STUDIES IN *FUZZINESS*
AND *SOFT COMPUTING*

Ronald Yager
Liping Liu
Editors

Arthur P. Dempster
Glenn Shafer
Advisory Editors

Classic Works of the Dempster-Shafer Theory of Belief Functions

Springer

Roland R. Yager · Liping Liu (Eds.)

Classic Works of the Dempster-Shafer Theory of Belief Functions

Studies in Fuzziness and Soft Computing, Volume 219

Further volumes of this series can be found on our homepage: springer.com

Vol. 205. Mika Sato-Ilic, Lakhmi C. Jain
Innovations in Fuzzy Clustering, 2006
ISBN 978-3-540-34356-1

Vol. 206. A. Sengupta (Ed.)
Chaos, Nonlinearity, Complexity, 2006
ISBN 978-3-540-31756-2

Vol. 207. Isabelle Guyon, Steve Gunn, Masoud Nikravesh, Lotfi A. Zadeh (Eds.)
Feature Extraction, 2006
ISBN 978-3-540-35487-1

Vol. 208. Oscar Castillo, Patricia Melin, Janusz Kacprzyk, Witold Pedrycz (Eds.)
Hybrid Intelligent Systems, 2007
ISBN 978-3-540-37419-0

Vol. 209. Alexander Mehler, Reinhard Köhler
Aspects of Automatic Text Analysis, 2007
ISBN 978-3-540-37520-3

Vol. 210. Mike Nachtegael, Dietrich Van der Weken, Etienne E. Kerre, Wilfried Philips (Eds.)
Soft Computing in Image Processing, 2007
ISBN 978-3-540-38232-4

Vol. 211. Alexander Gegov
Complexity Management in Fuzzy Systems, 2007
ISBN 978-3-540-38883-8

Vol. 212. Elisabeth Rakus-Andersson
Fuzzy and Rough Techniques in Medical Diagnosis and Medication, 2007
ISBN 978-3-540-49707-3

Vol. 213. Peter Lucas, José A. Gámez, Antonio Salmerón (Eds.)
Advances in Probabilistic Graphical Models, 2007
ISBN 978-3-540-68994-2

Vol. 214. Irina Georgescu
Fuzzy Choice Functions, 2007
ISBN 978-3-540-68997-3

Vol. 215. Paul P. Wang, Da Ruan, Etienne E. Kerre (Eds.)
Fuzzy Logic, 2007
ISBN 978-3-540-71257-2

Vol. 216. Rudolf Seising
The Fuzzification of Systems, 2007
ISBN 978-3-540-71794-2

Vol. 217. Masoud Nikravesh, Janusz Kacprzyk, Lofti A. Zadeh (Eds.)
Forging New Frontiers: Fuzzy Pioneers I, 2007
ISBN 978-3-540-73181-8

Vol. 218. Masoud Nikravesh, Janusz Kacprzyk, Lofti A. Zadeh (Eds.)
Forging New Frontiers: Fuzzy Pioneers II, 2007
ISBN 978-3-540-73184-9

Vol. 219. Roland R. Yager, Liping Liu (Eds.)
Classic Works of the Dempster-Shafer Theory of Belief Functions, 2007
ISBN 978-3-540-25381-5

Roland R. Yager · Liping Liu (Eds.)

Classic Works of the Dempster-Shafer Theory of Belief Functions

With 81 Figures 43 Tables

Springer

Dr. Roland R. Yager
Machine Intelligence Institute
Iona College
715 N. Avenue
New Rochelle, NY 10801-1890
USA
RYager@Iona.edu, yager@panix.com

Liping Liu
Department of Management
and Information Systems
University of Akron
College of Business Administration
351 Akron, OH 4325-4801
USA
liu@acm.org

Library of Congress Control Number: 2007933149

ISSN print edition: 1434-9922
ISSN electronic edition: 1860-0808
ISBN 978-3-540-25381-5 Springer Berlin Heidelberg New York

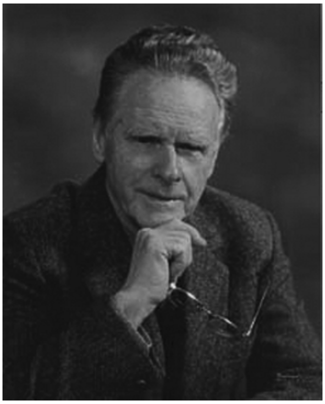
Springer is a part of Springer Science+Business Media
springer.com

Typesetting: Integra Software Services Pvt. Ltd., India
Cover design: WMX Design, Heidelberg

Printed on acid-free paper SPIN: 11352457 42/3100/Integra 5 4 3 2 1 0

This book is dedicated to the memory of
Philippe Smets (1938–2005),
a lifetime advocate for and contributor to
the Dempster-Shafer Theory of Belief Functions.

About the Founders



Arthur P. Dempster studied mathematics and physics at the University of Toronto, earning a B.A. in 1952 and an M.A. in 1953. He received a Ph.D. in mathematical statistics from Princeton University in 1956. From 1958 through 2005 he served on the active teaching faculty of the Department of Statistics at Harvard University, including 11 years as department chair. He has supervised about 50 Ph.D. theses.

His current title at Harvard is Research Professor of Theoretical Statistics, with interests ranging over the methodology and logic of statistical sciences, technical developments within the Dempster-Shafer framework, and the analysis of biological and physical phenomena. In addition to the theory of belief functions, he is known for the EM algorithm, and a range of contributions to multivariate statistical theory.

He is a fellow of the American Academy of Arts and Sciences, the American Statistical Association, and the Institute of Mathematical Statistics. He was awarded the Kampé de Feriet Award of the International Conference on Information Processing and Management of Uncertainty (IPMU) 2002.

Glenn Shafer is Board of Governors Professor at Rutgers University and a Professor in the Computer Learning Research Centre, Royal Holloway, University of London.

Glenn spent his childhood on a farm near Caney, Kansas. After earning a Ph.D. in mathematical statistics from Princeton in 1973, he taught at Princeton and the University of Kansas, moving to Rutgers in 1992. He and his wife Nell Painter, a distinguished historian, live in Newark, New Jersey, and Glenn serves on the board of the Newark Boys Chorus School.

Glenn's book on the Dempster-Shafer theory, *A Mathematical Theory of Evidence*, appeared in 1976. The most important of his more recent books is *Probability and Finance, It's Only a Game!* (2001, with Vladimir Vovk), which explains how mathematical probability can be based on game theory rather than measure theory. His other major books are *The Art of Causal Conjecture* (1996), concerning the relation between probability and causality, *Probabilistic Expert Systems* (1996), concerning the network structures that facilitate computation for Dempster-Shafer beliefs as well as for conventional probabilities, and *Algorithmic Learning in a Random World* (2005, with Gammerman and Vovk), concerning confidence intervals for machine-learning methods.

Glenn has also published in journals in statistics, philosophy, history, psychology, computer science, economics, engineering, accounting, and law. He was the 2004 recipient of the Daniel Gorenstein Memorial Award for research and service at Rutgers University. He was a Guggenheim fellow in 1983–84, a fellow at the Center for Advanced Study in the Behavioral Sciences in 1988–89, and a Fulbright fellow at the Free University of Berlin in 2001. He is a fellow of the Institute of Mathematical Statistics and the American Association for Artificial Intelligence.

Foreword

This volume is a welcome addition to the literature on the Dempster-Shafer theory. It may help turn the theory, which now enjoys a lively but fragmented existence, into a more coherent and better understood set of tools for probabilistic thinking in science and technology.

The volume's title suggests that the theory had a classical period extending from the 1960s through the 1980s. In its first two decades, it consisted of theoretical writings by the two of us: Dempster's work on upper and lower probabilities in the 1960s and Shafer's work on belief functions in the 1970s. Then interest in applications suddenly flowered. After Jeff Barnett introduced the name "Dempster-Shafer" in 1981 [1], the theory quickly acquired textbook status in artificial intelligence. By the end of the classical period, around 1990, the theory had acquired powerful computational tools, remarkably diverse applications, and the attention of many researchers interested in variations and generalizations.

By many measures, the theory continues to flourish in the 21st century. Internet searches for "Dempster-Shafer" produce ever more hits. The theory is used in many branches of technology, only a few of which are represented in this volume. Articles on the theory and its applications appear in a remarkable number of journals and recurring conferences. Books on the theory continue to appear.

In other important respects, however, the theory has not been moving forward. We still hear questions that were asked in the 1980s: How do we tell if bodies of evidence are independent? What do we do if they are dependent? We still encounter confusion and disagreement about how to interpret the theory. And we still find little acceptance of the theory in mathematical statistics, where it first began 40 years ago.

We have come to believe that three things are needed to move the theory forward.

- **A richer understanding of the uses of probability**. Some authors, including our departed friend Philippe Smets [6], have tried to distance the

Dempster-Shafer theory from the notion of probability. But we have long believed that the theory is best regarded as a way of using probability [2, 4, 5]. Understanding of this point is blocked by superficial but well entrenched dogmas that still need to be overcome.

- **A richer understanding of statistical modeling**. Mathematical statisticians and research workers in many other communities have become accustomed to beginning an analysis by specifying probabilities that are supposed known except for certain parameters. Dempster-Shafer modeling uses a different formal starting point, which may often be equally or more legitimate as a representation of actual knowledge [3].
- **Good examples**. The elementary introductions to the Dempster-Shafer theory that one finds in so many different domains are inadequate guides for dealing with the complications that arise in real problems. We need in-depth examples of sensible Dempster-Shafer analyses of a variety of problems of real scientific and technological importance.

Although neither of us has made the Dempster-Shafer theory our top priority in the last two decades, we plan to address these three challenges in the next few years. We hope that the current volume, by putting earlier contributions to the theory in some order, will encourage others, as it has encouraged us, to take stock of the theory's current state and think about how to address its current challenges.

References

1. J. A. Barnett. Computational methods for a mathematical theory of evidence. In *Proceedings of the 7th International Joint Conference on AI*, Vancouver, BC, pages 868–875, 1981. Reprinted in this volume.
2. A. P. Dempster. Belief functions in the 21st century: A statistical perspective. *Proceedings of Institute for Operations Research and Management Science Annual meeting (INFORMS 2001)*
3. A. P. Dempster. The Dempster-Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, to appear.
4. Glenn Shafer. Constructive probability. *Synthèses*, 48:1–60, 1981. Reprinted in this volume.
5. Glenn Shafer and Amos Tversky. Languages and designs for probability judgment. *Cognitive Science*, 9:309–339, 1985. Reprinted in this volume.
6. Philippe Smets and Robert Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994. Reprinted in this volume.

Arthur P. Dempster and Glenn Shafer

Preface

This year marks the 40th anniversary of the seminal publication by Arthur P. Dempster on upper and lower probabilities and the 30th anniversary of the classic monograph by Glenn Shafer. These pioneering works established a new theory for probabilistic reasoning based on a generalization of classic probability. Central to this theory is its ability to model imprecision as well as randomness. This capability often makes it superior to Bayesian approaches in modeling knowledge of uncertainty profiles. In the last 30 years, the concept of belief functions has penetrated into many scientific areas and been applied in many projects. The aim of this book is to bring together a collection of classic papers showcasing important theoretical advances and pioneering applications. The book intends to become an authoritative reference for those working in the field of evidential reasoning as well as an important archival reference for those working in a wide range of areas such as information fusion, reasoning under uncertainty, artificial intelligence and decision making in economics, engineering, and management.

The selection of these classic papers was made with the aid of many experts from this field. While the editors did not have a specific definition of what constitutes a classic paper they felt that the following three features should be present in a classic paper. First, the paper should have been published in highly regarded journal, collection, or conference proceedings. Second, the paper must be often recommended by professors to their graduate students as reading materials for research seminars or projects. Finally the paper initiated a stream of research that other scholars have followed leading to an impact on existing research and with a high prospect to continue to make an impact on the future development of the field.

The paper selection process roughly consisted of four stages:

Classic Paper Nominations: The editors elicited nominations through two channels. They first distributed a call for nominations in the academic news lists for Uncertainty in Artificial Intelligence, Operations Research, Association for Information Systems, and American Accounting Association, etc. They then wrote to a dozen prominent scholars in the field to request comments on specific references. They concluded the nomination period with over 100 nominations from over 50 researchers around the world.

Nomination Review: The editors carefully reviewed each nomination to assess its merit. Nominations by senior, well-known scholars were given careful consideration. As a result, the editors created a short list of 69 papers for further review.

Paper Review: The editors reviewed each paper in the short list to assess its overall contribution to the field. This reduces the number of papers under consideration to 40.

Citation Analysis: We made extensive citation analyses to ensure that all papers were influential ones in the field. In addition to eliminating some papers, this process brought two papers not previously nominated into the pool.

Eventually, the editors jointly selected 29 papers, each of which fall into at least one of the following categories: 1) major conceptual innovations that lead to the development of belief functions, establish a mathematical or semantic foundation, build connections with other scientific areas such as fuzzy logic and probabilistic reasoning, and extend the theory of evidence in significant ways; 2) major methodological developments that use belief functions as a tool for scientific research and general problem solving; 3) major computational developments that propose new algorithms or theories, or improve the efficiency of computation involving belief functions; and 4) groundbreaking applications that demonstrate the creative use of belief functions and contribute to the applied area in significant ways.

There is a large body of literature on belief functions and there are many truly outstanding publications that deserve recognition. We could not include all the publications we would like to include, particularly newer ones. Some notable streams of research include those on bridging belief functions to fuzzy logic and imprecise probabilities, those on learning belief functions from data, those on fast or approximate computations, those on belief functions in infinite or continuous frames of discernment, those on adapting Dempster's rule for integrating dependent or highly conflicting sources of information, and those on applying belief functions to a wide range of real problems.

The papers in this volume are in chronological order.

