the means across individuals at different levels of practice, and these means (points) decrease gradually, seemingly providing support for the gradual theory of learning. Estes, however, noted that this pattern does not necessarily imply that learning is gradual. Instead, learning might be all-at-once, but the time at which different individuals transition may be different. Figure 9.1C shows an example; for demonstration purposes, hypothetical data are shown without noise. If data are generated from the all-at-once model and there is variation in this transition time, then the mean will reflect the proportion of individuals in the unlearned state at a given level of practice. This proportion may decrease gradually, and consequently, the mean may decrease gradually even if every participant has a sharp transition from an unlearned state to a learned one. It is difficult to know whether the pattern of the means reflects a signature of cognitive processing or a signature of between-individual variation.

There are three critical elements in Estes' example: first, individuals' data are highly noisy, and this degree of noise necessitates combining data across people. Second, there is variability across individuals. For example, in the all-at-once

model, people differ in their transition times. Finally, the theories themselves are nonlinear,<sup>1</sup> and the all-at-once model in particular has a large degree of nonlinearity. It is the combination of these three factors – substantial variability within and across individuals that is analyzed with nonlinear models – that makes linking data to theory difficult. Unfortunately, the three elements that led to the difficulties in Estes' example are nearly ubiquitous in experimental psychology. Often, data are too noisy to draw conclusions from consideration of single individuals; there is substantial variability across participants; and realistic models of cognition are nonlinear. Note that the problem of nuisance variation is not limited to individuals. In memory and language studies, for instance, there is nuisance variation across items. For instance, in the learning example, it is reasonable to expect that if the all-at-once model held, the time to transition across different problems (items) would vary as well.

Several psychologists have noted that drawing conclusions from aggregated data may be tenuous. Estes' example in learning has been expanded upon by Haider and Frensch (2002), Heathcote *et al.* (2000), Myung *et al.* (2000), and Rickard (2004). The worry about aggregation over individuals has also been expressed in the context of multidimensional scaling (Ashby *et al.*, 1994), and the worry about aggregating over both individuals and items has been expressed in linguistics (Baayen *et al.*, 2002) and recognition memory (Rouder *et al.*, 2007b; Pratte *et al.*, 2010). Although the dangers of aggregation are widely known, researchers still routinely aggregate. For example, in studies of recognition memory, it is routine to aggregate data across individuals and items before fitting models. Even experienced researchers who fit sophisticated models to individuals routinely aggregate across some source of nuisance variation, typically items. The reason that researchers aggregate is simple – they do not know what else to do. Consider recognition memory tasks, where aggregation across items or individuals is seemingly necessary to form hit and false alarm rates. Without aggregation, the data for each item-by-individual combination is an unreplicated, dichotomous event. Our experience is that researchers would prefer to avoid aggregating data if alternatives are available.

In this chapter we present such an alternative: hierarchical modeling. In a hierarchical model, variability from the process of interest, as well as from nuisance sources such as from individuals and from items, are modeled simultaneously. The input to these models is the raw, unaggregated data, and the outputs are process-parameter estimates across individuals and items. In this regard, not only can the behavior of these process estimates be studied across conditions, but across individuals and items as well, and this later activity provides a process-model informed study of individual (and item) differences. Hence, hierarchical models turn a problem, how to account for nuisance variation that cloud our view of process, into a

<sup>1</sup> Linear models are those where the expected value of the data is a linear function of the parameters (Kutner *et al.*, 2004). Examples include ANOVA and regression. Nonlinear models violate this basic tenet: the expected value of the data cannot be expressed as a linear function of parameters.

strength. Hierarchical models provide both a clearer view of process and a means of exploring how these processes vary across populations of individuals or items. Not surprisingly, hierarchical linear models, models that extend ANOVA and regression to account for multiple sources of variance, are common in many areas of psychology, as well as across the social sciences (Raudenbush and Bryk, 2002).

Although hierarchical linear models are suitable in several domains, they rarely make good models of psychological process. Instead, models that account for psychological processes are typically nonlinear. The appropriate extensions for these cases are *hierarchical nonlinear models*. It is difficult, however, to analyze hierarchical nonlinear models in conventional frameworks. As a result, the field has been moving toward Bayesian hierarchical models because hierarchical models, including hierarchical nonlinear models, are far more conveniently and straightforwardly analyzed in the Bayesian framework than in conventional ones. It is for this reason that there has been much recent development of Bayesian hierarchical models in the mathematical psychology community, the psychological community most concerned with developing models of psychological process. Recent examples of applications in psychologically substantive domains include Anders and Batchelder (2012), Averell and Heathcote (2011), Karabatsos and Batchelder (2003), Kemp *et al.* (2007), Lee (2006), Farrell and Ludwig (2008), Merkle *et al.* (2011), Rouder *et al.* (2004, 2008), Vandekerckhove *et al.* (2010), and Zeigenfuse and Lee (2010). Tutorial articles and chapters covering hierarchical cognitive process models are becoming numerous as well (e.g., Busemeyer and Diederich, 2009; Kruschke, 2011; Lee and Wagenmakers, 2013; Rouder and Lu, 2005; Shiffrin *et al.*, 2008), and there is a special issue of the *Journal of Mathematical Psychology* (January 2011, Vol 55:1) devoted to the topic.

In the next section, we cover the basics of Bayesian probability. Included is a comparison of the basic tenets of frequentist and Bayesian probability, examples of using data to update prior beliefs, and an introduction to Markov chain Monte Carlo sampling. In Section 9.3, we show that the specification and analysis of hierarchical models is simple and natural with the Bayesian approach, and in Section 9.4 we provide a brief discussion of model comparison. Section 9.5 comprises our first example, and it is in the assessment of subliminal priming. Subliminal priming occurs when an undetectable stimulus nonetheless affects subsequent behavior. The methodological difficulty in establishing subliminal priming is proving that a set of participants cannot detect a stimulus at levels above chance. We show how previous approaches are woefully inadequate and demonstrate how a hierarchical approach provides a possible solution. We provide a second example of hierarchical modeling in Section 9.6. The example is from recognition memory, and shows how the estimation of parameters in Yonelinas' dual process model (Yonelinas, 1994) may be contaminated by aggregation bias. We develop a hierarchical model for uncontaminated assessment of the number of processes mediating recognition memory. Finally, in Section 9.7 we provide some advice on

choosing computer packages and receiving training to perform Bayesian hierarchical modeling.

## 9.2 Bayesian basics

In this paper we adopt a Bayesian rather than a conventional frequentist framework for analysis. One reason is pragmatic – the development of Bayesian hierarchical models is straightforward. Analysis of all Bayesian models, whether hierarchical or not, follows a common path. Bayesian techniques transfer seamlessly across different domains and models, providing a compact, unified approach to analysis. Because the Bayesian approach is unified, models that might be intractable in frequentist approaches become feasible with the Bayesian approach. The second reason we advocate Bayesian analysis is on philosophical grounds. The foundational tenets of Bayesian probability are clear, simple, appealing, and intellectually rigorous. In this section we review frequentist and Bayesian conceptualizations of probability. More detailed presentations may be found in Bayesian textbooks such as Gelman *et al.* (2007) and Jackman (2009).

### 9.2.1 Frequentist and Bayesian conceptualizations of probability

The frequentist conceptualization of probability is grounded in the Law of Large Numbers. Consider an event that may happen or not, and let  $Y$  be the number of occurrences out of  $N$  opportunities. The probability of an event is defined as the proportion when the number of opportunities is arbitrarily large; i.e.,

$$p = \lim_{N \rightarrow \infty} \frac{Y}{N}.$$

In this formulation, we may think of the probability as a physical property of the event. Consider, for example, the probability that a given coin results in a *heads* when flipped. This probability may be thought of as a physical property much like the coin's weight or chemical composition. And much like weight and chemical composition, the probability has an objective truth value even if we cannot measure it to arbitrary precision.

In both frequentist and Bayesian paradigms, useful models contain unknown parameters that must be estimated from data. For instance, if a participant performs  $N$  experimental trials on a task, we might model the resultant frequency of correct performance,  $Y$ , as a binomial random variable:

$$Y \sim \text{Binomial}(p, N),$$

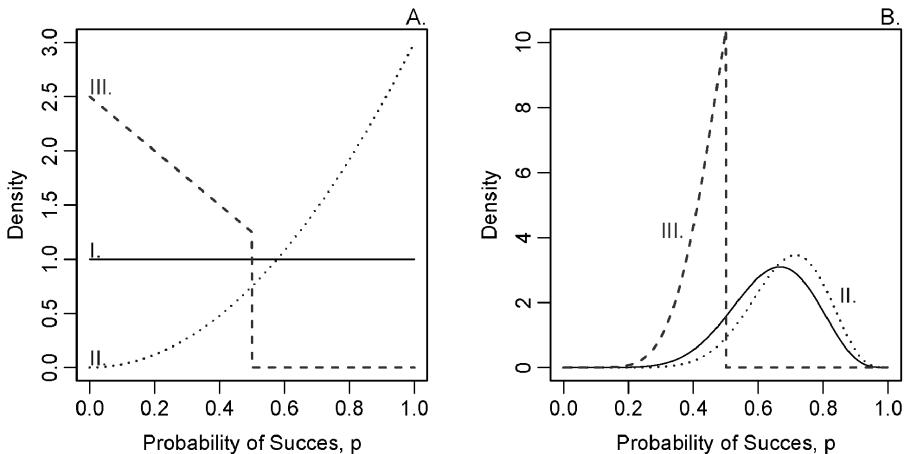
where  $p$  serves as a parameter and denotes the probability of a correct response on a trial. Another simple, ubiquitous model is the normal. For example,  $Y$  might denote the mean response time of a participant in a task and be modeled as

$$Y \sim \text{Normal}(\mu, \sigma^2),$$

where  $\mu$  and  $\sigma^2$  serve as free parameters that denote the mean and variance of the distribution of people's mean response times. Although it is well known that response times are not normals (Luce, 1986), the normal is a reasonable model of the distribution of mean RT across people. Consequently, the normal model is often useful for analyzing changes in mean RT as a function of experimental conditions or other covariates.

In the frequentist conceptualization, parameters are unknown fixed values which, in turn, are estimated from data. Because frequentist probability stresses the large-sample limit, the approach does not provide strict guidance on estimating parameters in finite samples sizes. Consequently, there are multiple approaches to estimation including finding estimates that maximize the likelihood (ML) or, alternatively, finding estimates that minimize squared error between predicted and observed data points (LS). These methods are not equivalent, and they may lead to different estimates in certain circumstances. For example, the ML estimator of  $\sigma^2$  in the normal model is  $\hat{\sigma}^2 = \sum(y_i - \bar{y})^2/N$ , while the LS estimator is  $\hat{\sigma}^2 = \sum(y_i - \bar{y})^2/(N - 1)$ . For frequentists, a minimal measure of acceptability of an estimator is its large-sample behavior. Principled estimators are *consistent*: they converge to true values in the large-sample limit. Both the ML and LS estimators of  $\sigma^2$  are consistent because they converge to the true value as  $N \rightarrow \infty$ .

The Bayesian conceptualization of probability differs substantially from the frequentist one. Probabilities are statements of subjective belief held by observers about the occurrences of events. In the Bayesian formulation, probabilities describe the analyst's belief rather than a physical property of the system under study. Analysts may express their beliefs compactly as distributions. Figure 9.2A shows the beliefs of three analysts about a certain coin, or more specifically about  $p$ , the probability that a flip of a coin will result in heads rather than tails. Analyst I believes that all values of  $p$  are equally likely. This belief is shown by the solid flat line. Analyst II believes heads is more likely than tails, and this belief is shown by the dotted line. Analyst III believes not only that tails are more likely than heads, but that there is no chance whatsoever that the coin favors heads. This belief is shown by the dashed line. These beliefs are called *prior* beliefs, because they are expressed before observing the data. After expressing these prior beliefs, the three analysts together flip the coin repeatedly and observe 8 heads in 12 flips. *The key tenet of Bayesian probability is that beliefs may be updated rationally in light of data.* To do so, one applies Bayes Rule, which is discussed subsequently. The rationally updated belief distributions, called the *posterior* beliefs, are shown in Figure 9.2B. There are three posterior distributions, one for each analyst. There are a few noteworthy points: first, the beliefs of all three analysts have been narrowed by the data; in particular, for Analyst I, the beliefs have updated from a flat distribution to one that is centered near the proportion of heads and with narrowed variance. Second, even though the prior beliefs of Analysts I and Analyst II diverged markedly, the posterior beliefs are quite similar. Third, Analyst III had ruled out certain values, all those for  $p > .50$  *a priori*. Indeed, because these have been ruled out, no result can make them probable, and the posterior has no density for  $p > .50$ .



**Figure 9.2** Prior and posterior beliefs from three analysts for the probability of heads. **(A).** Prior beliefs. Analyst I believes that all outcomes are equally plausible; Analyst II believes that heads are more likely than tails; and Analyst III not only believes that tails are more likely than heads, but that the coin has no chance of favoring heads. **(B).** The updated posterior beliefs after observing 8 heads and 4 tails.

In summary, Bayesian probability does not prescribe what beliefs analysts should hold. Instead, the emphasis is on how these beliefs should be updated in light of data. Posterior beliefs are still subjective even though they reflect data. For Bayesians, probabilities remain a construct of the observer rather than an objective property of the system, and this property holds regardless of how much data has been collected. However, because of the strong constraints imposed by Bayes Rule and their relationship to rational learning, Bayesian statistics offers a compelling, unified method for learning from data.

### 9.2.2 Bayes rule

The goal of Bayesian analysis is to update beliefs rationally with Bayes Rule. Consider again the model of  $Y$ , the number of heads out of  $N$  coin flips,  $Y \sim \text{Binomial}(p, N)$ , where  $p$  is a free parameter. Bayes Rule in this case is

$$\pi(p|Y) = \frac{\Pr(Y|p)}{\Pr(Y)}\pi(p). \quad (9.1)$$

The term  $\pi(p|Y)$  is the posterior distribution of beliefs, that is, beliefs about the parameter conditional on the data. The term  $\pi(p)$  is the prior distribution of beliefs. Three examples of prior and posterior beliefs are provided in Figure 9.2A and 9.2B, respectively. The term  $\Pr(Y|p)$  is the likelihood function and is derived from the model. For the binomial model,  $\Pr(Y|p) = \binom{N}{Y} p^Y (1-p)^{N-Y}$ . The remaining term  $\Pr(Y)$ , the probability of the data, may be re-expressed by the Law of Total

Probability as

$$Pr(Y) = \int_0^1 Pr(Y|p)\pi(p) dp.$$

Fortunately, it is unnecessary to compute  $Pr(Y)$  to express posterior beliefs. The distribution of posterior beliefs  $\pi(p|Y)$  must be proper, that is, the area under the curve must be 1.0. The term  $Pr(Y)$  is a normalizing constant on  $\pi(p|Y)$  such that  $\int_0^1 \pi(p|Y) dp = 1$ . Often, the expression for this normalizing constant is obvious from the form of  $Pr(Y|p)\pi(p)$  and need not be explicitly computed.

Let's use Bayes Rule to express the posterior beliefs for the prior in Figure 9.2A for Analyst II. This prior is  $\pi(p) = K_0 p(1-p)^3$ , where  $K_0$  is a constant that assures the prior integrates to 1.0. The data are 8 heads in 12 flips, and the likelihood  $Pr(Y|p)$  is  $\binom{12}{8} p^8 (1-p)^4$ . Multiplying the prior and likelihood yields the following:

$$\pi(p|Y=8) = Kp^9(1-p)^7,$$

where  $K$  is a constant of proportionality chosen such that  $\int_0^1 \pi(p|Y=8) dp = 1$ . The dashed line in Figure 9.2B is the evaluation of  $\pi(p|Y=8)$  for all values of  $p$ .

Bayes Rule is completely general, and may be extended to models with more than one parameter as follows. Let  $\mathbf{Y}$  denote a vector of data which is assumed to be generated by some model  $\mathcal{M}$  with a vector of parameters denoted by  $\boldsymbol{\theta}$ , i.e.,  $\mathbf{Y} \sim \mathcal{M}(\boldsymbol{\theta})$ . Then

$$\pi(\boldsymbol{\theta}|\mathbf{Y}) \propto Pr(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

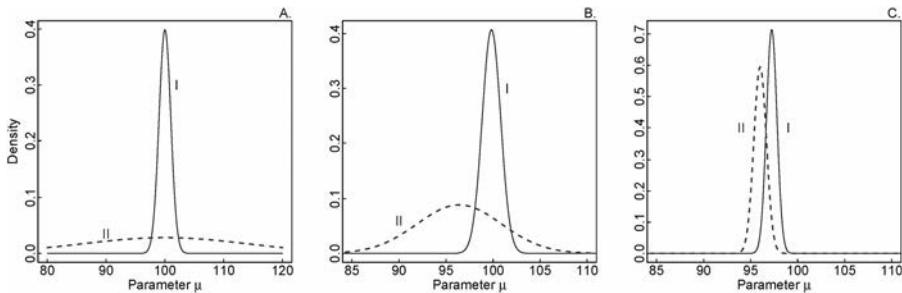
Once again,  $Pr(\mathbf{Y}|\boldsymbol{\theta})$  is the likelihood function in this context, and Bayes Rule may succinctly be stated as, “The posterior is proportional to the product of the likelihood and prior.” Bayesian updating, in contrast to frequentist parameter estimation, is highly constrained. There is only one Bayes Rule, and it may be followed consistently without exception. One of the appeals of Bayesian updating is its conceptual simplicity and universal applicability.

The binomial model is useful for modeling dichotomous outcomes such as accuracy on a given trial. It is often useful to model continuous data as normally distributed. For example, suppose we wished to know the effects of “Smarties,” a brand of candy, on IQ. Certain children have been known to implore their parents for Smarties with the claim that it assuredly makes them smarter. Let's assume for argument's sake that we have fed Smarties to a randomly selected group of school children, and then measured their IQ. Let  $\mathbf{Y} = (Y_1, \dots, Y_n)$  be a vector that denotes the IQ of the children fed Smarties. We model these IQ scores as

$$Y_i \stackrel{iid}{\sim} \text{Normal}(\mu, \sigma^2),$$

where *iid* indicates that each observation is *independent* and *identically distributed*.

The goal is to derive posterior beliefs about  $\mu$  and  $\sigma^2$  given prior beliefs and the data themselves. For now, we focus on  $\mu$  and, for simplicity, assume that  $\sigma^2$



**Figure 9.3** Prior and posterior beliefs on  $\mu$ , the center of a normal distribution. **(A)**. Prior beliefs of two analysts. **(B)**. Posterior beliefs conditional on a sample mean of  $\bar{Y} = 95$  and a small sample size of  $N = 10$ . **(C)**. Posterior beliefs conditional on a sample mean of  $\bar{Y} = 95$  and a larger sample size of  $N = 100$ .

is known to equal the population variance of IQ scores,  $\sigma^2 = 15^2 = 225$ . In Section 9.2.3, we relax this assumption, and discuss how to update beliefs on multiple parameters simultaneously.

An application of Bayes Rule to update beliefs about  $\mu$  yields

$$\pi(\mu|\mathbf{Y}) \propto L(\mu, \mathbf{Y})\pi(\mu),$$

where  $\mathbf{Y}$  is the vector of observations and  $L$  is the likelihood function of  $\mu$ . The likelihood for a sequence of independent and identically normally distributed observations is

$$\begin{aligned} L(\mu, \mathbf{Y}) &= f(Y_1; \mu, \sigma^2) \times f(Y_2; \mu, \sigma^2) \times \cdots \times f(Y_n; \mu, \sigma^2) \\ &= \prod_i f(Y_i; \mu, \sigma^2) \end{aligned}$$

where  $f(x; \mu, \sigma^2)$  is the density function of a normal with mean  $\mu$  and variance  $\sigma^2$  evaluated at  $x$ .

A prior distribution on  $\mu$  is needed. Consider prior beliefs to be distributed as normal:

$$\mu \sim \text{Normal}(a, b).$$

Constants  $a$  and  $b$  are the mean and variance of the prior, respectively, and must be chosen *a priori*. In this case, we consider two analysts with differing beliefs. Analyst I is doubtful that Smarties have any effect at all, and has chosen a tightly constrained prior with  $a = 100$  and  $b = 1$ . Analyst II, on the other hand, is far less committal in her beliefs, and chooses  $a = 100$  and  $b = 200$  to show this lack of commitment. These choices are shown in Figure 9.3A.

With this prior, the posterior beliefs may be expressed as

$$\pi(\mu|\mathbf{Y}) \propto \left( \prod_i f(Y_i; \mu, \sigma^2) \right) f(\mu; a, b).$$

The above equation may be expanded and simplified, and Rouder and Lu (2005) among many others show that

$$\pi(\mu|\mathbf{Y}) = f(cv, v),$$

where

$$c = \left( \frac{n\bar{Y}}{\sigma^2} + \frac{a}{b} \right), \quad (9.2)$$

$$v = \left( \frac{n}{\sigma^2} + \frac{1}{b} \right)^{-1}, \quad (9.3)$$

and  $n$  is the sample size and  $\bar{Y}$  is the sample mean.

The posterior beliefs about  $\mu$  follow a normal with mean  $cv$  and variance  $v$ , and this fact may equivalently be stated as

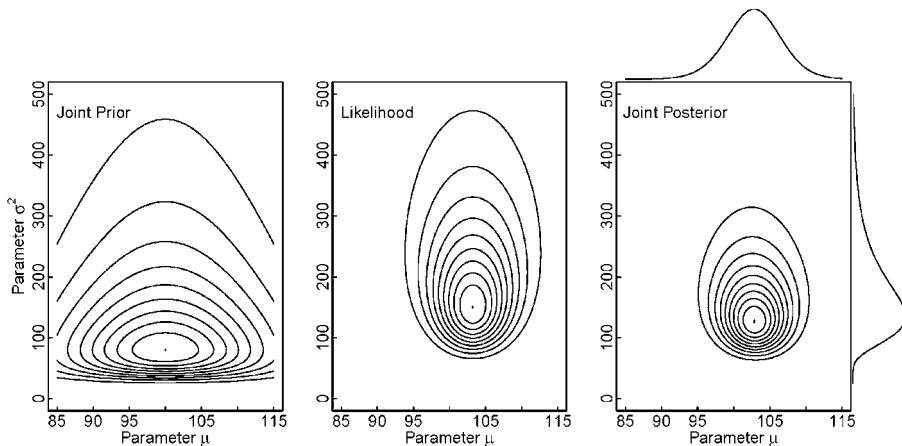
$$\mu|\mathbf{Y} \sim \text{Normal}(cv, v).$$

One property of the posterior is that it reflects information from both the prior and the data. Here, the posterior mean is a weighted average of the sample mean and the prior mean, with the number of observations determining the relative weight. If there are few observations the prior has relatively higher weight than if there are many observations. A second property of the posterior is that it is the same functional form as the prior – both are normally distributed. When a prior and posterior have the same functional form, the prior is termed *conjugate*. Conjugate priors are desirable because they are computationally convenient. A third notable property of the posterior is that it may be well localized even if the prior is arbitrarily variable. The prior variance  $b$  reflects the certitude of prior information, with larger settings corresponding to less certitude. In fact, it is possible to set  $b = \infty$ , and the resulting prior may be called flat as all values of  $\mu$  are equally plausible. This flat prior is *improper* – that is, it does not integrate to a finite value. Even though the prior is improper, the posterior in this case is proper and is given by

$$\mu|\mathbf{Y} \sim \text{Normal}(\bar{Y}, \sigma^2/N).$$

For the flat prior, the posterior for  $\mu$  corresponds to the frequentist sampling distribution of the mean.

Figures 9.3B and 9.3C show the role of sample size in posterior beliefs. Figure 9.3B shows the posterior beliefs of the two analysts for a very small set,  $N = 10$ , with a sample mean IQ score of  $\bar{Y} = 95$ . The data have slightly shifted and slightly widened the beliefs of Analyst I, the analyst who was *a priori* convinced there was little chance of an effect. It has more dramatically sharpened the beliefs of Analyst II, the less-committed analyst. Figure 9.3C shows the case with a larger set,  $N = 100$ , and  $\bar{Y} = 95$ . Here, the posterior beliefs are more similar because the data are sufficient in sample size to have a large effect relative to the prior. In the large-sample limit, these posterior distributions will converge to a point at the true value of  $\mu$ .



**Figure 9.4** Joint prior (left), likelihood (center), and joint posterior (right) distributions across normal-distribution parameters  $\mu$  and  $\sigma^2$ . Also shown, in the margins, are the marginal posterior distributions of  $\mu$  (top) and  $\sigma^2$  (right).

### 9.2.3 Sampling: an approach to Bayesian analysis with more than one parameter

In the previous example, we modeled IQ scores as a normal under the assumption that  $\sigma^2$  is known. Clearly such an assumption is too restrictive, and a more reasonable goal is to state posterior beliefs about both  $\mu$  and  $\sigma^2$ , jointly. An application of Bayes Rule yields

$$\pi(\mu, \sigma^2 | \mathbf{Y}) \propto L(\mu, \sigma^2, \mathbf{Y})\pi(\mu, \sigma^2).$$

The prior density, posterior density, and likelihood functions in this case are evaluated on a plane and take as inputs ordered pairs. Examples of a prior, likelihood, and posterior are shown in Figure 9.4 as two-dimensional surfaces. Because the posterior and prior in the above equation are functions of  $\mu$  and  $\sigma^2$  taken jointly, they are referred to as the *joint posterior* and the *joint prior*, respectively. Fortunately, deriving joint posteriors is straightforward as it is simply the result of Bayes Rule: the posterior is the product of the likelihood and the prior.

Expressing joint posterior beliefs as surfaces may be reasonable for models with two dimensions, but becomes unwieldy as the dimensionality increases. For instance, in models with separate parameters for individuals and items, it is not uncommon to have thousands of parameters. The expression of joint posterior distributions over high-dimensional parameter vectors is not helpful. Instead, it is helpful to plot *marginal posteriors*. The marginal posterior for one parameter, say  $\mu$ , is denoted  $\pi(\mu | \mathbf{Y})$ , and is obtained by averaging (integrating) the uncertainty in all other parameters. The two marginal posteriors for this model are

$$f(\mu | \mathbf{Y}) = \int_{\sigma^2} f(\mu, \sigma^2 | \mathbf{Y}) d\sigma^2$$

$$f(\sigma^2 | \mathbf{Y}) = \int_{\mu} f(\mu, \sigma^2 | \mathbf{Y}) d\mu$$

Marginal posteriors for the two parameters are shown in Figure 9.4, right panel. As can be seen, these provide a convenient expression of posterior beliefs.

Although marginal posteriors are useful for expressing posterior beliefs, they may be difficult to compute. In the two-parameter model above, the computation was straightforward because the integration was over a single dimension and could be solved numerically. In typical models, however, there may be hundreds or thousands of parameters. To express each marginal, all other parameters must be integrated out, and the resulting integrals span hundreds or even thousands of dimensions. This problem of high-dimensional integration was a major pragmatic barrier for the adoption of Bayesian methods until the 1980s, when new computational methods became feasible on low-cost computers.

A modern approach to the integration problem is sampling from the joint posterior distribution. We draw as many samples from the joint that is needed to characterize it to arbitrary precision. Each of these samples is a vector that has the dimensionality of the joint distribution. To characterize the marginal for any parameter, the corresponding element in the joint sample is retained. For example, if  $(\mu, \sigma^2)^{[m]}$  is the  $m$ th sample from the joint, then the value of  $\mu$ , which we denote as  $\mu^{[m]}$ , is a sample from the marginal posterior distribution of  $\mu$ , and the collection  $\mu^{[1]}, \mu^{[2]}, \dots$  characterize this distribution to arbitrary precision. So integrating the joint posterior may be reduced to sampling from it.

Directly sampling from high-dimensional distributions is often difficult. To mitigate this difficulty, alternative indirect algorithms have been devised. The most popular class of these algorithms is called Markov chain Monte Carlo (MCMC) sampling. These techniques are covered in depth in many textbooks (e.g., Jackman, 2009). Here, we cover the briefest outline. Those readers familiar with MCMC, or those who have no desire to learn about it, may skip this outline without loss, as the remainder of the chapter does not rely on understanding MCMC.

We focus here on the most common MCMC algorithm, the Gibbs sampler (Gelfand and Smith, 1990; Geman and Geman, 1984). When building a Gibbs sampler, researchers focus on *conditional posterior distributions*. The conditional posterior distributions are the beliefs about one parameter if all others were known. For the normal model, there are two full conditional posteriors denoted  $f(\mu|\sigma^2, \mathbf{Y})$  and  $f(\sigma^2|\mu, \mathbf{Y})$ . These are easily derived from an application of Bayes Rule:

$$\begin{aligned}\pi(\mu|\sigma^2, \mathbf{Y}) &\propto L(\mu, \sigma^2, \mathbf{Y})\pi(\mu|\sigma^2), \\ \pi(\sigma^2|\mu, \mathbf{Y}) &\propto L(\mu, \sigma^2, \mathbf{Y})\pi(\sigma^2|\mu).\end{aligned}$$

If the priors are independent of one another,

$$\begin{aligned}\pi(\mu|\sigma^2, \mathbf{Y}) &\propto L(\mu, \sigma^2, \mathbf{Y})\pi(\mu), \\ \pi(\sigma^2|\mu, \mathbf{Y}) &\propto L(\mu, \sigma^2, \mathbf{Y})\pi(\sigma^2).\end{aligned}$$

The reason researchers focus on the conditionals is that it is straightforward to analytically express these distributions. Moreover, and more importantly, it is often straightforward to sample from conditionals, which is the key to Gibbs sampling. For the normal-distribution case above, we denote samples of  $\mu$  as

$\mu^{[1]}|\sigma^2, \mu^{[2]}|\sigma^2, \dots, \mu^{[M]}|\sigma^2$ , where  $M$  is the total number of samples. Likewise, the samples of the conditional posterior distribution of  $\sigma^2$  may be denoted  $(\sigma^2)^{[1]}|\mu, \dots, (\sigma^2)^{[M]}|\mu$ . These samples, however, are conditional on particular values of  $\mu$  and  $\sigma^2$ , and, consequently, are not so interesting.

The goal is to obtain marginal samples of  $\mu$  and  $\sigma^2$ , rather than conditional ones. In our specific case, this goal may be achieved as follows: on the  $m$ th iteration,  $\mu$  is sampled conditional on the previous value of  $\sigma^2$ , i.e.,  $\mu^{[m]}|(\sigma^2)^{[m-1]}$ ; then  $\sigma^2$  is sampled conditional on the just-sampled value of  $\mu$ , i.e.,  $(\sigma^2)^{[m]}|\mu^{[m-1]}$ . In this manner, the samples are being conditioned on different values on every iteration, and if conditioning is done this way, the joint distribution of the samples approaches the true joint posterior as the number of samples grows infinitely large. If we have samples from the joint distribution, characterizing any marginal distribution is as easy as ignoring samples of all other parameters. Researchers new to Bayesian analysis can use modern tools such as JAGS (Plummer, 2003) and WinBUGS (Lunn *et al.*, 2000) to perform MCMC sampling without much special knowledge.<sup>2</sup> Those with more experience can write their own code in high-level languages such as R or Matlab. We discuss these options further in the concluding remarks.