Ronald R. Yager, New York, NY
Liping Liu, Akron, Ohio

Acknowledgement

This project was initially suggested by Arthur P. Dempster and received immediate support from Glenn Shafer, Smets Philippe, Jürg Kohlas, and many others. Since it took off in early 2003, it has involved many people in various ways and depths. First, we would like to thank our advisory editors, Art and Glenn, for their clear vision and guidance. Second, we would like to thank Jürg Kohlas and the late Philippe Smets for their careful analysis of the literature. Third, we would like to thank many scholars who participated in the nomination process and provided us with their favorite references. These include Jean-Yves Jaffray, George Klir, Henry Kyburg, Henri Prade, Enrique H. Ruspini, Prakash P. Shenoy, and fifty other distinguished scholars. Fourth, we would like to thank all the authors who have contributed to this volume. Some authors, including Jeffrey A. Barnett, John D. Lowrance, and Philippe Smets, went a long way ahead in providing invaluable help to us by retyping their articles in TeX. In particular, Philippe retyped one of his articles when he was seriously ill. We owe him a giant thank-you and would like to dedicate this volume to his memory. Fifth, we would like to thank Ms. Tong Bin, Ms. Jie Guan, and Mr. Venkata Nuli, who have performed the most tedious and laborious work of the project such as indexing and typesetting. Without them, this project would be impossible. Sixth, we would like to thank Hemalatha, D. and her project team at Integra Software Services Pvt Ltd for the beautiful production of this book.

Finally, the editors would like to thank the publishers and authors for permission to reprint copyrighted material in this volume:

Barnett, J. A., "Computational Methods for a Mathematical Theory of Evidence," ©1981. The American Association for Artificial Intelligence (The AAAI Press). Reprinted with the permission of the publisher and author from the Proceedings of the Seventh International Conference Artificial Intelligence, pp 868–875.

Chateauneuf, A. and Jaffray, J. Y., "Some Characterization of Lower Probabilities and Other Monotone Capacities through the Use of Möbius Inversion,

©1989. Reprinted with permission of the publisher and authors from *Mathematical Social Sciences*, Vol. 17, pp 263–283, Elsevier Publisher.

Dempster, A.P., “New Method for Reasoning toward Posterior Distributions Based on Sample Data,” ©1966. Reprinted with the permission of the publisher and author from *Annals of Probability*, Vol. 37, pp 355–374, Institute of Mathematical Statistics.

Dempster, A.P., “Upper and Lower Probability Induced by a Multivalued Mapping,” ©1967. Reprinted with the permission of the publisher and author from *Annals of Probability*, Vol. 38, pp 325–339, Institute of Mathematical Statistics.

Dempster, A.P., “A Generalization of Bayesian Inference,” ©1968. Reprinted with the permission of the publisher and author from *Journal of Royal Statistical Society Series B*, Vol. 30, No 2, pp 205–232, Blackwell Publishing.

Denœux, T., “A k-nearest Neighbor Classification Rule Based on Dempster-Shafer Theory,” ©1995 IEEE. Reprinted with the permission of the publisher and author from IEEE Transactions on Systems, Man and Cybernetics, Vol. 25, No 5, pp 804–813.

Dubois, D. and Prade, H., “A Set-Theoretic View of Belief Functions: Logical Operations and Approximations by Fuzzy Sets,” ©1986. Taylor & Francis Group Ltd. Reprinted with permission of the publisher and authors from the International Journal of General Systems, Vol. 12, pp 193–226.

Gordon, J. and Shortliffe, E. H., “A Method for Managing Evidential Reasoning in a Hierarchical Hypothesis Space,” ©1985. Reprinted with permission of the publisher and authors from *Artificial Intelligence*, Vol. 26, No 3, pp 323–357, Elsevier Publisher.

Jaffray, J. Y., “Bayesian updating and belief functions,” ©1992 IEEE. Reprinted with the permission of the publisher and author from IEEE Transactions on Systems, Man and Cybernetics, Vol. 22, pp 1144–1152.

Kohlas, J. and Monney, P.-A., “Representation of Evidence by Hints,” ©1994. Reprinted with the permission of the publisher and authors from *Advances in the Dempster-Shafer Theory of Evidence*, R. R. Yager, J. Kacprzyk and M. Fedrizzi (Ed.), pp 473–492, John Wiley.

Lowrance, J. D., Garvey, T. D. and Strat, T. M., “A Framework for Evidential-Reasoning Systems,” ©1986. The American Association for Artificial Intelligence (The AAAI Press). Reprinted with the permission of the publisher and author from the Proceedings of the Fifth National Conference on Artificial Intelligence, pp 869–901.

Nguyen, H.T., “On Random Sets and Belief Functions,” ©1978. Reprinted with the permission of the publisher and author from *Journal of Mathematical Analysis and Applications*, Vol. 65, pp 531–542, Elsevier Publisher.

Rogova, G., “Combining the Results of Several Neural-Network Classifiers,” ©1994. Reprinted with the permission of the publisher and author from *Neural Networks*, Vol. 7, No 5, pp 777–781, Elsevier Publisher.

Ruspini, E., "Epistemic Logics, Probability, and the Calculus of Evidence," ©1987. Reprinted with the permission of the publisher and author from *Proceedings of Tenth International Joint Conference on Artificial Intelligence* (IJCAI), pp 924–931, Elsevier Publisher.

Shafer, G., "Non-Additive Probabilities in the Work of Bernoulli and Lambert," ©1978. Reprinted with the permission of the publisher and author from *Archive for History of Exact Sciences*, Vol. 19, pp 309–370, Springer Publisher.

Shafer, G., "Allocations of Probability," ©1979. Reprinted with the permission of the publisher and author from *Annals of Probability,* Vol. 7, No5, pp 827–839, Institute of Mathematical Statistics.

Shafer, G., "Constructive Probability," ©1981. Reprinted with the permission of the publisher and author from *Synthese*, Vol. 48, pp 1–60, Springer Publisher.

Shafer, G., "Belief Functions and Parametric Models," ©1982. Reprinted with the permission of the publisher and author from *Journal of the Royal Statistical Society Series B*, Vol. 44, No 3, pp 322–339, Blackwell Publishing.

Shafer, G. and Tversky, A., "Languages and Designs for Probability Judgment," ©1985. Reprinted with the permission of the publisher and authors from *Cognitive Science Society*, Vol. 9, pp 309–339.

Shafer, G. and Logan, R., "Implementing Dempster's Rule for Hierarchical Evidence," ©1987. Reprinted with permission of the publisher and authors from *Artificial Intelligence*, Vol. 33, No 3, pp 271–298, Elsevier Publisher.

Shenoy, P. P. and Shafer, G., "Axioms for Probability and Belief-Function Propagation," ©1990. Reprinted with the permission of the publisher and the authors from R. D. Shachter, T. Levitt, J. F. Lemmer and L. N. Kanal (Ed.), *Uncertainty in Artificial Intelligence*, Vol. 4, pp 169–198, North-Holland, Amsterdam.

Smets, P., "Belief Functions: The Disjunctive Rule of Combination and the Generalized Bayesian Theorem," ©1993. Reprinted with the permission of the publisher and author from *International Journal of Approximate Reasoning*, Vol. 9, pp 1–35, Elsevier Publisher.

Smets, P. and Kennes, R., "The Transferable Belief Model", ©1994. Reprinted with the permission of the publisher and authors from *Artificial Intelligence*, Vol. 66, No 2, pp 191–234, Elsevier Publisher.

Srivastava, R. P. and Shafer, G., "Belief-Function Formulas for Audit Risk," ©1992. Reprinted with the permission of the publisher and authors from *The Accounting Review*, Vol. 67, No. 2, pp 249–283, American Accounting Association.

Yager, R., "Entropy and Specificity in *A Mathematical-Theory of Evidence*," ©1983. Taylor & Francis Group Ltd. Reprinted with permission of the publisher and author from the International Journal of General Systems, Vol. 9, No 4, pp 249–260.

Yager, R., "Decision Making under Dempster-Shafer Uncertainties," ©1992. Reprinted with the permission of the publisher and author from *International Journal of General Systems*, Vol. 20, pp 233–245, Taylor & Francis Group Ltd.

Yen, J., "Generalizing the Dempster-Shafer Theory to Fuzzy Sets," ©1990 IEEE. Reprinted with permission of the publisher and author from IEEE Transactions on Systems, Man, and Cybernetics, Vol. 20, No 3, pp 559–570.

Zhang, N.L., "Weights of Evidence and Internal Conflict for Support Functions," ©1986. Reprinted with the permission of the publisher and author from *Information Science*, Vol. 38, pp 205–212, Elsevier Publisher.

Contents

1

Classic Works of the Dempster-Shafer Theory of Belief Functions: An Introduction*

Liping Liu and Ronald R. Yager

Abstract. In this chapter, we review the basic concepts of the theory of belief functions and sketch a brief history of its conceptual development. We then provide an overview of the classic works and examine how they established a body of knowledge on belief functions, transformed the theory into a computational tool for evidential reasoning in artificial intelligence, opened up new avenues for applications, and became authoritative resources for anyone who is interested in gaining further insight into and understanding of belief functions.

1 Introduction

The Dempster-Shafer theory of belief functions was due to the seminal work of Glenn Shafer and its conceptual forerunner—lower and upper probabilities by Arthur P. Dempster. This year marks respectively the 30th and 40th anniversaries of these two important publications. In the last 30 years, belief functions have penetrated into many scientific areas, technological projects, and educational enterprises. By bridging fuzzy logic and probabilistic reasoning, the theory of belief functions has become a primary tool for knowledge representation and uncertain reasoning in expert systems. Thanks to the availability of powerful computers and user-friendly software, belief functions have been widely applied to business, engineering, and medical problems. The applications include auditing, process engineering, quality control, decision support, electronic commerce, financial asset evaluation, information fusion, information retrieval, knowledge management, medical diagnosis, mobile services, natural resource detection, network security, object classification, risk management, software engineering, target tracking, etc.

To celebrate the anniversaries, to showcase the achievements, and to assess the current state of knowledge, the editors bring together a volume of 29 classic papers on the theory of belief functions and its applications. The collection was

[*] The authors would like to thank Glenn Shafer for his invaluable comments on earlier versions of this chapter.

created from a pool of over 100 nominated contributions, which are regarded as classics with a high prospect to continue to influence the future development of the field.

In this chapter, we introduce the Dempster-Shafer theory and present its basic concepts and major results. The goal is to summarize Glenn Shafer's classic book [34] in a concise, comprehensive, and accessible manner so that the reader will gain sufficient conceptual background to pursue further readings. Then we sketch a brief history of the earlier conceptual development, from Ronald A. Fisher's fiducial arguments to Arthur P. Dempster's generalized Bayesian inference, and from Jakob Bernoulli's notion of pure evidence to Glenn Shafer's mathematical theory of evidence. The goal is to expose the origin of the concepts so that the reader will gain a broad perspective for understanding further development. Then we provide an overview of the classic works and point out their unique contributions in terms of how they established a body of knowledge on belief functions, transformed the theory into a computational tool for evidential reasoning in artificial intelligence, opened up new avenues for applications in business, engineering, and medicine, and became authoritative resources for anyone who is interested in gaining further insight into and understanding of the theory. Finally, we briefly discuss famous critiques by Lotfi A. Zadeh and Judea Pearl and point out a few open problems that need to be solved in future research.

2 Basic Concepts

The concept of belief functions may be formalized in various ways. In this section, we adopt the approach by Glenn Shafer in his seminal work—*A Mathematical Theory of Evidence* [34]—for exposition since its terminologies and notations are the standard in the literature.

Given a question of interest, let Θ be a finite set of possible answers to the question, called a *frame of discernment*, and 2^Θ be the set of all subsets of Θ:

$$2^\Theta = \{A \mid A \subseteq \Theta\}.$$

The subset A includes as special cases the empty set ϕ and the full set Θ. It represents a statement or proposition that the truth lies in A. A real function over the subsets $Bel : 2^\Theta \to [0,1]$ is called a *belief function* if and only if it satisfies the following three axioms:

Axiom 1 *$Bel(\phi) = 0$.*

Axiom 2 *$Bel(\Theta) = 1$.*

Axiom 3 *For any whole number n and subsets $A_1, A_2, .., A_n \subset \Theta$,*

$$Bel(\bigcup_{i=1}^{n} A_i) \geq \sum_{\substack{I \subset \{1,2,\ldots,n\} \\ I \neq \phi}} (-1)^{|I|+1} Bel(\bigcap_{i \in I} A_i).$$

In the case where $n = 2$ and $A_1 \cap A_2 = \phi$, Axiom 3 reduces to $Bel(A_1 \cup A_2) \geq Bel(A_1)+Bel(A_2)$. The student of probability theory may immediately recognize that these axioms are similar to those for a probability function with the inequality of Axiom 3 substituting for equality. When equality holds, $Bel(A_1 \cup A_2) = Bel(A_1)+Bel(A_2)$ if $A_1 \cap A_2 = \phi$. Thus, a probability function is additive whereas a belief function is generally not. The generalized axioms, however, indicate that belief functions include probability functions as special cases and may be equally or better used to express degrees of belief.

Additive probabilities are common sense. Are there any non-additive beliefs to justify an extension? The answer is affirmative. A modern example is from Bayesian statistics on how to represent ignorance, where the truth is in Θ but there is no information, probabilistic or logical, to justify the whereabouts of the truth. Thus, $Bel(\Theta) = 1$ but $Bel(A) = 0$ for any proper subset A of Θ. Clearly, this function fails to be additive. An ancient example was due to Jakob Bernoulli in his book *Ars Conjectandi*. Suppose a man was stabbed with a sword in a milling crowd and Gracchus was interrogated and turned pale. Since the sign of pallor betokens a finite number of reasons: melancholy, fear, cold, anger, amorous passion, etc., it proves Gracchus guilty if it arises from a guilty conscience, but does not prove his innocence if it arises from other reasons. Thus, $Bel(\{guilty\}) < 1$ and $Bel(\{innocent\}) = 0$, but $Bel(\Theta) = 1$. Again, this belief function is non additive.

The key to the concept of belief functions is limited division of belief. Whereas probability functions assume belief is apportioned to the points in the frame Θ, belief functions allow basic probability numbers (or mass numbers), to be assigned to whole sets of points in Θ without further subdivision. The basic idea is that a whole belief is divided into one or more basic probability numbers $m(A)$ and allocated to one or more subsets A, called *focal elements*, such that:

$$\sum\{m(A) \mid A \subseteq \Theta\} = 1. \tag{1}$$

The basic probability number $m(A)$ allocated to a focal element A is not further divided into smaller chunks allocated to proper subsets of A.

This suggests an alternative approach to the definition of a belief function. Given basic probability numbers $m(A)$, the belief $Bel(A)$ is defined by:

$$Bel(A) = \sum\{m(B) \mid B \subseteq A\}. \tag{2}$$

Logically, a portion of belief committed to one proposition is committed to any other proposition it implies. Thus, the total belief committed to a subset A is the sum of those that are committed to proper subsets of A and those to A itself.

Adding the boundary condition $m(\phi) = 0$ to (1), Shafer showed that the two definitions are equivalent, i.e., a function satisfies the three axioms if and only if it can be represented as the sum of basic probability numbers over focal elements. In fact, given a belief function, one may construct such a basic probability number for each $A \subseteq \Theta$ using a Möbius transformation:

$$m(A) = \sum\{(-1)^{|A-B|}Bel(B) \mid B \subseteq A\}, \tag{3}$$

where $|A - B|$ is the cardinality of $A - B$, or a recursive deduction:

$$m(\phi) = 0, m(A) = Bel(A) - \sum\{m(B) \mid B \subset A\}.$$

Despite the equivalence, however, one should note that the axiomatic definition allows the establishment of the theory of belief functions with no reference to probabilities.

Due to the limited divisibility, belief not committed to $\overline{A}$, the negation of A, is not automatically committed to A. But it does make A more credible or plausible. Thus, it is intuitive to define a plausibility function $Pl(A)$ as the sum of beliefs not committed to $\overline{A}$:

$$Pl(A) = 1 - Bel(\overline{A}). \tag{4}$$

Through (2), it is easy to see the interplay between basic probability numbers and plausibility numbers as follows:

$$Pl(A) = \sum\{m(B) \mid A \cap B \neq \phi\}. \tag{5}$$

For any proposition, its plausibility is no less than its committed belief, i.e., $Bel(A) \leq Pl(A)$. Thus, in his earlier works [8, 9, 10], Dempster called these functions respectively lower and upper probabilities. The terminology had caused some confusion and was abandoned by Shafer.

Belief functions are meant to be a representation of subjective beliefs. Unlike other alternative formalisms, however, belief functions represent the beliefs grounded on or supported by evidence. In fact, the idea of limited divisibility makes intuitive sense if one interprets a basic probability number $m(A)$ as a measure of evidential support to A. Given two subsets A and B in a frame of discernment, if B is a proper subset of A, then B represents a stronger proposition than A and requires stronger evidence to support it. Therefore, the evidence that supports A does not automatically support B and the belief $m(A)$ committed to A does not necessitate the commitment of a smaller number $m(B)$ to B.

Given a distinct piece of evidence, its support may be encoded as a list of mass numbers assigned to the corresponding focal elements. It may also be summarized as a belief or plausibility function over the frame of discernment.

When there exist multiple items of evidence, of course, it is necessary to combine them together. *Dempster's rule of combination* serves this purpose. In his original framework of a multivalued mapping that carries a probability measure into a system of upper and lower probabilities (see below), Dempster derived this rule of combining upper and lower probabilities based on the assumption that two probability measures were independent [9]. In the axiomatic framework, Shafer adopted the rule as a definition for combining distinct or independent bodies of evidence. Let m_1 and m_2 be the mass functions for two independent bodies of evidence. The combination via Dempster's rule follows a simple three-step process: intersection of focal elements, multiplication of corresponding basic probability numbers $m_1 m_2$, and normalization in accordance with (1). Each intersection, if it is not empty, becomes a new focal element of the combined belief function. The corresponding product of basic probability numbers contributes to the support to the new focal element. An empty intersection indicates a disagreement or conflict and is excluded from further consideration. Its corresponding product of basic probability numbers is subtracted from the whole belief mass for normalization. Mathematically, the new mass function over new focal elements is defined as follows:

$$m(A) = \frac{\sum\{m_1(B)m_2(C) \mid B \cap C = A\}}{\sum\{m_1(B)m_2(C) \mid B \cap C \neq \phi\}}. \tag{6}$$

Since empty intersection indicates a conflict, $\sum\{m_1(B)m_2(C) \mid B \cap C = \phi\}$ measures the total amount of conflict. Formally, we call the logarithm of the renormalization constant *the weight of conflict*:

$$W = \log(\frac{1}{\sum\{m_1(B)m_2(C) \mid B \cap C \neq \phi\}}). \tag{7}$$

Of course, two belief functions are *combinable* if and only if their weight of conflict is finite.

Example 1. Suppose, among three suspects, Tony (T), Smith (S), and Dick (D), we want to find out who committed a bank burglary. In the investigation, we questioned Mrs. Johnson, a witness who was living close to the bank. She said that she saw a big person near the bank around the time when the crime was committed. Assume Mrs. Johnson's testimony was 60% reliable based on her eyesight. If her testimony was reliable, the evidence pointed to Tony or Dick since they had big bodies. Thus, $m_1(\{T, D\}) = 0.6$. However, if she was not reliable, the testimony carried no information, i.e., $m_1(\{T, S, D\}) = 0.4$. Although the criminal wore a mask, a video camera recorded a fuzzy picture of the person's eyes, which were 4 times more likely to be black than to be gray. The second item of evidence suggested $m_2(\{S\}) = 0.8$ and $m_2(\{T, D\}) = 0.2$ since Smith had black eyes. To combine the two pieces of evidence, we can use a tabular form as in Table (1). For each cell, take the corresponding focal elements from each item of evidence, intersect them and multiply their corresponding basic probabilities. The weight of conflict between the two items of evidence is $\log(\frac{1}{1-0.48}) = 0.28$. In the combined evidence, there are two focal elements: {S} and {T, D}. The combined mass function is calculated as follows:

$$m(\{S\}) = \frac{0.32}{1 - 0.48} = 0.615,$$
$$m(\{T, D\}) = \frac{0.12 + 0.08}{1 - 0.48} = 0.385.$$

Table 1. An illustration of combination

	$m_2(\{S\}) = 0.8$	$m_2(\{T, D\}) = 0.2$
$m_1(\{T, D\}) = 0.6$	$\phi \rightarrow 0.48$	$\{T, D\} \rightarrow 0.12$
$m_1(\{T, S, D\}) = 0.4$	$\{S\} \rightarrow 0.32$	$\{T, D\} \rightarrow 0.08$

Thus Smith appeared to be more suspicious according to the combined evidence.

The combined belief and plausibility function may be symbolically expressed as $Bel_1 \oplus Bel_2$ and $Pl_1 \oplus Pl_2$, respectively. Unfortunately, there is no simple analytical expression for the orthogonal sum $\oplus$. To put the combination rule into multiplicative form as in the case for probability functions, Dempster introduced another function $Q(A)$, which Shafer called the *commonality function*, as follows:

$$Q(A) = \sum\{m(B) \mid B \supseteq A\}. \tag{8}$$

Let Q_1 and Q_2 be respectively the commonality functions for two independent items of evidence. Then the commonality function for the combined evidence is as follows:

$$Q(A) = \frac{Q_1(A)Q_2(A)}{\sum\{(-1)^{|A|+1}Q_1(A)Q_2(A) \mid A \neq \phi\}}. \tag{9}$$

Here the denominator is identical to that in (6).

Unlike belief and plausibility functions, a commonality function is not intuitive but Shafer interpreted $Q(A)$ as the total belief that is free to move to every element of A. According to (8) and (3), it is clear that the definition of a commonality function is opposite to that of a belief function in the sense that a belief for A sums all basic probability numbers committed to A and its proper subsets whereas a commonality number sums those that are committed to A and its proper supersets. Consequently, commonality functions are decreasing while belief (and plausibility) functions are increasing: for any two propositions A and B, if $A \supset B$, then $Q(A) \leq Q(B)$ but $Bel(A) \geq Bel(B)$ and $Pl(A) \geq Pl(B)$.

The four representations of evidence, namely, belief functions $Bel(A)$, mass functions $m(A)$, plausibility functions $Pl(A)$, and commonality functions $Q(A)$, are interrelated. Some of the relationships are shown below: for any non-empty set A,

$$\begin{aligned} Bel(A) &= \sum\{(-1)^{|B|}Q(B) \mid B \subseteq \overline{A}\}, \\ Q(A) &= \sum\{(-1)^{|B|}Bel(\overline{B}) \mid B \subseteq A\}, \\ Pl(A) &= \sum\{(-1)^{|B|+1}Q(B) \mid \phi \neq B \subseteq A\}, \\ Q(A) &= \sum\{(-1)^{|B|+1}Pl(B) \mid B \subseteq A\}. \end{aligned}$$

From any representation one can obtain another one through a series of additions and/or Möbius transformations. In this sense, all the representations are equivalent. Thus, one may start with any one model to encode evidence and end up with other representations for decision making or probable reasoning. The choice is purely based on convenience. Mass functions are often a more natural and superior device for encoding evidence, whereas belief and plausibility functions are a more intuitive summary of the impact of the evidence on propositions. After all, evidence often arises in the form of knowledge in a related domain that provides insights on or connections to propositions in the domain of interest. If the knowledge is probabilistic, it can then be carried over to the propositions of interest as basic probability numbers. For example [19], suppose I find a scrap of newspaper predicting a blizzard, which I regard as infallible. Also, suppose I am 75% certain that the newspaper is today's. Here the knowledge about the newspaper maps to tomorrow's weather as follows: if the newspaper is today's, then a blizzard is sure to come; if the newspaper is not today's, however, it provides no information on tomorrow's weather. Thus, we transfer 75% as a basic probability number to the focal element {blizzard}, i.e., $m(\{\text{blizzard}\}) = 0.75$, and 25% to Θ, i.e., $m(\Theta) = 0.25$. Of course, there are occasions when belief or plausibility functions become more convenient. For example, Srivastava and Shafer [40] interpret audit risks as the plausibility that a financial statement is not fairly stated or an audit objective is not met. Thus it is more convenient to use plausibility functions to encode audit evidence.

To illustrate the equivalence of the four representations, Table 2 shows the respective representations of three special cases of belief functions, namely vacuous belief functions, Bayesian belief functions, and simple support functions. A *vacuous belief function* represents full ignorance, i.e., evidence does not provide any support to or information on any specific proposition, i.e., any proper subset of a frame of discernment. Thus, Θ is the only focal element. *A Bayesian belief function* represents probabilistic knowledge that assigns a probability to each element of Θ. In other words, all focal elements are singletons. *A simple support function* represents a piece of homogeneous evidence that provides support to one and only one proposition that is a proper subset of Θ. In other words, there are two focal elements: S and Θ with $S \subset \Theta$.

Despite their simplicity, the three special cases play important roles in the theory of belief functions in the sense that: 1) they are the building blocks for more complex belief functions; and 2) they justify the superiority of belief functions to probability theory. Vacuous belief functions provide a simple solution to the problem of representing ignorance. Note that Bayesian statistics would represent full ignorance as a uniform distribution, which essentially mixes lack of belief with disbelief. For example, what is my belief that a coin will land a head? It is 50% if and only if I know the coin is fair. If I am ignorant, the most I can say is $Bel(\Theta) = 1$. However, Bayesian statistics will assign 50% as a prior probability regardless.

Table 2. Three special cases

	Mass function	Belief function
Vacuous belief functions	$m(\Theta) = 1$	$Bel(A) = \begin{cases} 0 & \forall A \subset \Theta \\ 1 & A = \Theta \end{cases}$
Bayesian belief functions	$\lvert A\rvert = 1$ for each focal element A	$Bel(A)$ is additive
Simple support functions	$m(A) = \begin{cases} s & A = S \\ 1-s & A = \Theta \\ 0 & \text{else} \end{cases}$	$Bel(A) = \begin{cases} s & A \supseteq S \\ 1 & A = \Theta \\ 0 & \text{else} \end{cases}$
	Plausibility function	Commonality function
Vacuous belief functions	$Pl(A) = 1\ \forall A \neq \phi$	$Q(A) = 1\ \forall A$
Bayesian belief functions	$Pl(A)$ is additive	$Q(A) = 0$ if $\lvert A\rvert > 1$
Simple support functions	$Pl(A) = \begin{cases} 1 & A \cap S \neq \phi \\ 1-s & A \cap S = \phi \end{cases}$	$Q(A) = \begin{cases} 1 & A \subseteq S \\ 1-s & \text{else} \end{cases}$

Bayesian belief functions are regular probabilities. They are the only case where beliefs and plausibilities are identical, i.e., $Bel(A) = Pl(A)$ for any $A \subseteq \Theta$, and additive as well, i.e., $Bel(A_1 \cup A_2) = Bel(A_1) + Bel(A_2)$ if $A_1 \cap A_2 = \phi$. Thus, belief functions include probability functions as a special case. It is also the only case that we have zero commonality number for any subset of cardinality 2 or larger.

The concept of simple support functions is the most important extension to Dempster's work on generalized Bayesian inference. It acts as the basis for defining the *weight of evidence*, by which Bernoulli meant probative force for a probability judgment. For a simple support function with $m(S) = s$ and $m(\Theta) = 1 - s$, the weight of evidence w is a nonnegative number in $[0, \infty)$ that maps to the support s in such a way that the sum of two weights maps to the combined support of the two items of evidence via Dempster's rule. This along with the following boundary condition:

$$s = \begin{cases} 0 & w = 0 \\ 1 & w \to \infty \end{cases}$$

leads to an analytical expression of the weight of evidence:

$$w = -\log(1 - s).$$

Since a simple support function uniquely determines a weight of evidence, it is tempting to decompose a general belief function into one or more simple support functions and then derive the weight of evidence underlying it. Toward this goal, Shafer defined the concept of *a separable support function* to be the orthogonal sum of one or more simple support functions. Unlike a simple support function, a separable support function may support multiple propositions that are proper subsets of Θ. Unlike a general belief function, it is distinct in that, for any two focal element A and B, if $A \cap B \neq \phi$, then $A \cap B$ is also a focal element.