### 9.3 Bayesian hierarchical models are simple and natural

There are several advantages of adopting a Bayesian perspective, and one of the most salient for cognitive modelers is the ease of building hierarchical models that may account for variation in real-world settings. Consider the following simple experiment where  $I$  individuals provide  $K$  replications in each of  $J$  conditions. To demonstrate the elegance and power of hierarchical modeling, we build a sequence of models, illustrating each with reference to an experiment where  $I = 20$  individuals provided  $K = 10$  replications in each of  $J = 2$  conditions. Figure 9.5A shows the overall mean for each of these conditions (bars) as well as the participant-by-condition means (points and lines). As can be seen, there is much participant variability as well as strong evidence for a condition effect.

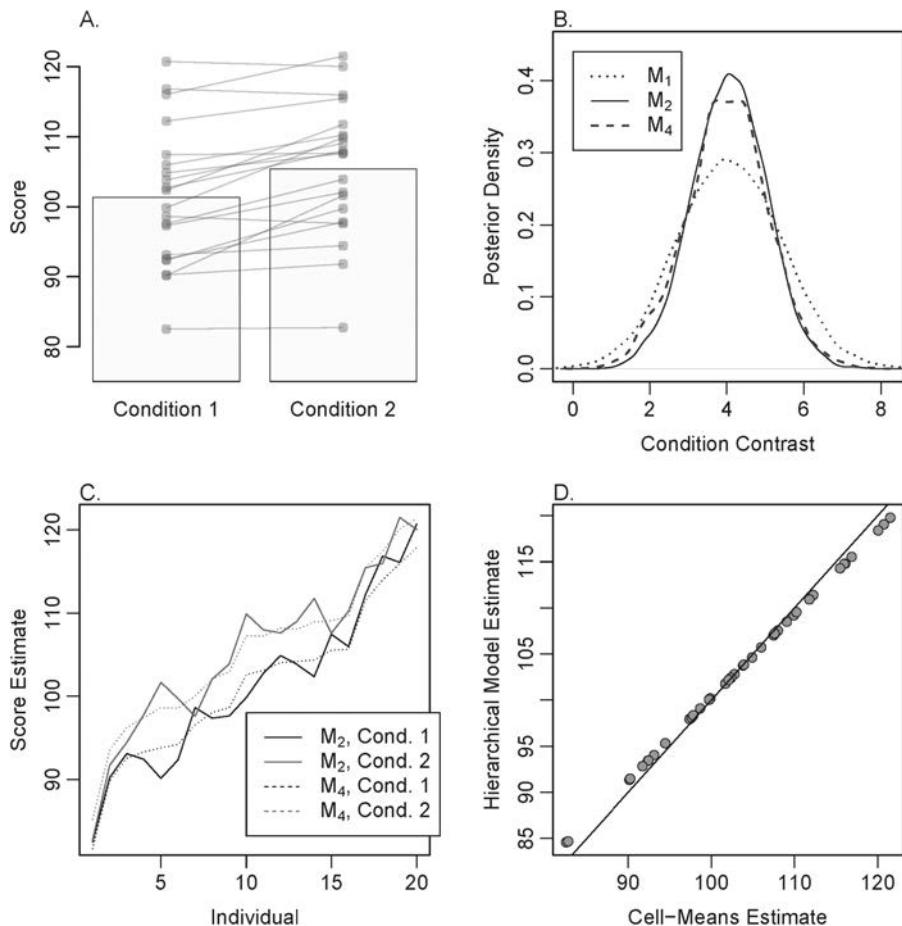
*Model  $M_1$ : an aggregation model.* One approach, which corresponds to aggregating, is to simply model condition means. Let  $Y_{ijk}$  denote the  $k$ th observation for the  $i$ th participant in the  $j$ th condition. The model is

$$Y_{ijk} \stackrel{iid}{\sim} \text{Normal}(\beta_j, \sigma^2), \quad (9.4)$$

where  $\beta_j$  is the condition effect. A prior is needed for each  $\beta_j$  and for  $\sigma^2$ , and we chose priors that makes no practical commitment to the location of these effects:

$$\begin{aligned} \beta_j &\stackrel{iid}{\sim} \text{Normal}(0, 10^6) \\ \sigma &\sim \text{Uniform}(0, 100). \end{aligned}$$

<sup>2</sup> JAGS may be obtained at <http://mcmc-jags.sourceforge.net>. WinBUGS and OpenBUGS (for non-Windows operating systems) may be obtained at <http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml> and <http://www.openbugs.info/w/>, respectively.



**Figure 9.5** The advantages of hierarchical modeling. (A). Hypothetical data from 20 individuals each providing observations in two conditions. The bars show overall condition means; the points and lines show individual's condition means. (B). Posterior distributions of the condition effect from Model  $M_1$ , the aggregation model (dotted line), Model  $M_2$ , the cell means model (solid line), and Model  $M_4$ , the hierarchical model with main effects and interactions (dashed line). Localization is worse for  $M_1$  because participant variability is not modeled. (C). Solid lines show participant-by-condition point estimates from  $M_2$ ; the dotted lines show the same from  $M_4$ . The shrinkage in  $M_4$  to main effects imposed by the hierarchical prior smooths these estimates. (D). A comparison of individual-by-condition estimates from the  $M_2$ , the cell-means model, and  $M_4$ , the hierarchical model with main effects and interactions. There is modest shrinkage for extreme estimates.

The model is designed to assess condition means, and the condition effect may be defined as the contrast  $\beta_2 - \beta_1$ .

We provide here the BUGS language code snippets for analysis of this and subsequent models, and these snippets may be used with WInBUGS, OpenBUGS, or

JAGS. Those researchers who are not familiar with these packages will benefit from the well-written documentation (see Footnote 2) as well as the tutorials provided in Kruschke (2011), Ntzoufras (2009), and Lee and Wagenmakers (2013). The following model statement defines Model  $\mathcal{M}_1$ :

```

model {
  #y is a vector of all observations
  #cond is a vector that indicates the condition
  #mu is a vector of J condition means

    # Model of Observations
  for (n in 1:N) {
    y[n] ~ dnorm(mu[cond[n], tau)
  }
    # note: BUGS uses precision to parameterize normal
    # note: tau is precision

  #Prior on mu
    for (j in 1:J){
      mu ~ dnorm(0, .0001)

    #Prior on precision (std. dev.)
  tau <- pow(sigma, -2)
  sigma ~ dunif(0, 100)
}

```

Posterior beliefs about this contrast are shown as the dotted line in Figure 9.5B.<sup>3</sup> This model is not hierarchical, as there is a single source of variability.

*Model  $\mathcal{M}_2$ : a cell means model.* A more useful approach is to model the combination of participant and condition effects:

$$Y_{ijk} \stackrel{iid}{\sim} \text{Normal}(\mu_{ij}, \sigma^2). \quad (9.5)$$

The parameters  $\mu_{ij}$  are the mean of the  $i$ th participant in the  $j$ th condition. In the example with 2 conditions and 20 participants, there are 40 of these effects, and each needs a prior. Again, we choose diffuse priors:

$$\begin{aligned} \mu_{ij} &\stackrel{iid}{\sim} \text{Normal}(0, 10^6) \\ \sigma &\sim \text{Uniform}(0, 100). \end{aligned}$$

The BUGS language snippet that defines this model is

```

mmodel{
  #y is a vector of all N observations
  #sub is a vector that indicates the participant

```

<sup>3</sup> Posterior beliefs may be computed by subtracting MCMC samples. For the  $m$ th iteration, let  $c^{(m)} = \beta_2^{(m)} - \beta_1^{(m)}$ . The dotted line in Figure 9.5B is the smoothed histogram of  $c^{(m)}$ .

```

#cond is a vector that indicates the condition
#mu is an I-by-J matrix

#Model of observations
for(n in 1:N){
    y[n] ~ dnorm(mu[sub[n],cond[n]], tau)
}

#Prior on mu
for (i in 1:I){
    for (j in 1:J){
        mu[i,j] ~ dnorm(0, .01)
    }
}

#Prior on precision (std. dev.)
tau <- pow(sigma, -2)
sigma ~ dunif(0, 100)
}

```

The posterior means for the cell mean parameters are shown in Figure 9.5C as solid lines. As can be seen, participants often have higher mean scores in Condition 2 than in Condition 1, providing evidence for the condition effect. We can construct a contrast for this comparison:  $\sum_i(\mu_{i2} - \mu_{i1})/I$ , and the posterior for this contrast is shown as the solid line in Figure 9.5B.<sup>4</sup> Note that this posterior is better localized than the comparable contrast from Model  $\mathcal{M}_1$ . The reason is simple: individual variation is subtracted off, leading to better parameter localization. It should be noted that these posterior beliefs, however, do not generalize to new participants. The reason is that people-by-condition effects are “fixed” in that they may vary arbitrarily and provide no information about a population of people, conditions, or their combination.

*Model  $\mathcal{M}_3$ : a first hierarchical model.* Although the interpretation of the cell means model is familiar and reasonable, we can make even more useful models. We start with the same data model:

$$Y_{ijk} \stackrel{iid}{\sim} \text{Normal}(\mu_{ij}, \sigma^2).$$

In the previous model, the priors on  $\mu_{ij}$  were very diffuse. Yet, it is unreasonable to think that these cell mean parameters will arbitrarily differ from one another. For example, if we were studying IQ, it is hard to believe that participant-by-condition means vary by even a factor of two, much less orders of magnitude. One way of adding information without undue influence is through a hierarchical prior. Consider the prior

$$\mu_{ij} \stackrel{iid}{\sim} \text{Normal}(\nu, \delta^2) \tag{9.6}$$

<sup>4</sup> The posterior for this contrast is computed in MCMC as  $c^{(m)} = (\sum_i(\mu_{i2}^{(m)} - \mu_{i1}^{(m)})/I$ , and the solid line in Figure 9.5B is the smoothed histogram.

where  $\nu$  and  $\delta$  describe the center and dispersion of the population of cell means. These values need not be fixed *a priori*. Instead, they may be treated as parameters upon which we may place priors and compute posteriors. Consider the following priors:

$$\begin{aligned}\nu &\sim \text{Normal}(0, 10^6) \\ \delta &\sim \text{Uniform}(0, 100).\end{aligned}$$

Here, we bring little if any *a priori* information about the population center and dispersion of effects. All we commit to is that the effects themselves are samples from this parent distribution. Hierarchical models are therefore implemented as hierarchical priors. Of course, a prior is still needed on  $\sigma^2$ , and we again use a diffuse prior:

$$\sigma \sim \text{Uniform}(0, 100).$$

The hierarchical nature of model  $\mathcal{M}_3$  is embedded in the relationships between parameters. The data  $Y_{ijk}$  are only explicitly dependent on the mean  $\mu_{ij}$  and variance  $\sigma^2$ . If we know these two parameters, then the population from which  $Y_{ijk}$  is drawn is completely determined. Conversely, having observed  $Y_{ijk}$ , we constrain our beliefs about the parameters governing this population distribution. The hierarchy in  $\mathcal{M}_3$  reflects the treatment of the collection  $\mu$  parameters. These parameters are also treated as draws from a population. If we could observe the  $\mu$  parameters directly, we could learn about  $\nu$ , which is a parent parameter for this population of mean parameters. However,  $\nu$  is one step removed from the data: we can only learn about  $\nu$  through learning about the  $\mu$  parameters. Bayes Rule, the unifying rule for Bayesian inference, gives us a natural way of representing the way the information passes from level to level through the simple fact from probability theory that  $p(a, b) = p(a|b)p(b)$ . The posterior  $p(\mu, \nu|\mathbf{Y})$  is then proportional to

$$p(\mu, \nu|\mathbf{Y}) \propto (\mathbf{Y}|\mu, \nu)p(\mu, \nu) = p(\mathbf{Y}|\mu)p(\mu|\nu)p(\nu)$$

(the parameters  $\sigma^2$  and  $\delta$  are assumed known for clarity). The right-hand side of the equation shows how knowledge about parameters is passed up through the hierarchy from the data to the higher-level parameters: the data  $\mathbf{Y}$  and parameter  $\mu$  are connected through the likelihood, and  $\mu$  and  $\nu$  are connected through the hierarchical prior on  $\mu$ . Likewise, constraint from  $\nu$  is passed down through the hierarchy from higher-level level parameters to the lower-level ones.

Figure 9.5D shows the effects of the constraint passed from the higher-level parameters. As can be seen, extreme cell mean values for this hierarchical model are somewhat moderated; that is, they are modestly pulled toward the population mean. This effect is often termed *hierarchical shrinkage*, and it leads to posterior estimates that have lower root-mean-squared error than nonhierarchical estimates. The effect here is modest because the data were generated with low noise for demonstration, but shrinkage can be especially pronounced in nonlinear models.

The use of hierarchical models has an element that is counterintuitive: one adds parameters to the prior to add constraint. In most models, adding parameters is adding flexibility, and more parameters implies a more flexible account of data. In

hierarchical models, the opposite may hold when additional parameters are added to the prior. For instance, the cell means model has 40 cell mean parameters and a variance parameter; the hierarchical model has these 41 parameters and additional population mean and variance parameters. Yet, the cell means model is more flexible as the 40 cell mean parameters are free to vary arbitrarily. In the hierarchical model, no one cell mean can stray arbitrarily from the others, and this behavior is a form of constraint even though it comes with more rather than less parameters. In Bayesian hierarchical modeling, flexibility is not a matter of the number of parameters; it is, instead, a matter of constraint or the lack thereof in the priors. Principled Bayesian model comparison methods such as Bayes factors capitalize on this fact.

In addition to more accurate estimation of individual effects through shrinkage, hierarchical models offer two other benefits. First, posterior beliefs about group parameters,  $\nu$  and  $\delta^2$  in the above example, can be generalized to other participant-by-condition combinations. These parameters, therefore, provide a means of applying the results more broadly. Second, more advanced models may be placed on  $\nu$  that incorporate covariates.

Hierarchical models are straightforward to code in BUGS:

```

model{
  #y is a vector of all N observations
  #sub is a vector that indicates the participant
  #cond is a vector that indicates the condition
  #mu is an I-by-J matrix

  #Model of Observations
  for(n in 1:N){
    y[n] ~ dnorm(mu[sub[n],cond[n]], tau)
  }

  #Level 1: Prior on mu
  for (i in 1:I){
    for (j in 1:J){
      mu[i,j] ~ dnorm(nu,tauI)
    }
  }

  #Level 1: Prior on precision (std. dev.)
  tau <- pow(sigma, -2)
  sigma ~ dunif(0, 100)

  #Level 2: Prior on nu, tauI
  nu~dnorm(0,.000001)
  tauI <- pow(deli, -2)
  deli ~ dunif(0, 100)
}

}

```

*Model  $\mathcal{M}_4$ : a hierarchical model with main effects and interactions.* The shrinkage in Model  $\mathcal{M}_3$  shrinks estimates toward the overall mean. Yet, there is clearly structure from participants and items. We add this structure into the prior as follows:

$$\mu_{ij} \stackrel{iid}{\sim} \text{Normal}(\alpha_i + \beta_j, \delta^2).$$

Priors are then needed for  $\alpha_i$ , the effect of the  $i$ th participant,  $\beta_j$ , the effect of the  $j$ th condition, as well as  $\delta^2$  which now describes the variability of participant-by-condition interactions. We use the following vaguely-informative priors:

$$\begin{aligned}\alpha_i &\sim \text{Normal}(0, \delta_\alpha^2) \\ \beta_j &= \text{Normal}(0, 10^6) \\ \delta &\sim \text{Uniform}(0, 100) \\ \delta_\alpha &\sim \text{Uniform}(0, 100).\end{aligned}$$

This new model treats participant effects as random effects drawn from a population distribution. Generalization to new people is possible through the inclusion of population variability parameter  $\delta_\alpha^2$ . This model also treats conditions as fixed effects, that is, conditions may differ from each other without constraint. Here, there is no concept of a population of conditions and, consequently, the results apply only to these two conditions. Finally, the interaction term reflects an asymptotic interaction between people and conditions; that is, it is the interaction that remains even in the limit that the number of replicates,  $K$ , increases without bound. We include this term hesitantly, because if it is too large, it is difficult to interpret the participant and condition effects. In these cases, we recommend that this interaction become more a target of inquiry, and models of patterned interactions be proposed and compared.

Even though this model is even more heavily parameterized than the previous hierarchical model, it is straightforward to estimate with the following BUGS snippet:

```
model{

  #Model of observations
  for(n in 1:N){
    y[n] ~ dnorm(mu[sub[n],cond[n]], tau)

  #Level 1: Prior on mu
  for (i in 1:I){
    for (j in 1:J){
      mu[i,j] ~ dnorm(alpha[i]+beta[j],tauI)
    }}}
```

```

#Level 1: Prior on tau
tau <- pow(sigma, -2)
sigma ~ dunif(0, 100)

#Level 2: Prior on alpha, beta
for (i in 1:I){ alpha[i]~dnorm(0,tauA)}
for (j in 1:J){ beta[j]~dnorm(100,.001)}

#Level 2: Prior on tauI, scale of interactions.
tauI <- pow(deli, -2)
deli ~ dunif(0, 100)

#Level 3: Prior on tauA, variability of individuals
tauA <- pow(dela, -2)
dela ~ dunif(0, 100)
}

```

The resulting values for the cell means, which are now treated hierarchically, are shown as dotted lines in Figure 9.5C. Notice that these are smoothed versions of the cell means models. The shrinkage to main effects has smoothed away the interaction, making it easy to interpret the condition and participant effects. In fact, in this model, the standard deviation of these interactions ( $\delta \approx 1.2$ ) is considerably less than the standard deviation of participant effects ( $\delta_\alpha \approx 9.9$ ) or the difference between condition effects ( $\approx 4.0$ ). The posterior beliefs about the condition effect is shown as the dashed line in Figure 9.5B.

*Model  $M_5$ : a hierarchical main-effects model.* In many cases, it is desirable to remove the asymptotic interaction terms from the models. Not only do these make interpretation difficult, they may be unidentifiable, and this is certainly the case when there is a single replicate per cell ( $K = 1$ ). Instead of modeling  $\mu_{ij}$  as a random variable, we assume it is a deterministic sum:

$$\mu_{ij} = \alpha_i + \beta_j.$$

Missing is the parameter  $\delta$ , which effectively is set to zero. The prior on the main effects is retained. This additive approach is taken in both applications in Sections 9.4 and 9.5. The following snippet is used for analysis:

```

model{

  #Model of observations
  for(n in 1:N){
    y[n] ~ dnorm(alpha[sub[n]]+beta[cond[n]], tau)}

  #Level 1: Prior on alpha and beta
  for (i in 1:I){ alpha[i]~dnorm(0,tauA)}
  for (j in 1:J){ beta[j]~dnorm(100,.001)}
}

```

```

#Level 1: Prior on tau
tau <- pow(sigma, -2)
sigma ~ dunif(0, 100)

#Level 2: Prior on tauA, variability across participants
tauA <- pow(deltaA, -2)
deltaA ~ dunif(0, 100)
}

```

Analysis of  $\mathcal{M}_5$  results in essentially the same posterior beliefs as Model  $\mathcal{M}_4$  for main effects of people ( $\alpha_i$ ), conditions ( $\beta_j$ ) and their combinations ( $\mu_{ij}$ ). These results are omitted for clarity. Deciding whether asymptotic interactions are needed is not easy, and the topic of model comparison is discussed next.

## 9.4 Comparing hierarchical models

In most applications, it is useful to define a set of models and compare the relative evidence from the data for each. In the above case, for example, we might assess the evidence for interactions between people and condition by considering the relative evidence for Model  $\mathcal{M}_4$ , the model with interactions, and  $\mathcal{M}_5$ , the model without interactions. The condition main effect may be assessed if we compare Model  $\mathcal{M}_5$  to a model without condition effects, specified by the constraint  $\mu_{ij} = \alpha_i$ . The critical question is how the relative evidence for such models may be stated.

Model comparison is a broad and expansive topic about which there is substantial diversity in the statistical and psychological communities. The chapter by Myung in this volume provides an overview of some of this diversity. Even though there is much diversity, we believe one model comparison method, comparison by Bayes factor (Jeffreys, 1961), is superior because it (a) directly provides a measure of evidence for models, and (b) is the unique, logical resultant of applying the Bayes Rule to model comparison. Although Bayes factors are ideal, they are associated with two issues. First, the Bayes factor is sometimes difficult to compute, especially in hierarchical settings. Second, the Bayes factor is integrally dependent on the prior. We and others have argued that this dependency is necessary for valid model comparison (Gallistel, 2009; Jeffreys, 1961; Lindley, 1957; Rouder *et al.*, 2009; Rouder and Morey, 2012; Wagenmakers, 2007). Nonetheless, it often is not obvious how to structure priors to compare different nonlinear accounts of the same phenomena. In the following, (i) we define Bayes factors; (ii) discuss some of the difficulties in implementation including computational difficulties; (iii) mention some of the methods of circumventing these difficulties; and (iv) describe an alternative, Deviance Information Criterion (DIC, Spiegelhalter *et al.*, 2002), a less desirable but more computationally feasible approach that may be used as a last resort.

The Bayesian interpretation of probability as subjective belief licenses the placing of probabilities (beliefs) on models themselves. To compare two models, denoted generically as  $\mathcal{M}_A$  and  $\mathcal{M}_B$ , we may place the model probabilities in ratio. The ratios  $Pr(\mathcal{M}_A)/Pr(\mathcal{M}_B)$  and  $Pr(\mathcal{M}_A | \mathbf{Y})/Pr(\mathcal{M}_B | \mathbf{Y})$  are the prior and posterior odds of the models, respectively. Bayes Rule for updating these prior odds is

$$\frac{Pr(\mathcal{M}_A)}{Pr(\mathcal{M}_B)} = \frac{f(\mathbf{Y} | \mathcal{M}_A)}{f(\mathbf{Y} | \mathcal{M}_B)} \times \frac{Pr(\mathcal{M}_A)}{Pr(\mathcal{M}_B)}. \quad (9.7)$$

The term  $f(\mathbf{Y} | \mathcal{M}_A)/f(\mathbf{Y} | \mathcal{M}_B)$  is called the *Bayes factor*, and it describes the updating factor from the data (Kass and Raftery, 1995). We denote the Bayes factor by  $B_{AB}$ , where the subscripts indicate which two models are being compared. The term  $f(\mathbf{Y} | \mathcal{M}_A)$  may be expressed as:

$$f(\mathbf{Y} | \mathcal{M}_A) = \int_{\theta \in \Theta_A} L_A(\theta, \mathbf{Y}) \pi_A(\theta) d\theta, \quad (9.8)$$

where  $L_A$  is the likelihood of  $\mathcal{M}_A$ , and  $\theta$  and  $\Theta_A$  are the parameters and parameter space, respectively, of the model. This term is called *the marginal likelihood*, and it is the weighted average of the likelihood over all possible parameter values. We use  $m_A$  and  $m_B$  to denote the marginal likelihoods of models  $\mathcal{M}_A$  and  $\mathcal{M}_B$ , respectively. The Bayes factor is

$$B_{AB} = \frac{m_A}{m_B}.$$

A Bayes factor of  $B_{AB} = 10$  means that prior odds should be updated by a factor of 10 in favor of model  $\mathcal{M}_A$ ; likewise, a Bayes factor of  $B_{AB} = .1$  means that prior odds should be updated by a factor of 10 in favor of model  $\mathcal{M}_B$ . Bayes factors of  $B_{AB} = \infty$  and  $B_{AB} = 0$  correspond to infinite support of one model over the other, with the former indicating infinite support for model  $\mathcal{M}_A$  and the latter indicating infinite support for model  $\mathcal{M}_B$ .

Bayes factors provide a principled method of inference, and advocacy in psychology is provided by Edwards *et al.* (1963), Gallistel (2009), Myung and Pitt (1997), Morey and Rouder (2011), Rouder *et al.* (2009), and Wagenmakers (2007), among others. There are two critical issues in use: (i) the choice of priors  $\pi_A$  and  $\pi_B$ , and (ii) the evaluation of the integrals in (9.8), and we discussed these issues in turn. First, the choice of priors: the choice of priors will vary from situation to situation. In the case of linear models, such as those underlying *t*-test, linear regression, and ANOVA, researchers already know the range of plausible effect sizes. For instance, rarely do effect sizes exceed 5 or 10 in value, and we do not run experiments to search for effect sizes less than say .05 in value. These constraints may be capitalized upon to form reasonable and broadly acceptable priors for comparisons within the linear model (see Rouder *et al.*, 2012; Rouder and Morey, 2012). The case for nonlinear models, however, is neither as straightforward nor as well explored. It is an area for future work.

The second issue is the evaluation of the integral in computing the marginal likelihoods in (9.8). Here, the parameter space is often of high dimension, especially in hierarchical models where there are several parameters for each participant and item. For example, in the subsequent recognition memory example in Section 9.6, there are over 2000 parameters. To compute the Bayes factor, the likelihood must be integrated across the whole of the parameter space, and integration across high dimensional spaces is in general challenging. To make matters worse, the likelihood is often concentrated, and the integrand is highly peaked. The integration often becomes a matter of finding a needle in a multidimensional haystack. The topic of computing Bayes factors, especially in hierarchical models, remains topical in Bayesian analysis. Fortunately, Bayes factor computations for linear models underlying the *t*-test, ANOVA, and regression are well established. Seminal work was provided by Jeffreys (1961) and Zellner and Siow (1980). The key innovation from Zellner and Siow was specifying the problem in a manner so that the integration over most dimensions could be done analytically in closed form. The modern implementation of this work is provided among several others by Bayarri and Garcia Donato (2007) and Liang *et al.* (2008). Our group has translated and refined this approach, and we provide Bayes factor replacements for *t*-tests (Rouder *et al.*, 2009), statistical-equivalence tests (Morey and Rouder, 2011), linear regression (Rouder and Morey, 2012), and ANOVA (Rouder *et al.* 2012). We have also provided development of meta-analytic Bayes factors so researchers can assess the totality of evidence across several experiments (Rouder and Morey, 2011; Rouder *et al.* 2013).

Although this Bayes factor development covers a majority of statistical models used in psychology, current computational development does not cover a bulk of the psychological process model which tend to be nonlinear. There are a handful of advanced techniques that are potentially applicable, and we mention them in passing. Perhaps the most relevant is the *Laplace approximation*, where the likelihood is assumed to approach its asymptotic normal limits, and its center and spread are well approximated by classical statistical theory. Sarbanés Bové and Held (2011) use the Laplace approximation to provide a general Bayes factor solution for the class of generalized linear models. An alternative technique is to perform the integration by Monte Carlo sampling, and there has been progress in a number of sampling based techniques including bridge sampling (Meng and Wong, 1996), importance sampling (Doucet *et al.*, 2001), and a new variation on importance sampling termed direct sampling (Walker *et al.*, 2011). These techniques assuredly will prove useful for future Bayes factor development in psychology. The final advanced technique in our survey is Bayes factor computation by means of Savage–Dickey density ratio (Dickey and Lientz, 1970; Verdinelli and Wasserman, 1995), which has been imported into psychology by Morey *et al.* (2011), Wagenmakers *et al.* (2010), and Wetzels *et al.* (2010). Under appropriate circumstances, this ratio is the Bayes factor and is convenient to calculate (see Morey *et al.*, 2011). Wagenmakers *et al.* (2010) and Rouder *et al.* (2012) show how the Savage–Dickey ratio can be used in the comparison

of hierarchical models of psychological process, and Rouder *et al.* use it to discriminate between the power law and exponential law of learning in hierarchical settings.

Even though there has been notable progress in developing Bayes factor solutions, there are several cases without such development, and, at present, Bayes factors are simply not available. For these cases, we have a backup, inference by *deviance information criterion* (DIC, Spiegelhalter *et al.*, 2002). DIC is a Bayesian analog to AIC designed for hierarchical models. Unlike AIC (and BIC), DIC accounts for the flexibility of priors, and penalizes models with more flexible priors more heavily than those with more constrained priors. Such behavior is useful for hierarchical models where increased prior constraint is often accompanied by an increased number of parameters. The main advantage of DIC is computational ease; it is often computed in the same MCMC chain used to compute posterior beliefs about the parameters. The disadvantage is one of principle and calibration. DIC shares a calibration with AIC, and like AIC, tends to penalize flexibility too lightly (Rouder *et al.*, 2009), especially for large sample sizes. The argument in favor of BIC over AIC by Raftery can be applied to favor Bayes factor over DIC. Unlike Bayes factor, which is a principled direct result of Bayes Rule, DIC is best viewed as a heuristic motivated by out-of-sample concerns. We use DIC only as a matter of last resort, and recommend Bayes factors be used without qualification when they are available.

## 9.5 Hierarchical models for assessing subliminality

It is widely believed that a large portion of human cognition is unconscious (Greenwald, 1992). This unconscious cognition can manifest itself in many ways: for instance, we may have unconscious goals and motivations; we may be unaware of the effects of stimuli on these goals and motivations; we may even perceive and be affected by stimuli of which we are unaware. One example of this last category is the popular myth of subliminal advertisements in movie theaters: advertisement images were purportedly presented so quickly as to be consciously imperceptible; nonetheless, these images supposedly changed the subsequent behavior of movie-goers by causing them to buy expensive snacks. This myth has been debunked (Rogers, 1992).

The fact that subliminal advertising was debunked does not mean that under controlled circumstances psychologists could not observe similar (if smaller) effects. In fact, many such claims have been made with demonstrations of subliminal priming (for examples, see Dehaene *et al.*, 1998; Finkbeiner, 2011; Greenwald *et al.*, 1995; Merikle *et al.*, 2001; Naccache and Dehaene, 2001). A subliminal prime is one that cannot be perceived, and yet has an effect on subsequent behavior. To answer the question of whether subliminal priming exists, one needs to show both that a prime cannot be identified at a rate greater than chance, and that this prime nonetheless affects behavior.

The priming task we model is a numerosity decision task. Participants are shown target numerals between 2 and 8 and judge whether the target is greater than or less than 5 in value. Preceding these targets are quickly-presented-and-subsequently masked prime numerals. When the prime has the same status as the target – that is, both are less than five or both are greater than five – responses are known to be speeded relative to the case where the prime and target do not have the same relation to five (e.g., Dehaene *et al.*, 1998; Koechlin *et al.*, 1999; Naccache and Dehaene, 2001; Pratte and Rouder, 2009). The critical question is whether this priming persists even for presentations that are so fast that participants' ability to assess the prime's relation to five is at chance level.

We focus here on the difficult part of assessing subliminal priming: the assessment of whether a prime is identified at chance or above chance. Let  $p_i$  denote the true probability correct for the  $i$ th participant. Primes are subliminal for the  $i$ th participant if true performance is at chance, that is, if  $p_i = .5$ . One approach to assessing subliminality is to perform a null hypothesis significance test on the observed proportions against the null hypothesis that average performance across participants is at chance. If  $y_i$  and  $N_i$  are the number correct and the total sample size for participant  $i$ , and  $q_i = y_i/N_i$ , we might test the hypothesis that  $\mu_q = .5$ . If the sample sizes  $N_i$  are reasonably large and approximately the same, then  $\bar{q}$  will be approximately normal, and we can apply a  $t$ -test against  $\mu_q = .5$ . If the  $t$ -test is not statistically significant, we conclude that performance is at chance. This logic has been used in several influential studies in the subliminal priming literature (Dehaene *et al.*, 1998; Murphy and Zajonc, 1993).