As an example of special importance, *consonant support functions* are separable support functions. A belief function is called consonant if its focal elements are nested, i.e., for any two focal elements A and B, either $A \subset B$ or $B \subset A$. Thus, all focal elements may be arranged in an order of increasing precision, pointing in a single direction. A consonant support function Bel has the following distinct features:

$$\begin{aligned} Bel(A \cap B) &= \min(Bel(A), Bel(B)) \ \forall A, B \subseteq \Theta, \\ Pl(A \cup B) &= \max(Pl(A), Pl(B)) \ \forall A, B \subseteq \Theta, \\ Q(A) &= \min\{Q(\theta) \mid \theta \in A\} \ \forall A \neq \phi. \end{aligned}$$

Those familiar with fuzzy logic may recognize that the *possibility* and *necessity functions* introduced by Zadeh [42] are the same as consonant plausibility and support functions. A function f is a consonant support function if and only if it satisfies: $f(\phi) = 0$, $f(\Theta) = 1$, and $f(A \cap B) = \min(f(A), f(B))$ for any $A, B \subseteq \Theta$. These are the axioms used for developing the theory of possibility.

There is no unique way to decompose a separable support function into simple support functions. For example, one simple support function may be further represented as the orthogonal sum of two or more simple support functions that support the same proposition. If no component has infinite weight of evidence, however, this non-uniqueness does not cause any trouble because the total weight of evidence focused on each subset will be the same no matter which decomposition is used. Let S_i be the proposition supported by the ith component and w_i be the corresponding weight of evidence. If w_i is finite for all i, then the *total weight of evidence* focused on any non-empty proper subset A of Θ is

$$w(A) = \sum\{w_i \mid S_i = A\},$$

with $w(\phi) = 0$ and $w(\Theta) = \infty$.

Through a weight function $w(A)$, one may define two related concepts: the impingement function $v(A)$ and the weight of internal conflict W. The impingement function $v(A)$ is defined as the sum of the weights of evidence focused on the propositions not containing A:

$$v(A) = \sum\{w(B) \mid A \cap \overline{B} \neq \phi\}. \tag{10}$$

Each weight $w(B)$ impugns all propositions not in its focus B. Thus $v(A)$ is the total weight of evidence not favoring A. Given an impingement function, one may recover the weight function using a Möbius transformation, i.e., for each non-empty proper subset A of Θ,

$$w(A) = \sum\{(-1)^{|B-A|}v(B) \mid A \subseteq B\}.$$

The internal conflict of a separable support function refers to the conflict among the simple support functions that make up the separable support function. Its weight can be defined as in (7) with a straightforward extension to multiple belief functions. Since decomposition may not be unique, the weight of conflict in general varies from decomposition to decomposition. The *weight of internal conflict* is actually defined as the minimum of the weights of conflict for all possible decompositions. The weight of internal conflict can be expressed in terms of the impingement function $v(A)$ or the commonality function $Q(A)$:

$$\begin{aligned} W &= -\log(\textstyle\sum\{(-1)^{|A|+1}\exp(-v(A)) \mid A \neq \phi\}, \\ W &= -\textstyle\sum\{(-1)^{|A|}\log Q(A) \mid A \subseteq \Theta\}. \end{aligned}$$

The above equations give another way to express the commonality function $Q(A)$ for a separable support function as follows:

$$\log Q(A) = W - v(A). \tag{11}$$

The total weight of evidence determines the impingement function, which in turn determines the weight of internal conflict. Thus, it determines a commonality function, from which one can recover a mass function, a belief function, and a plausibility function. Therefore, for separable support functions, the total weight of evidence provides a sufficient assessment of evidence.

Equation (11) shows an intuitive association of smaller commonality numbers with greater degrees of impingement. Formally, suppose v_1 and v_2 are two impingement functions and Q_1 and Q_2 are the corresponding commonality functions. If $Q_1(A) \leq Q_2(A)$ for all $A \subseteq \Theta$, then $v_1(A) \geq v_2(A)$ for all $A \subseteq \Theta$. This association can easily be derived from the following still unproven *weight-of-conflict conjecture*: if Q_1 and Q_2 are the commonality functions for two separable support functions and W_1 and W_2 are their corresponding weights of internal conflicts, then

$$Q_1(A) \leq Q_2(A)\ \forall A \subseteq \Theta \implies W_1 \geq W_2. \tag{12}$$

In the axiomatic approach, any function that satisfies Axioms 1–3 is a belief function. A whole body of mathematical theory of belief functions could have been built based on these axioms. However, Shafer was interested in building a theory of evidence as a science of probable reasoning. Thus, his central theme was to investigate which subclasses of belief functions could be useful for the representation of evidence. As we have seen, both simple and separable support functions were proposed for such a purpose; they are or can be decomposed into components each of which precisely and homogeneously supports a given proposition.

Toward the same goal, the concept of support functions was proposed. A *support function* is a belief function that can be derived from the marginalization of a separable support function to a coarser framer of discernment. Unlike a *sample space* in probability theory, a frame of discernment is epistemic in nature and is constructed for probable reasoning. It can be refined or coarsened as needed. For example, suppose we are interested in whether tomorrow's weather will be raining (r), snowing (s), or normal (n). The frame of discernment is $\Theta = \{r, s, n\}$. This frame may be coarsened into $\Theta' = \{n, \overline{n}\}$ if we just want to know whether the weather is normal or not. The coarsening combines fine elements r and s into a coarse element $\overline{n}$. Thus, we call Θ' a coarsening of Θ or Θ a refinement of Θ'. A more refined frame is able to represent more details than its coarsenings and so a proposition discerned by a coarsening is also discerned by a refinement. The converse is not true.

Each coarse element in a coarse frame maps to a subset of fine elements in a refined frame. If a belief function Bel is defined on a refined frame Θ, it can be carried over to a coarse frame Θ' as a *marginal belief function* as follows. A focal element of the marginal is a set of coarse elements that map to subsets, all of which intersect with the same set of focal elements of Bel. The basic probability number is the sum of the corresponding basic probability numbers of the intersecting focal elements. On the other hand, if a belief function Bel' is defined on a coarse frame Θ', it can also be carried over to a refined frame Θ by using the same probability numbers but replacing each focal element by the union of corresponding mapped subsets. The resulting belief function is called a *vacuous extension*.

Both vacuous extension and marginalization can be easily expressed in the special case when a refined frame is the Cartesian product of two or more independent frames [22]. Suppose $\Theta_1, \Theta_2, ...$ are independent frames. Let I be a set of indices. Then $\Theta(I) = \prod\{\Theta_i \mid i \in I\}$ will be a refinement for all Θ_i ($i \in I$) so that each element $\theta_i \in \Theta_i$ maps to subset $\{\theta_i\} \times \Theta(I - \{i\})$ in Θ. Given any belief function on $\Theta(I)$ with a mass function $m(A)$, its marginal on $\Theta(J)$, $J \subset I$, is a belief function with mass function $m^{\downarrow J}$: for any $B \subseteq \Theta(J)$,

$$m^{\downarrow J}(B) = \sum\{m(A) \mid A \cap (B \times \Theta(I - J) \neq \phi\}. \tag{13}$$

On the other hand, if a belief function with mass function $m(A)$ is defined on $\Theta(J)$, its vacuous extension to $\Theta(I)$ ($I \supset J$) is a belief function with mass function $m^{\uparrow I}$: for any $B \subseteq \Theta(J)$,

$$m^{\uparrow I}(B \times \Theta(I - J)) = m(B). \tag{14}$$

It is easy to see that if a belief function Bel is a separable support function, its vacuous extension will also be separable. However, the converse is not true, i.e., the marginal of a separable support function may not be separable. For this reason, Shafer calls such a marginal belief function a *support function*. So a belief function is a support function if it can be extended to a separable support function. Since any frame is a refinement of itself, a separable support function is itself a support function. Thus, we have four nested classes of belief functions:

$$\left\{\begin{array}{c}\textit{simple}\\ \textit{support}\\ \textit{functions}\end{array}\right\} \subset \left\{\begin{array}{c}\textit{separable}\\ \textit{support}\\ \textit{functions}\end{array}\right\} \subset \left\{\begin{array}{c}\textit{support}\\ \textit{functions}\end{array}\right\} \subset \left\{\begin{array}{c}\textit{belief}\\ \textit{functions}\end{array}\right\}.$$

As it turns out, a belief function is a support function if and only if the union of all of its focal elements is also a focal element. Thus, not all belief functions are support functions. Moreover, not all support functions are separable. For example, assume $m(\{r, n\}) = 0.2$, $m(\{s, n\}) = 0.5$, and $m(\Theta) = 0.3$. This is a support function since $\{r, n\} \cup \{s, n\} \cup \Theta = \Theta$ is a focal element. However, this is not a separable support function since $\{r, n\} \cap \{s, n\} = \{n\}$ is not a focal element.

3 A Brief History of Concepts

Einstein once said [14], "...creating a new theory is not like destroying an old barn and erecting a skyscraper in its place. It it rather like climbing a mountain, gaining new and wider views, discovering unexpected connections between our starting point and its rich environment." The theory of belief functions arose first from Dempster's attempt in understanding and perfecting Fisher's fiducial approach to probability inference and then from Shafer's elaboration of Dempster's work toward a general theory of reasoning based on evidence.

In the 1960s, due to the work of Leonard J. Savage [32], Bayesian statistics was showing renewed vigor and gaining popularity but, at the same time, was in growing conflict with a school of thought led by Ronald A. Fisher and, increasingly, Jerzy Neyman.

The general statistical inference problem is that, given a sample observation x from a parametric distribution $f(x, \theta)$ with parameter θ, how one could obtain a probability distribution of θ. When reduced to its mathematical essentials, Bayesian inference means starting with a prior probability distribution $p(\theta)$, observing the value x, and computing the conditional distribution of θ given x using Bayes theorem:

$$p(\theta \mid x) = \frac{p(\theta) f(x, \theta)}{\int p(\theta) f(x, \theta) d\theta}. \tag{15}$$

In theory, there is nothing wrong with this formulation. In practice, however, one often finds the conception of prior probabilities vague, arbitrary, or controversial, lacking the spirit of objectivity required by a scientific method.

To overcome the difficulty with prior probabilities, Fisher announced the possibility of obtaining posterior distributions with no need for priors (see [17]), and called his method the *fiducial argument* to emphasize its differences from the Bayesian argument. In the nutshell, assume $F(x, \theta)$ is a parametric cumulative distribution. Besides x and θ, the fiducial method introduces a so-called *pivotal variable* u, which is assumed to follow the uniform distribution $U(0, 1)$, so that

$$u = F(x, \theta). \tag{16}$$

Suppose, for each value x, $F(x, \theta)$ is monotonic in θ. Equation 16 will admit a unique solution

$$\theta = \theta(u, x) \tag{17}$$

for each $u \in (0, 1)$. Assuming no prior probabilities, Fisher defined the fiducial distribution of θ, given the observed value x, as the distribution of θ implied by (17) when x is regarded as fixed and u is uniformly distributed.

The fiducial method was poorly understood and often led to inconsistencies [6]. The concept of pivotal variables was highly confusing, restrictive, and controversial [7]. Dempster devoted much of his early research career at Harvard to clarifying, extending, and perfecting the method. For example, he once proposed the concept of *direct probabilities* as his interpretation of the fiducial argument [5]. First, to make the derivation of fiducial probabilities explicit, he introduced an arbitrary function $v = V(x)$ so that it along with (16) implied a smooth one-to-one function from x and θ to u and v, and therefore ensured the existence of the following Jacobian:

$$\left| \frac{\partial(u, v)}{\partial(x, \theta)} \right| . \tag{18}$$

Second, in addition to Fisher's assumption that u is uniform in $(0, 1)$, he assumed that v follows an arbitrary distribution $p(v)$ and u is independent of x (and so of v) so that the joint density function of u and v is $p(v)$. Finally, according to the Jacobian formula, the joint density function of x and θ is $p(V(x))$ multiplied by the Jacobian in (18). From this joint distribution, of course, one can compute the conditional probability distribution of θ given x, which is the fiducial (or direct) probability distribution.

Like Bayesian priors, functions $V(x)$ and $P(v)$ are arbitrary and meant to compose a joint distribution, from which a conditional distribution can be obtained. Although $P(v)$ does not enter the final result, a fiducial distribution is generally not free from the choice of $V(x)$. In fact, as Dempster showed, it is independent of $V(x)$ if and only if $F(x, \theta)$ can be transformed into a location parameter family.

The direct probability method did not fully demystify the fiducial argument. Although it explicated the process of deriving fiducial probabilities, it left the concept of pivotal variables unexplained. Some regard the uniform distribution $U(0, 1)$ as analogous to a Bayesian prior. Most importantly, like the fiducial argument, the method works only if there exists a smooth one-to-one mapping Γ: $u \to \theta$ so that a probability measure for u can be carried to θ by the familiar Jacobian formula.

A breakthrough led to a new theory that unified Bayesian and fiducial arguments. It was first exposited in a paper published in 1966 [8] and republished here as Chap. 2. In this paper, Dempster abandoned Fisher's controversial pivotal variable and replaced it with the concept of a population. Instead of considering u as a pivotal variable, uniformly distributed in $(0, 1)$, he construed u to be a sample individual randomly drawn from a population with probability measure m governing the random sampling operation. Here, m is not necessarily a uniform distribution as in the case of the fiducial argument. Second, instead of (16), Dempster proposed a new model for constructing the mapping from u to θ as follows. Assume each sample individual u corresponds to an observable characteristic x. Assume further that the probability measure m for u induces a probability distribution $f(x, \theta)$ for x with an unknown parameter θ. Thus, one may construct a mapping $u \to x \times \theta$. When the observation x is fixed, it determines a conditional mapping Γ: $u \to \theta$, from which m induces a probability distribution for θ. Interestingly, when Γ is multivalued, the induced distribution for θ is no longer unique. Instead, Γ carries a unique probability measure m to a system of upper and lower probabilities for θ.