There are at least two major flaws with this approach. First, there is the issue of acceptance of the null hypothesis. The  $t$ -test essentially assumes that all participants are performing at chance unless there is sufficient evidence against that hypothesis. Thus, researchers who wish to show subliminality have an incentive to underpower their designs; after all, with sufficiently small sample sizes, even very good average performance can be claimed to be subliminal simply because there is not enough evidence against it. For this reason, a null result from a null hypothesis significance test cannot be used to argue for the null hypothesis itself. We will use a Bayesian approach to overcome this fundamental limitation.

The second major flaw with this  $t$ -test approach is a failure to properly separate between-participant and across-participant variability. Consider the sources of variability in estimated performances  $q_i$ : the statistic can vary due to natural sampling variability in the task, but also because people vary in their performance. We can reduce the first source of variability by increasing  $N_i$ , but not the second. Consider the extreme case where we have two participants, and they perform an arbitrarily large number of trials. Suppose that  $q_1 = .6$  and  $q_2 = .9$ . We know that both participants are above chance with near-perfect certainty as  $N_i \rightarrow \infty$ , yet, we will always conclude that all participants are at chance because with two participants, the  $t$ -test will not lead to a rejection of the null.

The failure to account for variability across participants leads not only to spurious acceptances of the null, but to spurious rejections as well. For example,

suppose that to avoid the power problem outlined above, we obtain a large sample of participants. Suppose 99% truly perform at chance, and 1% of the population performs above chance at  $p = .75$ . Although 99% of our population is appropriate for assessing subliminal priming, we are guaranteed to reject all participants as we increase our sample size, because the true *average* performance is above 0.5.

To make these problems concrete, we consider data from a subliminal priming experiment reported in Rouder *et al.* (2007a). In this experiment, 27 participants performed 288 trials in a prime identification task. The primes were displayed briefly, only 22 ms, and were forward and backward masked. Performance was generally quite poor, with an average proportion correct of .53. A classical analysis of the accuracies reveals that average accuracy is significantly different from .5 ( $t_{26} = 2.7$ ,  $p = 0.011$ ) with a 95% CI of (0.507, 0.551). Yet, a more complex story unfolds when participant variability is examined (see Figure 9.6A). Although the majority of participants' observed accuracies are clustered around .5, there are two who score substantially higher than the rest. Under the logic outlined in the previous paragraphs, we would throw out the entire sample, even though the majority of participants' observed accuracies are concordant with chance performance. Instead, in the next section we present a hierarchical approach that overcomes this issue by modeling participant variability in assessing the subliminally of primes.

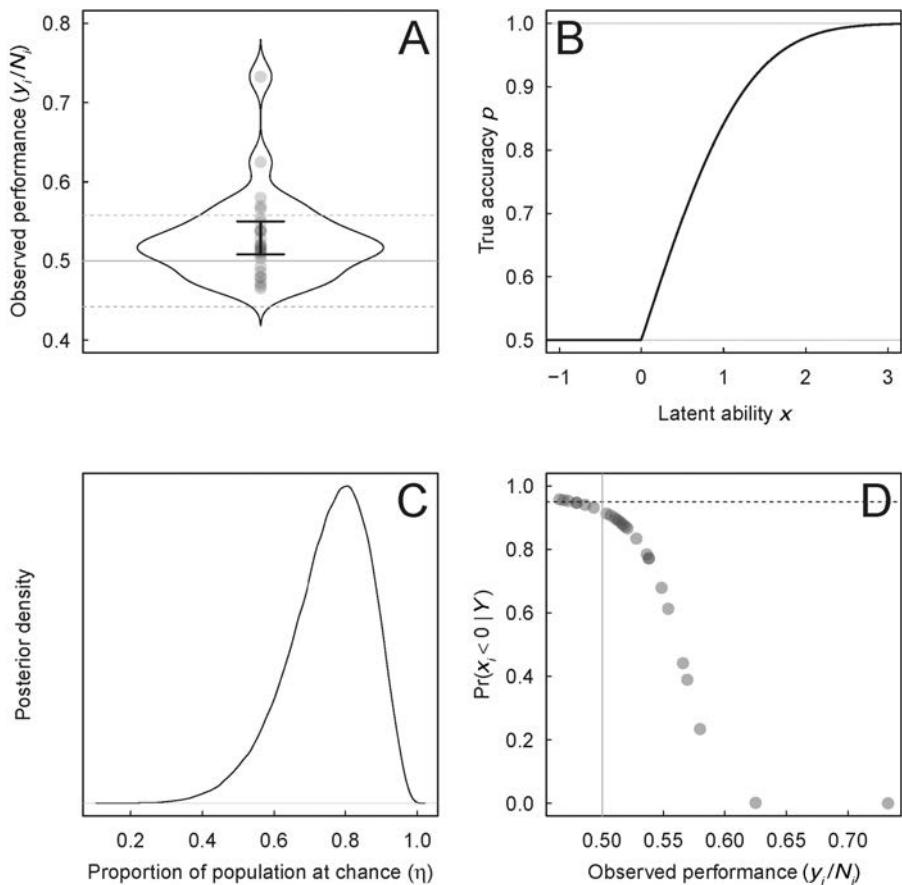
### 9.5.1 A hierarchical model

Our goal is to specify models of accuracy that include a psychological threshold. If activation from the stimulus is lower than this threshold, then performance is at chance. Conversely, if activation exceeds this threshold, then performance is above chance. The hierarchical model presented below is from Rouder *et al.* (2007a). At the first level of the model, we link the observed number correct  $y_i$  for each participant with an underlying true parameter,  $p_i$ :

$$y_i \stackrel{iid}{\sim} \text{Binomial}(p_i, N_i).$$

A hierarchical model is developed by specifying distributions on the individual performance parameters. In our case, we must carefully consider the parent population for the  $p_i$ s. Because  $p$  is restricted to  $[0, 1]$ , it is inappropriate for a traditional normal population. Logit and probit models specify transformations of  $p$  into  $(-\infty, \infty)$ , making normal population distributions possible. For our purposes, however, these transformations are inappropriate because they allow true accuracy to be below  $p = .5$ . Instead, we use a half-probit transformation that restricts true accuracy to  $p \geq .5$ :

$$p_i = \begin{cases} \Phi(x_i) & x_i \geq 0 \\ .5 & x_i < 0 \end{cases}$$



**Figure 9.6** (A). Violin plot of 27 participants' performance in a prime identification task. The confidence interval within the violin plot is the 95% CI on the mean accuracy; the horizontal line at 0.5 represents chance performance, and the horizontal dashed lines bound the interval within which we would expect 95% of participants to perform if they were truly at chance. (B). The mass-at-chance link function. (C): Posterior distribution of the population proportion at chance. (D): Posterior probability that individuals are at chance as a function of their observed performance.

where  $\Phi$  is the CDF of the standard normal distribution. Rouder *et al.* (2007a) called this function the mass-at-chance (MAC) link, due to the fact that it allows participants to have true performance  $p = .5$ . We call  $x_i$  a “latent ability” because it indexes a person’s ability even when  $p_i = .5$ . Figure 9.6B shows the relationship of latent ability to true accuracy. Consider two participants whose latent abilities are  $x_1 = -.01$  and  $x_2 = -2$ . Participant 1 is very near the threshold of  $x_i = 0$ ; perhaps a small increase in the duration of the prime stimuli would lead this participant to discriminate its less-than-five status more often than chance. Participant 2,

however, is far below the threshold, and may need a larger increase in duration than Participant 1 to achieve above-chance performance.

The second level of the hierarchical model may be specified by placing a population distribution on the latent ability parameters:

$$x_i \stackrel{iid}{\sim} \text{Normal}(\mu, \sigma^2).$$

The parameters  $\mu$  and  $\sigma^2$  together define the proportion of the participants whose performance is at chance. At the first level of the hierarchical model, we linked the observations with individuals' parameters; at the second level of the hierarchical model, we described how the individuals' parameters were distributed in a population. At the third and top level of the hierarchical model, we specify prior distributions for the parameters of the population of participants. There is some information for this specification from the context. In subliminal priming experiments, the goal is to make the primes difficult to see. It is therefore reasonable to place an informative prior on  $\mu$  that is centered on the value of 0:

$$\mu \sim \text{Normal}(0, 1).$$

There is also natural constraint from the experimental context on parameter  $\sigma$ . If  $\sigma$  is too large, then bimodal distributions on  $p_i$  are likely with modes at chance and at ceiling. To avoid bimodal distributions on performance, but still allow substantial variability across participants we choose a uniform prior on  $\sigma$ :

$$\sigma \sim \text{Uniform}(0, 1).$$

With all levels of the hierarchical model specified, joint and marginal posterior distributions may be computed. Of particular interest is the marginal posterior probability that the  $i$ th participant is performing at chance:

$$\omega_i = \Pr(x_i \leq 0 | \mathbf{Y}).$$

If this posterior probability is sufficiently high, then we should retain this participant to assess whether the primes truly influence judgments about the target. Also of interest are the marginal posteriors of the population-level parameters  $\mu$  and  $\sigma$ . A convenient statistic is the probability that any participant drawn from the population-level distribution is at chance. We denote this probability  $\eta$ , and it is

$$\eta = \Phi\left(-\frac{\mu}{\sigma}\right).$$

For example, if  $\eta = .8$  for a given stimulus duration, then we expect that 80% of people will be at chance.

We can compute the marginal posterior distributions in a variety of ways: Rouder *et al.* (2007a) derived full conditional distributions and implemented a Gibbs sampler in R. Here, we present BUGS code model specification:

```

model {
  for( i in 1:M ){
    # Level 1: Binomial
    y[i] ~ dbin( p[i], N[i] )

    # Transformation between p and x, level 1
    p[i] <- phi( x[i] * step( x[i] ) )

    # Level 2: population on the latent abilities
    x[i] ~ dnorm( mu, precision )
  }

  # Level 3: Prior parameters
  mu ~ dnorm( 0, 1 )
  sig ~ dunif( 0, 1 )
  # BUGS uses precision, not std dev, to define normal
  precision <- 1 / ( sig * sig )
}

```

We fit the hierarchical model to the data of Rouder *et al.* (2007a) that is shown in Figure 9.6A. Figure 9.6C shows the resulting posterior distribution of the proportion of population judged to be performing at chance ( $\eta$ ). Most of the mass is above .5, indicating that well over half of the population has performance at chance. Of particular interest are the posterior probabilities that the  $i$ th participant performs at chance ( $\omega_i$ ). Figure 9.6D shows the relationship between each participant's observed accuracy  $y_i/N_i$  and the corresponding posterior mean of  $\omega_i$ .

One approach to selecting participants for subliminal priming analysis is to choose a criterion  $c$  such that if  $\omega_i > c$ , participant  $i$  is categorized as "at chance." The horizontal line in Figure 9.6D at .95 shows one possible criterion. The three points above the horizontal line represent participants whose priming effects we might examine; if we found evidence of priming for those participants, it could be used as evidence for subliminal priming.

The hierarchical model outlined above is quite simple, and allowed us to categorize by the plausibility that their ability correspond to at-chance levels given the assumptions of the model. Perhaps from a broader perspective, it may be viewed as a psychometric model of performance. The key innovation is the use of a half-probit link that accounts for a true psychological threshold. This threshold, unlike usual operationalizations in psychophysics, describes the point on latent ability where performance first rises above chance. One reasonable concern is the role of parametric assumptions, and the most salient is the half-probit mapping from latent ability to probability. To model the threshold, it seems necessary to have a link that maps many latent ability values to chance performance, but there are many

alternatives to the half-probit link, such as the CDF of a Weibull which meet this requirement. We chose the half-probit for computational convenience, but there remains the question of whether this link is reasonable. Moreover, it is a somewhat open question of whether different links, such as that from the Weibull, will lead to different assessments of which participants are at chance.

Unfortunately, it seems difficult to assess the fit of the half-probit and the dependence of conclusions of subliminality to parametric assumptions in typical priming studies. The reason for this difficulty is that in typical studies, many participants perform at near chance levels, and thus their performance offers little in the way of information to determine whether the link is reasonable. A better approach may be to change the paradigm to allow for a greater range of performance across individuals. In the current paradigm, stimulus difficulty reflects the duration of presentation, which was set to 22 ms. In subsequent experiments (Morey *et al.*, 2008b), we asked participants to identify stimuli presented at durations from 17 ms to 167 ms. The model extension to this case is covered next.

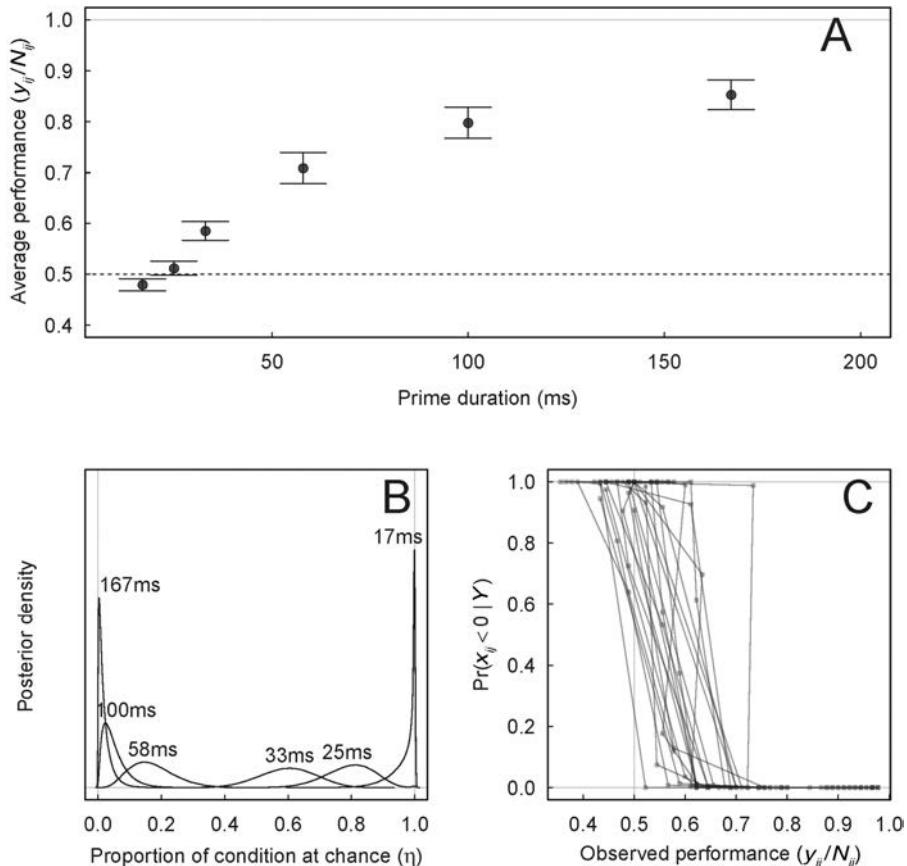
### 9.5.2 Extending the hierarchical model

Extending the paradigm and model to multiple stimulus durations affords several advantages, including the ability to state evidence that participants are at chance at specific durations. Participants who are particularly good at identifying the primes may require very short durations for chance performance, whereas participants who are not as good may be at chance to a wider range of prime durations. A second advantage is that the extended model covers the full range of performance of individuals across stimulus duration, and in this regard, may be treated as a psychophysical and psychometric model. The question of how good the probit link is in accounting for performance may be assessed.

Morey *et al.* (2008b) and Morey *et al.* (2009) developed several models that allow for multiple prime duration conditions. To demonstrate how hierarchical models can be naturally extended, we present the model of Morey *et al.* (2008b) here, which is the simplest of the set. Consider an experiment in which  $J$  participants attempt to identify masked primes in  $I$  stimulus-duration conditions. In condition  $i$ , participant  $j$  performs  $N_{ij}$  prime identification trials, of which  $y_{ij}$  are correct. Figure 9.7A shows average accuracies in a prime identification task with  $I = 6$  conditions (17 ms, 25 ms, 33 ms, 58 ms, 100 ms, and 167 ms). Identification for the shortest prime duration was extremely poor at 48%; the longest duration prime, however, was correctly identified an average of 85% of the time.

The first level of the extended hierarchical model is essentially the same as before, with the exception that now we index both participants and conditions. Observed accuracy for the  $j$ th participant in the  $i$ th condition,  $y_{ij}$ , is distributed as a binomial:

$$y_{ij} \stackrel{iid}{\sim} \text{Binomial}(p_{ij}, N_{ij}),$$



**Figure 9.7** (A): Mean performance by duration condition in a prime identification task. Error bars are standard errors of the mean. (B): The posterior distribution, for each duration condition, of the proportion of the population that would perform at chance in that condition. (C): Posterior probability that individuals are at chance as a function of their observed performance. Lines represent participants, and each point a condition. The top/left-most point for each participant is the briefest duration condition, and subsequent points along the lines are increasingly higher-duration conditions.

and, as before, we link true accuracy  $p$  with latent ability  $x_{ij}$  through the half-probit transformation:

$$p_{ij} = \begin{cases} \Phi(x_{ij}) & x_{ij} \geq 0 \\ .5 & x_{ij} < 0 \end{cases}.$$

In the previous model, all latent abilities  $x$  were drawn from a normal parent distribution. In this case, however, it is desirable to have this parent distribution depend systematically on the duration condition. We place an additive model on

latent ability at the second level:

$$x_{ij} = \mu_i + \alpha_j,$$

where  $\mu_i$  is the average ease with which primes in condition  $i$  are identified, and  $\alpha_j$  is the identification ability of the  $j$ th participant. We have thus reduced the number of parameters underlying latent ability from  $ij$  to  $i + j$ . This type of reduction in complexity is one of the strengths of hierarchical modeling.

We assume that the participant ability parameters  $\alpha_j$  are drawn from a normal population:

$$\alpha_j | \sigma_\alpha^2 \stackrel{iid}{\sim} \text{Normal}(0, \sigma_\alpha^2).$$

$\alpha_j$  can thus be interpreted as the random effect of participant  $j$ . We place an inverse gamma prior on the unknown variance  $\sigma_\alpha^2$ :

$$\sigma_\alpha^2 | a, b \sim \text{Inverse Gamma}(a, b).$$

When parameters  $a$  and  $b$  are chosen to be small (e.g., 0.01), this prior is less constraining than the uniform prior on  $\sigma$  in the Rouder *et al.* (2007a) model. We can use a less constraining prior here because the data extend across multiple conditions, including those where performance is definitively above chance.

The condition effect parameters  $\mu$  can be interpreted as fixed effects; we thus place independent priors on each  $\mu$ :

$$\mu_i \stackrel{iid}{\sim} \text{Normal}(0, 1),$$

where the prior parameters were selected to be similar to those for  $\mu$  in the previously presented model.

The full model can be described in the BUGS language:

```
model {
  for( j in 1:J ){
    for( i in 1:I ){
      # Binomial, level 1
      y[i,j] ~ dbin( p[i,j], N[i,j] )

      # Transformation between p and x, level 1
      p[i,j] <- phi( x[i,j] * step( x[i,j] ) )

      # latent ability is now a linear combination
      # of mu and alpha
      x[i,j] <- mu[i] + alpha[j]
    }
  }

  # Level 2 - define alpha
```

```

for( j in 1:J ){
  alpha[j] ~ dnorm(0, precisionAlpha)
}

# Level 2 - define mu
for(i in 1:I ){
  mu[i] ~ dnorm(0, 1)
}

# Prior parameters, level 3
precisionAlpha ~ dgamma(aAlpha, bAlpha)
}

```

Figure 9.7, panels B and C, show the results of fitting the model to the data shown in panel A. panel B shows the posterior probability on the proportion of participants in the six conditions who perform at chance. At the shortest duration, nearly all participants are predicted to be at chance; at the longest, the proportion at chance is surely less than 10%. Panel C shows the posterior probability that true performance is at chance for each participant by condition combination, that is,  $Pr(x_{ij} < 0 | \mathbf{Y})$ . Each line represents a participant, and each point on the line represents a condition.

As one would expect, the posterior probability of chance performance decreases for all participants as the prime duration increases. For most participants, the decrease occurs in a graded way. Interestingly, there are several participants whose curve is nonmonotonic; that is, for some posterior probability by observed performance pairs, posterior probability increases as performance increases, which is the opposite of what one would expect. This is due to the fact that the additive nature of the model enforces the ordering of true performance to be the same for all participants across conditions. The model does not allow, for instance, one participant to improve their true performance as duration is increased from 25 ms to 33 ms, and another to get worse. However, because observed performance is subject to binomial noise, differences in performance across conditions may, for some participants, be the opposite of what one would expect. The hierarchical model allows us to use information from all participants to infer what the ordering should be, and enforce it for all participants.

The extension of the hierarchical model – specifically, the addition of multiple conditions – allows for a more constrained, more easily tested model. In the simple hierarchical model, most participants were expected to be near chance performance, leaving us very little information by which to falsify the model. The extended model predicts a pattern of data for each participant across stimulus durations, and when this pattern is violated, it will be apparent.

Consider the participant represented by the right-most line in Figure 9.7C. In one condition, this participant is performing at an accuracy of .72, but the model says that this participant is almost surely at chance in that condition. This strange

result led Morey *et al.* (2009) to further extend the model to allow for individual participant slopes:

$$x_{ij} = \theta_j(\mu_i + \alpha_j).$$

This model allows participants to improve at different rates as the stimulus intensity is changed, which improves model fit for some participants. Given the previous development, such an extension is conceptually straightforward. It requires an additional prior for  $\theta$ ; Morey *et al.* (2009) chose a normal distribution truncated below at 0, to require that  $\theta$  be positive, and thus all participants must have the same ordering of true performance across conditions. The model can be easily defined in the BUGS language and fit with WinBUGS or JAGS.

The current set of hierarchical models are based on item response theory (IRT) type formulation with a novel link to account for thresholded behavior. Unlike IRT models, however, the effect of a person is modeled with two parameters while the effect of items (stimulus durations in this case) is modeled with just one. In this sense, these models may be considered perhaps the first set of *hierarchical psychophysical models*. We believe that such models may be of great use: they allow researchers to measure a truly at-chance threshold level of performance across a large number of individuals with a limited number of experimental trials.

Subliminal priming remains a controversial topic. To assess whether it exists, Morey (2008) asked participants to both identify primes, and then identify primed targets. He first used the hierarchical model in Morey *et al.* (2009) to select participant-by-duration combinations for which it was more likely that latent ability was below rather than above chance. For these combinations, however, there was about 5-to-1 evidence by Bayes factor for a null priming effect on the time to identify primed targets. Hence, once one is somewhat sure that prime identification is at chance, the priming effect disappears! One notable study that contradicts this claim, however, comes from Finkbeiner (2011). Finkbeiner used two stimulus durations in a word priming experiment, and used the above extended hierarchical model to select participant-by-duration combinations as being at chance. With these combinations, Finkbeiner found about 10-to-1 evidence by Bayes factor for a priming effect. The approaches used by Morey and Finkbeiner provide for more rigorous assessment of subliminality and subliminal priming than previous methods, and further research with them will be a valuable part of unraveling the puzzle if and when subliminal priming occurs

## 9.6 Hierarchical models for signal-detection experiments

In this section, we demonstrate how hierarchical modeling strengthens the inferential link between theory and data in understanding human memory. We focus on recognition memory, and a prevailing theoretical question is whether recognition memory is mediated by a single strength process or by the two processes of recollection and familiarity (Mandler, 1980). Aggregation, unfortunately,

is the norm in recognition memory experiments. In these experiments, the basic unit of data is a dichotomous outcome. Either a participant indicates a test item is old or new, and to form hit and false alarm rates, these outcomes seemingly must be aggregated across individuals or items. In the following section, we show how this aggregation may gravely distort conclusions about processing. We then introduce a hierarchical model that simultaneously accounts for participant and item variability, mitigating the need for aggregation. This hierarchical model provides for more valid assessment of processing, and we highlight our findings about the number and nature of processes underlying recognition memory.

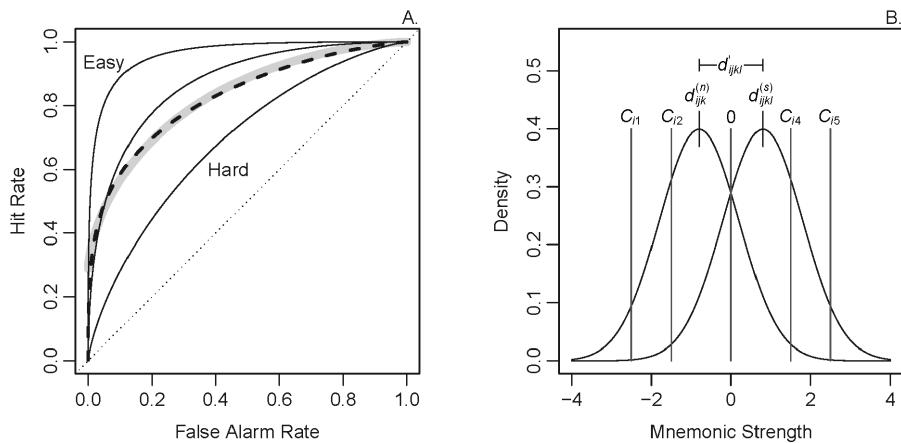
### 9.6.1 Consequences of aggregation in memory experiments

Recognition memory data have traditionally been modeled using the theory of signal detection (Green and Swets, 1966; Kintsch, 1967). Each tested item is assumed to generate some amount of mnemonic *strength*, which is graded and varies from trial to trial. This strength is compared to a criterion; an “old” response is produced if this strength is greater than the criterion, and a new response is produced otherwise. In the most conventional approach, called *equal variance signal detection*, the strength distribution for new items is a standard normal with a mean of 0 and variance of 1, and the strength of new items is shifted by an amount  $d'$ , which serves as a sensitivity parameter. The corresponding hit and false alarm rates are given by

$$\begin{aligned} h &= \Phi(d' - C), \\ f &= \Phi(-C), \end{aligned}$$

where  $\Phi$  denotes the cumulative distribution function (CDF) of the standard normal distribution, and  $C$  denotes the criterion. If hit rates are plotted as a function of false alarm rates for many values of the criterion, the resulting receiver operating characteristic (ROC) curve can be used to assess the veracity of the signal detection model of memory. In particular, this model predicts ROC curves that are curvilinear, as has now been observed in many recognition memory experiments. In addition, this model predicts that the ROC curve will be symmetric about the negative diagonal. The solid black lines in Figure 9.8A correspond to equal-variance signal detection ROCs for  $d' = 0.7, 1.6$ , and  $2.5$ .

The symmetric ROCs in Figure 9.8 are not characteristic of empirical ROC curves observed in recognition memory tasks. In almost all studies, observed ROCs are asymmetric with higher hit rates than expected for small values of false alarms (see Glanzer *et al.*, 1999, for a review). This asymmetric pattern can be seen in the dashed line in Figure 9.8. There have been several models proposed to account for this asymmetry, including signal detection models with strength distributions of unequal variance across new and studied items (e.g., Ratcliff *et al.*, 1992), and signal detection models that assume nongaussian strength distributions (e.g., DeCarlo, 1998; Pratte and Rouder, 2009). Alternatively, Kellen *et al.* (2013) argue that this asymmetry, indeed the curvature in general, is a result of



**Figure 9.8** (A). ROC curves from the equal variance signal detection model (solid black lines), the distorted data from this model averaged over participants (dashed line), and the dual-process model fit to these distorted data (thick gray line). (B). The signal detection component of the hierarchical dual-process model.

aggregation and the true underlying curves are straight lines in accordance with a discrete-state model. Perhaps the most influential account, however, is a *dual-process* model proposed by Yonelinas (1994) and Yonelinas and Parks (2007). This model assumes that the recognition of a previously studied item can come about by one of two separate processes: the item can be explicitly recollected in an all-or-none fashion, or failing recollection, it may be recognized based on its level of familiarity. Familiarity for both new and studied items follows the equal-variance signal detection model presented above. The hit and false alarm rates for this mixture model are given by:

$$h = R + (1 - R) \times \Phi(d' - C), \\ f = \Phi(-C),$$

where  $d'$  and  $C$  are parameters of the signal detection process governing familiarity, and  $R$  is the probability of explicit recollection. The light, thick line in Figure 9.8A shows a typical ROC prediction for this model ( $R = 0.29$ ,  $d' = 1.0$ ). If  $R = 0$ , then the model reduces to the equal-variance signal detection model, and the resulting ROC is symmetric. Recollection,  $R$ , has a one-to-one correspondence with the degree of asymmetry in the ROC curve. Accordingly, the ubiquitous finding of asymmetry in ROC curves is consistent with the presence of two processes mediating recognition memory.

We show here the potential distortions from aggregation in measuring the symmetry of ROC curves. Let's consider the role of item variability, as items are typically aggregated across to form hit and false-alarm rates. Suppose, for demonstration, that there is no recollection. That is, the data from each item follow an

equal-variance signal detection model. Let's also suppose for demonstration that there are two items: an easy item with a true  $d' = 2.5$ , and a harder one, with true  $d' = 0.7$ . The ROCs for these items are shown as the solid lines labeled "Easy" and "Hard" in Figure 9.8A. Now, suppose the hit and false alarm events are averaged over these items. It is hoped that the resulting ROC would reflect the underlying structure, and perhaps be the middle solid line, which is the signal detection model with  $d' = 1.6$ , the average  $d'$  of the easy and hard items. Unfortunately, this ROC does not result from aggregating data. Instead, the dashed line occurs, and this line has a substantial degree of asymmetry. This asymmetry is distortion; an artifact of aggregation, and is not at all a signature of cognitive processing. Perhaps most unsettling is that this distortion is asymptotic – it will remain regardless of how much data are collected (Rouder and Lu, 2005). The dashed line is alarmingly close to the ROC prediction for the two-process model, and researchers who fit models to data aggregated across items run the risk of concluding that there are two processes with substantial recollection, when in fact there is only one process.

The question of whether the data are better described by the dual-process model or by simpler models is important and topical. It cannot be answered with data aggregated across items or individuals, as this aggregation may gravely distort the ROC patterns. To assess whether the asymmetry in ROC curves is a true signature of cognitive process or an artifact of aggregation, we have constructed a series of hierarchical models (Morey *et al.*, 2008a; Pratte *et al.*, 2010; Pratte and Rouder, 2011). In this chapter, we use a hierarchical dual-process model (Pratte and Rouder, 2012) based on Yonelinas' model to assess ROC asymmetry. The degree of asymmetry in this model is indexed by the recollection parameter  $R$ , with  $R = 0$  corresponding to the symmetric curves and greater values of  $R$  corresponding to greater degrees of asymmetry. The main feature of the model is that it accounts for variability across individuals and items, and there is no need to aggregate data for analysis. Consequently, estimates of recollection, which index asymmetry, are not distorted by these nuisance sources of variation.

### 9.6.2 A hierarchical dual-process model of recognition memory

Consider an experiment in which each of  $i = 1, \dots, I$  participants is tested on each of  $j = 1, \dots, J$  items. For each participant, some of these items were indeed studied, while the rest are novel. The participant responds by endorsing one of  $K$  confidence ratings options. In the signal detection approach, the multiple ratings options are modeled with multiple criteria: there are  $K - 1$  criteria as shown in Figure 9.8B. In constructing the hierarchical model, it is useful to reparameterize the signal detection model such that one of the criteria is set to 0, and the center of the new-item distribution is free. We let  $d^{(s)}$  and  $d^{(n)}$  denote the centers of studied and novel-item distributions, respectively.

The hierarchical model is constructed by specifying parameters for each participant-by-item combination. Let  $R_{ij}$  be the participant-by-item recollection value, and let  $d_{ij}^{(s)}$  and  $d_{ij}^{(n)}$  be the participant-by-item values of the centers of the

familiarity distribution for studied and novel items, respectively. The resulting hit and false alarm probabilities for each participant by item combination are

$$h_{ijk} = R_{ij} + (1 - R_{ij}) \times \Phi(d_{ij}^{(s)} - C_{ik}),$$

$$f_{ijk} = \Phi(d_{ij}^{(n)} - C_{ik}),$$

where  $f_{ijk}$  and  $h_{ijk}$  are the false alarm and hit rates for the  $i$ th person responding to the  $j$ th item in the  $k$ th confidence rating. Individual criteria parameters  $C_{ik}$  are also free to vary across participants, reflecting individuals' response biases for particular confidence responses. The familiarity component of the model is depicted in Figure 9.8B.

In this model there are separate parameters for every participant by item combination for novel-item familiarity, studied-item familiarity and recollection. However, because each participant is tested on each item only once, there are no participant-by-item replicates in the data, and thus some constraint is needed. We assume that parameters are additive combinations of person and item effects in order to provide this constraint. The new-item familiarity follows:

$$d_{ij}^{(n)} = \mu^{(n)} + \alpha_i^{(n)} + \beta_j^{(n)},$$

where  $\mu^{(n)}$  denotes a grand mean,  $\alpha_i^{(n)}$  denotes participant effects, and  $\beta_j^{(n)}$  denotes item effects. Rather than place participant and item effects on the mean of studied-item familiarity  $d_{ij}^{(s)}$ , we place them on  $d'_{ij}$ , the difference between the studied-item and new-item distributions:

$$\log(d'_{ij}) = \mu^{(d)} + \alpha_i^{(d)} + \beta_j^{(d)}.$$

Placing an additive model on the log of  $d'_{ij}$  constrains the increase in sensitivity due to study to be positive for all participant by item combinations. Finally, the probability of recollection for each person and item is given by:

$$\Phi^{-1}(R_{ij}) = \mu^{(R)} + \alpha_i^{(R)} + \beta_j^{(R)},$$

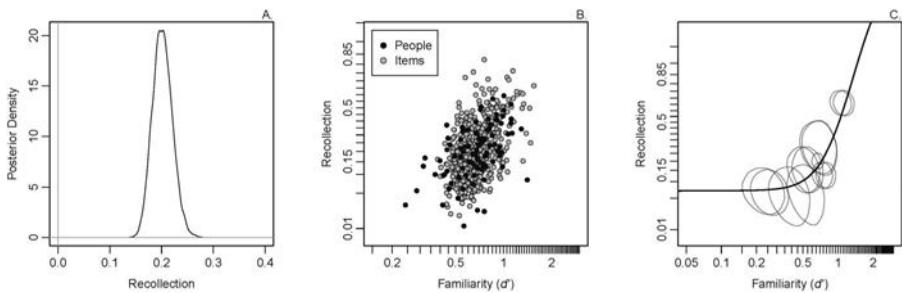
where the inverse of the normal CDF (quantile) function is used to constrain the sum of participant and item effects to be between 0.0 and 1.0, as recollection is a probability.

Although these additive structures greatly simplify the model, there are still a large number of parameters to be estimated. Further constraint is achieved by placing hierarchical structures on participant and item effects. For example, new-item familiarity values follow:

$$\alpha_i^{(n)} \sim \text{Normal}(0, \sigma_\alpha^2)$$

$$\beta_j^{(n)} \sim \text{Normal}(0, \sigma_\beta^2),$$

where the variance parameters are estimated from the data, and provide measures of participant and item variability in new-item familiarity. Similar hierarchical structures are placed on participant and item effects in studied-item familiarity and



**Figure 9.9** (A). Posterior distribution of mean recollection, estimated with the hierarchical dual-process model. (B). Participant and item effects in recollection plotted as a function of effects in familiarity. (C). Joint posterior distributions of recollection and familiarity for 13 experimental conditions. The line is a nonparametric fit, highlighting the monotonic relationship between recollection and familiarity estimates.

recollection, providing for efficient parameter estimation even with small numbers of participants and items.

### 9.6.3 Applications of the hierarchical memory model

The hierarchical model allows for the estimation of underlying mnemonic processes from recognition memory data without recourse to aggregation and the accompanying distortions. The presented model is discussed in detail in Pratte and Rouder (2011, 2012), and estimation may be performed with the R package *HBMEM*, available on CRAN.

One of the main questions is whether the asymmetry in ROC curves is truly the result of cognitive processing, such as all-or-none recollection, or reflects distortion that results from averaging data over participants or items, as is demonstrated in Figure 9.8. This question can be answered by consideration of the mean recollection parameter ( $\mu^{(R)}$ ), a measure of ROC asymmetry that in the hierarchical model is uncontaminated by participant and item variability. If posterior beliefs are centered far from zero, then the ubiquitous ROC asymmetry is indeed a cognitive signature rather than an artifact. We applied the model to Experiment 1 in Pratte *et al.* (2010), a large recognition memory experiment in which 94 participants were tested on 480 items. The resulting posterior distribution for the mean recollection parameter ( $\mu^{(R)}$ ) is shown in Figure 9.9A. All of the posterior mass is substantially above zero, implying that ROC asymmetry is present even when participant and item variability are modeled. This asymmetry is seemingly a cognitive signature rather than an artifact of aggregation (see Morey *et al.*, 2008a; Pratte *et al.*, 2010), and should be treated as an important benchmark in theory construction.

Although both the aggregated and hierarchical analysis of these data provide the same qualitative conclusion of ROC asymmetry, aggregation nonetheless leads to

distorted parameter estimates and a dramatic overestimation of precision of these estimates. Familiarity, for example, is overestimated by 15% when data are aggregated over both participants and items, compared to mean familiarity in the hierarchical model. More alarming is the dramatic overestimation of precision from aggregation. For example, the 95% credible interval on mean recollection in Figure 9.9A is 2.7 times larger than the 95% confidence interval resulting from the aggregated analysis. This overstatement of precision from aggregation is a direct result of mismodeling multiple sources of variation and is well known (Clark, 1973; Raudenbush and Bryk, 2002; Rouder and Lu, 2005). Conversely, the wider credible intervals from the hierarchical estimates directly represent the uncertainty from accounting for multiple sources of variation. The differing degrees of precision has dramatic effects on assessing whether mean recollection or familiarity changes with condition variables, and it is possible that previous demonstrations of effects with aggregated data are overstatements of the true significance of condition effects (Pratte and Rouder, 2012).

The above assessment shows that ROC asymmetry is a signature of the cognitive processes subserving recognition memory, but does not necessarily imply that recognition memory is mediated by recollection and familiarity. The hierarchical model provides additional insights because it provides for separate assessment of recollection and familiarity for each individual and for each item. If recollection and familiarity are statistically independent processes, then the recollection and familiarity across individuals should be uncorrelated; likewise, recollection and familiarity across items should be uncorrelated. The dark points in Figure 9.9B show the relationship between recollection and familiarity for people ( $\alpha_i^{(r)}$  vs.  $\alpha_i^{(d)}$ ), and the light points show the relationship for items ( $\beta_j^{(r)}$  vs.  $\beta_i^{(d)}$ ). Two trends are evident. First, there is substantial variability in both individuals' mnemonic abilities and how easily items are remembered. Second, there is substantial correlation: people with high recollection also have high familiarity ( $r = .48$ ), and items with high recollection also have high familiarity ( $r = .49$ ). Pratte and Rouder (2011) found that the degree of correlation is statistically significant, but, nonetheless, a model with only shared variability does not do as well as a model with both shared and unique variability for recollection and familiarity.

One of the main sources of evidence for two separable processes has been the demonstrations of dissociations across experimental conditions. One classic dissociation is between a levels-of-processing manipulation and a perceptual-feature manipulation. Deep levels of study, such as producing a related word to an item at study, should lead to an increase in recollection over shallow levels of study, such as counting vowels in study items. Conversely, changing perceptual features between study and test, such as font or color, should attenuate familiarity rather than recollection. Some researchers have had success in generating these dissociations, but they seemingly occur only under special circumstances. In particular, perceptual effects are difficult to obtain (Hockley, 2008; Mulligan *et al.*, 2010; Murnane and Phelps, 1995), and tend to occur only in experiments with poor overall performance (e.g. Boldini *et al.*, 2004).

In Pratte and Rouder (2012), we used the hierarchical model to assess recollection and familiarity across 13 conditions in 4 experiments. Our manipulations produced effects that were as large or larger than previous ones in the literature. Figure 9.9C shows joint posterior distributions of mean recollection as a function of mean familiarity across the conditions. Each ellipse is a 95% credible region. If there was evidence for two distinct processes, then these conditions should lie in a plane rather than on a monotonic curve (see Bamber, 1979, and Newell and Dunn, 2008, for an overview of the logic in interpreting such *state-trace* plots). Note that the curve is not incompatible with a double dissociation: some pairs of points differ more in familiarity than recollection (see the poorest performing points), whereas others differ more in recollection than familiarity (see the best-performing points). Yet, all of the condition effects can be connected by an increasing curve, suggesting that a single factor, the location on the curve, is needed to account for these data. We think the relative attenuation of recollection in conditions with poor performance reflects the nature of ROC space. When performance is poor, the ROCs are near the diagonal and it is easier to detect small overall sensitivity effects (familiarity) than to detect small changes in asymmetry (recollection). Hence, even though our data have degrees of dissociation as large as any in the comparable literature, they are more compatible with a single-process than a dual-process approach.

## 9.7 Concluding remarks

In this chapter, we have shown that while experimental psychologists have a rich theoretical and experimental tradition, the link between theory and data often presents difficulties in real-world contexts. These difficulties arise because theories are nonlinear, and there is often substantial nuisance variation across individuals and items. If these sources of nuisance variation are not appropriately modeled, they will distort the assessment of the underlying cognitive signatures, and lead to erroneous conclusions about theory. These potential problems occur across psychology, and here we have presented examples in assessing learning, subliminal priming, and recognition memory.

We advocate a Bayesian hierarchical approach for linking theory and data. These models provide for the simultaneous assessment of nuisance variation and variation from the target cognitive process of interest. They not only allow researchers to uncover the rich cognitive structure in their data without aggregation artifacts, but allow for an understanding of how this structure varies across individuals and items.

In this chapter, we have tried to focus on the types of problems hierarchical modeling can solve, as well as an introduction to Bayesian probability. We have avoided the nuts and bolts of estimation, and this avoidance leaves open the question of how interested researchers can develop and analyze their own models. There are now several excellent texts on Bayesian modeling that include development of Bayesian hierarchical models, and advanced texts include Gelman *et al.* (2004)

and Jackman (2009). More recently there have been tutorials and texts specific for psychology, including Rouder and Lu (2005), Kruschke (2011), and the book by Lee and Wagenmakers (2013). Here, we tackle more global questions about how researchers should learn Bayesian hierarchical modeling.

One question that arises is about software: which language and packages should researchers use? We think researchers should invest in three classes of languages. At the highest level, there are specialty languages developed especially for Bayesian hierarchical modeling, of which JAGS (Plummer, 2003) and WinBUGS (Lunn *et al.*, 2000) are the most popular. These languages allow researchers to specify models and priors as input in a natural random-variable notation, and provide samples from posterior distributions as output. When they work, they often work well and save much development time. Therefore, these specialty languages serve as an excellent first option, and, importantly, require little special knowledge above and beyond the skills needed to specify models. Unfortunately, as general-purpose sampling solutions, they sometimes do not work well in specific situations: they may lack a feature necessary to define a model, or take an exceedingly long time to sample.<sup>5</sup> Determining whether a specialty language such as JAGS or WinBUGS will work is often fast and should be a first step for most researchers.

In cases where the general-purpose solutions fail, researchers may need to derive conditional posterior distributions, develop sampling routines, and implement them. Data-analytic languages such as R (R Development Core Team, 2009) and MATLAB (MATLAB, 2010) are ideal for implementation, and often contain useful routines for MCMC sampling. Sometimes, however, the speed of R and MATLAB can be improved by implementing the sampling in a fast, low-level language such as C or Fortran. We use JAGS as our high-level specialty language, R as our mid-level data-analytic language, and C as our fast, low-level language, and we routinely move between these three as dictated by the model we wish to analyze. The hierarchical normal model and the mass-at-chance model in this chapter are both implementable in JAGS; analysis of the hierarchical dual-process model, however, was more convenient using a combination of R and C routines for efficiency. Our hope is that as more researchers use hierarchical models, they will develop the skills to go beyond WinBUGS or JAGS implementations as needed.

Perhaps the most important question is how young scholars should be trained so that they may use Bayesian hierarchical models. In our view, it is hard to overstate the usefulness of solid training in statistics including courses in calculus-based mathematical statistics, linear algebra, and Bayesian analysis. We realize that many talented students will not have the aptitude or time for such study, and so it is worthwhile to consider alternatives. A good course would be one that stresses the logic of modeling. This course would focus on the basics of probability and statistics, and promote a deep understanding of conditional probability. Course objectives would include the ability to specify models, and write down and visualize likelihoods,

<sup>5</sup> Fortunately, these general-purpose samplers are extensible (Lunn, 2003) and have improved greatly in recent years. In addition, newcomers such as Stan (Stan Development Team, 2013) show promise.

and would provide an overview of the issues in model comparison. We hope the appeal of Bayesian hierarchical models will motivate more rigorous general statistical training in psychology.

## References

- Anders, R. and Batchelder, W. H. (2012). Cultural consensus theory for multiple consensus truths. *Journal of Mathematical Psychology*, **56**, 452–469.
- Ashby, F. G., Maddox, W. T. and Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, **5**, 144–151.
- Averell, L. and Heathcote, A. (2011). The form of forgetting curve and the fate of memories. *Journal of Mathematical Psychology*, **55**, 25–35.
- Baayen, R. H., Tweedie, F. J. and Schreuder, R. (2002). The subjects as a simple random effect fallacy: subject variability and morphological family effects in the mental lexicon. *Brain and Language*, **81**, 55–65.
- Bamber, D. (1979). State trace analysis. A method of testing simple theories of causation. *Journal of Mathematical Psychology*, **19**, 137–181.
- Bayarri, M. J. and Garcia-Donato, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, **94**, 135–152.
- Boldini, A., Russo, R. and Avons, S. E. (2004). One-process is not enough! A speed-accuracy tradeoff study of recognition memory. *Psychonomic Bulletin and Review*, **11**, 353–361.
- Busemeyer, J. R. and Diederich, A. 2009. *Cognitive Modeling*. Thousand Oaks, CA: Sage.
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, **12**, 335–359.
- DeCarlo, L. M. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, **3**, 186–205.
- Dehaene, S., Naccache, L., Le Clech, G., *et al.* (1998). Imaging unconscious semantic priming. *Nature*, **395**, 597–600.
- Dickey, J. M. and Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, **41**, 214–226.
- Doucet, A., de Freitas, N. and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. New York, NY: Springer.
- Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193–242.
- Estes, W. K. (1956). The problem of inference from curves based on grouped data. *Psychological Bulletin*, **53**, 134–140.
- Farrell, S. and Ludwig, C. J. H. (2008). Bayesian and maximum likelihood estimation of hierarchical response time models. *Psychonomic Bulletin and Review*, **15**, 1209–1217.
- Finkbeiner, M. (2011). Subliminal priming with nearly perfect performance in the prime-classification task. *Attention, Perception, & Psychophysics*, **73**, 1255–1265.

- Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, **116**, 439–453.
- Gelfand, A. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis (2nd edition)*. London: Chapman and Hall.
- Gelman, A., Shor, B., Bafumi, J. and Park, D. (2007). Rich state, poor state, red state, blue state: what's the matter with connecticut? *Quarterly Journal of Political Science*, **2**, 345–367.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.
- Glanzer, M., Kim, K., Hilford, A. and Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **25**, 500–513.
- Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York, NY: Wiley.
- Greenwald, A. G. (1992). New look 3: unconscious cognition reclaimed. *American Psychologist*, **47**, 766–779.
- Greenwald, A. G., Klinger, M. R. and Schuh, E. S. (1995). Activation by marginally perceptible (“subliminal”) stimuli: dissociation of unconscious from conscious cognition. *Journal of Experimental Psychology: General*, **124**, 22–42.
- Haider, H. and Frensch, P. A. (2002). Why aggregated learning follows the power law of practice when individual learning does not: comment on Rickard (1997, 1999), Delaney *et al.* (1998), and Palmeri (1999). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **28**, 392–406.
- Heathcote, A., Brown, S. and Mewhort, D. J. K. (2000). The power law repealed: the case for an exponential law of practice. *Psychonomic Bulletin and Review*, **7**, 185–207.
- Hockley, W. E. (2008). The effects of environmental context on recognition memory and claims of remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **34**, 1412–1429.
- Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. Chichester: John Wiley & Sons.
- Jeffreys, H. (1961). *Theory of Probability (3rd edition)*. New York, NY: Oxford University Press.
- Karabatsos, G. and Batchelder, W. H. (2003). Markov chain estimation methods for test theory without an answer key. *Psychometrika*, **68**, 373–389.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.
- Kellen, D., Klauer, K. and Broder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin & Review*, **20**, 1–27.
- Kemp, C., Perfors, A. and Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, **10**, 307–321.
- Kintsch, W. (1967). Memory and decision aspects of recognition learning. *Psychological Review*, **74**, 496–504.

- Koechlin, E., Naccache, L., Block, E. and Dehaene, S. (1999). Primed numbers: exploring the modularity of numerical representations with masked and unmasked semantic priming. *Journal of Experimental Psychology: Human Perception and Performance*, **25**, 1882–1905.
- Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, **6**, 299–312.
- Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. (2004). *Applied Linear Statistical Models*. Chicago, IL: McGraw-Hill/Irwin.
- Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, **30**, 1–26.
- Lee, M. D. and Wagenmakers, E.-J. (2013). *Bayesian Modeling for Cognitive Science: A Practical Course*. Cambridge: Cambridge University Press.
- Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, **103**, 410–423.
- Lindley, D. V. (1957). A statistical paradox. *Biometrika*, **44**, 187–192.
- Luce, R. D. (1986). *Response Times*. New York, NY: Oxford University Press.
- Lunn, D. (2003). WinBUGS development interface (WBDev). *IBSA Bulletin*, **10**.
- Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.
- Mandler, G. (1980). Recognizing: the judgment of previous occurrence. *Psychological Review*, **87**, 252–271.
- MATLAB. (2010). *Version 7.10.0 (R2010a)*. Natick, MA: The MathWorks Inc.
- Meng, X. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, **6**, 831–860.
- Merikle, P. M., Smilek, D. and Eastwood, J. D. (2001). Perception without awareness: perspectives from cognitive psychology. *Cognition*, **79**, 115–134.
- Merkle, E., Smithson, M. and Verkuilen, J. (2011). Hierarchical models of simple mechanisms underlying confidence in decision making. *Journal of Mathematical Psychology*, **55**, 57–67.
- Morey, R. D. and Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, **16**, 406–419.
- Morey, R. D., Rouder, J. N., Pratte, M. S. and Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, **55**, 368–378.
- Morey, R. D., Pratte, M. S. and Rouder, J. N. (2008a). Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology*, **52**, 376–388.
- Morey, R. D., Rouder, J. N. and Speckman, P. L. (2008b). A statistical model for discriminating between subliminal and near-liminal performance. *Journal of Mathematical Psychology*, **52**, 21–36.
- Morey, R. D. (2008). *Item Response Models for the Measurement of Thresholds*. PhD thesis, University of Missouri.
- Morey, R. D., Rouder, J. N. and Speckman, P. L. (2009). A truncated-probit item response model for estimating psychophysical thresholds. *Psychometrika*, **74**, 603–618.

- Mulligan, N. W., Besken, M. and Peterson, D. (2010). Remember–know and source memory instructions can qualitatively change old-new recognition accuracy: the modality-match effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **36**, 558–566.
- Murnane, K. and Phelps, M. P. (1995). Effects of changes in relative cue strength on context-dependent recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **21**, 158–172.
- Murphy, S. T. and Zajonc, R. B. (1993). Affect, cognition, and awareness: affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology*, **64**, 723–739.
- Myung, I.-J. and Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychonomic Bulletin and Review*, **4**, 79–95.
- Myung, I.-J., Kim, K. and Pitt, M. A. (2000). Toward an explanation of the power law artifact: insights from response surface analysis. *Memory & Cognition*, **28**, 832–840.
- Naccache, L. and Dehaene, S. (2001). Unconscious semantic priming extends to novel unseen stimuli. *Cognition*, **80**, 215–229.
- Newell, B. R. and Dunn, J. C. (2008). Dimensions in data: testing psychological models using state–trace analysis. *Trends in Cognitive Sciences*, **12**, 285–290.
- Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. Hoboken, NJ: Wiley.
- Plummer, M. (2003). JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.
- Pratte, M. S. and Rouder, J. N. (2009). A task-difficulty artifact in subliminal priming. *Attention, Perception, & Psychophysics*, **71**, 276–283.
- Pratte, M. S. and Rouder, J. N. (2011). Hierarchical single- and dual-process models of recognition memory. *Journal of Mathematical Psychology*, **55**, 36–46.
- Pratte, M. S. and Rouder, J. N. (2012). Assessing the dissociability of recollection and familiarity in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **38**, 1591–1607.
- Pratte, M. S., Rouder, J. N. and Morey, R. D. (2010). Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **36**, 224–232.
- R Development Core Team. (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Ratcliff, R., Sheu, C. F. and Grondlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, **99**, 518–535.
- Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods. Second Edition*. Thousand Oaks, CA: Sage.
- Reder, L. M. and Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**, 435–451.
- Rickard, T. C. (2004). Strategy execution in cognitive skill learning: an item-level test of candidate models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **30**, 65–82.

- Rogers, S. (1992). How a publicity blitz created the myth of subliminal advertising. *Public Relations Quarterly*, **37**, 12–17.
- Rouder, J. N. and Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin and Review*, **12**, 573–604.
- Rouder, J. N. and Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, **18**, 682–689.
- Rouder, J. N. and Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, **47**, 877–903.
- Rouder, J. N., Morey, R. D., Cowan, N. and Pfaltz, M. (2004). Learning in a unidimensional absolute identification task. *Psychonomic Bulletin and Review*, **11**, 938–944.
- Rouder, J. N., Morey, R. D., Speckman, P. L. and Pratte, M. S. (2007a). Detecting chance: a solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin and Review*, **14**, 597–605.
- Rouder, J. N., Lu, J., Sun, D., *et al.* (2007b). Signal detection models with random participant and item effects. *Psychometrika*, **72**, 621–642.
- Rouder, J. N., Tuerlinckx, F., Speckman, P. L., Lu, J. and Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin & Review*, **15**(1201–1208).
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. and Iverson, G. (2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, **16**, 225–237.
- Rouder, J. N., Morey, R. D., Speckman, P. L. and Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, **56**, 356–374.
- Rouder, J. N., Morey, R. D. and Province, J. M. (2013). A Bayes-factor meta-analysis of recent ESP experiments: a rejoinder to Storm, Tressoldi, and Di Risio (2010). *Psychological Bulletin*, **139**, 241–247.
- Sarbanés Bové, D. and Held, L. (2011). Hyper-*g* priors for generalized linear models. *Bayesian Analysis*, **6**, 1–24.
- Shiffrin, R. M., Lee, M. D., Kim, W. and Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, **32**, 1248–1284.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **64**, 583–639.
- Stan Development Team. (2013). *Stan: A C++ Library for Probability and Sampling, Version 1.1*.
- Vandekerckhove, J., Verheyen, S. and Tuerlinckx, F. (2010). A cross random effects diffusion model for speeded semantic categorization decisions. *Acta Psychologica*, **133**, 269–282.
- Verdinelli, I., and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *Journal of the American Statistical Association*, **90**, 614–618.
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of *p* values. *Psychonomic Bulletin and Review*, **14**, 779–804.

- Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H. and Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the Savage–Dickey method. *Cognitive Psychology*, **60**, 158–189.
- Walker, S. G., Laud, P. W., Zanterdeschi, D. and Damien, P. (2011). Direct samping. *Journal of Computational and Graphical Statistics*, **20**, 692–713.
- Wetzels, R., Grasman, R. P. P. P. and Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage–Dickey density ratio. *Computational Statistics and Data Analysis*, **54**, 2094–2102.
- Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **20**, 1341–1354.
- Yonelinas, A. P. and Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: a review. *Psychological Bulletin*, **133**, 800–832.
- Zeigenfuse, M. D. and Lee, M. D. (2010). Finding the features that represent stimuli. *Acta Psychologica*, **133**, 283–295.
- Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In: Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds), *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*. Valencia: University of Valencia, pp. 585–603.

# 10 Model evaluation and selection

Jay Myung, Daniel R. Cavagnaro, and Mark A. Pitt

10.1	Introduction	553
10.2	Basic ideas	555
10.3	Model estimation	556
10.3.1	How is a model specified?	556
10.3.2	Formal definition of a model	558
10.3.3	Parameter estimation	560
10.4	Model evaluation	562
10.4.1	How should a model be evaluated?	562
10.4.2	Goodness-of-fit and the overfitting problem	563
10.4.3	Model complexity	565
10.4.4	Generalizability	566
10.4.5	Relationship among goodness-of-fit, complexity, and generalizability	567
10.5	Model selection	568
10.5.1	Penalized-likelihood model selection	569
10.5.2	Cross-validation and accumulative prediction error	571
10.5.3	Bayesian model selection	573
10.5.4	Illustrated example	574
10.5.5	Summary	576
10.6	Design optimization	577
10.6.1	Further improving model selection through design optimization	577
10.6.2	Design optimization	579
10.6.3	Adaptive design optimization	581
10.6.4	Illustrative example	582
10.6.5	Limitations	584
10.7	General discussion	585
Acknowledgments		588
Appendix: Source code		588
Matlab code		588
R code		592
References		595

## 10.1 Introduction

The study of cognition is challenged by the difficulty of inferring representation and processes in such a complex system as the brain. The field of cognitive science has met this challenge by borrowing and developing research tools with which to study the brain. *Tools* are meant broadly to include not just hardware (e.g., computers, eye trackers, imaging equipment) that is used for data collection, but also the quantitative tools used to guide inference, including statistical methods (frequentist and Bayesian) and cognitive modeling.

Cognitive modeling assists in scientific inference by, among other things, assessing the plausibility of an explanation (e.g., theory, process). It achieves this by instantiating a version of the explanation in some quantitative form (i.e., the mathematical model), and thereby demonstrating its plausibility (Polk and Seifert, 2002; Busemeyer and Diederich, 2010; Lewandowsky and Farrell, 2011; Lee and Wagenmakers, 2014). However, as with theories, all quantitative explanations are not equally good or convincing, so what criteria should be used to evaluate models? What makes a model a good explanation from which it is reasonable to draw inferences, and what signs indicate the model is poor? These questions are the focus of this chapter. Like modeling itself, the field is still very much in its infancy. Progress has been made, but many challenges remain. Before reviewing the state-of-the-art, we first provide a broader context in which to situate the enterprise of model evaluation.

Although cognitive modeling has been around since the 1950s, its popularity increased once computers became cheap and fast. Also, user-friendly software has accelerated its adoption, to the point where more and more researchers recognize the value of models and their usefulness for knowledge discovery (Shiffrin and Nobel, 1997; Fum *et al.*, 2007; McClelland, 2009). Theories in much of the field tend to be broad claims about foundational issues in cognition (e.g., representations are distributed rather than local; grammar acquisition is probabilistic rather than rule-based; category learning is Bayesian). By instantiating these claims in a model, the theory becomes more viable, and as a consequence more persuasive, especially when its performance is shown to mimic that of individuals. In addition, it can be difficult to develop a theory with much depth without formalizing it quantitatively in some way. The reason for this stems from the challenge in understanding how the many parts of the theory (e.g., constructs, processes) operate and interact to lead to a specific outcome. The more complex the theory, the more difficult it is to make predictions with certainty (a version of this problem haunts models as well, which we will discuss at the end of the chapter). Quantification of a theory provides a framework in which these interactions can be explored systematically so that they are understood, and then tested in future experiments.

For example, a theory of memory should include an explanation of forgetting. Initially it might posit only that memory decays as a function of the time between studying a list of words and being asked to recall them. Such a claim is specific enough to evaluate it experimentally, and might then lead to the further assertion

that the relationship between recall and test delay is best described by a power function, with recall performance decreasing quickly over the first few time delays and then gradually leveling off over longer delays. At this point a simple model starts to emerge, and multiple roads can be taken. One is to demonstrate that other quantitative relationships (e.g., exponential, hyperbolic) between test delay and recall provide poorer descriptions of the data. A more ambitious goal, and one that is of more interest to the researcher, is to expand the model to specify the memory processes and how they operate to yield retention that decays according to a power function. Although the first path of exploration might constrain solutions to the second, it can be nontrivial to accomplish even the more modest goal of determining the shape of the retention function (e.g., Rubin and Wenzel, 1996; Navarro *et al.*, 2004). The reasons for this are discussed later in the chapter.

The preceding example is intended to draw attention to a few aspects of cognitive modeling and their implications for model evaluation. One is that modeling requires the researcher to specify a mathematical formulation of the theory, however simple or tenuous it might be. This initial implementation then becomes a stand-in for the theory, serving as a surrogate that is evaluated on its own, with the theory lurking somewhere in the distance. A second implication is that no matter how thoughtful or careful a researcher is in developing an implementation, it is instructive to remember that it will always be wrong, exemplified by the famous quote, “All models are wrong, but some are useful” (G. E. P. Box, 1976). For a model to be even in the ballpark of possibilities, one would have to have a vast amount of very informative and precise (noise-free) data. Even then, the task of inferring the true form of the underlying model seems so daunting that one might wonder whether the approach is misguided. The tools in cognitive science are rarely able to provide the strong constraints needed to achieve such a lofty objective.

It is for these reasons that models are most productively viewed as tools for studying cognition. A quantitative framework serves to direct inquiry toward particular issues, whether testing basic model assumptions, implementational choices, or predictions about variable interactions. If a model serves as a useful explanatory device and has advanced understanding in a field, then it has done its job, even if it is ultimately abandoned in favor of an alternative.

Again, how do we determine if a model is useful? An answer to this question requires careful evaluation of the model itself and can be nontrivial to obtain, but it behooves the researcher to scrutinize the design choices made in creating the model. What is needed for this purpose is yet another set of tools to evaluate the models themselves. More are being developed, but they lag behind modeling itself. Although many of the methods themselves were developed in the context of mathematical models, the fundamental issues in model evaluation are pertinent to all types of models.

Of course, the types of questions about model performance that one asks can depend on the modeling framework adopted, whether it be cognitive architectures, parallel distributed processing (PDP) systems, or Bayesian models. That so many

styles of modeling flourish simultaneously in the cognitive sciences underscores the challenge of modeling all of cognition using a single framework. Instead, each style of modeling seems best suited for a particular content domain (McClelland, 2009).