Chapter 2 was a milestone, representing not only an advancement of the fiducial argument but also the inception of the idea for a new theory of belief functions. At this point, the basic concepts had already emerged, including the basic probability assignment m, the multivalued mapping Γ, and a device for deriving upper and lower probabilities from m. In Chap. 3, first published in 1967 [9], Dempster abstracted these concepts from the fiducial argument, envisioning a fundamental method of reasoning with imprecise probabilities, based on the idea of obtaining a degree of belief for one event from probabilities for related events. He proposed a general model (S, m, Γ, T) for such reasoning, where S is a source space, m is a probability measure over S, T is a target space, and Γ is a mapping from S to T. If Γ is a one-to-one or many-to-one mapping, it is well known that the probability measure m carries over to T as $p(t) = \sum\{m(s) \mid t = \Gamma(s)\}$. In mathematical essence, Chap. 3 extended the familiar result to the case when Γ is a one-to-many or many-to-many mapping and derived a system of upper and lower probabilities for T based on a probability measure m. Its real thrust, of course, is to view a probability measure as defining degrees of belief, which quantifies a state of partial knowledge arising from a source of imprecise information. Since information is imprecise, it does not always pinpoint a unique value of the variable of interest. Thus, a multivalued mapping is a necessary representation for imprecise information. Since there may be multiple independent sources of information, a mechanism for combining such sources becomes a necessity for a general calculus oriented toward statistical inference and probabilistic reasoning. Therefore, besides the formal definitions of upper and lower probabilities, distributions, and expectations, Chap. 3 presented a rule for deriving upper and lower conditional probabilities and further generalized it into a rule of combining independent sources of information, which was later called Dempster's rule of combination by Shafer [34].

The concept of upper and lower probabilities can be traced back to Boole [2]. Before Dempster, there were already other approaches to the concept [16, 18, 38, 39]. Dempster's multivalued mappings provides a rigorous device for generating these probabilities. As Chap. 3 showed, however, Dempster's concept is not the same as alternative ones. For example, the set of probabilities compatible with Dempster's upper and lower probabilities is smaller than alternatives. The unique feature of Dempster's concept is to map upper and lower probabilities to a single probability measure, allowing for a more rigorous logic for defining conditioning. The resulting upper and lower conditionals are, of course, not same as upper and lower bonds of conditionals. Using standard notations, let $Bel(A)$ and $Pl(A)$ be Dempster's lower and upper probabilities. Then, given a subset E with $Pl(E) > 0$, Dempster's conditional is

$$Bel(A \mid E) = \frac{Bel(A \cup \overline{E}) - Bel(\overline{E})}{1 - Bel(\overline{E})}. \tag{19}$$

In contrast, let $\mathbf{P}$ be the set of probability measures compatible with Bel: $\mathbf{P} = \{P \mid P(A) \geq Bel(A)\}$. Given E with $P(E) > 0$, we can take Bayesian conditioning of P in $\mathbf{P}$: $P_E(A) = P(A)/P(E)$. Let $\mathbf{P}_E$ be the set of resulting conditionals: $\mathbf{P}_E = \{P_E \mid P \in \mathbf{P}\}$. Then, the lower envelope of $\mathbf{P}_E$ exists when $Bel(E) > 0$: $\forall A \subset E$,

$$\underline{P}(A \mid E) = \frac{Bel(A)}{Bel(A) + 1 - Bel(A \cup \overline{E})}. \tag{20}$$

In general, we have $Bel(A \mid E) \geq \underline{P}(A \mid E)$. Therefore, Chap. 3 not only established the mathematical foundation for the theory of belief functions but also clarified many confusions that later arose in the literature [29]. In fact, to avoid these confusions, Shafer [34] renamed Dempster's upper and lower probabilities into respectively plausibility and belief functions.

Being a statistician, Dempster first explicitly applied his rule of combination to statistical inference. He did this in Chap. 4, first published in 1968 [10]. Although Chap. 4 derived upper and lower probabilities for the same parameters, it did so without explicitly invoking the rule of combination. Chap. 4 framed the inference problem using a formal model (S, m, Γ, T), where S is a population, m is a probability measure governing how each individual may be sampled from the population, and $T = X \times \Theta$ is the product of the set of all possible observations x and the set of all possible parameter values θ. A multivalued mapping $\Gamma : S \rightarrow T$ was then used to derive a restricted mapping $\Gamma_\theta : S \rightarrow X$ when the parameter θ is fixed or $\Gamma_x : S \rightarrow \Theta$ when an observation x is made. Thus, one could obtain two restricted models: (S, m, Γ_θ, X) and (S, m, Γ_x, Θ). The former may be used to derive upper and lower probability for future observations x, and the latter to derive the same for the parameter θ. When there are multiple independent observations, one can produce one restricted model for each observation and then combine these models using Dempster's rule to derive combined upper and lower probabilities for θ. When a prior distribution $p(\theta)$ is available, it can be regarded as yet another restricted model (Θ, p, I, Θ), where I is the identity mapping, which can be also combined with the restricted models based on sample observations. Therefore, Chap. 4 consolidated the fiducial arguments and Bayesian inference and brought them under the same umbrella of belief functions. It not only showed the feasibility of probabilistic inference without priors but also re-expressed Bayesian inference as the combination of independent sources of information, including priors and sample observations.

As side products of its application of belief functions, Chap. 4 made additional theoretical contributions. The first was the concept of total ignorance and its representation via upper and lower probabilities. This provided a simple resolution to the old controversy about the representation of ignorance via a probability distribution, and led to the concept of vacuous belief function [34] that showcased the superiority of belief functions for subjective judgments. The second was the idea of viewing prior knowledge as a source of information similar to other sources such as sample observations to be combined via Dempster's rule. This idea led to the concept of Bayesian belief functions [34] and embraced Bayesian probabilities as a special case of belief functions. The third was the idea of viewing a multivalued mapping as a random set. This idea led not only to an alternative formalization of the theory of belief functions but also to an alternative perspective on belief functions as the extension of probability distributions over random variables. It also allowed for a rigorous mathematical foundation for belief functions.

Chapters 2–4 established most of the basic ideas and concepts for a new theory of probable reasoning. Without extensions, refinements, and reinterpretations by Glenn Shafer, however, these elements would still have been in the narrow statistical confines of random sampling. While studying for his Ph.D. at Harvard, Shafer got acquainted with Dempster's work. Later he was asked to make a presentation on Dempster's upper and lower probabilities at Princeton. His book—*A Mathematical Theory of Evidence* [34]—was a result of his ensuing effort. Characterized by Shafer's intellectual boldness, the book announced the establishment of a new mathematical theory for probable reasoning as a genuine generalization of or superior alternative to subjective Bayesian theory. To distinguish the theory from theories of imprecise probability, the book renamed Dempster's lower and upper probabilities respectively as belief and plausibility functions. Whereas Dempster had emphasized the derivation of lower and upper probabilities from S, m and Γ, Shafer regarded belief functions as a fundamental concept—an alternative to subjective probabilities. Following Andrei Kolmogorov, who built probability theory on three mathematical axioms, Shafer built his theory of belief functions on three similar axioms, with the additivity of probabilities being replaced by the super-additivity of belief functions. He showed that a belief function satisfied the three axioms if and only if it was as derived from a basic probability assignment $m(2)$. It was this connection that allowed Shafer to simplify Dempster's four-element model (S, m, Γ, T) into a two element model (m, T), which assigned probabilities m directly to subsets of the target space T while keeping S and Γ implicit. It was also this connection that allowed belief functions to express partial beliefs for probable reasoning using the two basic ideas due to Arthur Dempster: the idea of obtaining degrees of belief for one question from subjective probabilities for a related question (evidence), and Dempster's rule for combining degrees of belief when they were based on independent items of evidence.

Besides providing new terminologies, notations, and the axiomatization, Shafer also greatly extended Dempster's mathematical results. Most notable are the concepts of support functions and weights of evidence. These concepts served two purposes. First, they showed how weights of evidence might be converted into degrees of belief and combined using Dempster's rule, and thus showed how the theory of belief functions could be rebuilt and applied around these concepts. Second, they justified the theory of belief functions from works of Jakob Bernoulli and other ancient scholars on probabilities. It was probably from these works Shafer generated his idea of re-interpreting Dempster's work as a theory of probable reasoning through the combination of evidence.

The notion of weights of evidence can be traced back to Jakob (James, Jacques) Bernoulli in his book *Ars Conjectandi.* Jakob Bernoulli died in 1705. His book was given to the printer by his nephew Nicholas Bernoulli, under the pressure of mathematicians. After it was published in 1713 by the Thurneysen Brothers Press in Basel, *Ars Conjectandi* became the founding document of mathematical probability, replacing *Calculating in Games of Chance* by Christian Huygens, which was the first ever printed book on probability and served as the standard text for over 50 years after 1657. *Ars Conjectandi* consisted of four parts. Part 1 was an improved version of Huygens' book on games of chance with annotations. This part made many well-known contributions in elementary probability theory. For example, the notion of Bernoulli trials, the multiplication rule for independent events, and the Bernoulli distribution were all presented in this part. Part 2 offered a thorough treatment of the mathematics of combinations and permutations, including the numbers known as "Bernoulli numbers." Part 3 solved some complicated problems of games of chance using combinatorics. The final part manifested Bernoulli's crowning achievement in mathematical probability. For example, he proved what we now know as the weak law of large numbers. A complete English translation of the book was done only recently by Edith Dudley Sylla [1]. Part 4 was translated into English by Bing Sung [41] with a preface by Arthur Dempster.

Part 4 of *Ars Conjectandi* envisioned the application of probability theory to economics, morality, and politics. Bernoulli did not in fact make such practical applications. But he did succeed in formulating a concept of mathematical probability that went beyond the application to games of chance. He characterized probability as a degree of certainty that differs from absolute certainty as a part differs from a whole. The art of conjecture was to measure as exactly as possible the probabilities of things. With respect to games of chance, the symmetry of physical devices suggested we could calculate the probability of a specified outcome as the number of favorable cases divided by the total number of cases. In many other situations, however, such symmetry could not be relied upon and the classical procedure could not be applied. Thus, probability was a measure of imperfect knowledge and was personal in the sense that it varied from person to person according to his knowledge. This statement has credited Bernoulli today as the father of subjective probability theory. Nevertheless, it is instructive to compare Bernoulli's notion with several distinct modern ones of subjective probability. In the personalist theory of Bruno de Finetti, Frank P. Ramsey, and Leonard J. Savage, probabilities may be unknown only insofar as one "fails to know one's own mind" and are measured by the betting ratio at which the person in question is willing to bet on the truth of the statement. In the logical theory of John M. Keynes and Harold Jeffreys, probabilities may be unknown by failure to do logic but no experiment will help check up on logical probability. In the subjective theory of Werner Heisenberg, probability contains the objective element of tendency and the subjective element of incomplete knowledge. An observation cannot predict a result with certainty; what can be predicted is the probability of a certain result, and this probability can be checked by repeating the experiment many times. In contrast, Bernoulli believed that everything was governed by God and causal mechanism. As long as we knew the causes, what could seem to be to one person at one time an uncertain event might be at another time to another person (indeed, to the very same person) a deterministic event. From this comparison, Hacking [19] concluded that Bernoulli's subjectivism was less like the personalist or logical point of view, and more like that of the physicists.

Because he wanted to measure probabilities, Bernoulli was concerned with how to combine evidence of different sorts. He stated that probabilities are estimated by the number of cases and the weight of evidence.[1] His first scheme of combination followed the Port Royal logic of Pascal and distinguished internal versus external evidence. *Internal evidence* arises from the topics—cause, effect, subject, sign, circumstance, or anything that directly connected to the question of interest. *External evidence* appeals to human authority or testimony. His second scheme descended from Gottfried Wilhelm Leibniz's notion of pure and mixed evidence. *Pure evidence* proves a thing with a certain probability without giving a positive probability to the opposite thing, whereas *mixed evidence* proves a thing with a certain probability and proved the opposite with the complementary probability. That Gracchus turned pale

[1] Bernoulli used *argumentum* instead of evidence. This Latin word has a broad sense emcompassing the meanings of the modern English words "evidence" and "argument."

when interrogated is an example of pure evidence. To assess the probability of a thing, one can list all pieces of evidence. If all pieces are mixed, then the probability is the number of favorable cases divided by the total number of cases. The resulting probabilities are additive and complementary. However, if all or some pieces of evidence are pure, Bernoulli formulated a rule of combination, which Shafer [34] showed was a special case of Dempster's rule. The resulting probabilities may not be additive and complementary.

Clearly, Bernoulli's notion of non-additive probabilities was the ancestor of what we now call belief functions. This explains why Shafer reinterpreted lower probabilities as epistemic probabilities or degrees of belief while abandoning the term of lower probability, which can arise as lower bounds over classes of Bayesian probabilities. It was also clearly Bernoulli's idea of probability assessment through combining weights of evidence that motivated Shafer to recast Dempster's theory of random sampling into a theory of evidence and to represent evidence using support functions.

4 Classic Works

Although they are presented chronologically, the classic contributions in this volume can be grouped, at least roughly, by their content and emphasis into seven categories: conceptual foundations, philosophical perspectives, theoretical extensions, alternative interpretations, and applications to artificial intelligence, decision making, and statistical inference.

4.1 Conceptual Foundations

Four chapters may be said to have established the conceptual foundation of belief functions presented in Shafer's book [34]: Chaps. 2–4 by Dempster and Chap. 7 by Shafer.

The previous section has given a detailed account of Chaps. 2–4. In brief, Chap. 2 proposed the multivalued mapping approach to deriving upper and lower probabilities to replace posterior distributions in the absence of Bayesian priors. It was the first belief-function treatment of Fisher's fiducial method. Chapter 3 envisioned the problem of obtaining degrees of belief for one question from a probability measure of a related question through a multivalued mapping. It introduced Dempster's role of combination and a corresponding notion of commonality functions. Chapter 4 explicitly applied Dempster's rule to statistical inference and marked the birth of generalized Bayesian theory or a theory of belief functions.

Chapter 7 extended the concept of belief functions defined in [34] to continuous frames of discernment. Following the approach by Gustave Choquet [4], the chapter considered a subset in a continuous frame as the limit of a sequence of finite subsets, and proposed the concepts of continuity and condensability. Continuity was defined in the same way as the continuity of a Lebesgue measure. Condensability was a key assumption for the extension: a belief function is *condensable* if its plausibility function Pl satisfies

$$Pl(A) = \sup\{Pl(B) \mid B \subset A \text{ and } B \text{ is finite}\}.$$

The chapter then showed how to extend a continuous or condensable belief function on an algebra of (finite) subsets of Θ—a set of subsets that is closed under both set union and complement operations—to a continuous or condensable belief function on the power set 2^{Θ}. The main tool used for such an extension was Choquet's integral representation theorem, which implies that every belief function can be represented by an *allocation of probability.* Technically, for every belief function Bel on an algebra of subsets of Θ, there exists a homomorphic mapping ρ into a probability algebra with a positive and additive probability measure m such that $\rho(A \cap B) = \rho(A) \cap \rho(B)$ and $Bel(A) = \int \rho(A) dm(\rho)$.