In this chapter, we focus on the evaluation of and selection among mathematical models, which are defined as models that can be expressed in algebraic form and have a likelihood function. They include models of decision making, memory retention, psychophysical models, and some models of categorization. In each content area, models differ in the number of parameters (never more than four) and functional form (how the parameters and input are combined in the model equation). You might think that such simple models should be easy to choose among, but as we discuss in the following pages there are challenges at every turn. Because these models are similar to those developed in fields such as statistics and engineering, model evaluation methods developed in those domains have been imported into cognitive science.

We devote the majority of this chapter to methods that assess a model's suitability in accounting for data collected in an experiment. These methods are well-known, and include measuring a model's fit to data as well as its flexibility. The final part of the chapter introduces computational methods that improve model selection by optimizing the design of the experiment used to discriminate the models. That is, knowledge of model behavior is used to identify experimental designs (e.g., choices of stimuli) that have the highest likelihood of discriminating models. We begin with a discussion of the fundamental problems in model evaluation and selection.

## 10.2 Basic ideas

In order to develop appropriate tools for evaluating models and selecting among them, we must first answer the questions of what makes a model good, and what makes one model better than another. A common misconception about modeling is that the goal is to “fit” the data as well as possible. This misconception probably stems from the fact that when people are first introduced to mathematical modeling, usually in the form of simple linear regression, the focus is on model fitting (i.e., finding the parameters of the model that best fit the data; see Section 10.3). Model fitting (e.g., parameter estimation) is an important aspect of mathematical modeling, but modeling is about much more than just finding a set of best-fitting parameters. A model entails assumptions about the structure of data and the relationships between variables. For example, a simple linear regression model assumes a linear relationship between the independent and dependent variables, and a normal distribution of the dependent variable at each level of the independent variable. If our goal were to fit the data as well as possible, why limit ourselves to a linear relationship with just two parameters (slope and intercept) when a better fit could be obtained by assume a more complex

relationship such as a polynomial with three, for, or even five parameters? Why not 50 parameters?

John Von Neumann famously said, “With four parameters, I can fit an elephant, with five I can make him wiggle his trunk.” By this he meant that one should not be impressed when a complex model fits a data set well. With enough parameters, you can fit any data set. A model with a lot of parameters is said to be “complex” because it can fit complex patterns of data. While data fitting is important, a complex model can fit for the wrong reasons, by fitting noise instead of regularities. Although we aim to precisely control the conditions of our experiments, real data are awash with idiosyncrasies due to individual differences, quirks, and nuance that cannot be controlled. These idiosyncrasies are commonly referred to as “noise.” All other things being equal, simpler models are more attractive because they are sufficiently constrained to make them easily falsifiable. The issue of model complexity will be addressed in Section 10.4.3.

What is desired for model evaluation is a yardstick that measures a model’s ability to capture the underlying regularity only, not idiosyncratic noise. This requires a balance between goodness of fit and complexity, key concepts in model evaluation that are elaborated in the following sections.

## 10.3 Model estimation

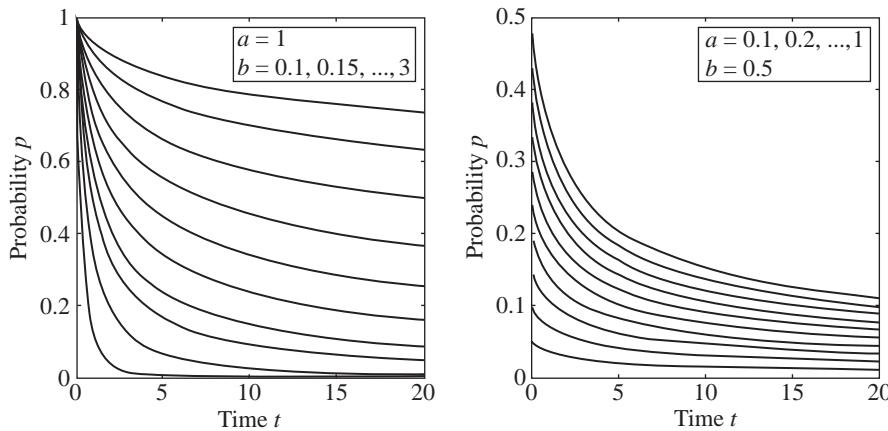
### 10.3.1 How is a model specified?

As noted above, models are quantitative stand-ins for theories. A formal model expresses the theorized relationships between latent (not directly observable) processes (e.g., memory, attention) and observable responses using the language of mathematics. A mathematical mode is defined in terms of a mathematical equation that specifies the range of data patterns it predicts by varying the values of model parameters. As a concrete example, consider the following power model of memory retention (e.g., Wixted and Ebbesen, 1991):

$$\text{Power model : } p(a, b|t) = a(t + 1)^{-b} \quad (10.1)$$

where  $p(a, b|t)$  denotes the model’s prediction of the probability of correct response at retention interval  $t$ , and  $a$  ( $0 < a < 1$ ) and  $b$  ( $b > 0$ ) are the model’s two parameters. Shown on the left panel of Figure 10.1 are 10 different power curves created by varying the value of parameter  $b$  with parameter  $a$  fixed to 1. The right panel of the same figure shows another 10 power curves generated by varying the value of parameter  $a$  with parameter  $b$  set to 0.5. Note the diversity of memory decay patterns that the power model predicts for different choices of its parameters  $a$  and  $b$ .

Writing down the model equation such as in Equation (10.1) is an important first step of model specification, but is not its end. This is because the *deterministic* equation, as it is, represents an idealistic but unrealistic view of the mental and



**Figure 10.1** Sample power curves generated by varying values of model parameters ( $a, b$ ) in Equation (10.1).

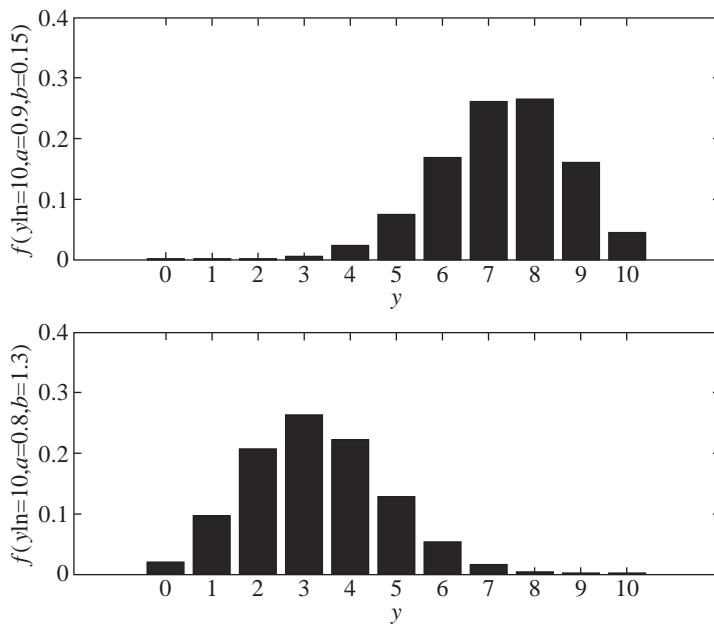
behavioral processes that the model is trying to capture. Experimental data are inevitably corrupted by random variability, whether it is due to the variability of experimental procedures or due to the imprecision of the measurement instrument. The same experimental stimulus does not necessarily invoke the same behavioral response from participants, often not even from the same participant. It is therefore important for a model to also specify how the random variability in the observed data is accounted for, in addition to the theorized regularity (memory decay rate) underlying the data.

To show how this is done, let us consider a retention memory experiment in which a participant first studies a list of  $n$  words in a study phase and then in a test phase is asked to recall the studied words. The time lag between the study and test phase defines the retention interval  $t$ . The power function in Equation (10.1) specifies the probability of correct recall of a word at time  $t$  given a parameter vector  $(a, b)$ . For example, for  $t = 1$  and  $(a = 0.9, b = 0.15)$ , the power model predicts the probability to be equal to  $p(a, b|t) = 0.9(1 + 1)^{-0.15} = 0.731$ . Assuming that the participant recalls each word independently of the others, the number of correctly recalled words, denoted by  $y$ , out of the  $n$  words in the studied list would follow the binomial probability distribution:

$$f(y|n, (a, b)) = \frac{n!}{(n-y)!y!} p(a, b|t)^y (1 - p(a, b|t))^{n-y} \quad (10.2)$$

where  $y = 0, 1, \dots, n$ .

Figure 10.2 depicts two binomial probability distributions corresponding to different choices of the parameter vector,  $(a = 0.9, b = 0.15)$  (top panel) and  $(a = 0.8, b = 1.3)$  (bottom panel), given  $n = 10$  and  $t = 1$ . The top panel shows the distribution of the number of correct responses when the probability of correct recall of one word is equal to 0.731. While there is a relatively high probability of



**Figure 10.2** Example probability distributions. Both distributions are obtained from the binomial probability distribution in Equation (10.2) for two different values of the parameter vector  $(a, b)$  indicated on the y-axis of each graph.

observing 7 or 8 correct responses out of 10, as one might expect, it is still possible to observe lower or higher numbers, and even much lower (e.g., 3) correct responses. This “random” variability is due to the binomial sampling plan used during data collection in an experiment; that is, only the overall number of correct responses, not the identity of individual words that are correctly recalled, is collected and recorded on each trial of the experiment.

To summarize, a mathematical model can be viewed as made of two sub-components: (1) a deterministic component that describes the theorized functional relationships between latent and observable variables; and (2) a stochastic noise component that accounts for random variability, often due to sampling error and/or imprecise measurements. In what follows, we formalize this notion in statistical language.

### 10.3.2 Formal definition of a model

The goal of modeling is to capture the essential features of complex behavior with a simplified mathematical model. The model is usually motivated and developed from behavioral data. From a statistical standpoint, the data consist of a set of observations, denoted by a vector  $y = (y_1, \dots, y_n)$ , as a random sample drawn

from an unknown population, or equivalently, the probability distribution that specified the probability of observing a value of  $y$ . The underlying probability distribution is to be inferred from an actual observed value of  $y$ , often referred to as just “the data.”

Suppose that we have a model  $M$  with  $k$  free parameters denoted by a parameter vector  $\theta = (\theta_1, \dots, \theta_k)$ . Formally, a model consists of a parametric collection of probability distributions indexed by the parameter  $\theta$ . That is, associated with each value of  $\theta$  is a unique probability distribution such that as the parameter changes in value, different probability distributions are identified, as illustrated in Figure 10.2. It is assumed that one of the model’s probability distributions corresponds to the population underlying the data.

To capture the functional role of the data variable  $y$ , the probability distribution associated with each parameter value is called the *probability density function* (PDF), denoted by  $f(y|\theta)$ , which specifies the probability of observing the particular value of  $y$  given a fixed parameter  $\theta$ . By definition, the total probability must be equal to one,  $\sum_y f(y|\theta) = 1$  or  $\int f(y|\theta) dy = 1$  for all  $\theta$ , depending upon whether the random variable  $y$  is discrete or continuous, respectively. Under the assumption that the  $n$  observations are independently distributed, the PDF of the data vector  $y = (y_1, \dots, y_n)$  can be rewritten as a product of individual PDFs

$$f(y = (y_1, \dots, y_n)|\theta) = \prod_{i=1}^n g_i(y_i|\theta) \quad (10.3)$$

where each  $g_i(y_i|\theta)$  is the PDF of each  $y_i$ . The further assumption that the observations are *identically* distributed leads to a simpler expression of  $g_i(y_i|\theta) = g(y_i|\theta)$ ,  $i = 1, \dots, n$ .

As a concrete example of Equation (10.3), consider a lexical decision experiment in which participants are asked to decide as quickly as possible whether a string of letters is a word or a nonword, and the time taken to correctly classify words is measured as a function of word frequency. Suppose that the results suggest that the higher the word is in frequency, the faster the response time is to that word. Many different forms of mathematical functions can capture this qualitative relationship, including the following exponential model equation:

$$y_i = ae^{-bx_i} + c + e_i \quad (10.4)$$

where  $y_i$  is the lexical decision time in seconds for word  $i$ ,  $x_i$  is the word frequency measured in an appropriate unit,  $a$ ,  $b$  and  $c$  are positive parameters, and finally,  $e_i$  is a normal error term with zero mean and variance  $\sigma^2$ . The PDF for observation  $y_i$  is then given by

$$g_i(y_i|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - ae^{-bx_i} - c)^2} \quad (10.5)$$

where  $\theta = (a, b, c, \sigma)$ . Putting these individual PDFs together according to Equation (10.3), we obtain the overall PDF for the data consisting of  $n$  observations as

$$f(y = (y_1, \dots, y_n) | \theta) = \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - ae^{-bx_i} - c)^2} \quad (10.6)$$

with the parameter vector  $\theta = (a, b, c, \sigma)$ .

### 10.3.3 Parameter estimation

Once we have collected data and have specified a model for the data, the next step is to assess the model's descriptive adequacy: how well does the model fit the data? A good model should, minimally, replicate the essential characteristics of observed data. Specifically, the goal is to find the model's parameter value that best fits the data in a properly defined sense. This procedure is called *parameter estimation* in statistics (e.g., Casella and Berger, 2002; Myung, 2003).

Statisticians usually employ one of two generally accepted methods of parameter estimation. They are the *least squares estimation* (LSE) and the *maximum likelihood estimation* (MLE). In LSE, the goal is to identify the parameter value that minimizes the difference between observations and model predictions. The goal of MLE is to identify the probability distribution that is most likely to have generated the observed data. We discuss each in turn below.

Formally, LSE seeks the parameter value that minimizes the sum of squares error (SSE) between observations and predictions defined as:

$$SSE(\theta) = \sum_{i=1}^n (y_i - y_{i,prd}(\theta))^2 \quad (10.7)$$

where  $y_{i,prd}(\theta)$  is a model's prediction for observation  $i$  given parameter  $\theta$ . The value of the parameter that minimizes the above SSE is called the *least squares estimate* denoted by  $\theta_{LSE}$ . Note that LSE does not require a probabilistic specification of a model so the model does not have to be defined in terms of its PDFs, insofar as the model makes predictions that can be compared against observations. LSE results are often summarized in terms of the *root mean square error* (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - y_{i,prd}(\theta_{LSE}))^2}{n}}. \quad (10.8)$$

LSE, while intuitive and easy-to-interpret, is primarily a descriptive method of parameter estimation that is developed to provide a summary of the data at hand, as opposed to make inferences about the regularity behind the data, thus gaining insights into the underlying processes. MLE is designed to serve this latter purpose.

MLE is an inferential method of parameter estimation that requires a probabilistic specification of a model in terms of PDFs. MLE provides a formal basis for many statistical methods, including chi-square goodness-of-fit testing, missing data analysis, and model selection.

Formally, MLE seeks the parameter value that maximizes the *likelihood function* of the model given observed data defined as follows:

$$\text{Likelihood function : } L(\theta) = f(y_{obs}|\theta). \quad (10.9)$$

Note that the likelihood function  $L(\theta)$  is a function of model parameter  $\theta$  and is obtained from the model's PDF by substituting the observed data vector  $y_{obs}$  for  $y$ . Also note that the likelihood function itself is defined over parameter space and is not a probability distribution. As such, it does not normalize to one, i.e.,  $\int L(\theta) d\theta \neq 1$ . The parameter value that maximizes the likelihood function in equation (10.9) is called the *maximum likelihood estimate* denoted by  $\theta_{MLE}$ , whose value of course depends upon the observed data so we often denote it by  $\theta_{MLE}(y_{obs})$ . The particular PDF associated with the maximum likelihood estimate, that is,  $f(y|\theta_{MLE})$ , is called the *maximum likelihood distribution* associated with the data. It is this distribution that is most likely to have generated the data in the MLE sense.

To illustrate MLE, let us revisit the power model of memory retention discussed earlier. Equations (10.1) and (10.2) describe the model equation and the corresponding PDF, respectively. Suppose that we conducted an experiment with  $n = 50$  Bernoulli trials and eight time intervals of  $t = (0.5, 1, 2, 4, 8, 12, 16, 18)$ , and recorded the number of correct responses out of 50 trials at each time interval. The observed data vector was obtained as  $y_{obs} = (44, 34, 27, 26, 19, 17, 20, 11)$ , or equivalently, the observed proportion correct as  $p_{obs}(= y_{obs}/n) = (0.88, 0.68, 0.54, 0.52, 0.38, 0.34, 0.40, 0.22)$ . The proportion data are shown as solid circles in Figure 10.3. The desired likelihood function to be maximized is given by

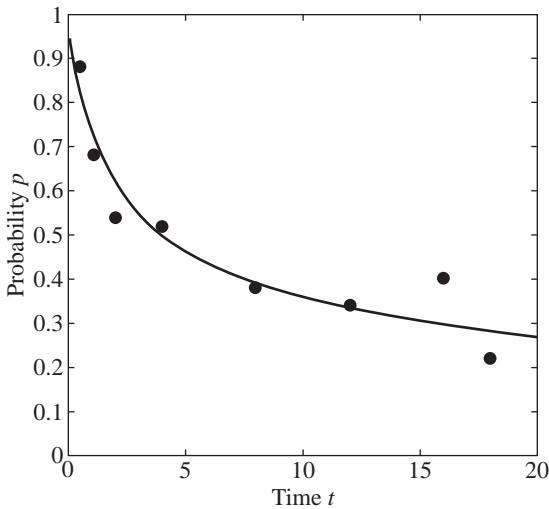
$$L(\theta = (a, b)) = \prod_{i=1}^8 \frac{n!}{(n - y_{obs,i})! y_{obs,i}!} p(\theta|t_i)^{y_{obs,i}} (1 - p(\theta|t_i))^{n-y_{obs,i}} \quad (10.10)$$

where  $p(\theta = (a, b)|t_i) = a(t_i + 1)^{-b}$ ,  $n = 50$ , and  $\theta$  is a variable. The maximum likelihood estimate is obtained as  $\theta_{MLE} = (0.985, 0.424)$ , and the resulting best-fit power curve as  $p = 0.985(t + 1)^{-0.424}$ , as shown as the solid line in Figure 10.3. The corresponding maximum likelihood distribution is then obtained as

$$f(y|\theta_{MLE}) = \prod_{i=1}^8 \frac{n!}{(n - y_i)! y_i!} p(\theta_{MLE}|t_i)^{y_i} (1 - p(\theta_{MLE}|t_i))^{n-y_i} \quad (10.11)$$

where  $\theta_{MLE} = (0.985, 0.424)$ ,  $n = 50$ , and  $y_i (= 0, 1, \dots, n)$  is now a random variable.

We close this section with a brief mention of how to find LSE and MLE parameter estimates. It is generally not possible to find these estimates in analytic closed-form expressions when the model is nonlinear in its parameters. Instead, the solutions must be sought numerically by computer using optimization search methods,



**Figure 10.3** Illustration of parameter estimation. The filled circles are eight simulated observations, and the solid curve represents the MLE best-fit power curve,  $p = 0.985(t + x)^{-0.424}$ , obtained by maximizing the likelihood function in Equation (10.10).

which perform “smart” searches of the parameter space iteratively until a solution is reached. For a discussion of the technical details, the reader is directed to a tutorial article by Myung (2003).

## 10.4 Model evaluation

### 10.4.1 How should a model be evaluated?

Once a model has been fitted to data and the best-fit parameter values have been found, the questions that may arise naturally to the modeler are: is the model any good? If so, in what sense? How persuasive is a good fit? These are the questions of *model evaluation*. A model may fit the data well, and even better than its competitor models. It does not, however, necessarily lead to the conclusion that the model successfully captures the underlying process (Roberts and Pashler, 2000). It may or may not because a good fit is only a necessary, but not sufficient, condition for drawing that conclusion (e.g., Myung, 2003, p. 97). How should we then evaluate a model?

One can think of several criteria with which to evaluate a model (Myung *et al.*, 2005). First of all and minimally, a model must be *falsifiable*, or equivalently, must satisfy the *testability* criterion. By this we mean that there must exist potential data patterns that the model cannot account for. Obviously, there would be no point of testing an unfalsifiable model that is simply a re-description of the data. As a rule of thumb, if the number of a model’s parameters is equal to or greater than the

number of observations in a data set, so the model has zero or negative degrees of freedom, then the model would be unfalsifiable. For example, for the retention data in Figure 10.3 consisting of eight observations, the following 8-parameter power model would be unfalsifiable:  $p = a(t + b)^{-c} + de^{-et}\sin(ft + g) + h$ , where  $\theta = (a, b, c, d, e, f, g, h)$ . It turns out, however, that the aforementioned counting rule does not always work, especially for nonlinear models, for which more sophisticated rules must be used to determine model falsifiability (Bamber and van Santen, 1985, 2000).

*Explanatory adequacy* is another criterion of model evaluation. A model should provide insight into the underlying process of interest that is generally not possible to gain otherwise. The model should also be *plausible* in that its assumptions make sense and are consistent with established biological and psychological findings. Relatedly, the model should be *interpretable* as well so that each of its parameters permits interpretation in terms of a known psychological process or construct. Further, the model should be *faithful* in that its ability to account for the underlying process derives from the theoretical principles the model substantiates but not from the subsidiary assumptions the model makes in its computational implementation (Myung *et al.*, 1999). While it is important to consider these four criteria (*explanatory adequacy*, *plausibility*, *interpretability*, and *faithfulness*) in model evaluation, given the qualitative (as opposed to quantitative) nature of their definitions, the modeler will have to apply them in a subjective and as sensible as possible manner in assessing the viability of the model under consideration.

On the other hand, there exist other criteria that are quantifiable and thus entail quantitative metrics by which the model can be evaluated. They are (1) *goodness-of-fit* (the extent to which a model fits observed data); (2) *model complexity* (a model's inherent flexibility to fit a wide variety of data patterns); and (3) *generalizability* (a model's ability to predict new observations). In what follows, we discuss each of these three quantitative measures in turn and their interrelationships with one another in greater depth.

### 10.4.2 Goodness-of-fit and the overfitting problem

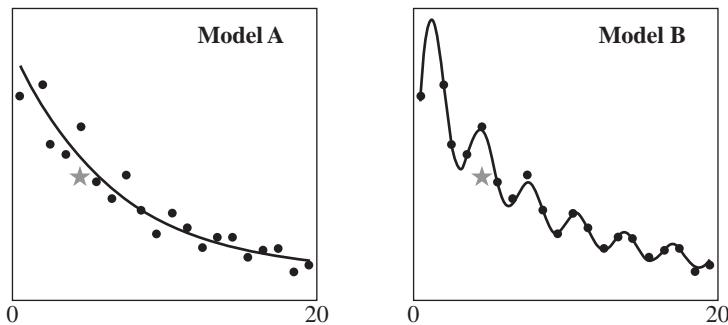
A common but misleading practice of modeling is to evaluate a model solely on the basis of its *goodness-of-fit* (GOF), that is, how well the model fits the data at hand. GOF measures that are commonly used include *root mean square error* (RMSE) in Equation (10.8), *percent variance accounted for* (PVAF), and *maximized likelihood* (ML). The latter two are defined as

$$PVAF = 100(1 - SSE(\theta_{LSE})/SST) \quad (10.12)$$

$$ML = L(\theta_{MLE})$$

where  $SSE(\theta_{LSE})$  is the maximized sum of squares error in Equation (10.7) and  $SST$  is the sum of squares total defined as  $SST = \sum_i(y_i - y_{mean})^2$ .

Off-hand, GOF measures may seem like good metrics by which to measure the model's ability to describe the underlying process being modeled. This would make



**Figure 10.4** Schematic illustration of the tradeoff between goodness-of-fit and model complexity. The complex model (Model B) provides a superior fit to the data (solid circles) than the simple model (Model A), but does a poor job in predicting a new observation (star).

sense if observations are free of random error. In practice, however, behavioral measurements are invariably corrupted with noise and artifacts of various kinds that have little to do with the regularity of interest. In other words, conceptually, GOF contains two unrelated components:

$$GOF = \text{Fit to regularity} + \text{Fit to noise}. \quad (10.13)$$

It is the first component we are interested in, but not the second one. It is, however, generally not possible to disentangle the separate effects of the two factors. Consequently, a GOF measure gives the overall value of their sum. What makes matters worse is the fact that there is a property of a model that enables it to fit random noise to an arbitrary extent, independent of its ability to fit the regularity. The case in point is a complex model with many parameters and a highly nonlinear model equation that may absorb noise rather easily, but without actually capturing the underlying regularity. As a result, a complex model often provides a better fit than a simpler model that has actually generated the data. This is why a good fit can be misleading and even bad (Pitt and Myung, 2002).

The problem of a model fitting random noise over and above the underlying relationship is known as *overfitting* in statistics. This is illustrated in Figure 10.4. The solid circles are observed data and the curves represent best-fits by two hypothetical models that differ in the number of parameters. Model A is a simple model with fewer parameters than Model B, and does a good (although not perfect) job of describing the variation in the data including the general trend, but also predicts well a new data point (red star symbol). In contrast, Model B, with its extra parameters, fits much better than Model A, but this superior goodness-of-fit is achieved at the cost of failing to predict the new data point. Despite its extra complexity, Model B predicts the new data point no better than, and perhaps worse than, Model A.

In conclusion, assessing the adequacy of a model based solely on goodness-of-fit (GOF) can result in overfitting. The overfitting occurs because GOF can be

improved arbitrarily by increasing the complexity of a model. What is model complexity, then? How do we measure it? This is the topic of the next section.

### 10.4.3 Model complexity

Intuitively, *model complexity*, or flexibility, refers to the property of a model that enables it to fit a wide range of data patterns, regardless of whether they represent the regularity of interest or the idiosyncratic random noise. Model complexity comes in at least two dimensions: (1) the number of model parameters; (2) the functional form of model equation. The first dimension of complexity (number of parameters) is well established in statistics: a model with many parameters is more complex than the one with fewer parameters.

On the other hand, the second dimension of complexity is less obvious. This *functional form complexity* refers to the way in which the parameters of a model are combined in its model equation (Myung and Pitt, 1997). As an example, the power and exponential models of retention memory,  $p = a(t + 1)^{-b}$  and  $p = ae^{-b}$ , respectively, have the same number of parameters (two) but differ in functional form, and they may therefore have different complexity. Pitt *et al.* (2002, figure 4) presents a striking example of variation in model complexity in which three one-parameter models differ widely in their ability to describe a range of data patterns.

In the literature, several approaches to quantifying model complexity in numerical metrics have been proposed. In what follows, we introduce and discuss two such measures.

Perhaps the most sophisticated measure of model complexity (MC) is the one derived from *normalized maximum likelihood* (NML; Rissanen, 2001),<sup>1</sup> which takes the following form:

$$MC_{NML} = \ln \int L(\theta_{MLE}(z))(d)z \quad (10.14)$$

where  $L(\theta_{MLE}(z))$  denotes the maximum likelihood (ML) given a data vector  $z$ . Note in the above equation that the model complexity is expressed as an integration of the ML over the entire data space. Accordingly, the NML model complexity is defined conceptually as the *sum of all best-fits* the model can provide collectively for all potential (not just actually observed) data that can be realized in an experimental setting. The larger the sum is, the more complex the model is. As such, the complexity measure represents a formalization of what is referred to intuitively and informally as a model's ability to fit a wide range of data patterns.

The NML complexity in Equation (10.14), although conceptually elegant, does not clearly reveal what constitutes model complexity, that is, what dimensions of complexity are captured in the measure. This is revealed in an asymptotic approximation of NML complexity that is known as the *Fisher information approximation*

<sup>1</sup> The NML is a method of model selection to which we return later in this chapter.

(FIA) model complexity (Rissanen, 1996; Su *et al.*, 2005),<sup>2</sup>

$$MC_{FIA} = \frac{k}{2} \ln \frac{n}{2\pi} + \ln \int \sqrt{\det(I(\theta))} d\theta \quad (10.15)$$

where  $k$  is the number of parameters,  $n$  is the sample size (i.e., number of independent identically distributed observations),  $I(\theta)$  is the Fisher information matrix of sample size 1,<sup>3</sup>  $\det$  denotes the determinant of a matrix, and finally,  $\ln$  is the natural logarithm of base  $e$ . The first term of the right-hand side of the equation captures the effects of complexity due to the number of parameters ( $k$ ), and the second term captures the functional form effects of complexity through  $I(\theta)$ . Both dimensions of complexity are therefore reflected in  $MC_{FIA}$ . We note further that the magnitude of the first term increases logarithmically with the sample size  $n$ , whereas the second term does not depend upon  $n$ . An implication of this observation is that in the limit of large sample sizes, the functional form effects of complexity become negligible relative to the effects due to the number of parameters, thereby effectively making the former the sole contributor to model complexity, as conventionally conceptualized.

To summarize, we have learned so far that goodness-of-fit (GOF) is a necessary but not sufficient criterion of model adequacy. We also learned that the model must be sufficiently complex to capture the regularity in the data, but not too complex to overfit the data by capitalizing on random error. It seems, then, that a good model is one that strikes the “right” balance between GOF and model complexity, both of which can be objectively measured in quantifiable terms. How much right is right? Answering this question is tied directly to the goal of modeling that in turn leads us to the third quantifiable criterion of model evaluation, *generalizability*.

#### 10.4.4 Generalizability

The goal of modeling is to deduce the model that generated the observed data. In reality, however, this is not possible because of two fundamental limitations: (1) *finiteness of data*. Observations in a data set may never be sufficient to exactly and uniquely identify the ground truth; and (2) *complexity of truth*. The truth may be quite complex, well beyond anyone’s imagination and thus not among the models under consideration. This second limitation is another reminder of G. E. P. Box’s quote that all models are wrong. Given these challenges, a more realistic, and perhaps achievable, goal is to identify a model that is deemed the “as best as possible” approximation to the underlying truth in a defined sense. The current consensus in the field is that the best (and so most useful) approximate truth is the model with highest generalizability.

*Generalizability* (GN) is defined as a model’s ability to fit not only the observed data at hand, but also new data from the same process that has generated the

2 The FIA is another method of model selection that is discussed later in this chapter.

3 The Fisher information, defined in terms of the covariances of the second-order partial derivatives of the log likelihood function,  $\ln L(\theta)$ , with respect to the parameters, in essence measures the amount of information in data about model parameters (e.g., Schervish, 1995).

current data. Put another way, this criterion, often known as *predictive accuracy*, refers to how accurately the model predicts future observations. GN is indeed the ultimate yardstick and gold standard of model evaluation by which all models with varying degrees of goodness-of-fit and complexity are to be judged for their usefulness. Revisiting the models in Figure 10.4, model A clearly generalizes better and would therefore be judged as a closer approximation to the underlying process being modeled.

To reiterate, the central creed of model evaluation is good generalizability. This is achieved by striking the right balance between goodness-of-fit and model complexity by trading off one for the other. In other words, the model should be no more complex than what is necessary to extract the underlying regularity. It is in this sense that the generalizability criterion can be regarded as a formal embodiment of the principle of *Occam's razor*, “Entities should not be multiplied beyond necessity” (William of Occam, 1288–1348).

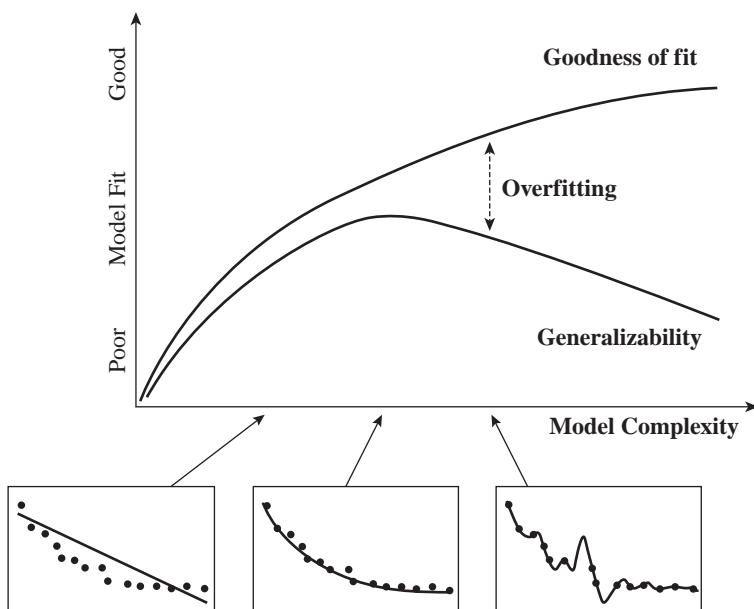
Generalizability can be made more precise and rigorous in formal terms. In doing so, let us first define the notion of discrepancy between two probability distributions. Specifically, a discrepancy function  $D(f, g)$  between two distributions,  $f$  and  $g$ , is a continuous real-valued positive function that satisfies the condition of  $D(f, g) > D(f, f) = 0$  for all  $f \neq g$  (e.g., Linhart and Zucchini, 1986). The well-known Kullback–Leibler information divergence is an example of such a *discrepancy function*. The smaller the  $D(f, g)$ , the more similar the two distributions are to each other, and thus, the better the distribution  $f$  approximates the distribution  $g$ , and vice versa. As such,  $D(f, g)$  is a kind of “distance” measure between two distributions, although the discrepancy itself does not necessarily satisfy the symmetric condition, i.e.,  $D(f, g) = D(g, f)$ . In terms of the discrepancy function, a formal definition of the generalizability for model  $M$  is given as (e.g., Su *et al.*, 2005, p. 413)

$$E[D(f_T, f_M)] = \int D(f_T, f_M(\theta_{MLE}(y)))f_T(y) dy \quad (10.16)$$

where  $f_M(\theta_{MLE}(y))$  denotes the model’s *maximum likelihood distribution* given a data vector  $y$  and  $f_T(y)$  is the probability distribution that generates observed data (i.e., ground truth). As shown above, generalizability is defined as a mean discrepancy or “distance” between the true model distribution and the best-fitting distribution of the model family of interest, averaged over all possible data under the true model. It is in this sense that the model with highest generalizability can be regarded as the “best possible” approximation to the truth.

#### 10.4.5 Relationship among goodness-of-fit, complexity, and generalizability

Figure 10.5 illustrates the relationships among the three quantitative criteria discussed so far: GOF, model complexity, and generalizability. There are a few main points to make from the figure. First, as indicated by the upper curve in the figure, GOF can always be improved by increasing model complexity. Second, increasing



**Figure 10.5** An illustration of the relationships among GOF, model complexity, and generalizability as a function of model complexity. The vertical axis represents a model fit index, where a larger value indicates a better fit (e.g., percent variance accounted for). Reprinted from Pitt and Myung (2002).

complexity only improves generalizability up to a certain point, as shown by the lower curve. Third, the most adequate model given observed data is the one with the complexity corresponding to the peak of the generalizability curve. This is the model whose complexity best matches the complexity in the data. The model is sufficiently complex to capture the regularity but not overly complex to start absorbing random noise in the data. Fourth, as indicated in the figure, overfitting manifests itself as the difference between the goodness of fit and generalizability curves beyond and above the optimal point of model complexity.

The three smaller graphs in the bottom of Figure 10.5 provide concrete examples. In the left graph, the model (line) is not complex enough to match the complexity of the data (dots). The model and data are well matched in complexity in the middle graph, which is why this occurs at the peak of the generalizability function. In the right graph, the model is more complex than the data, capturing incidental variation due to random error.

## 10.5 Model selection

The preceding section describes the dimension on which models should be evaluated. How are they implemented in practice to guide researchers in choosing one model over the others? In this section, we review methods of model

selection that are currently in use. For a more thorough treatment of the topic, the reader is directed to two *Journal of Mathematical Psychology* special issues (Myung *et al.*, 2000; Wagenmakers and Waldorp, 2006), and an excellent tutorial article on the topic (Shiffrin *et al.*, 2008).

The central tenet of model selection, as discussed earlier, is to choose, among a set of candidate models being compared, the one that generalizes best, or equivalently, the one that provides the closest approximation to the truth. One may then use the generalizability measure in Equation (10.16) for the purpose of identifying the best generalizing model. Unfortunately, however, this measure cannot be directly computed as it is defined in terms of the true distribution  $f_T(y)$ , which is unknown or unknowable. Consequently, the generalizability measure must be *estimated* from observed data. Virtually all methods of model selection including the ones we discuss here can be seen as generalizability estimates of one kind or another.

### 10.5.1 Penalized-likelihood model selection

What is wanted in model selection is, again, a method that estimates a model's generalizability by taking into account the effects of model complexity on model fit. In other words, model selection is in essence about achieving a balance between two opposing forces, model complexity on one side and GOF on the other. In each of the four methods of model selection we introduce in this section, this is instantiated by penalizing the model under consideration for excessive and unnecessary complexity, that is, the portion of its complexity that is more than what is needed to capture the regularity in the data, thereby putting all the models on an equal footing so to speak.

The four methods are the *Akaike information criterion* (AIC; Akaike, 1973), the *Bayesian information criterion* (BIC; Schwarz, 1978), the *Fisher information approximation* (FIA; Rissanen, 1996; Su *et al.*, 2005), and the *normalized maximum likelihood* (NML; Rissanen, 2001). They are defined as

$$\begin{aligned} AIC &= -2\ln L(\theta_{MLE}(y)) + 2k \\ BIC &= -2\ln L(\theta_{MLE}(y)) + k \ln(n) \\ FIA &= -\ln L(\theta_{MLE}(y)) + \frac{k}{2} \ln \frac{n}{2\pi} + \ln \int \sqrt{\det(I(\theta))} d\theta \\ NML &= -\ln L(\theta_{MLE}(y)) + \ln \int L(\theta_{MLE}(z)) dz. \end{aligned} \quad (10.17)$$

In the above equation,  $y$  is a vector of observed data,  $L(\theta_{MLE}(y))$  ( $= f(y|\theta_{MLE})$ ) is the maximum likelihood of the data,  $z$  is a vector variable of potential data, and finally,  $k$ ,  $n$  and  $I(\theta)$  are defined earlier in Equations (10.14) and (10.15).

Each criterion above consists of the first term representing a lack-of-fit measure and the second and remaining terms representing a model complexity measure. Combined, they estimate a model's generalizability such that a lower value of the

overall criterion indicates better generalizability. Accordingly, the criterion prescribes that among a set of competing models, the one with the lowest criterion value should be selected as the best-generalizing model.

Notice two counteracting forces at work in each criterion: An increase in the second and remaining terms (i.e., increasing complexity) generally results in a decrease in the first term (i.e., better goodness-of-fit), and vice versa. A logical corollary of this asymmetry is that the criterion implicitly penalizes the model with excess complexity. That is to say, a model may provide a superior GOF over other models being considered, but that alone does not necessarily make it a better generalizing model. This is because that model's complexity may be too large to the extent of causing a net positive increase in the overall criterion value, thereby in essence creating an overfitting situation. In short, the aim is to achieve the “optimal” tradeoff between the two forces so as to avoid overfitting as well as underfitting. The notion of optimality is conceptualized and defined differently for different model section methods, which we discuss below one at a time.

The AIC criterion is historically the first method of model selection that has been introduced for choosing among nonlinear and nonnested models, and is rooted in information theory (Cover and Thomas, 1991). Specifically, AIC is derived as a large sample (i.e., asymptotic) approximation of the generalizability measure defined in Equation (10.16) in which the discrepancy function  $D(f, g)$  is the Kullback–Leibler information divergence between the true probability distribution and the maximum likelihood distribution of the model under consideration. Accordingly, AIC selects one model, among a set of competing models, that has the closest distance to the truth in an information theoretic sense. One shortcoming, however, is that from the AIC standpoint, the number of model parameters ( $k$ ) is the sole contributing factor to model complexity, thereby ignoring other relevant and potentially significant factors such as sample size and function form.<sup>4</sup>

The BIC criterion is a Bayesian statistical criterion and is derived as an asymptotic approximation of the Bayesian model selection (BMS), which is introduced later in this chapter. The basic idea of BMS as well as BIC is to identify the model that is most likely to have generated observed data in the sense of Bayesian probability theory (e.g., Gelman *et al.*, 2013). Notice the model complexity term of the BIC that includes the contribution of the sample size ( $n$ ) as well as that of the number of free parameters ( $k$ ). As such, as the sample size, or the number of observations, increases, BIC tends to favor models with fewer parameters, unlike AIC that does not take into account the sample size factor.

Both FIA and NML criteria are methods of model selection derived from the principle of *minimum description length* (MDL; Grünwald *et al.*, 2005; Myung *et al.*, 2006; Grünwald, 2007) in computer science, with FIA being an asymptotic approximation of NML. According to the principle of MDL, the goal of modeling is to compress the data as tightly as possible without loss of information; the model

<sup>4</sup> To be fair, the original derivation of AIC (Akaike, 1973) does include higher-order terms that reflect the effects of sample size and functional form, but the terms are subsequently dropped at the later stages of the asymptotic expansion for the sake of simplicity of computation.

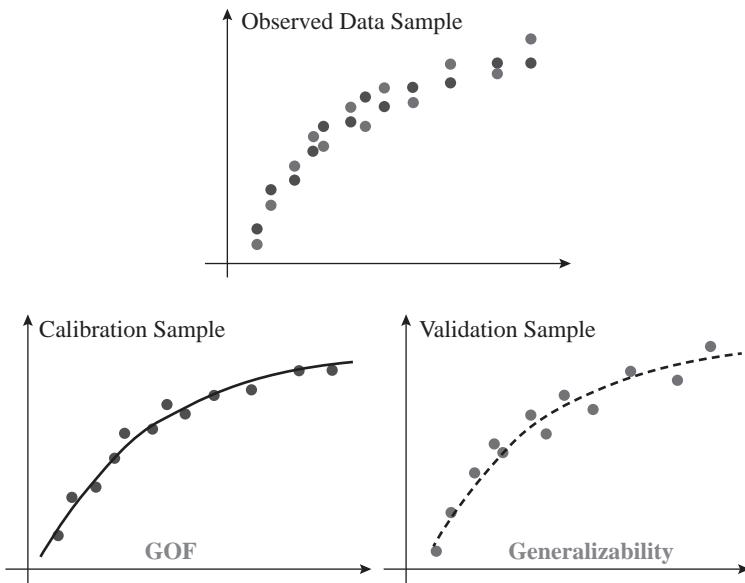
is viewed as a code with which to encode the data, and finally, the best model is the one that provides the shortest description length of the data in bits. To elaborate further, the more the model is able to extract regularities in data structure, the better the model can compress the raw data with the help of the uncovered regularities, thus providing a shorter description length of the data. This in turn leads to better generalization because the model can now use the extracted regularities to predict future data accurately. In short, regularity extraction, data compression, and generalization are three variations of essentially the same idea. Regarding how model complexity is conceptualized and implemented in FIA and NML, as discussed earlier in this chapter, the complexity terms of both criteria are sensitive to the number of parameters, the sample size, and importantly, the functional form. Of particular note is that the NML complexity in Equation (10.14) is not only intuitively appealing but also represents a “full and complete” view of model complexity. For a technically rigorous treatment of this and related material, the reader is directed to an excellent tutorial article by Grünwald (2005). Example applications of FIA and NML to model selection in cognitive psychology can be found in Wu *et al.* (2010), Klauer and Kellen (2011), and Singman and Kellen (2013), for example.

### 10.5.2 Cross-validation and accumulative prediction error

Each of the four methods of model selection discussed in the preceding section estimates a model’s generalizability through an equation that specifies explicitly how goodness-of-fit should be traded off for model complexity. In contrast, *cross-validation* (CV; Stone, 1974; Browne, 2000) and the *accumulative prediction error* (APE; Dawid, 1984; Wagenmakers *et al.*, 2006) that we introduce in this section estimate generalizability directly from the observed data by simulating the data-generation and model-prediction steps, but without relying upon a formulaic equation with an explicit measure of model complexity.

Both CV and APE are implemented in the following three-step procedure. First, the data sample is split into two nonoverlapping subsamples, the calibration sample denoted by  $y_{cal}$  and the validation sample denoted by  $y_{val}$ . Second, the model of interest is fit to the calibration sample  $y_{cal}$  and the model’s MLE estimate,  $\theta_{MLE}(y_{cal})$ , is obtained. Third, the model is then fit to the validation sample  $y_{val}$  directly with its parameter values *fixed to*  $\theta_{MLE}(y_{cal})$ . The resulting fit, or prediction error, to the validation sample is taken as the model’s generalizability estimate. The two methods, CV and APE, differ from each other in how the data are split into calibration and validation samples. Even within CV, there exist different variations of this general scheme depending upon how the calibration-validation split is defined.

In what is known as the *split-half CV*, the observed data are divided randomly into two subsamples of equal size, one half for the calibration and the other half for the validation. This split-half CV method is illustrated in Figure 10.6 for a hypothetical model. As shown in the figure, the data of 24 observations in the top panel are split into a calibration sample consisting of 12 blue (darker) filled circles



**Figure 10.6** Illustrated scheme of split-half cross-validation.

in the lower left panel and a validation sample consisting of 12 red (lighter) filled circles in the lower right panel. The solid curve in the lower left is the best-fit MLE curve to the calibration sample. How well this solid curve fits the calibration sample defines the model's GOF. The same curve, now denoted by the dotted curve in the lower right, is fitted directly to the validation sample without further parameter tuning. How well this dotted curve fits the validation sample defines the model's generalizability estimate. One drawback of split-half CV is that its generalizability estimate depends upon the particular way the data are divided into two equal halves and further, that there are practically an infinite number of ways to do the splitting. Another CV procedure we discuss next gets around this problem by adopting an unequal splitting rule.

In *leave-one-out-cross-validation* (LOOCV), the data sample of  $n$  observations, denoted by a vector  $y = (y_1, \dots, y_n)$ , is split into a calibration sample of  $(n - 1)$  observations and a validation sample of the remaining one observation, and the model's generalizability is then estimated according the three-step procedure mentioned above. This  $(n - 1)$ -vs.-1 split process is repeated for all possible  $n$  splits. The model's final generalizability is obtained as the arithmetic mean of  $n$  individual generalizability estimates, formally expressed as

$$LOOCV = - \sum_{i=1}^n \ln f(y_i | \theta_{MLE}(y_{\neq i})). \quad (10.18)$$

APE is similar to LOOCV in spirit, but differs in implementation such that generalizability is estimated in a sequential and accumulative manner, instead of the  $(n - 1)$ -vs.-1 split between calibration and validation samples. To elaborate, given

the data of  $n$  observations and a model with  $k$  parameters, the data are split into a calibration sample consisting of the *first*  $(k + 1)$  observations and a validation sample of the  $(k + 2)$ th observation. The model's generalizability with respect to this particular  $(k + 1)$ -vs.-1 split is estimated following the same three-step procedure. The calibration sample is then increased in size by one observation by taking in the next  $(k + 2)$ th observation, and the model's generalizability with respect to the new  $(k + 2)$ -vs.-1 split is again estimated. This successive and accumulative process continues until there is only one observation left in the validation sample. The model's *final* generalizability is obtained as the arithmetic mean of  $(n - k - 1)$  individual generalizability estimates, formally expressed as

$$APE = - \sum_{i=k+2}^n \ln f(y_i | \theta_{MLE}(y_{1,2,\dots,i-1})). \quad (10.19)$$

Both cross-validation and the accumulative prediction error prescribe that among a set of competing models, the one with the lowest value of the given criterion should be selected as the best generalizing model. There are several characteristics that make these methods appealing alternatives to the penalized-likelihood methods of model selection in Equation (10.17). One is their ease of computation. CV and APE can easily be implemented; all that is needed is the calculation of maximum likelihood estimation. This is unlike FIA and NML, which involve high-dimensional integration, which can be nontrivial to compute numerically. Another appealing feature of CV and APE is that they supposedly (albeit implicitly) take into account the effects of all three dimensions of model complexity, that is, the number of parameters, the sample size, and the functional form. Accordingly, the performance of CV and APE should generally be superior to that of either AIC or BIC, which does not consider these dimensions of complexity.

### 10.5.3 Bayesian model selection

*Bayesian model selection* (BMS; Kass and Raftery, 1995; Wasserman, 2000) is the standard, state-of-the-art method of model selection for Bayesian inference and is defined as the minus logarithm of the *marginal likelihood* of the model of interest,

$$BMS = -\ln \int L(\theta(y))p(\theta) d\theta \quad (10.20)$$

where  $y$  is a vector of observed data,  $L(\theta(y))$  ( $= f(y|\theta)$ ) is the likelihood function of the data defined in Equation (10.9), and  $p(\theta)$  is the parameter prior distribution. BMS prescribes that the model with the lowest BMS value should be preferred. It is worth noting that BMS is closely related to the Bayes factor, which is defined as the ratio of two marginal likelihoods between a pair of competing models, in such a way that either criterion, BMS or Bayes factor, always leads to the same model choice.

Note in Equation (10.20) that the marginal likelihood,  $\int L(\theta(y))p(\theta) d\theta$ , is nothing but the mean likelihood obtained by averaging the likelihood across all parameter values and weighted by the parameter prior. This Bayesian averaging is exactly how BMS avoids overfitting, that is, by selecting the model with the highest *mean* likelihood value, instead of the one with the highest *maximum* likelihood value. The latter would necessarily result in overfitting. In other words, model complexity is automatically adjusted in BMS through the built-in averaging operation. In so doing, the method considers all three dimensions of complexity; this can be seen more clearly in an asymptotic approximation of BMS (Balasubramanian, 1997), which turns out to be just the same as FIA in Equation (10.17)! This points to a potentially intriguing connection between minimum description length and Bayesian model selection, despite the fact that they are rooted in divergent theoretical and philosophical foundations. A further approximation of BMS leads to one-half of BIC so that the latter can be considered as a quick and rough version of the former. Finally, the computation of BMS would be in general nontrivial as it involves numerical integration of high dimensions. We noted earlier that similar computational difficulties plague the routine use of FIA and NML.

#### 10.5.4 Illustrated example

In this section we present and discuss an illustrated example of the four model selection methods (AIC, BIC, LOOCV, APE).<sup>5</sup> Two goodness-of-fit measures, PVAF and ML in Equation (10.12), are also included for the comparison purpose.

Four retention models are compared in terms of their generalizability estimated by each of the four selection criteria. The four models are as follows:

$$\begin{aligned} POW &: p = a(t + 1)^{-b} \\ POW2 &: p = a(t + 1)^{-b} + c \\ POW3 &: p = a(t + 1)^{-b} + c + d \cdot \sin(e \cdot t) \\ EXP &: p = ae^{-bt}. \end{aligned} \tag{10.21}$$

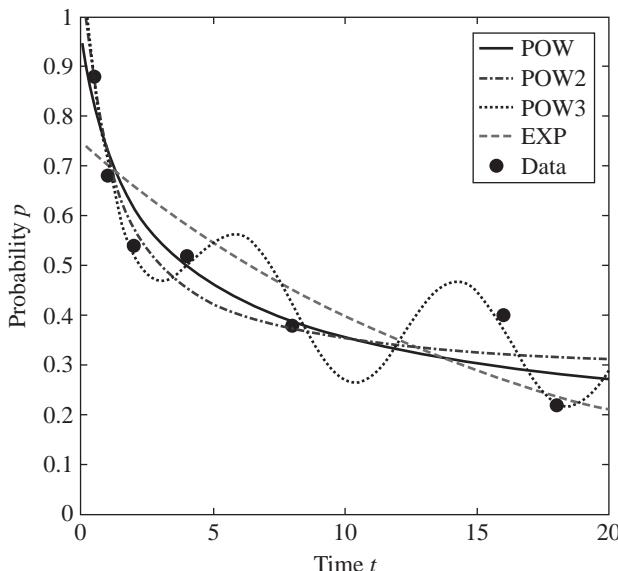
Note that POW and EXP have two parameters, POW2 has three parameters, and POW3, the most complex among the four, has five parameters. Each of these models was fitted to the data shown as the solid circles in Figure 10.3 with the binomial likelihood function in Equation (10.10). The source code of this simulation is included in the Appendix.

The fitted data and best-fit curves of the four models are shown in Figure 10.7. The model selection results are reported in Table 10.1. Let us first examine GOF performance of the four models. All three power models fit the data generally well. As expected, the more parameters a model has, the better the model fits the data. Not surprisingly, POW3, the most complex model among the four, provided the

<sup>5</sup> FIA and NML are not included in the example due to the computational challenges to implement them.

Table 10.1 *Model selection results for four retention models defined in Equation (10.21). The LogLik stands for the log maximum likelihood. The best-fit parameters were sought by maximizing the likelihood function in Equation (10.10) without the constant, parameter-independent term,  $n!/(n - y_i)! y_i!$ .*

Model	POW	POW2	POW3	EXP
Number of params	2	3	5	2
PVAF	91.2	92.6	<b>98.7</b>	79.0
LogLik	-247.34	-246.68	<b>-244.48</b>	-252.06
AIC	<b>498.67</b>	499.35	498.96	508.11
BIC	<b>502.50</b>	505.09	508.53	511.94
LOOCV	31.41	31.52	<b>30.80</b>	32.53
APE	32.63	31.94	<b>30.64</b>	35.57



**Figure 10.7** Illustration of model selection for four models of memory retention in Equation (10.21). The filled circles are the data and the curves represent the best-fit model predictions.

best fit, capturing 98.7% of the total variance, which is perhaps an overfitting. The two-parameter exponential model (EXP) fared poorly, only capturing 79% of the variance. Accordingly, it seems this model can be ruled out safely for further consideration. The log maximum likelihood results based on *LogLik* values lead to essentially the same conclusion.

On the other hand, the generalizability results lead us to different conclusions. Both AIC and BIC preferred the simplest power model (POW) as the best-generalizing model among the four under consideration. In other words, according

to AIC and BIC, the other power models, POW2 and POW3, apparently overfit the data, and when their extra parameters were penalized by Occam's factors, they both lose out to POW. Interestingly enough, however, LOOCV and APE results draw a different picture. Both of these methods select POW3, the most complex of the four, as the best-generalizing model! This model choice is, obviously, in direct contraction with that based on AIC and BIC, and is thus somewhat of a surprise. Because the underlying truth is unknown in this case and also given the relatively small size of the data, it is difficult to discern the possible causes and reasons for these conflicting results so we would have to take them at face value. That is, different methods of model selection can sometimes lead to differing interpretations of the same data.

To summarize, we demonstrated application of the model selection procedure to the problem of choosing among a set of models that differ not only in the number of parameters, but also in functional form. The reader should not over-generalize the particular results from this example application, which was simply intended to serve as an illustration of model selection – but no more. We conclude this section with a quote from Myung and Pitt (2004, p. 365) on the importance of viewing model selection as a statistical inference problem:

*Model selection is an inference problem. The quality of the inference depends strongly on the characteristics of the data (e.g., sample size, experimental design, type of random error) and the models themselves (e.g., model equation, parameters, nested vs. nonnested). For this reason, it is unreasonable to expect a selection method to perform perfectly all the time.*

### 10.5.5 Summary

Computational modeling has become an important tool for advancing the study of mind and brain and has contributed substantially to theorizing and experimentation in cognitive psychology. The success of modeling depends upon the availability of theoretically sound methods for comparing and selecting among computational models. Often, a number of models (or theoretical explanations) can account for a given set of empirical data, and it is not clear how one can best choose among these competing models. At its most basic level, this is a problem of uncertainty of inference from data to the model, and model selection (comparison) methods help reduce this uncertainty by using sound statistical methods.

In this section we reviewed quantitative approaches that guide model selection and thus improve scientific inference. The main take-home messages from this review can be summarized into the following four steps.

- Step 1 (Goodness-of fit): evaluate each model's fit, among a set of candidate models being evaluated, to observed data to assess its descriptive adequacy, or goodness-of-fit.
- Step 2 (Model complexity): consider the model's inherent flexibility to fit other potential data that could be collected.

- Step 3 (Generalizability): estimate the model's generalizability by properly trading-off goodness-of-fit for model complexity using a given method of model selection.
- Step 4 (Model choice): choose the model with the best generalizability.

To reiterate, models should be evaluated based on generalizability, not on goodness of fit, as echoed by the following statement: "Thou shall not select the best-fitting model but shall select the best-generalizing model."

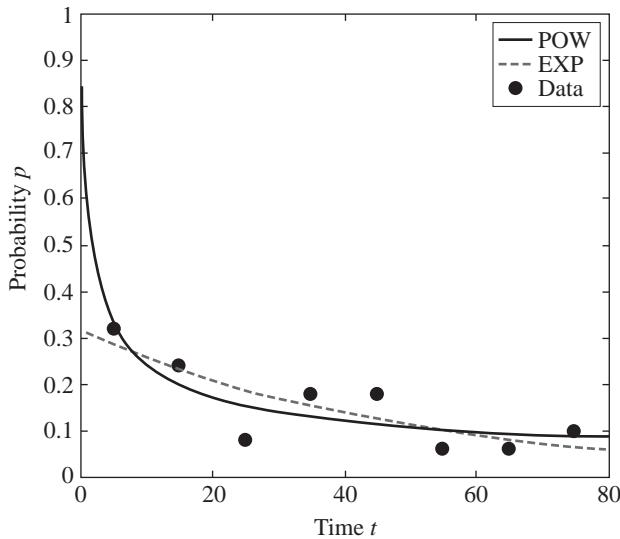
## 10.6 Design optimization

### 10.6.1 Further improving model selection through design optimization

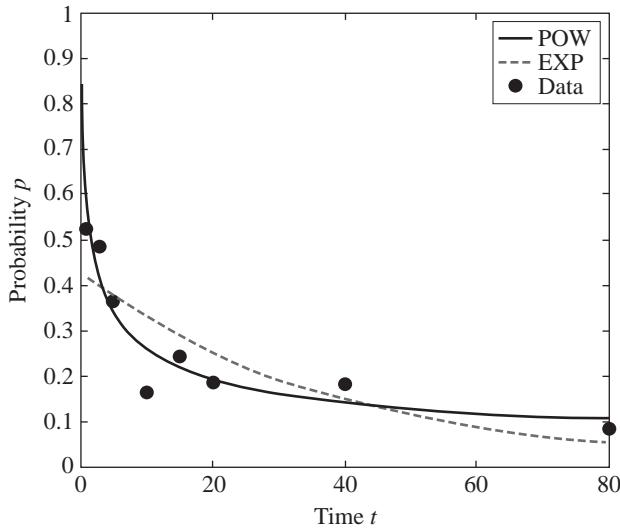
Statistical tools for model selection that we have discussed in the preceding sections are developed to assist researchers in making inferences about models given data samples collected in experiments. However, because such tools are applied *after* data have been collected, their potential to yield definitive conclusions is limited by the quality of the empirical data that they have to work with; sometimes the data simply do not provide clear differentiation between models. As mentioned previously, part of the problem stems from the fact that different models can mimic one another. That is, many sets of data can be explained just as well by one model as by another.

For example, consider the hypothetical set of retention data shown in Figure 10.8, which were generated by computer from the model POW, which is defined in Equation (10.21) with  $a = 0.8$ ,  $b = 0.5$ , and 50 Bernoulli trials at each of 8 different retention intervals: 5, 15, 25, 35, 45, 55, 65, and 75 seconds. Suppose, for the purposes of illustration, that one did not know the true data-generating model and wished to determine whether these data were more likely to have been generated by POW or a competing model, EXP, also as defined in Equation (10.21). The BIC, introduced in the preceding section, is an appropriate criterion for this task, as the model with the lower BIC is more likely to have generated the observed data in the sense of Bayesian probability theory. However, for these data, the BICs for POW and EXP are quite similar: 49.1 for EXP and 48.3 for POW. This difference between these two values,  $\Delta\text{BIC} = 0.8$ , means that POW is slightly more likely to have generated the data, but the margin is razor thin, so the result is inconclusive.

Now consider a different set of hypothetical retention data, shown in Figure 10.9. These data were generated from the same model as in the preceding example, with the same number of Bernoulli trials at each retention interval, but at a different set of 8 retention intervals: 1, 3, 5, 10, 15, 20, 40, and 80 seconds. If we perform the same model selection analysis on these data, we get a BIC of 57.8 for EXP and 49.0 for POW. This result,  $\Delta\text{BIC} = 8.8$ , means that POW is much more likely to have generated the data than EXP. In other words, these data strongly identify POW as the generating model.



**Figure 10.8** Example of a poor experimental design in discriminating between power (POW) and exponential (EXP) models of retention. The design consists of eight equally spaced time intervals  $\{5, 15, 25, \dots, 75\}$  for which both models provide equally good descriptions of the data ( $BIC_{POW} = 48.3$  vs.  $BIC_{EXP} = 49.1$ ) and thus are not much discriminable.



**Figure 10.9** Example of a good experimental design with a different set of eight time intervals  $\{1, 3, 5, 10, \dots, 80\}$ , than the one shown in Figure 10.8. Note that in terms of BIC, model POW provides a better description of the data than model EXP:  $BIC_{POW} = 49.0$  vs.  $BIC_{EXP} = 57.8$ .

The preceding examples illustrate that our ability to successfully discriminate between models can sometimes hinge on decisions that are made *before* data are collected. For instance, in collecting retention data, one must first choose which retention intervals at which to test memory. In the first example, memory was tested at retention intervals that were roughly evenly spaced between zero and 80 seconds, while in the second example, testing was concentrated at intervals near zero seconds, with far fewer observations in the range of 20–80 seconds. It turned out that the latter choice of retention intervals improved the informativeness of the data for discriminating between the models.

In general, many such decisions about the design of an experiment must be made before data collection can begin, including the number of treatment groups, the sample size in each treatment group, and the timing and order of stimuli, among others. The precise set of relevant design variables will vary from experiment to experiment. The settings of these variables are referred to as the experimental design, and the experimenter's choices regarding the experimental design affect not only the potential statistical value of the results, but also the cost of the experiment in terms of time, money, and participants. Therefore, an optimal experimental design can be regarded as one that maximizes the informativeness of the experiment while being cost effective for the experimenter.

### 10.6.2 Design optimization

To optimize design decisions and maximize the chances of discriminating between models, *design optimization* (DO) applies statistical inference to evaluate design decisions at the front-end of an experiment (i.e., before data have been collected), in order to make model evaluation easier at the back end of the experiment (i.e., after data have been collected). There is a rich literature in statistics on the problem of design optimization dating back to the 1950s (e.g., Kiefer, 1959; Atkinson and Federov, 1975; Atkinson and Donev, 1992; Chaloner and Verdinelli, 1995). Here we introduce the reader to the conceptual framework of DO as it relates to the optimal design of psychological experiments. For in-depth details of the theoretical and computational aspects of DO, the reader is directed to other publications from our lab (Myung and Pitt, 2009; Cavagnaro *et al.*, 2010; Myung *et al.*, 2013).