4.2 Philosophical Perspectives

We place in this group Chaps. 6 and 9 by Glenn Shafer, Chap. 13 by Glenn Shafer and Amos Tversky, and Chap. 30 by Arthur P. Dempster. All these chapters justify the theory of belief functions from broader perspectives.

Chapter 6 provided a historical account of non-additive probabilities as well as rules of combining evidence. It focused on the work of Jakob Bernoulli and its extension by Johann Heinrich Lambert, a 18th century scholar. It related these ancient concepts of non-additive probabilities to the modern concept of belief functions and showed that both Bernoulli and Lambert's rules of combination are special cases of Dempster's rule.

Chapter 9 systematically examined the critiques by Bayesian or imprecise probability theorists. Both Bayesian and lower probability theories can appeal to the betting interpretation or the Dutch-Book argument for the semantics of its degrees of belief. What is the semantics of belief for a belief function? In the literature, some authors appeal to the probability of provability [28, 31, 37] or the support of arguments [23, 21]. Nevertheless, Chap. 9 argued that Bayesian, imprecise probability, and belief functions are all constructive theories for probability judgment. They need not rely for their meaning and justification on any behavioral interpretation. Instead, the degree of belief is the result of comparing evidence to knowledge about chances governing the truth. The chapter proposed the randomly coded message as a scale for such a comparison: suppose someone chose a code at random from a list of codes and we knew the probability of each code being chosen. Then $m(A)$ is the sum of probabilities of codes, by which the decoded message is A.

Furthering the idea of constructive probability, Chap. 13 dealt with human judgments of probabilities and belief functions. It illustrated that both Bayesian theory and the theory of belief functions were formal languages for one to analyze evidence and express his degrees of belief; they had the usual components of a language, including vocabularies, semantics, and syntax. It suggested that making a probability judgment was a process of conducting a mental experiment and hence the quality of the experimental design affected the quality of the judgment. The chapter offered some alternative designs for using the languages of Bayesian probabilities and belief functions. For example, the total-evidence design often used with Bayesian theory is distinguished from the belief function that emphasizes the decomposition of evidence. The chapter emphasized that theories of subjective probability (including belief functions) were not psychological models, either normative or descriptive, for making judgments. An experimental design for using such a theory (or its semantics and syntax) must guide the process of making probabilistic judgments.

Chapter 30 is a new contribution based on the 1998 R.A. Fisher Memorial Lecture.[2] The theory of belief functions arose from the need for a new scientific method unifying various statistical methods, including fiducial and Bayesian methods. As opposed to Bayesian, Fisherian, or frequentist statistics, Dempster proposed logicist statistics as a unified way to study principled and explicit reasoning about uncertainty. The key concept was formal subjective probability, which interprets each numerical probability as a degree of certainty reflecting specific formalized evidence and information within a formal mathematical model. Dempster showed that this concept encompasses both modern Bayesian and traditional Fisherian thinking, and he interpreted frequentist theory in a way that gives appropriate weights to both science and mathematics, and to both subjective and objective elements. He also suggested that the Dempster-Shafer theory embodies a more suitable paradigm for logicist statistical inference than Bayesian inference and is logicist in a fundamental way because it integrates nonprobabilistic "propositional" logic with probabilistic reasoning.

4.3 Theoretical Extensions

This group contains Chap. 5 by Hung T. Nguyen, Chap. 11 by Ronald R. Yager, Chap. 15 by Nevin L. Zhang, Chap. 19 by Alain Chateauneuf and Jean-Yves Jaffray, and Chap. 21 by John Yen. These five chapters extended the theory of belief functions in various ways.

Chapter 5, by Nguyen, was the first research work on belief functions published by someone other than Dempster and Shafer. It carried out the idea in Chap. 3 by Dempster that a multivalued mapping might be considered a random set and established the connection between belief functions and random sets. It showed that, in finite cases, the probability distribution

[2] The Fisher Lectureship and Award was established in 1963 by the Committee of Presidents of Statistical Societies to recognize the importance of statistical methods for scientific investigations.

of a random set is a basic probability assignment and a belief function is deduced from the probability distribution of the random set. It characterized the condensability of belief functions of Chap. 7 using the notion of regularity of probability measures. It showed that a plausibility function is condensable if and only if the corresponding probability distribution of a random set is regular.

In probability theory, entropy is a measure of the disorder and randomness present in a distribution. In fuzzy logic, specificity is an overall measure of how much a possibility distribution points to one and only one element as the manifestation of a fuzzy variable. A belief function has both randomness and non-specificity components. Thus, Chap. 11, by Yager, developed similar concepts for belief functions. For a belief function with mass function m and plausibility function Pl, its entropy is

$$E = -\sum\{m(A)\log(Pl(A)) \mid A \subseteq \Theta\}.$$

This formula reduces to Shannon entropy for Bayesian belief functions. It attains zero entropy for consonant belief functions and the maximum entropy when focal elements are disjoint and when the belief mass is equally distributed among all focal elements. The specificity of a belief function with mass function m is defined as

$$S = \sum\{\frac{m(A)}{|A|} \mid \phi \neq A \subseteq \Theta\}.$$

This measure reduces to the specificity of a fuzzy variable for a consonant belief function. It reaches the minimum value for a vacuous belief function and the maximum value for Bayesian belief functions. Chapter 11 led to many studies on the measurement of total uncertainty encompassing both randomness and nonspecificity. One noteworthy contribution [27] uses a set of reasonable axioms to derive measures such as

$$H = \sum\{m(A)\log(\frac{|A|}{m(A)}) \mid \phi \neq A \subseteq \Theta\}.$$

This measure has many desirable features, including additivity for independent belief functions and reduced computational complexity.

Chapter 15, by Zhang, was one of few contributions that directly improved the classic book by Shafer [34]. Note that the weight of evidence provides a full assessment of evidence for simple and separable support functions. Can a similar concept be extended to support functions that may not be separable? Shafer [34] approached the problem indirectly through notions of internal conflict and impingement. For any separable support function T, let W_T and v_T be respectively its weight of internal conflict and impingement function. For any support function S over Θ, let ϵ_S be the set of all its extensions that are separable support functions over some refinements of Θ. Then the weight of internal conflict for S was defined as the minimum weight of internal conflict among all separable support functions in ϵ_S:

$$W = \inf\{W_T \mid T \in \epsilon_S\}. \tag{21}$$

Similarly, the impingement function of S was derived from those of all its separable extensions: for any subset $A \subset \Theta$,

$$v(A) = \inf\{v_T(\omega(A)) \mid T \in \epsilon_S,\ \omega \textit{ is a refinement mapping}\}. \tag{22}$$

Since a separable support function itself is a support function, the definitions in (21) and (22) should also apply to separable support functions and the result should be consistent, i.e., if S is a separable support function, then $W = W_S$ and $v = v_S$. Shafer [34] proved the consistency by assuming the weight-of-conflict conjecture, which has not been proved to be true yet. This chapter proved the consistency without the conjecture.

As we see in Chaps. 5 and 7, a belief function is a monotone capacity of infinite order whereas a mass function is the Möbius inversion of the capacity. Chapter 19, by Chateauneuf and Jaffray, studied the properties of capacities of all orders, whose relationship is that, for any $K \geq 2$, if a capacity is K-monotone, then it is also L-monotone for $K \geq L \geq 2$ and 1-monotone (or monotonic in usual sense) if $f(\theta) \geq 0$ for any $\theta \in \Theta$. A capacity is defined as ∞-monotone if it is K-monotone for any $K \geq 2$. The chapter obtained some useful results characterizing the capacities through Möbius transformations. For example, it showed that, capacity f is K-monotone ($K \geq 2$) if and only if, for any A and $C \subset \Theta$ with $2 \leq |C| \leq K$, its Möbius inversion m satisfies:

$$\sum_{C \subset B \subset A} m(B) \geq 0.$$

The chapter also characterized probability distributions that dominate (or "are compatible with" in terms of Chap. 2) a belief function. It showed that if the probability distribution P satisfies $P(A) \geq f(A)$ for any A, then P is the weighted average of the Möbius inversions of f:

$$P(x) = \sum_{x \in B} \lambda(B, x) m(B).$$

It generalized a result in Chap. 3 by Dempster and showed f is ∞-monotone if and only if every probability distribution dominating f is the weighted average of Möbius inversions.

Many scholars in the area of fuzzy logic consider Chap. 21, by Yen, an outstanding paper. It is a favorite reference on the fuzzification of belief functions. It studied the computation of beliefs and plausibilities for fuzzy sets and extended Dempster's rule to fuzzy logic. It significantly improved other approaches by Zadeh, Ishizuka, Yager, and Ogawa while maintaining the semantics of the Dempster-Shafer theory of belief functions as well as possibility theory. It brought together belief functions and fuzzy logic into a hybrid approach to reasoning under various kinds of uncertainty in intelligent systems. The chapter started with a novel viewpoint, from which the computation of $Bel(A)$ was formulated as a linear programming problem:

$$\min \sum_{x \in A} \sum_{B} m(x, B)$$
$$s.t. \;\; m(x,B) \geq 0; m(x,B) = 0 \;\forall x \notin B; \sum_{x} m(x,B) = m(B),$$

here $m(x, B)$ denoted the probability mass allocated to x from $m(B)$. Then, when A was a fuzzy set, the chapter proposed to extend the problem into one of minimizing the extended objective function:

$$\sum_{x \in A} \sum_{B} m(x, B) \mu_A(x),$$

here $\mu_A(x)$ denoted the membership of x in A. If all focal elements were crisp (non-fuzzy), then the solution to the generalized problem is

$$Bel(A) = \sum m(B) \inf_{x \in B} \mu_A(x).$$

If any focal element B is fuzzy, it will be broken into one or more crisp focal elements, each of which is an $\alpha - cut$ of B:

$$B_\alpha = \{x \mid x \in B, \;\; \mu_B(x) \geq \alpha\},$$

with a basic probability mass

$$m(B_\alpha) = (\alpha_i - \alpha_{i-1}) m(B),$$

here $\alpha_0, \alpha_1, \alpha_2, ..., \alpha_n$ is a series of membership degrees of increasing order with $\alpha_0 = 0$ and $\alpha_n = 1$. For example, if focal element $B = \{$*(yound, 0.4), (old, 0.7)*$\}$ with $m(B) = 0.8$, then we get two *α–cuts* as follows: $B_{0.4} = \{yound, old\}$ and $B_{0.7} = \{old\}$ with basic probability masses $m(B_{0.4}) = (0.4 - 0) \times m(B) = 0.32$ and $m(B_{0.7}) = (0.7 - 0.4) \times m(B) = 0.24$. Then, $Bel(A)$, for any fuzzy set A, is

$$Bel(A) = \sum_{B} m(B) \sum_{i} (\alpha_i - \alpha_{i-1}) \inf_{x \in A_{\alpha_i}} \mu_A(x).$$

The approach to extending Dempster's rule was also novel. It considered a multivalued mapping $S \rightarrow T$ as a compatibility relation $S \times T$ and generalized it to a fuzzy relation $C : 2^{S \times T} \rightarrow [0, 1]$, which is a joint possibility distribution. It considered Dempster's rule as the combination of compatibility relations and generalized it as the combination of fuzzy relations, which in turn is equivalent to the multiplication of noninteractive possibility distributions.

This led to the generalized rule for combining fuzzy belief functions. Let m_1 and m_2 be two fuzzy mass functions. Then,

$$m_1 \oplus m_2(C) = \frac{\sum_{A \cap B = C} \max_x \mu_{A \cap B}(x) m_1(A) m_2(B)}{1 - \sum_{A,B} (1 - \max_x \mu_{A \cap B}(x)) m_1(A) m_2(B)}.$$

4.4 Artificial Intelligence

Five chapters apply belief functions to uncertain reasoning in artificial intelligence. Chapter 8 by Jeffrey Barnett was the first paper dealing with computational issues in implementing Dempster's rule of combination. It proposed an algorithm based on the very strong assumption that each piece of evidence either confirms or denies a single proposition, i.e., all focal elements are singletons or their negations. Chapter 12 by Jean Gordon and Edward Shortliffe proposed an improved algorithm capable of handling hierarchical evidence, where focal elements and their negations could be arranged in a tree-like structure. To avoid the exponential explosion in computations, the algorithm employed approximation to combine evidence. The approximation was usually reasonable but did give unsatisfactory results in the case of highly conflicting evidence. In addition, the approach did not produce the degrees of belief for all focal elements involved in the computation except for those in the tree. Chapter 18 by Glenn Shafer and Roger Logan presented a further improvement that is at least equally efficient while removing all the above limitations. These chapters built upon each other technically but are all included here because they made the history in distinct ways. Chapter 8 coined the name "Dempster-Shafer theory" and introduced it to the AI community. It was clearly one of the initial sources that led Edward Shortliffe to realize the relevance and applicability of belief functions to the issues addressed by the certainty factor model implemented in the medical advising program MYCIN. Because of their role in MYCIN, Gordon and Shortliffe were probably the most influential of the authors who made belief functions widely known as "the Dempster-Shafer theory" to AI researchers.

Chapter 16 by John D. Lowrence, Thomas D. Garvey, and Thomas M. Strat proposed a formal framework based on belief functions for knowledge representation and uncertainty reasoning in expert systems, setting belief functions up as an alternative to rules, frames, and semantic networks. It introduced the new term "evidential reasoning" for the framework and demonstrated its application in the Gister project at SRI International. Stemming from the application of belief functions to Navy intelligence problems, Chap. 16 was very practical in nature. Its approach to knowledge representation, i.e., modeling compatibility relations, provided a perfect example to illustrate the applicability of belief functions to real problems.