The implementation of the DO approach in practice requires that one must first define the design space, which consists of the set of all possible values of design variables that are controlled by the experimenter. The problem of design optimization is then to search that design space and identify the design that has the greatest potential to successfully discriminate among the models under consideration.

But how does one measure the potential of each design? Bayesian decision theory offers a principled approach to this problem. In Bayesian design optimization, each potential design is treated as a gamble whose payoff is determined by the outcome of an experiment carried out with that design. The idea is to estimate the “utilities” of hypothetical experiments carried out with a given design, so that an expected utility of that design can be computed. This is done by considering

every possible observation that could be obtained from an experiment with a given design and then evaluating the relative likelihoods and statistical values of these observations. The design with the highest expected utility (i.e., the one that yields the most informative data, on average) is then chosen as the optimal design.

For example, in a retention experiment, the set of retention intervals at which to test memory is a design variable that could be optimized. In the example above, we considered two different sets of eight retention intervals, but many other sets are possible. To calculate the expected utility of a given set of retention intervals, simulated experiments must be conducted by computer (i.e., Bernoulli trials with the probabilities specified by a hypothesized generating model and parameters). The utility of those simulated data depend on how conclusively they identify the generating model and parameters (according to some model selection statistic, such as the Bayes factor). The expected utility is obtained by repeating this process across numerous iterations with different generating models and parameters, and taking an average of resulting utilities. Thus, the expected utility measures how conclusively the data are expected to identify the generating model, whatever that model may be.

In quantitative terms, the utility function to be optimized, denoted by  $U(d)$  as a function of design  $d$ , is expressed in the following form (e.g., Chaloner and Verdinelli, 1995):

$$U(d) = \sum_m p(m) \iint u(d, \theta_m, y_m) p(y_m | \theta_m, d) p(\theta_m) dy_m d\theta_m. \quad (10.22)$$

In the above equation,  $y_m$  denotes the data outcome under a hypothesized generating model  $m$  in a simulated experiment,  $\theta_m$  denotes the model parameter,  $p(y_m | \theta_m, d)$  is the model's probability distribution (often called the likelihood function),  $p(m)$  is the model's prior probability, and finally,  $p(\theta_m)$  is the parameter's prior distribution. The function,  $u(d, \theta_m, y_m)$ , measures the "local" utility of design  $d$  given the parameter value  $\theta_m$  and the data outcome  $y_m$ . Note that "global" utility function  $U(d)$  is defined as an average of the local utility  $u(d, \theta_m, y_m)$  over the models under consideration, model parameters, and data outcomes, with respect to the model prior  $p(m)$ , the parameter prior  $p(\theta_m)$ , and the likelihood function  $p(y_m | \theta_m, d)$ , respectively.

The choice of the local utility function used is derived by the specific goal of the experiment, whether the goal is parameter estimation, i.e., the estimation of a model's parameters, or alternatively, whether the goal is model discrimination, i.e., the identification of the underlying data-generating model among a set of competing models. Just to give an example, for parameter estimation, one may consider  $u(d, \theta_m, y_m) = \log \frac{p(\theta_m | y_m, d)}{p(\theta_m)}$ , which has an information theoretic interpretation (see, e.g., Myung *et al.*, 2013, p. 58).

The design optimization problem entails identifying an optimal design  $d^*$  that maximizes the utility function  $U(d)$  in Equation (10.22). In practice, however, solving the problem presents computational challenges that make standard solution methods impractical or impossible. First, evaluating the expected utility of a

given design entails high-dimensional numerical integration (over possible models, parameters, and observed data patterns), which requires Monte Carlo simulation. Moreover, spaces are often high-dimensional, requiring an intelligent search algorithm to ensure convergence to the global optimum. For example, if the design space to be searched consisted of all possible sets of eight retention intervals, even a very sparse grid search would entail evaluating millions, or perhaps billions of expected utilities.

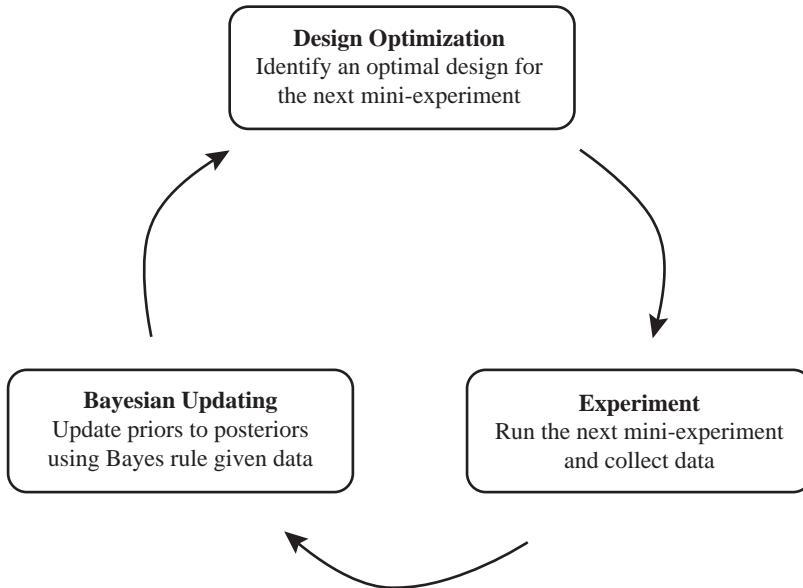
Recent breakthroughs in stochastic optimization have made this problem tractable (Müller *et al.*, 2004; Amzal *et al.*, 2006), but they are beyond the scope of this chapter. Alternatively, the problem of searching a high-dimensional design space can be mitigated by turning to a related method called *adaptive design optimization* (ADO). In ADO, which we discuss in the following section, the full optimization problem is broken into pieces that can be evaluated sequentially, as the experiment progresses.

### 10.6.3 Adaptive design optimization

ADO is an adaptive search algorithm that combines design optimization with real-time Bayesian updating of parameter estimates and model probabilities. The idea of *adaptive design optimization* is to treat the full experiment as a sequence of mini-experiments, with the design of the mini-experiments optimized on the fly as the experiment progresses. *Adaptive* in adaptive design optimization refers to the fact that the design of each mini-experiment is adapted based on the results of the preceding mini-experiments. By using all available information about the models and how the participant has responded, ADO collects data intelligently, making it well-suited for evaluating computational models.

ADO has two main advantages over fixed (nonadaptive) DO. The first advantage stems from the fact that ADO optimizes the designs of the mini-experiments sequentially, as the experiment progresses, rather than optimizing them jointly before experimentation begins. This reduces the dimensionality of the search space, thereby greatly reducing the overall computational load. For example, consider a retention experiment in which memory is to be tested at eight different retention intervals. This experiment could be partitioned into a sequence of eight mini-experiments in which memory is to be tested at one retention interval. Finding an optimal design for the entire experiment would entail searching an eight-dimensional Euclidean space, whereas finding an optimal design for each mini-experiment would only entail searching a one-dimensional Euclidean space.

The second advantage pertains to flexibility and data quality. The adaptive nature of ADO, by construction, controls for individual differences and thus makes it well-suited for studying the most common and often largest source of variance in experiments. When participants are tested individually using ADO, the algorithm adjusts (i.e., optimizes) the design of the experiment to the performance of that participant, thereby maximizing the informativeness of the data at an



**Figure 10.10** Schematic diagram illustrating the three-step procedure of adaptive design optimization (ADO).

individual participant level. Response strategies and group differences can be readily identified.

Specifically, an ADO experiment cycles through three basic steps: (1) design optimization (DO); (2) experiment; and (3) Bayesian updating. The relationship between these steps is depicted in Figure 10.10. The process begins with a design optimization step, in which the optimal design for the first mini-experiment is sought (top box in Figure 10.10). This step amounts to solving Equation (10.22) with prior information about the models, i.e., the parameter prior  $p(\theta_m)$  and the model prior  $p(m)$ . Once an optimal design  $d^*$  is identified, the first mini-experiment is carried out with that design (experiment step, lower-right in Figure 10.10). After data have been collected in the mini-experiment, they are used to update the parameter and model priors of each model using Bayes rule to the corresponding posteriors (lower-left in Figure 10.10). Model evaluation and comparison statistics such as the MLE and BIC of each model can also be computed in this step. The updated parameter estimates and model probabilities then become the priors for the next design optimization step, in which the design for the second mini-experiment is identified, and the full cycle repeats. This adaptive and sequential process continues until all of the mini-experiments have been completed.

#### 10.6.4 Illustrative example

In this section we provide an example application of the ADO methodology using a simulated experiment. We illustrate its application to demonstrate that ADO

can successfully identify not only the data-generating model underlying simulated data between two competing models, but also the true parameter values of the model.

Recent findings from developmental studies on how children represent numbers, such as the location of integers on number lines and amount of money, suggested that their numerical estimates can be highly inaccurate and warped (e.g., Opfer and Siegler, 2007). To give an example, children often perceive the difference between number 100 and number 1 as being greater than the difference between 1000 and 901, thereby indicating a compressed, logarithm-like scale representation, instead of the correct, linear scale representation.

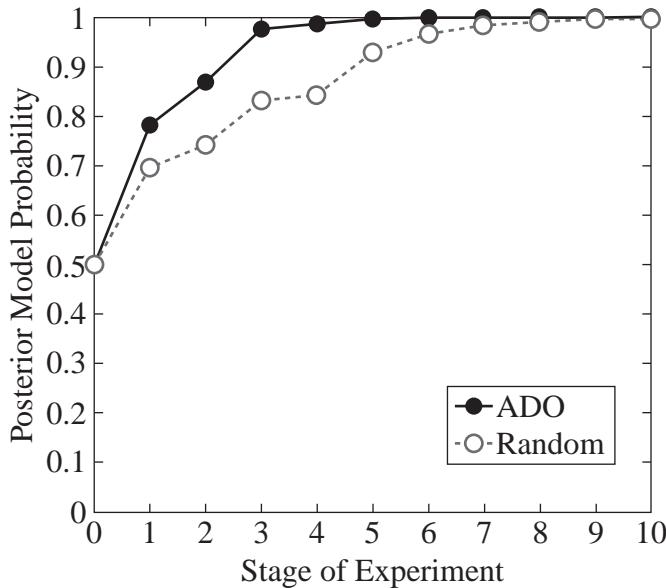
Suppose that you as a developmental psychologist wish to identify the exact form of a child's numerical representation; in particular, whether the representation is linear or logarithmic, with a number-line experimental task. In the task, the child is presented by computer with an integer between 1 and 1000 and asked to indicate how large that number is by placing a vertical hatch with a mouse on a horizontal number line labelled with 1 on the left most end and 1000 on the right most end (Opfer and Siegler, 2007). The two competing models to be discriminated in this experiment are defined as

$$\begin{aligned} LIN : y &= ax + b + e \\ LOG : y &= a \ln(x) + b + e, \end{aligned} \tag{10.23}$$

where  $x$  is the stimulus value between 0 and 1000,  $y$  is the observed response given  $x$ , and  $e$  is a normal random error with zero mean and standard deviation  $s$ . So each model has three parameters,  $\theta = (a, b, s)$ .

To illustrate how ADO works, we conducted a simulated number-line experiment using ADO to select the optimal design (value on the number line), i.e.,  $d = x$ , on each stage of a mini-experiment. The optimal design is the one that "best" discriminates the two models, LIN and LOG in some defined sense. Simulated responses were generated from model LIN with its parameter values of  $\theta = (1.5, -0.5, 0.1)$ . The three-step procedure of ADO shown in Figure 10.10 was repeated for 10 stages of the experiment, each stage consisting of one trial of the experiment. In seeking an optimal design  $d^*$  that maximizes the global utility function  $U(d)$  in Equation (10.22), we employed a local utility function  $u(d, \theta_m, y_m)$  defined as the ratio of two marginal likelihoods in Equation (10.20) of the two competing models, which is known as the Bayes factor.

A summary of the results from the ADO simulation is presented in Figure 10.11. As shown in this figure, ADO clearly outperformed non-ADO, random selection in which the stimulus value was selected randomly and uniformly between 0 and 1000 on each stage, independent of either the observed response or the parameter and model priors. Note that ADO needed just three stages to identify the data-generating model (LIN) with over 0.95 probability, whereas non-ADO random selection required six stages, twice as many, to reach the same level of



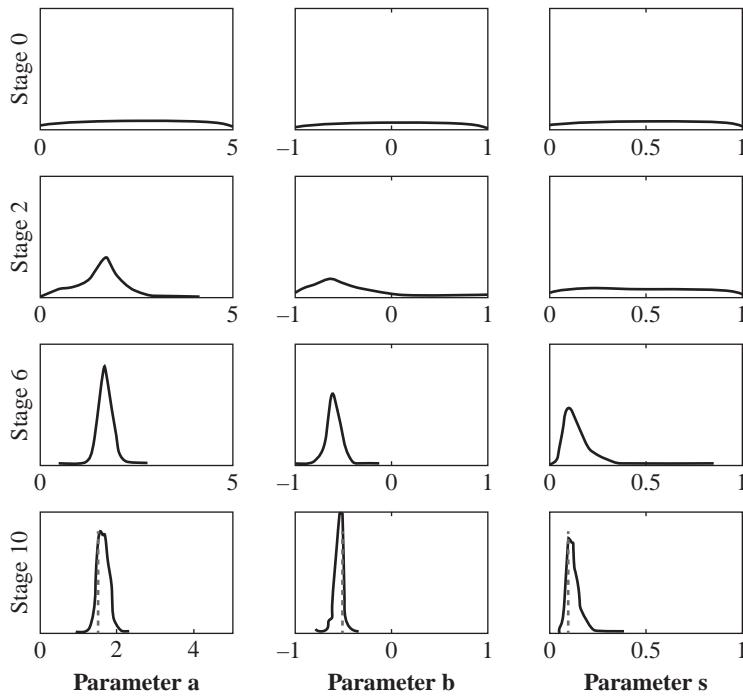
**Figure 10.11** Posterior model probability curves of model LIN as a function of stages of experiment under ADO (solid filled) and random (open broken) conditions. Each curve represents the mean of ten independent replications. The simulated data were generated from model LIN.

performance. Figure 10.12 is another summary of the simulation that shows the posterior distributions of the data-generating model at four selected stages under the ADO condition. Shown in the top row are uniform distributions used for all three parameters at stage 0. Note that as the stage of experiment progresses, all three posterior distributions becomes more peaked and narrower around the true parameter values. In short, these results clearly demonstrate the superior efficiency of the ADO procedure in identifying the underlying data-generating model as well as its parameter values compared to non-ADO design selection.

For additional application examples of ADO in cognitive psychology, including ADO-based experiments with human participants, the reader is referred to Cavignaro *et al.* (2010, 2011, 2013a, 2013b), and Zhang and Lee (2010).

### 10.6.5 Limitations

DO and ADO can increase the efficiency of data collection for the purposes of evaluating and comparing mathematical models, but it is important to be aware of the limitations and assumptions of the methodologies. First, not all design variables in an experiment can be optimized computationally. The variables must be quantifiable in such a way that the likelihood function depends explicitly on the values of the design variables being optimized. Consequently, neither DO nor ADO is applicable to nominal variables (e.g., task modality: words vs. pictures).



**Figure 10.12** Posterior distributions of the LIN model's parameters, shown for four selected stages of experiment. The distributions are approximated using kernel smoothing densities with the normal kernel function. The vertical dotted red (lighter) lines in the bottom row indicate the parameter values with which simulated responses were generated by model LIN, i.e.,  $\theta = (a, b, s) = (1.5, -0.5, 0.1)$ .

Another important limitation is the assumption that one of the models under consideration is the true, data-generating model. This assumption, obviously, is likely to be violated in practice, given that our models are merely approximations of the cognitive process under study. Ideally, one would like to optimize an experiment for an infinite array of models representing a whole spectrum of realities. However, no implementable methodology currently exists that can handle a problem of this scope.

## 10.7 General discussion

Cognitive models are useful to the extent that we understand how they work. The model evaluation tools described in this chapter are meant to provide this understanding, elucidating the performance characteristics of a model and how it performs relative to other models. The pairing of optimal experimental design with post-experiment model selection methods forms a potent combination of tools that can maximize inference about the structure and function of the cognitive process

under study. Having devoted the preceding pages to a description of these tools, we close this chapter by placing them in the context of broader issues in model evaluation and comparison so that readers can understand the strengths and weaknesses of the approach.

It should always be remembered that we have presented a purely quantitative approach to model evaluation. Qualitative criteria, such as model plausibility or interpretability, are not considered, yet they are extremely important to assess a model, so much so that they should be satisfied to a reasonable degree before applying quantitative criteria. If the assumptions underlying the formulation of a model are not sensible or are incompatible with one another, there is no reason to pursue the model further, regardless of how well it might perform on quantitative measures.

The majority of quantitative criteria involve a consideration of a model's fit to the data. As fundamental as it may seem, it has its drawbacks. On the one hand, data are our only link to the cognitive process being studied, so the model must describe the data to some level of accuracy. As discussed in Section 10.4.3, model complexity must be considered to properly interpret the quality of a fit. However, the perils of overfitting can be mitigated by relaxing the objective and making GOF a more qualitative-like criterion, relying instead on a more subjective evaluation as to whether model behavior resembles human behavior.

Most behavioral data do not exhibit complex patterns, which is one reason why there can be such a crowded field of competing models. Too many functions can fit simple data. Of primary importance is whether the model captures the main pattern in the data, whether it be a linear trend of some sort (e.g., Figure 10.5) or an ordinal relation across nominal conditions. If the model mimics human performance sufficiently well, when is it overkill to split hairs regarding whether a competing model absorbs slightly more or less variance in the data? At some point, data fitting can become an unproductive and distracting exercise that no longer advances knowledge.

We are not suggesting that GOF be abandoned, but rather it be used thoughtfully. For example, it is an easy means of weeding out models that fail even to describe the salient trends in the data. It is equally important to recognize when it is ineffective in discriminating among the remaining contenders, especially if repeated experiments will likely fail to identify a clear winner. In fields where there are many contenders (e.g., decision making), a saturation point can be reached at which the models are no longer discriminable, at least given current technology. Because behavioral data are never noise-free, a point will be reached at which competitors mimic each other closely, and are thus similarly good. When confronted with this situation, it could be more productive to compare models on other criteria instead of continuing to design (even optimal) experiments with the hope of discriminating them. In short, avoid obsessing over GOF. It is a mistake to focus on any single criterion.

A situation in which model evaluation tools can be particularly valuable is in deciding whether and how to expand a model. This is a thorny problem. A

mathematical model is usually introduced in a very narrow topic area to explain a particular phenomenon (e.g., recognition memory). To the extent that the model proves valuable, a natural next step is to expand its scope and extend it to related phenomena (e.g., recall). There is little guidance on how to do this, so it is left to the researcher's ingenuity. Model evaluation tools can be instrumental in informing the development process. In particular, they can aid in justifying how the model is revised. What are the consequences of adding more parameters, which can be necessary when additional psychological constructs must be incorporated into the model? How should this new parameter be added, as a multiplicative or exponential term? Extensive simulations are necessary to answer these question. Even then, it can be extremely difficult to decide on the proper decision because of the many constraints that must be satisfied. It is not just that the new model must fit a new type of data or a broader range of data, but that these modifications must also provide explanatory value. That is, they must provide new insight into the operation of the cognitive process. For example, if a common memory mechanism is postulated to be responsible for recognition and recall, its operation in both tasks must be described at a level of detail to understand how the new model differs from and is an improvement over its predecessor. Without this additional yet crucial step, the process of model expansion turns into an exercise in statistical modeling (i.e., data fitting), not cognitive modeling.

A more common and similar endeavor is model revision, whereby an existing model is altered in light of new data that show its performance to be inferior to competing models in some way, usually a poorer fit. The same issues of justifying the change on theoretical grounds as well as on performance apply here. Although revision is a natural part of the scientific process, an overemphasis on quantitative performance such as GOF can eventually yield models that are indistinguishable. They mimic each other closely because they have all, over time, been tuned to fit the same collection of data generated from years of experiments. The plus side of this scenario is that the models are together converging on the underlying cognitive model of interest, it is just being expressed in different forms. Although perhaps wishful thinking, some creative inquiry could identify an overarching (superordinate) model that includes each model as a special case. Short of this, the field is left with competitors that could well prove very difficult to distinguish given that they all grew to be so similar.

A few final words are in order about the evidence justifying the modification or expansion of a model. Model development generally proceeds from the simple to the complex. The higher the bar for justifying the more complex model, the more likely simple models (explanations) will prevail. In this regard, it is unclear how wise it is to abide by Occam's razor, because it is not always clear whether one has truly multiplied entities beyond necessity. A considerable amount of new data collection, not just a single data-fitting exercise or model simulation, is required to determine this.

The bias to favor simple models might also reflect a limitation of cognitive modeling itself. As a model is expanded to account for ever more phenomena, its

behavior can become intractable. When too many parameters combine in complex ways, an understanding of behavior becomes elusive, even if the model mimics human performance impressively well (*Bonini's paradox*). It is for this reason that many of the tools described in this chapter are most productively applied to relatively simple models (<8 parameters). The exception is cross-validation, which is probably the simplest and most versatile model selection tool.

In conclusion, science is driven more by technological advances than theoretical ones, even though technology works in the service of theory. The model evaluation tools discussed in this chapter provide a means of advancing model development. They are not fool-proof, nor are they alone sufficient, but when used with other criteria can assist the researcher in making informed decisions about model design and model choice.

## Acknowledgments

This research is supported in part by National Institute of Health Grant R01-MH093838 to JIM and MAP. The sections of this chapter on model evaluation and model selection draw upon the work of Myung *et al.* (2009). We thank Henrik Singmann for kindly providing an R-code version of the Matlab code used in the simulation study reported in Table 10.1.

## Appendix: Source code

For the convenience of the reader, the source code for model selection simulation in Table 10.1 is provided in both Matlab and R programming languages. It should be noted that running the two programs can possibly yield slightly different numerical results, due to differences in optimization algorithms implemented in Matlab and R.

### Matlab code

```
%%%%+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
%% MATLAB Code for Model Selection Simulation
%%
%% Author: Jay Myung (August 2014), myung.1@osu.edu
%% Distribution: Public & Unlimited
%+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
%+++ Main Function ===
function NHMPtableALL
%-- Initialization
clear;
global n;
```

```

opts=optimset('DerivativeCheck','off','Display','off','TolX',...
    1e-7,'TolFun',1e-7,'Diagnostics','off','MaxIter',500,...
    'LargeScale','on');

n=50;% sample size, i.e., number of binomial trials
t=[0.5 1 2 4 8 12 16 18];t=t';% time intervals
ycnt=[44 34 27 26 19 17 20 11];% observed correct responses
y=ycnt/n;y=y';
x=n*y;

%--- MLE, AIC & BIC
[am1,loglik1]=fmincon('power_mle',rand(2,1),[],[],[],[],...
    zeros(2,1),1*ones(2,1),[],opts,t,x);
[am2,loglik2]=fmincon('power2_mle',rand(3,1),[],[],[],[],...
    zeros(3,1),1*ones(3,1),[],opts,t,x);
yprd1=am1(1,1)*(t+1).^( -am1(2,1));
r2(1,1)=1-sum((yprd1-y).^2)/sum((y-mean(y)).^2);
yprd2=am2(1,1)*(t+1).^( -am2(2,1))+am2(3,1);
r2(2,1)=1-sum((yprd2-y).^2)/sum((y-mean(y)).^2);
%
pinit=[rand(5,1)];
pinit=[.906 .836 .263 -.094 .769]';% initial seed values for MLE
plower=[0 0 0 -.3 0];
pupper=[1 1 1 1 10]';
[am3,loglik3]=fmincon('power3_mle',pinit,[],[],[],plower,',...
    pupper,[],opts,t,x);
yprd3=am3(1,1)*(t+1).^( -am3(2,1))+am3(3,1)+am3(4,1)*sin(am3(5,1)*t);
r2(3,1)=1-sum((yprd3-y).^2)/sum((y-mean(y)).^2);

[am4,loglik4]=fmincon('expo_mle',rand(2,1),[],[],[],[],...
    zeros(2,1),1*ones(2,1),[],opts,t,x);
yprd4=am4(1,1)*exp(-am4(2,1)*t);
r2(4,1)=1-sum((yprd4-y).^2)/sum((y-mean(y)).^2);

%-----
logml=[loglik1 loglik2 loglik3 loglik4]';logml=(-1)*logml;
% Log ML
aic=[2*loglik1+2*2 2*loglik2+2*3 2*loglik3+2*5 2*loglik4+2*2]';
% AIC
bic=[2*loglik1+2*log(n) 2*loglik2+3*log(n) 2*loglik3+5*log(n) ...
    2*loglik4+2*log(n)]'; % BIC

disp('--R2 LogML AIC BIC -----');
disp(num2str([r2 logml aic bic], '% 10.3f'));
disp('-- MLE estimates -----');

```

```

disp(num2str([am1'],'% 10.3f'));
disp(num2str([am2'],'% 10.3f'));
disp(num2str([am3'],'% 10.3f'));
disp(num2str([am4'],'% 10.3f'));

%--- Plot the results
tt=(0.1:1:20)';
yPow=am1(1,1)*(tt+1).^( -am1(2,1));
yPow2=am2(1,1)*(tt+1).^( -am2(2,1))+am2(3,1);
yPow3=am3(1,1)*(tt+1).^( -am3(2,1))+am3(3,1)+am3(4,1)*sin(am3(5,1)*tt);
yExp=am4(1,1)*exp(-am4(2,1)*tt);

clf;
plot(tt,yPow,'k-',tt,yExp,'b--',tt,yPow2,'r-',tt,yPow3,'g-',...
      'LineWidth',3);hold on;grid on;
xlim([0 20]);ylim([0 1]);xlabel('Time t', 'FontSize', 20);
ylabel('Probability p', 'FontSize', 20);
plot(t,y,'ko','MarkerFaceColor','k','MarkerSize',10);

%--- LOOCV
bm1=am1;bm2=am2;bm3=am3;bm4=am4;

tCV=zeros(7,1);xCV=zeros(7,1);
loocv=zeros(8,4);
for jj=1:8
if jj==1; tCV=t(2:8,:);xCV=x(2:8,:);
elseif jj==8;tCV=t(1:7,:);xCV=x(1:7,:);
else tCV=[t(1:jj-1,:);t(jj+1:8,:)];xCV=[x(1:jj-1,:);x(jj+1:8,:)];
end;

[am1]=fmincon('power_mle',bm1,[],[],[],[],zeros(2,1),...
    1*ones(2,1),[],opts,tCV,xCV);
[am2]=fmincon('power2_mle',bm2,[],[],[],[],zeros(3,1),...
    1*ones(3,1),[],opts,tCV,xCV);
pinit=[.906 .836 .263 -.094 .769]';% PVAF 98.8
pLower=[0 0 0 -.3 0];
pUpper=[1 1 1 1 10]';
[am3]=fmincon('power3_mle',bm3,[],[],[],[],pLower,pUpper,... 
    [],opts,tCV,xCV);
[am4]=fmincon('expo_mle',bm4,[],[],[],[],zeros(2,1),...
    1*ones(2,1),[],opts, tCV,xCV);

loglik1=power_mle(am1,t(jj,1),x(jj,1));
loglik2=power2_mle(am2,t(jj,1),x(jj,1));

```

```

loglik3=power3_mle(am3,t(jj,1),x(jj,1));
loglik4=expo_mle(am4,t(jj,1),x(jj,1));
loocv(jj,:)=[loglik1 loglik2 loglik3 loglik4];
end;% jj
disp('--- LOOCV -----');
disp(num2str([mean(loocv)' ], '% 10.3f'));

%--- APE
bm1=am1;bm2=am2;bm3=am3;bm4=am4;

apepow=zeros(5,1);apeexp=zeros(5,1);
for jj=1:5;
    tape=t(1:2+jj,:);xape=x(1:2+jj,:);
    [am1]=fmincon('power_mle',bm1,[],[],[],[],zeros(2,1),...
        1*ones(2,1),[],opts,tape,xape);
    [am4]=fmincon('expo_mle',bm4,[],[],[],[],zeros(2,1),...
        1*ones(2,1),[],opts,tape,xape);
    loglik1=power_mle(am1,t(jj+3,1),x(jj+3,1));
    loglik4=expo_mle(am4,t(jj+3,1),x(jj+3,1));
    apepow(jj,1)=loglik1;
    apeexp(jj,1)=loglik4;
end;% jj

apepow2=zeros(4,1);
for jj=1:4;
    tape=t(1:3+jj,:);xape=x(1:3+jj,:);
    [am2]=fmincon('power2_mle',bm2,[],[],[],[],zeros(3,1),...
        1*ones(3,1),[],opts,tape,xape);
    loglik2=power2_mle(am2,t(jj+4,1),x(jj+4,1));
    apepow2(jj,1)=loglik2;
end;% jj

apepow3=zeros(2,1);
for jj=1:2;
    tape=t(1:5+jj,:);xape=x(1:5+jj,:);
    pinit=[.906 .836 .263 -.094 .769]';% PVAF 98.8
    plower=[0 0 0 -.3 0];
    pupper=[1 1 1 1 10]';
    [am3]=fmincon('power3_mle',bm3,[],[],[],[],plower,pupper, ...
        [],opts,tape,xape);
    loglik3=power3_mle(am3,t(jj+6,1),x(jj+6,1));
    apepow3(jj,1)=loglik3;
end;% jj

```

```

disp('--- APE -----');
disp(num2str([mean(apepow) mean(apepow2) mean(apepow3) ...
    mean(apeexp)]', '% 10.3f'));

%+++ End of Main Function +++

function loglik = power_mle(a,t,x)
global n
[mc,mr]=size(x);
p=a(1,1)*(t+1).^( -a(2,1));
p=(p < ones(mc,1)).*p+(p >= ones(mc,1))* .999999;
loglik=(-1)*(x.*log(p)+(n-x).*log(1-p));
loglik=sum(loglik);

function loglik = power2_mle(a,t,x)
global n
[mc,mr]=size(x);
p=a(1,1)*(t+1).^( -a(2,1))+a(3,1);
p=(p < ones(mc,1)).*p+(p >= ones(mc,1))* .999999;
loglik=(-1)*(x.*log(p)+(n-x).*log(1-p));
loglik=sum(loglik);

function loglik = power3_mle(a,t,x)
global n
[mc,mr]=size(x);
p=a(1,1)*(t+1).^( -a(2,1))+a(3,1)+a(4,1)*sin(a(5,1)*t);
p=(p < ones(mc,1)).*p+(p >= ones(mc,1))* .999999;
loglik=(-1)*(x.*log(p)+(n-x).*log(1-p));
loglik=sum(loglik);

function loglik = expo_mle(a,t,x)
global n
[mc,mr]=size(x);
p=a(1,1)*exp(-a(2,1)*t);
p=(p < ones(mc,1)).*p+(p >= ones(mc,1))* .999999;
loglik=(-1)*(x.*log(p)+(n-x).*log(1-p));
loglik=sum(loglik);