In the framework of Chap. 16, each piece of knowledge is represented by a belief function. Making inferences boils down to combining all component belief functions and marginalizing the joint belief function into a subframe of discernment (see definition in (13)). Of course, such a straightforward approach would be very inefficient, if not infeasible, when the size of the joint frame is large. A creative solution to the problem is so-called *local computation* that computes marginals without computing the joint. The basic idea is to arrange all the frames of discernment into a tree-structured graph, called *a join-tree* or *Markov tree*, and propagate knowledge by sending and absorbing messages step-by-step in the tree. Each step involves sending a message from a node to a neighbor and thus involves only a small number of frames that are near each other in the join-tree.

Scholars in belief functions, including Glenn Shafer, Prakash P. Shenoy, Augustine Kong, and Khaled Mellouli, pioneered the local computation method. Later they demonstrated the applicability of this method to other calculi, including Bayesian probabilities and fuzzy logics. Chapter 20, by Shenoy and Shafer, presented an abstract framework that covered diverse local computation models as special cases. It characterized many types of computational problems as one of applying two operators: combination and marginalization, where combination corresponds to the integration of two or more factors into a joint model and marginalization corresponds to the projection of a model to a subset of variables. The chapter showed that local computation was applicable to such problems if the two operators satisfied four axioms. For belief functions, for example, these axioms can be represented as follows:

Axiom 4 *Combination operator $\oplus$ is commutative: for any Bel_1 and Bel_2,*

$$Bel_1 \oplus Bel_2 = Bel_2 \oplus Bel_1.$$

Axiom 5 *Combination operator $\oplus$ is associative: for any Bel_1, Bel_2, and Bel_3,*

$$Bel_1 \oplus (Bel_2 \oplus Bel_3) = (Bel_1 \oplus Bel_2) \oplus Bel_3.$$

Axiom 6 *Marginalization is consonant: for any Bel on the frame $\Theta(I)$ and $K \subset J \subset I$. Then*

$$(Bel^{\downarrow J})^{\downarrow K} = Bel^{\downarrow K}.$$

Axiom 7 *Marginalization is distributive over combination: for any Bel_1 and Bel_2 and I,*

$$(Bel_1 \oplus Bel_2)^{\downarrow I} = (Bel_1)^{\downarrow I} \oplus (Bel_2)^{\downarrow I}.$$

Chapter 20 also presented the Shenoy-Shafer architecture for carrying out local computation over a Markov tree, and demonstrated the algorithm using an example of probability propagation. Compared with other similar approaches (e.g., [24]), this architecture gains some efficiency by avoiding divisions, which are required by other methods for obtaining conditional probabilities.

4.5 Decision Making

The theory of belief functions is not meant to be a normative or descriptive theory for decision making. Thus, it does not provide normative axioms or behavioral predictions on how to make decisions and judgments. Because of its expressive power in encoding evidence or modeling uncertainty, however, it has exceptional prescriptive value as a decision support tool. Here we review four chapters demonstrating creative use of belief functions for the purpose, including Chap. 23 by Rajendra P. Srivastava and Glenn Shafer, Chap. 24 by Ronald R. Yager, Chap. 27 by Galina Rogova, and Chap. 29 by Thierry Denoeux.

Chapter 23, by Srivastava and Shafer, applied belief functions to audit decision-making. The chapter derived analytical expressions of the audit risk at three levels: the financial statement level, the account level, and the audit objective level. It made a distinct contribution to the field by showing how to interpret and use plausibility numbers to encode accounting evidence. It also proposed a hierarchical network for evidential reasoning and dealt with belief propagation through the "AND" gates, which were inherent in business decision problems.

There have been numerous attempts to incorporate belief functions into expected utility theory to take advantage of their flexibility in uncertainty modeling. Chapter 24, by Yager, showcased such attempts. It is included here because it is theoretically sound and computationally feasible. Whereas other work reduced Dempster-Shafer degrees of belief to probabilities for use as decision weights, this chapter proposed deriving decision weights from a mathematical programming model. Once we set a pessimism level—a necessary concept for decision making under uncertainty—entropy maximization problem gives weights to be assigned to each outcome within a focal element. The weights then determine the weighted average value of outcomes in the focal element, which along with the corresponding basic probability numbers determine an overall value for each choice. It is shown that the formalism unifies several common decision models for decision-making under risk, uncertainty, and ignorance. Its ordered weighting mechanism is also consistent with psychological findings that have led decision theorists to generalize *expected utility theory* to so-called *rank-dependent utility* [25, 26, 30].

Chapter 27, by Rogova, is a real application with real results. The topic is very timely. In machine learning, the idea of boosting, i.e., combining simple poor learners to form an ensemble that outperforms individual single ensemble members while avoiding overfitting, is gaining a lot of interest in the last decade. In theory, it is known that learners, each performing only slightly better than random, can be combined to form an arbitrarily good ensemble hypothesis [20]. Schapire [33] was the first to provide a provably polynomial time boosting algorithm. He and his colleagues [13] applied boosting to a real-world optical character recognition by using neural networks as base learners. Chapter 26 demonstrated the application of Dempster's rule to the same problem. Interestingly, it also used neural networks as base learners. It showed that the proposed approach allowed 15–30% reduction of misclassification error compared to the best individual classifier. The method made Eastman Kodak one of the small group of the leaders in an industrial competition for the best optical recognition algorithm.

Chapter 29, by Denoeux, is considered an outstanding application of belief functions to decision making. It proposed a new approach to pattern classification that considered each of the k-nearest neighbors as an item of evidence and used Dempster's rule of combination to pool all evidence together to form a judgment concerning the class membership of a new incoming pattern. Simulation results showed that the proposed approach outperformed the classic voting k-nearest neighbor approach as well as its distance-weighted variant.

4.6 Statistical Inference

Parametric statistical inference is not only the source of motivation for the theory of belief functions but also one of its most important application domains. Chapter 4 demonstrated the potential of belief functions for unifying the traditional fiducial argument and modern Bayesian inference. Here we review three additional chapters revisiting the problem of parametric inference using belief functions, including Chap. 10 by Glenn Shafer, Chap. 22 by Jean-Yves Jaffray, and Chap. 25 by Philippe Smets.

In his book [34], Shafer suggested translating each observation into a consonant belief function on a parameter based on the normalized likelihood. He recognized that this approach does not possess the desirable property that the result using a set of n independent observations be equal to the combination of the n belief functions obtained from the individual observations. Chapter 10, by Shafer, discussed three alternative approaches, including the fiducial argument, the generalized Bayesian method of Chap. 4, and the conditional embedding method of Chap. 25 (see below). It showed that these methods produce coherent results when the nature of the evidence establishing the parametric model is taken into account.

Chapter 22, by Jaffray, studied the effect of Bayesian conditioning when a belief function is (mis)understood as the lower envelope of compatible probability measures. It obtained two important results. First, it reproved the result by Fagin and Halpern [15] that the lower envelope of all Bayesian conditionals is still a belief function, and going beyond Fagin and Halpern, it developed an explicit expression for the mass function for the lower envelope. Second, it showed that the resulting lower envelope does not characterize the set of all conditionals. Let $\mathbf{Q}_E$ be the set of Bayesian conditionals that dominate $\underline{P}(A \mid E)$ (see (20)). Then, $\mathbf{P}_E \subset \mathbf{Q}_E$ if and only if there exist subsets A and B such that $Bel(A \cap B) > 0$, $Bel(A \cup B) < 1$, and $Bel(A \cup B) > Bel(A) + Bel(B) - Bel(A \cap B)$ (see Sect. 3 for the definition of $\mathbf{P}_E$). Also, $\mathbf{P}_E \subset \mathbf{Q}_E$ if and only if there exist E and F with $F \subset E$ and $Bel(F) > 0$ such that the lower envelopes of Bayesian conditionals do not satisfy $\underline{P}((A \mid E) \mid F) = \underline{P}(A \mid F)$, which is observed by both Bayesian and Dempster's conditioning.

A belief function $Bel(A)$ may be re-expressed in a conditional form as $Bel(A \mid E)$ given evidence E. Then Dempster's rule may be called the *conjunctive rule of combination*, because $Bel_1(A \mid E_1) \oplus Bel_2(A \mid E_2)$ is the combined belief function when both E_1 and E_2 are true. Chapter 25, by Smets, proposed the *disjunctive rule of combination* that allows the combination of two belief functions induced by two pieces of evidence, of which only one can be true. The disjunctive rule is intuitive when applied to parametric inference problems. Suppose B is a set of possible parameter values, one of which is true. For each $\theta \in B$, let us assume there is a belief function $Bel(A \mid \theta)$ representing the likelihood that the true value of X is in A when the parameter is θ. Then the combination of these belief functions follow the disjunctive rule as

$$Bel(A \mid B) = \prod\nolimits_{\theta \in B} Bel(A \mid \theta).$$

The disjunctive rule corresponds the multiplication of belief functions whereas the conjunctive rule corresponds to the multiplication of commonality functions. Based on the disjunctive rule, the chapter derived the generalized Bayesian theorem, where conditional probabilities are replaced by belief functions and prior probabilities by vacuous belief functions. Let B be a set of possible parameter values and A be a set of observations. Let $Bel(A \mid \theta)$ be the likelihood that X is in A given parameter θ. Then, the generalized Bayesian formula represents the posterior belief of B given A as follows:

$$Bel(B \mid A) = \prod\nolimits_{\theta \in \overline{B}} Bel(\overline{A} \mid \theta) - \prod\nolimits_{\theta \in \Theta} Bel(\overline{A} \mid \theta).$$

Some results in Chap. 25 were initially developed in an unpublished dissertation [36]. The generalized Bayesian theorem permits the induction of a belief function for parameters from an observation, leading to a new statistical method, called *conditional embedding*, which was extensively discussed in Chap. 10. Here the author represented them in his framework of transferable belief functions (see below) and attempted to develop a new approach for belief function propagation in a directed belief network.

4.7 Alternative Interpretations

Besides theoretical foundations, perspectives, advances, and applications, there have been tens of studies targeting alternative formalisms and interpretations of belief functions. Here we review four representative ones: Chap. 14 by Didier Dubois and Henri Prade, Chap. 17 by Enrique H. Ruspini, Chap. 26 by Jürg Kohlas and Paul-André Monney, and Chap. 28 by Philippe Smets and Robert Kennes.

There are many connections between *fuzzy logic* and belief functions. As we have seen earlier, possibility and necessity functions are consonant plausibility and support functions that have nested focal elements. Chapter 14, by

Dubois and Prade, exposed another connection between bodies of evidence and fuzzy sets. The classic concept of a set is simply a collection of elements, e.g., $A = \{x, y, z\}$. The concept of a *fuzzy set* extends it to include a *membership function* $m \to [0, 1]$ describing a graded assessment of the membership of elements in relation to a set. For example, $A = \{(x, 0.3), (y, 0.7), (z, 1)\}$ is a fuzzy set consisting of elements x, y, and z with membership grades 0.3, 0.7, and 1. Chapter 14 viewed a belief function as a further generalization of fuzzy logic and interpreted a body of evidence to be an extended fuzzy set, where an element was replaced by a focal element and a membership grade was replaced by a basic probability number. For example, $A = \{(\{x, y\}, 0.2), (\{z\}, 0.5), \{x, y, z\}, 0.3)\}$ is a body of evidence representing a belief function with $m(\{x, y\}) = 0.2$, $m(\{z\}) = 0.5$, and $m(\{x, y, z\}) = 0.3$. Chapter 14 studied belief functions using this formalism and introduced the notions of extended set operations such as union, intersection, and complementation to bodies of evidence. It discussed and compared four alternative definitions of set inclusion on bodies of evidence. Since it was easier to deal with consonant plausibility and support functions, the chapter applied the notions of inclusion, and pioneered the research on *possibilistic approximation* of bodies of evidence.

Recall that Chap. 9, by Shafer, interpreted belief functions as a constructive theory for probability judgment, and proposed the randomly coded message as the metaphor for understanding the semantics of belief functions. There was another popular interpretation that understood a degree of belief as the *probability of provability* [37, 29]. Formally, suppose we are given a set of logical theories, each logical theory is characterized by a set of axioms, and each theory is assigned a probability such that the probabilities add up to 1. The belief in a proposition A is then the sum of the probabilities of the theories from which A follows as a logical consequence. Chapter 17, by Ruspini, presented a similar interpretation based on the probabilities of a modal proposition toward developing a formal theoretical foundation for evidential reasoning as proposed by Lowrance, Garvey, and Strat in Chap. 16. In particular, it extended Carnap's notion of the *epistemic universe* [3] by including all possible combined descriptions of not only the state of the real world but also the state of knowledge that certain rational agents have about it. It showed that the probabilities defined over a *sigma algebra* of subsets of the epistemic universe have the properties of belief and mass functions and can represent the effect of evidence on the state of knowledge of the rational agents. The epistemic probabilities also induces lower and upper probabilities in the truth algebra that are identical to the interval bounds derived in Chap. 3. Finally, the chapter applied the epistemic logic approach to the problem of knowledge integration and obtained an *additive combination formula* for integrating a wide variety of knowledge of both dependent and independent sources. Under the assumptions of probabilistic independence, the formula is reduced to Dempster's rule of combination.

Chapter 26, by Kohlas and Monney, presented the theory of hints, another interpretation or formalism of the Dempster-Shafer theory of belief functions based on multivalued mapping Γ from a probability space (Ω) to another space of interest ($\varTheta$). As we explained earlier, Dempster's original model was $(\Omega, P, \Gamma, \varTheta)$, which in fact was exactly the same as *the model of hints.* The difference lies at the interpretation of Ω, which Fisher called the sample space of a pivotal variable, Dempster called the population of sample individuals, but here Kohlas and Monney called the space of *arguments.* Note that in his axiomatic approach, Shafer made the elements Ω and Γ implicit and assigned basic probability numbers directly to subsets of $\varTheta$. Chapter 26 argued that the model of a hint contains more information than its derived belief function does, and allows for a straightforward and logical derivation of Dempster's rule for combining independent and dependent bodies of information.