```

### R code

```

#####
## R Code for Model Selection Simulation
##
## Author: Henrik Singmann, (July 2014), singmann@gmail.com

```

```

## Distribution: GPL v2
#####
#### initialization
n <- 50 # sample size
t <- c(0.5, 1, 2, 4, 8, 12, 16, 18) # time intervals
ycnt <- c(44, 34, 27, 26, 19, 17, 20, 11)
y <- ycnt/n

#### models
pow1 <- function(par, t) par[1]*(t+1)^(-par[2])
pow2 <- function(par, t) par[1]*(t+1)^(-par[2]) + par[3]
pow3 <- function(par, t) par[1]*(t+1)^(-par[2]) + par[3] +
    par[4]*sin(par[5]*t)
expl <- function(par, t) par[1]*exp(-par[2]*t)

#### wrapper functions for optimization:
log_lik <- function(par, model, t, d) {
  pred <- model(par, t = t)
  -sum(d*log(pred)+(n-d)*log(1-pred)) #returns likelihood
}

r_squared <- function(par, model, t) {
  pred <- model(par, t = t)
  1-sum((pred-y)^2)/sum((y-mean(y))^2) #returns PVAF
}

#### obtain MLEs and PVAF

set.seed(1) # seed ensures that MLEs are obtained for full data.

loglik_pow <- nlminb(runif(2), log_lik, model = pow1, t = t, d = ycnt)
loglik_pow2 <- nlminb(runif(3), log_lik, model = pow2, t = t, d = ycnt)
# initial values for pow3 (to make sure initial values work):
initial <- c(.906, .836, .263, -.094, .769)
loglik_pow3 <- nlminb(initial, log_lik, model = pow3, t = t, d = ycnt)
loglik_exp <- nlminb(runif(2), log_lik, model = expl, t = t, d = ycnt)

pvaf_pow <- r_squared(loglik_pow$par, model = pow1, t = t)
pvaf_pow2 <- r_squared(loglik_pow2$par, model = pow2, t = t)
pvaf_pow3 <- r_squared(loglik_pow3$par, model = pow3, t = t)
pvaf_exp <- r_squared(loglik_exp$par, model = expl, t = t)

# prepare output
n_pars <- c(2, 3, 5, 2)

```

```

round(pvaf <- c(pvaf_pow, pvaf_pow2, pvaf_pow3, pvaf_exp), 3)
round(loglik <- -1*c(loglik_pow$objective, loglik_pow2$objective,
                      loglik_pow3$objective, loglik_exp$objective), 2)
round(aic <- -2*loglik+2*n_pars, 2)
round(bic <- -2*loglik+n_pars*log(n), 2)

# plot of results
par(lwd = 2)
new_x <- seq(-5, 25, 0.1)
plot(new_x, pow1(loglik_pow$par, t = new_x), type = "l", lty = 1,
     ylim = c(0, 1), xlim = c(0, 20))
lines(new_x, pow2(loglik_pow2$par, t = new_x), lty = 2)
lines(new_x, pow3(loglik_pow3$par, t = new_x), lty = 3)
lines(new_x, expl(loglik_exp$par, t = new_x), lty = 4)
points(t, y, pch = 16, cex = 1.5)
legend("topright", c("POW", "POW2", "POW3", "EXP", "Data"),
       lty = c(1:4, -1), pch = c(rep(-1, 4), 16), cex = 1.2)

### LOOCV
get_loocv <- function(x, model, initial) {
  tmp_par <- nlminb(initial, log_lik, model = model, t = t[-x],
                     d = ycnt[-x])$par
  log_lik(tmp_par, model = model, t = t[x], d = ycnt[x])
}

loocv_pow <- mean(sapply(seq_along(t), get_loocv, model = pow1,
                          initial = loglik_pow$par))
loocv_pow2 <- mean(sapply(seq_along(t), get_loocv, model = pow2,
                           initial = loglik_pow2$par))
loocv_pow3 <- mean(sapply(seq_along(t), get_loocv, model = pow3,
                           initial = loglik_pow3$par))
loocv_exp <- mean(sapply(seq_along(t), get_loocv, model = expl,
                           initial = loglik_exp$par))
round(loocv <- c(loocv_pow, loocv_pow2, loocv_pow3, loocv_exp), 2)

### APE
get_ape <- function(x, model, initial) {
  tmp_par <- nlminb(initial, log_lik, model = model, t = t[1:x],
                     d = ycnt[1:x])$par
  log_lik(tmp_par, model = model, t = t[(x+1)], d = ycnt[(x+1)])
}

ape_pow <- mean(sapply(3:7, get_ape, model = pow1,
                       initial = loglik_pow$par))

```

```

ape_pow2 <- mean(sapply(4:7, get_ape, model = pow2,
                        initial = loglik_pow2$par))
ape_pow3 <- mean(sapply(6:7, get_ape, model = pow3,
                        initial = loglik_pow3$par))
ape_exp <- mean(sapply(3:7, get_ape, model = expl,
                        initial = loglik_exp$par))
round(ape <- c(ape_pow, ape_pow2, ape_pow3, ape_exp), 2)

### Final output:
round(data.frame(n_pars, pvaf, loglik, aic, bic, loocv, ape,
                 row.names = c("POW", "POW2", "POW3", "EXP")) , 2)

```

## References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N. and Caski, F. (eds), *Proceedings of the Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, pp. 267–287.
- Amzal, B., Bois, F. Y., Parent, E. and Robert, C. P. (2006). Bayesian-optimal design via interacting particle systems. *Journal of the American Statistical Association*, **101**, 773–785.
- Atkinson, A. C. and Donev, A. N. (1992). *Optimum Experimental Designs*. Oxford: Oxford University Press.
- Atkinson, A. C. and Federov, V. V. (1975). Optimal design: experiments for discriminating between several models. *Biometrika*, **62**, 289.
- Balasubramanian, V. (1997). Statistical inference, Occam's razor and statistical mechanics on the space of probability distributions. *Neural Computation*, **9**, 349–368.
- Bamber, D. and van Santen, J. P. H. (1985). How many parameters can a model have and still be testable. *Journal of Mathematical Psychology*, **29**, 443–473.
- Bamber, D. and van Santen, J. P. H. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology*, **44**, 20–40.
- Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, **71**, 791–799.
- Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, **44**, 108–132.
- Busemeyer, J. R. and Diederich, A. (2010). *Cognitive Modeling*. Thousand Oaks, CA: Sage Publications.
- Casella, G. and Berger, R. L. (2002). *Statistical Inference* (2nd edition). Pacific Grove, CA: Duxbury.
- Cavagnaro, D. R., Myung, J. I., Pitt, M. A. and Kujala, J. V. (2010). Adaptive design optimization: a mutual information based approach to model discrimination in cognitive science. *Neural Computation*, **22**, 887–905.
- Cavagnaro, D. R., Pitt, M. A. and Myung, J. I. (2011). Model discrimination through adaptive experimentation. *Psychonomic Bulletin & Review*, **18**, 204–210.
- Cavagnaro, D. R., Pitt, M. A., Gonzalez, R. and Myung, J. I. (2013a). Discriminating among probability weighting functions using adaptive design optimization. *Journal of Risk and Uncertainty*, **47**, 255–289.

- Cavagnaro, D. R., Gonzalez, R., Myung, J. I. and Pitt, M. A. (2013b). Optimal decision stimuli for risky choice experiments: an adaptive approach. *Management Science*, **59**, 358–375.
- Chaloner, K., and Verdinelli, I. (1995). Bayesian experimental design: a review. *Statistical Science*, **10**, 273–304.
- Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Hoboken, NJ: John Wiley & Sons, Inc.
- Dawid, A. P. (1984). Statistical theory: the prequential approach. *Journal of the Royal Statistical Society, Series A*, **147**, 278–292.
- Fum, D., Del Missier, F. and Stocco, A. (2007). The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,1000 words. *Cognitive Systems Research*, **8**, 135–142.
- Gelman, A., Carlin, J. B., Stern, H. S., *et al.* (2013). *Bayesian Data Analysis* (3rd edition). Boca Raton, FL: Chapman & Hall/CRC.
- Grünwald, P. D. (2005). A tutorial introduction to the minimum description length principle. In Grünwald, P., Myung, I. J. and Pitt, M. A. (eds), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.
- Grünwald, P. D. (2007). *The Minimum Description Length Principle*. Cambridge, MA: MIT Press.
- Grünwald, P. D., Myung, I. J. and Pitt, M. A. (eds). (2005). *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.
- Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society: Statistical Methodology B*, **21**, 272–319.
- Klauer, K. C. and Kellen, D. (2011). The flexibility of models of recognition memory: an analysis by the minimum-description length principle. *Journal of Mathematical Psychology*, **55**, 430–450.
- Lee, M. D. and Wagenmakers, E-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge: Cambridge University Press.
- Lewandowsky, S. and Farrell, S. (2011). *Computational Modeling in Cognition: Principles and Practice*. Thousand Oaks, CA: Sage Publications.
- Linhart, H. and Zucchini, W. (1986). *Model Selection*. New York, NY: John Wiley & Sons.
- McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, **1**, 11–38.
- Müller, P., Sanso, B. and De Iorio, M. (2004). Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association*, **99**, 788–798.
- Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, **47**, 90–100.
- Myung, I. J. and Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79–95.
- Myung, I. J., Brunsman, A. E. and Pitt, M. A. (1999). True to thyself: assessing whether computational models of cognition remain faithful to their theoretical principles. In: Hahn, M., and Stoness, S. C. (eds), *Proceedings of the 21st Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 462–467.

- Myung, I. J., Forster, M. and Browne, M. W. (2000). Special issue on model selection. *Journal of Mathematical Psychology*, **44**, 1–2.
- Myung, I. J., Pitt, M. A. and Kim, W. (2005). Model evaluation, testing and selection. In: Lambert, K., and Goldstone, R. (eds), *The Handbook of Cognition*. Thousand Oaks, CA: Sage Publications, pp. 422–436.
- Myung, I. J., Navarro, D. J. and Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, **50**, 167–179.
- Myung, J. I. and Pitt, M. A. (2004). Model comparison methods. *Methods in Enzymology*, **383**, 351–366.
- Myung, J. I. and Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, **58**, 499–518.
- Myung, J. I., Tang, Y. and Pitt, M. A. (2009). Evaluation and comparison of computational models. *Methods in Enzymology*, **454**, 287–304.
- Myung, J. I., Cavagnaro, D. R. and Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology*, **57**, 53–67.
- Navarro, D. J., Pitt, M. A. and Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, **49**, 47–84.
- Opfer, J. and Siegler, R. (2007). Representational change and children's numerical estimation. *Cognitive Psychology*, **55**, 165–195.
- Pitt, M. A. and Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, **6**, 421–425.
- Pitt, M. A., Myung, I. J. and Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, **190**, 472–491.
- Polk, T. A. and Seifert, C. M. (eds). (2002). *Cognitive Modeling*. Cambridge, MA: MIT Press.
- Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, **42**, 40–70.
- Rissanen, J. (2001). Strong optimality of the normalized ML models as universal code and information in data. *IEEE Transactions on Information Theory*, **42**, 1712–1717.
- Roberts, S. and Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, **107**, 358–367.
- Rubin, D. C. and Wenzel, A. E. (1996). One hundred years of forgetting: a quantitative description of retention. *Psychological Review*, **103**, 734–760.
- Schervish, M. J. (1995). *The Theory of Statistics*. New York, NY: Springer-Verlag.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.
- Shiffrin, R. M. and Nobel, P. A. (1997). The art of model development and testing. *Behavior Research Methods, Instruments, & Computers*, **29**, 6–14.
- Shiffrin, R. M., Lee, M. D., Kim, W. and Wagenmakers, E-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, **32**, 1248–1284.
- Singman, H. and Kellen, D. (2013). MPTinR: analysis of multinomial processing tree models in R. *Behavioral Research Methods*, **45**, 560–575.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, **36**, 111–147.
- Su, Y., Myung, I. J., Pitt, M. A. and Kim, W. (2005). Minimum description length and cognitive modeling. In: Grünwald, P. D., Myung, I. J., and Pitt, M. A. (eds), *Advances*

- 
- in *Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press, pp. 411–433.
- Wagenmakers, E-J. and Waldorp, L. (2006). Editors' introduction. *Journal of Mathematical Psychology*, **50**, 99–100.
- Wagenmakers, E-J., Grünwald, P. D. and Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, **50**, 149–166.
- Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, **44**, 92–107.
- Wixted, J. T. and Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, **2**, 409–415.
- Wu, H., Myung, J. I. and Batchelder, W. H. (2010). Minimum description length model selection of multinomial processing tree models. *Psychonomic Bulletin & Review*, **17**, 276–286.
- Zhang, S. and Lee, M. D. (2010). Optimal experimental design for a class of bandit problems. *Journal of Mathematical Psychology*, **54**, 499–508.

# Index

- $\sigma$ -algebra, 8
- ABO blood group model, 472, 479, 489
- accessible (knowledge structure), 279
- accumulative prediction error, 571
- adapted QUERY routine, 311
- adaptive design optimization, 581
- additive Cauchy equation, 153
- additive representation, 152
- adjacency matrix, 198
- adjusted QUERY routine, 312
- AIC, 527
- Akaike information criterion, 569
- almost hanging (state), 310
- antimatroid, 279
- antisymmetric relation, 283
- assortativity,  $r$ , 262
- asymptotically stable, 344
- atom, 284
- attractor, 345
- attribution function, 285
- base, 283
- basin of attraction, 345
- Bayes factor, 525, 573
- Bayes rule, 510
- Bayesian hierarchical modeling, 504, 545
- Bayesian information criterion, 569
- Bayesian probability, 509
- Bernoulli
  - graph model, 252
  - model, 249
- BIC, 527
- big-O notation,  $O(g(n))$ , 201
- binary (preference) relations, 386
  - lexicographic, 388
  - orders
    - interval, 387
    - semi-, 387
    - strict linear, 387
    - strict partial, 387
    - strict weak, 387
    - weak, 387
- binary multinomial processing tree, 466, 469, 486
- binocular vision, 173
- bipolar coordinates, 174
- birth–death process, 364
- Bishop–Cannings theorem, 329
- Bonini’s paradox, 588
- BUGS, 517
- careless error probability, 300
- child, 299
- choice
  - binary, 383
  - Luce model, 431
  - multinomial logit (MNL) model, 431
- probabilities
  - attraction effect, 440
  - best, 383
  - best choice induced by rankings, 399
  - best-worst, 383
  - best-worst choice induced by rankings, 400
  - binary, 383
  - binary choice induced by rankings, 397
  - compromise effect, 441
  - constant error model, 390
  - context effects, 439
  - distance-based specification, 393
  - independence of irrelevant alternatives, 431
  - regularity of, 429
  - similarity effect, 441
  - tremble model, 390
  - worst, 383
- probability distributions, 382
- clause, 285
- clique, 206, 247
- cognitive modeling, 553
- competing risks model
  - hazard based, 30
  - latent-failure times, 32
  - Weibull, 30
- complementation, 246
- completely monotone function, 61
- computational complexity, 199
- concatenation (of a word and an item), 289
- conditional posterior distributions, 515

- conditionally  
   independent, 244  
   uniform models, 259  
 conjugate prior, 513  
 connected, 205  
   component, 205  
 constant ratio condition, 155  
 context-free grammar, 490  
 convex  
   hull, 395  
   polytope, 396  
   set, 395  
 convex dimension, 294  
 convolution, 20  
 copula, 56  
   Archimedean  
     generator, 61  
   Archimedean, 60  
 bivariate extreme value, 57  
 Clayton, 62  
 co-, 60  
 dual of, 60  
 $n$ -dimensional, 57  
 density, 58  
 Fréchet bound, 60  
 independence, 58  
 survival, 59  
 vine, 58  
 coupling, 44, 108, 119  
   event, 48  
   inequality, 48  
 i.i.d., 46  
 identity, 119  
 independent, 108, 147  
 maximal, 49  
 quantile, 46  
 reduced, 117–119, 122, 124, 129,  
   136  
 self-, 46  
 covered (state covered by), 278  
 covering diagram, 278  
 cross-validation, 571  
  
 d'Alembert's functional equation, 169  
 degree  
   digraph, 202  
   graph, 203  
   sequence, 203  
 dependence, 63  
   Kendall's tau, 54  
   negative, 66  
   negative lower orthant (*NLOD*), 66  
   negative upper orthant (*NUOD*), 66  
   positive, 63  
   positive lower orthant (*PLOD*), 64  
   positive quadrant, 77  
  
   positive upper orthant (*PUOD*), 64  
   Spearman's rho, 67  
 dependency graph, 247  
 derivable operation, 62  
 design optimization, 579  
 determines (a word determines), 289  
 Deviance Information Criterion (DIC), 524, 527  
 diameter, 204  
 digraph, 469, 485  
 discrepancy function, 567  
 discriminative (knowledge structure), 276  
 discriminative (surmise system), 286  
 distribution for degrees  
    $\beta$ -negative binomial, 218  
   negative binomial, 219  
   Poisson, 218  
   power law, 218  
   Zipf's law, 222  
 distribution function, 11  
   Beta, 75  
   bivariate Marshall–Olkin, 15  
   exponential, 12  
   extreme value, 37  
     Fréchet type, 40  
     Gumbel type, 40  
     Weibull type, 40  
   inverse Gauss (or Wald), 20  
   Kumaraswamy, 75  
   normal, 20  
   of the same type, 38  
   sub-, 30  
   uniform, 27  
   Weibull, 12  
 domain, 276  
 downgradable (knowledge structure), 279  
 dyad independence model, 254  
  
 eigenvalue, 226  
   dominant, 232  
   second smallest Laplacian, 238  
 embedding, 207  
 empty word, 289  
 encoded (learning space), 294  
 entropies of degree  $\alpha$ , 170  
 equality  
   in distribution, 10  
   point-wise, 10  
 equilibrium evolutionarily stable set, 335  
 error function, 171  
 evolutionarily stable set, 334  
 evolutionarily stable state, 348  
 evolutionarily stable strategy  
   Maynard Smith and Price definition, 326  
   Taylor–Jonker definition, 327  
 exchangeability (of random variables), 24  
 explanatory adequacy, 563

- 
- exponential random graph models (ERGMs),  
     244, 258  
 extended distribution, 300  
 extra problem, 314  
 facet  
     defining, 396  
 factor, 111  
 failure rate, *see* hazard (rate) function  
 faithfulness, 563  
 falsifiable, 562  
 Fechner's theory, 159  
 finite (knowledge structure), 276  
 Fisher information approximation, 565, 569  
 fixed point  
     definition, 344  
     hyperbolic, 353  
 Fréchet–Hoeffding bound, 50  
     comonotonicity, 51  
     countermonotonicity, 51  
 frequentist probability, 508  
 full binary tree, 469, 493  
 functional equation, 151  
 functional form, 565  
 fundamental equation of information measures,  
     153  
 fundamental theorem of natural selection, 358  
 gambles, 183  
     branches of, 408  
     first-order  
         of gains, 408  
     idempotence, 409  
     mixed, 415  
     risky, 409  
     uncertain, 409  
     utility representations  
         configural weighted, 411  
         cumulative prospect theory, 410  
         ranked additive, 407  
         ranked weighted, 407  
         TAX, 411  
 game  
     Battle of the sexes, 342, 360  
     coordination, 342  
     doubly symmetric, 358  
     hawk–dove, 343  
     Prisoners' Dilemma, 341, 367  
     Rock–Paper–Scissors, 328, 344  
 generalizability, 567  
 Gibbs sampler, 515, 531  
 global identification, 468, 478  
 goodness-of-fit, 563  
 gradation, 312  
 granular (knowledge space), 284  
 graph, 202  
     bipartite, 204  
     connected, 205  
     connectivity,  $\kappa$ , 206  
     cycle, 202,  $C_N$ , 204  
     path, 202,  $P_N$ , 204  
     star,  $S_N$ , 204  
     subgraph, 205, induced —,  $< V >$ , 205  
 groundedness, 57  
 half-split rule, 305  
 Hammersley–Clifford (H-C) theorem, 244, 248  
 hanging (state), 310  
 hazard (rate) function, 27  
     cause-specific, 30  
     cumulative, 28  
     sub-, 30  
 hidden Markov chain, 455, 458, 460  
 hierarchical dual-process models, 540  
 hierarchical linear models, 507  
 hierarchical model, 499  
 hierarchical modeling, 506  
 hierarchical nonlinear models, 507  
 hierarchical prior, 520  
 hierarchical shrinkage, 520  
 hierarchical single-process models, 544  
 horse race model  
     context-dependent, 430  
     context-free, 429  
     distribution-free, 428  
     independent random utility, 429  
     linear ballistic accumulator model  
         additive, 438  
         drift rate variability, 438  
         multiattribute, 441  
         multiplicative, 434  
         of choice and response time, 439  
         parallel, 439  
         start point variability, 434  
     of choice, 427  
 hyperbolae of Hillebrand, 173  
 identity  
     Hoeffding's, 19  
 imitate-the-best, 367  
 improper prior, 513  
 inner fringe, 288  
 input, 85, 110, 111  
 instance, 275  
 interaction neighborhood, 367  
 interpretability, 563  
 item, 276  
 JAGS, 516, 545  
 joint posterior, 514  
 joint prior, 514  
 joint receipt, 183

- 
- knowledge space, 279  
 knowledge state, 276  
 knowledge structure, 276  
 Laplace approximation, 526  
 Laplace transform, 61  
 Laplacian  
     eigenvalue, algebraic connectivity, 236, 238,  
         262  
     eigenvector, Fiedler, 238  
     matrix, 236  
 latent parameters, 244  
 latent state, 306  
 learning diagram, 292  
 learning path, 312  
 learning sequence, 289  
 learning space, 278  
 learning word, 289  
 least squares estimation, 509, 560  
 leave-one-out-cross-validation, 572  
 length (of a word), 289  
 likelihood function, 497, 511, 512, 561  
 linear ballistic accumulator model, 433  
 linear ordering  
     model  
         of best choice probabilities, 399  
         of best-worst choice probabilities, 400  
     polytope, 398  
         of best choice probabilities, 399  
         of best-worst choice probabilities, 402  
         of worst choice probabilities, 399  
 local identification, 468  
 local interaction model, 366  
 lucky guess probability, 300  
 Lyapunov stability, 344  
 Lyapunov stability theorem, 347  
 Lyapunov's indirect method, 353  
*m*-dimensional fundamental equation of  
     information, 171  
 marginal likelihood, 525, 573  
 marginal posterior, 514  
 marginal selectivity, 125  
 Markov  
     chain, 249  
     chain Monte Carlo (MCMC), 256  
     graph, 255  
 Markov chain, 458  
 Markov chain Monte Carlo (MCMC), 515  
 mass-at-chance (MAC) link, 530  
 maximized likelihood, 563  
 maximum likelihood distribution, 567, 570  
 maximum likelihood estimation, 509, 560  
 measurable  
     function, 9  
     space, 8  
 measurable set, *see* sigma-algebra  
 minimum description length, 570  
 model complexity, 565  
 model evaluation, 562  
 model revision, 587  
 model selection, 569  
 model validity, 478  
 morphism  
     auto—, 207  
     homeomorphic, 208  
     iso—, 207  
 multinomial distribution, 496  
 multiplicative updating rule, 306  
 Nash equilibrium  
     components of, 335  
     definition, 324  
     existence of, 328  
     inadequate for evolutionary stability,  
         325  
 relation between strict and ESS, 326  
 rest point of continuous replicator dynamics,  
     350  
 solution concept, 324  
     strict, 326  
 neutrally stable, 344  
 normalized maximum likelihood, 565, 569  
 normalizing constant, 258  
 nuisance variation, 506  
 numerosity decision task, 528  
 Occam's razor, 567  
 OpenBUGS, 516, 517  
 optimal design, 577  
 order  
     convex, 71  
     dilation, 71  
     dispersive, 72  
     hazard rate, 69  
     likelihood ratio, 70  
     Lorenz, 72  
     lower orthant, 76  
     positive dependence, 76  
     quantile spread, 73  
     stochastic, 68  
         univariate, 68  
     strong stochastic, 75  
     upper orthant, 76  
     variability, 71  
 order statistics, 34  
 ordinal (space), 282  
 outer fringe, 288  
 output, 85, 110, 112  
     dependence on inputs, 110  
     selectivity in, *see* selective influence  
 overfitting, 564

- 
- pair-clustering model, 474  
 parameter estimation, 560  
 parametric order constraint, 481  
 parametrized (probabilistic knowledge structure), 300  
 parent (structure), 299  
 partial order, 282  
 partially ordinal (space), 282  
 payoff function, 324  
 penalized-likelihoods, 569  
 percent variance accounted for, 563  
 Pexider equation, 152  
 Pexiderized version, 159  
 Plateau's experiment, 164  
 plausibility, 563  
 Poisson process, 22  
     non-homogeneous, 23  
 polytope  
     face of a, 396  
     facet description of, 396  
     facet of a, 396  
     facet-defining inequalities of a, 396  
     full-dimensional, 396  
     hypercube, 384  
     minimal description of, 396  
     vertex description, 396  
 posterior, 510  
 posterior beliefs, 511  
 posterior distribution, 513  
 power set, 278  
 preference order, 183  
 preferences  
     mixture of, 394  
     random, 395  
     revealed, 376  
     stated, 376  
 preferential attachment, 219  
 prefix (of a word), 289  
 preorder traversal, 493  
 prior, 510  
 prior beliefs, 511  
 prior distribution, 512  
 probabilistic (knowledge structure), 300  
 probabilistic digraph model (PDM), 241, 252  
 probabilistic graph model (PGM), 241, 252  
 probabilistic projection (of a probabilistic knowledge structure), 300  
 probability  
     measure, 8  
     space, 8  
 probability density function, 559  
 probability redistribution vector, 332  
 probability simplex, 467  
 probit transformation, 529  
 product multinomial, 497  
 product space, 19  
 projected distribution, 300  
 projection (of a knowledge structure), 297  
 psychological measurement, 152  
 psychophysical invariances, 176  
 quantile function, 26  
     density, 26  
     hazard, 29  
     quantile spread, 73  
 quasi order, 282  
 quasi ordinal (space), 282  
 quasiarithmetic mean, 181  
 query, 308, 382  
 race model, 51  
     inequality, 51  
 radius, rad, 204  
 random  
     advantage model, 444  
     element, 46  
     function model, 422  
     preference, 395  
     relation model, 395  
     utility model, 420  
         distribution-free, 420  
         noncoincident, 420  
         unidimensional, 420  
 utility representation  
     EBA, 430  
     of best choice, 424  
     of best-worst choice, 424  
     of interval orders, 421  
     of semiorders, 421  
     of strict linear orders, 421  
     of strict weak orders, 421  
     of weak orders, 421  
 variable, 10  
 vector, 13  
 vector  
     associated random vectors, 64  
     negatively associated random vectors ( $NA$ ), 66  
 random variable, 92, 100  
     as measurable function, 105  
     continuous, 101  
     discrete, 100  
     distribution of, 92  
     equality, 109  
     functions of, 102  
     identity of, 94, 98, 109  
     jointly distributed, 95  
     marginal, 98, 100, 105  
     sameness, *see* identity of  
     stochastically independent, 97  
     stochastically unrelated, 107  
 rank, 387

- 
- rank-dependent utility representation, 184  
 receiver operating characteristic (ROC), 538, 543  
 recognition memory, 456, 476, 479, 495, 538,  
     540  
 record, 40  
     counting process, 41, 43  
     inter-times, 41  
     time, 41  
     value sequence, 41  
 relational structure, 198, 199  
 replicator dynamics  
     continuous, 337  
     discrete, 360  
     two populations, 361  
 replicator–mutator dynamics, 362  
 representation  
     configural weighted, 411  
     cumulative prospect theory, 410  
     full-dimensional, 391  
     order-preserving, 408  
     ranked additive, 407  
     ranked weighted, 407  
     separable, 416  
     stakes sensitive, 415  
     TAX, 411  
 root mean square error, 560
- Savage–Dickey ratio, 526  
 selective influence, 116  
     canonical form, 114  
     diagram of, 112  
     invariance to transformations, 139  
     nestedness, 124  
     test, *see* tests of selective influence  
 semiorder, 387  
     simple lexicographic, 389  
 sequence, 289  
 Shannon’s entropy, 153  
 sigma-algebra, 92  
     Borel, 93, 103  
     discrete, 93, 102  
     Lebesgue, 94, 98, 100, 101, 103  
     product, 95–97, 99, 100  
 signal detection models, 537  
 software  
     *Mathematica*®, 209  
     nauty, 267  
     Pajek, 209  
     RSIENNA, 267  
     Traces, 267  
 source monitoring, 476, 501  
 span, 283  
 split-half CV, 571  
 splitting representation, 49  
 Stan, 545  
 star, 204, 255
- state, 276  
 straight (parametrized probabilistic knowledge structure), 300  
 strategic form game, 324  
 strategy  
     completely mixed, 330  
     locally superior, 331  
     mixed, 324  
     neutrally stable, 334  
     pure, 323  
     strictly dominated, 327  
     support of, 329  
     weakly dominated, 327  
 strategy profile, 324  
 strict weak order model  
     of ternary paired-comparison probabilities, 405  
 string language, 486  
 structurally unstable point, 345  
 subliminal priming, 507, 527, 528, 537  
 surmise function, 285  
 surmise system, 286  
 survival function, 27  
     sub-, 30  
 symmetric difference, 279
- take-the-last heuristic, 340  
 ternary paired-comparison probabilities, 385  
 testability, 562  
 tests of selective influence, 86, 124, 128  
     cosphericity test, 139, 141, 143  
     distance test, 135  
     joint distribution criterion, 117, 118  
     linear feasibility test, 116, 129, 132  
     marginal selectivity test, 129
- theorem  
     de Finetti, 25  
     Sklar, 57  
     Strassen, 47  
     Tsiatis, 33  
 theory of signal detection, 538  
 total positivity, 65  
     multivariate ( $MTP_2$ ), 65  
     of order 2 ( $TP_2$ ), 65  
 total variation distance, 50  
 trace, 297  
 transitive closure, 309  
 treatment, 111  
     allowable, 111, 115  
 trial, 301  
 trial number, 305  
 triangle inequalities, 397  
 trivial (child), 299
- uniform extension (of a probabilistic knowledge structure), 300

- 
- uniform invasion barrier, 330
  - uniform random graph, 260
  - union-closed (knowledge structure), 279
  - uniqueness theorem, 153
  - unstable point, 345
  - update neighborhood, 367
  - utility
    - configural weighted, 411
    - cumulative prospect theory, 410
    - multiplicative separability of, 418
    - of gambling, 409
    - order-preserving, 408
    - parametric forms, 416
    - rank-dependent, 410
    - ranked weighted, 407
    - representation
      - of gambles, 409
      - of interval orders, 406
      - of semiorders, 406
      - of strict linear orders, 406
      - of strict weak orders, 406
    - TAX, 411
    - weak, 394
  - variability across items, 506
  - variability across participants, 506
  - vector, 224
    - eigen—, 226
    - eigen— centrality, 232
    - inner product of, 224
    - normal, 224
    - orthogonal, 228
    - Perron, 232
  - vertex, 384
    - representation, 391
  - Vieth–Müller circles, 173
  - Weber’s law, 160
  - well-graded (knowledge structure), 279
  - WinBUGS, 516, 545
  - word, 289