Chapter 28, by Smets and Kennes, presented the transferable belief model (TBM), a subjectivist and non probabilistic view of the Dempster-Shafer theory of evidence. In response to the need for integrating belief functions into a normative decision theory such as expected utility theory, the TBM distinguished clearly the *credal level*, where beliefs are entertained, from the *decision level* where standard utility theory applies, the belief functions being converted into probabilities using the *pignistic transformation.* Another main idea underlying the TBM is the notion of unnormalized belief function and unnormalized conjunctive rule of combination, and the interpretation of the mass $m(\emptyset)$ assigned to the empty set, under the *open-world assumption*, as a degree of belief in the event that the frame of discernment does not contain the true value of the variable of interest.

5 Conclusion

In this chapter, we reviewed the basic concepts and major results presented in Glenn Shafer's book, provided a brief history of the conceptual development, and summarized the major contributions of the selected classic works.

In this volume we deliberately did not include any papers that involve misunderstandings of basic concepts. This includes well known papers by Lotfi A. Zadeh [43] and Judea Pearl [29]. Zadeh criticized the normalization procedure in Dempster's rule of combination. He used an example to show that, in the case of combining two highly conflicting pieces of evidence, the result is not intuitive, although Shafer thought otherwise [35]. Because of this criticism, many authors introduced the "open world" hypothesis and assigned a non-zero basic probability number $m(\emptyset)$ to the empty set (see Chap. 27). Judea Pearl [29] was mainly concerned with the inability of belief functions to represent imprecise probabilities. This concern was addressed 40 years ago by Dempster (see Chap. 2). A belief function was never meant to replace or represent an imprecise probability, which involves a larger set of compatible probability functions than a belief function does. Instead, it is meant to be a faithful representation of knowledge based on evidence and to combine the knowledge obtained from multiple independent pieces of evidence for making provable probable inferences.

With respect to future research on belief functions, Dempster [11] called for more realistic applications of belief functions to complex systems. He stressed the critical need for credible and tractable models to represent the details of complex systems where quantified uncertainties cannot be obtainable through more traditional routes. He suggested the development of Fisher pivotals and efficient inference algorithms, in particular two-stage MC and MCMC methods, in conjunction with simplification from local computation with graphical structures. In order to improve public awareness of belief functions, Dempster [12] recently suggested a new semantics whereby every proposition A is associated with a triple (p, q, r), where p is the probability "for" A, i.e., $Bel(A)$, q is the probability "against" A, i.e., $Bel(\overline{A})$, and r is the probability of "don't know", i.e., $Pl(A) - Bel(A)$. He showed how this semantics can coherently interpret the notion of p-value, which is often misconstrued as a Bayesian probability "for" the null hypothesis. Theoretically, open problems still remain. For example, in earlier chapters Dempster left some questions on asymptotic properties of the combined belief function when the number of pieces of evidence approaches infinity. In this chapter, we reviewed Shafer's the weight-of-conflict conjecture that is still unsolved, although Chap. 14 showed that it was not needed for justifying the concepts of weight of internal conflict and impingement for a support function. Another problem is posed by Bayesians who seek behavioral justifications of belief functions. Formally, is there a set of behavioral axioms that justifies the existence of a belief function? In other words, are there any necessary and sufficient conditions in terms of how people make choices or judgments in the face of uncertainty underlying a class of belief functions appropriate for the representation of the uncertainty?

References

[1] Bernoulli, J. *The Art of Conjecturing: Together with His Letter to a Friend on Sets in Court Tennis.* Johns Hopkins University Press, 2006. Translated by Edith Dudley Sylla.

[2] Boole, G. *An Investigation into the Laws of Thought.* Walton and Maberly, London, 1854. Reprinted 1951, Dover, NY.

[3] Carnap, R. *Meaning and Necessity.* University of Chicago Press, Chicago, Illinois, 1956.

[4] Choquet, G. Theory of capacities. *Ann. Inst. Fourier 5* (1953), 131–295.

[5] Dempster, A. P. On direct probabilities. *Journal of the Royal Statistical Society Series B 25* (1962), 100–110.

[6] Dempster, A. P. Further examples of inconsistencies in the fiducial argument. *Annals of Mathematical Statistics 34* (1963), 884–891.

[7] Dempster, A. P. On the difficulties inherent in Fisher's fiducial argument. *J. Amer. Statist. Assoc. 59* (1964), 56–66.

[8] Dempster, A. P. New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics 37* (1966), 355–374.

[9] Dempster, A. P. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics 38* (1967), 325–339.

[10] Dempster, A. P. A generalization of Bayesian inference (with discussion). *Journal of the Royal Statistical Society Series B 30* (1968), 205–247.

[11] Dempster, A. P. Belief functions in the 21st century: A statistical perspective. In *Proceedings of Insitute for Operations Research and Management Science Annual Meeting (INFORMS-2001)* (Miami Beach, FL, 2001).

[12] Dempster, A. P. The Dempster-Shafer calculus for statisticians. *International Journal of Approximate Reasoning* (2006), in press.

[13] Drucker, H., Schapire, R. E., and Simard, P. Y. Boosting performance in neural networks. *International Journal of Pattern Recognition and Artificial Intelligence 7* (1993), 705–719.

[14] Einstein, A., and Infeld, L. *The Evolution of Physics.* Simon and Schuster, New York, 1961.

[15] Fagin, R., and Halpern, J. Y. A new approach to updating beliefs. In *Uncertainty in Artificial Intelligence 6*, P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, Eds. Morgan Kaufmann, San Mateo, CA, 1991, pp. 317–325.

[16] Fishburn, P. *Decision and Value Theory.* Wiley, New York, 1964.

[17] Fisher, R. A. Inverse probability. *Proc. Camb. Phil. Soc. 26* (1930), 154–57, 172–173. Reprinted in Bennett, J. H. (1971). *Collected Papers of R. A. Fisher* 2, Univ. of Adelaide.

[18] Good, I. The measure of a non-measurable set. In *Logic, Methodology and Philosophy of Science*, E. Nagel, P. Suppes, and A. Tarski, Eds. Stanford University Press, Stanford, 1962, pp. 319–329.

[19] Hacking, I. *The Emergence of Probability.* Cambridge University Press, New York, 1975.

[20] Kearns, M., and Valiant, L. Cryptographic limitations on learning Boolean formulae and finite automata. *Journal of the ACM 41* (1994), 67–95.

[21] Kohlas, J., and Monney, P.-A. *A Mathematical Theory of Hints.* Springer, 1995.

[22] Kong, A. *Multivariate Belief Functions and Graphical Models.* PhD thesis, Department of Statistics, Harvard University, Cambridge, MA, 1986.

[23] Laskey, K. B., and Lehner, P. E. Assumption, belief and probabilities. *Artificial Intelligence 41* (1989), 65–77.

[24] Lauritzen, S. L., and Spiegelhalter, D. J. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society Series B 50* (1988), 157–224.

[25] Liu, L. A note on Luce-Fishburn axiomatization of rank-dependent utility. *Journal of Risk and Uncertainty 28*, 1 (2004), 55–71.

[26] Luce, R. D., and Fishburn, P. C. A note on deriving rank-dependent linear utility using additive joint receipts. *Journal of Risk and Uncertainty 11* (1995), 5–16.

[27] PAL, N. R., BEZDEK, J. C., AND HEMASINHA, R. Uncertainty measures for evidential reasoning II: A new measure of total uncertainty. *International Journal of Approximate Reasoning 8* (1993), 1–16.

[28] PEARL, J. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, CA, 1988.

[29] PEARL, J. Reasoning with belief functions: An analysis of compatibility. *International Journal of Approximate Reasoning 4* (1990), 363–389.

[30] QUIGGIN, J. A theory of anticipated utility. *Journal of Economic Behavior and Organization 3* (1982), 323–343.

[31] RUSPINI, E. H. The logical foundations of evidential reasoning. Tech. rep., SRI International, Menlo Park, California, 1986.

[32] SAVAGE, L. J. *The Foundations of Statistics.* Wiley, New York, NY, 1954.

[33] SCHAPIRE, R. E. The strength of weak learnability. *Machine Learning 5* (1990), 197–227.

[34] SHAFER, G. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ, 1976.

[35] SHAFER, G. Belief functions and possibility measures. In *The Analysis of Fuzzy Information*, J. Bezdek, Ed., vol. 1. CRC Press, Boca Raton, FL, 1987, pp. 51–84.

[36] SMETS, P. *Un modle mathmatico-statistique simulant le processus du diagnostic mdical.* PhD thesis, Universit Libre de Bruxelles, Bruxelles, Belgium, 1978.

[37] SMETS, P. Probability of provability and belief functions. *Logique et Analyse 133-134* (1993), 177–195.

[38] SMITH, C. A. B. Consistency in statistical inference and decision (with discussion). *Journal of the Royal Statistical Society Series B 23* (1961), 1–25.

[39] SMITH, C. A. B. Personal probability and statistical analysis (with discussion). *Journal of the Royal Statistical Society Series A 128* (1965), 469–499.

[40] SRIVASTAVA, R. R., AND SHAFER, G. Belief-function formulas for audit risk. *The Accounting Review 67*, 2 (1992), 249–283.

[41] SUNG, B. *Translations from James Bernoulli (with a preface by A. P. Dempster).* Department of Statistics, Harvard University, Cambridge, Massachusetts, 1966.

[42] ZADEH, L. A. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems 1* (1978), 3–28.

[43] ZADEH, L. A. Review of *A Mathematical Theory of Evidence. AI Magazine 5* (1984), 81.

2

New Methods for Reasoning Towards Posterior Distributions Based on Sample Data*

Arthur P. Dempster

Abstract. This paper redefines the concept of sampling from a population with a given parametric form, and thus leads up to some proposed alternatives to the existing Bayesian and fiducial arguments for deriving posterior distributions. Section 2 spells out the basic assumptions of the suggested class of sampling models, and Sect. 3 suggests a mode of inference appropriate to the sampling models adopted. A novel property of these inferences is that they generally assign upper and lower probabilities to events concerning unknowns rather than precise probabilities as given by Bayesian or fiducial arguments. Sections 4 and 5 present details of the new arguments for binomial sampling with a continuous parameter p and for general multinomial sampling with a finite number of contemplated hypotheses. Among the concluding remarks, it is pointed out that the methods of Sect. 5 include as limiting cases situations with discrete or continuous observables and continuously ranging parameters.

1 Introduction

Consider an observable x, a parameter θ, and a specified family of distributious $\mathcal{F}_\theta$ over x-space. A conventional way of thinking about sample observations $x_1, x_2, \cdots, x_n$ from an unknown member of the family of distributions $\mathcal{F}_\theta$ is roughly as follows. First, a specific θ is determined by a process which need not be specified. Then, using this θ, the observations $x_1, x_2, \cdots, x_n$ are drawn independently at random each with the distribution $\mathcal{F}_\theta$. I believe that this attitude is held almost universally, where the schools of Fisher and Neyman usually think rather vaguely about θ as "chosen by Nature," while the Bayesian school specifies a prior distribution governing the random choice of θ. Some Bayesians prefer to think of θ as not fixed at all while $x_1, x_2, \cdots, x_n$ are governed by their joint marginal distribution. I do not see any operational importance in this distinction, since I assume that a parameter value may be

* Research supported in part by the Office of Naval Research through Contract Nonr 1866(37).

fixed and still legitimately be assigned a probability distribution, as long as the fixed value remains unknown.

The inference methods of this paper rest on a weaker definition of sample than that of the conventional model. The revised model gives up the idea that $x_1, x_2, \cdots, x_n$ are independently distributed according to $\mathcal{F}_\theta$ for fixed θ while retaining the feature that any observed sample $x_1, x_2, \cdots, x_n$ shall appear consistent with a distribution $\mathcal{F}_\theta$ for some θ regardless of the size n of the sample. Thus, a single observed sample can never be used to distinguish between the more relaxed model and the conventional model.

A trivial example will serve here to illustrate the new approach, the general theory being defined in Sect. 2. Suppose that x and θ take values on the real line. Suppose that $\mathcal{F}_\theta$ is the normal distribution $N(\theta, 1)$ with mean θ and variance unity. In contrast to the conventional approach of fixing θ and drawing $x_1, x_2, \cdots, x_n$ independently from the corresponding fixed $N(\theta, 1)$ distribution, an example of the new model is provided by asserting that $x_1 - \theta, x_2 - \theta, \cdots, x_n - \theta$ are governed by the law of n independent $N(0, 1)$ random variables, *but asserting no further laws whether deterministic or probabilistic about the variables* $x_1, x_2, \cdots, x_n, \theta$. Such an assumption no doubt appears artifical as stated here, but the discussion of Sect. 2 will provide a general foundation for it. The immediate purpose is to remark that, however one may think of determining θ, whether from a known process or from a black box, and whether dependent on $x_1, x_2, \cdots, x_n$ or not, the observed sample should in no way look unlike repeated drawings from some normal distribution with variance unity. In the absence of further empirical data involving repeated choices of θ, I do not see why the conventional model should be preferred over the new model.

The new model was first introduced in Dempster (1963), but with a further assumption. In the earlier paper it would have been assumed, for example, that $\theta, x_1, x_2, \cdots, x_n$ were jointly distributed random variables, i.e., that there existed a probability law simultaneously governing all of the variables $\theta, x_1, x_2, \cdots, x_n$. This joint distribution would have been specified only to the extent that $x_1 - \theta, x_2 - \theta, \cdots, x_n - \theta$ were asserted to be independently $N(0, 1)$ distributed while the conditional distribution of θ given $x_1 - \theta, x_2 - \theta, \cdots, x_n - \theta$ was not specified in any way. I now find it more satisfying to avoid extraneous complications due to assuming the existence of unknown laws. According to the present approach, it is correct to regard variables, such as parameters or yet-to-be-observed sample variables, as having existing but unknown real-world values. But it is seen as intellectually wasteful and possibly deceptive to assume the existence of probability laws governing such variables, unless these laws may be specified. This change has in turn suggested the more satisfying methods of defining posterior probabilities given in this paper.

An underlying motivation for this work is to be found in the need to break the serious deadlock between those statisticians who prefer Bayesian formulations and those who prefer formulations relying on the repeated sampling aspects of probability laws. These two traditions have a longer history of conflict than is generally realized. Todhunter, writing *circa* 1865, traced what would now be called a confidence or fiducial argument about binomial p to J. Bernoulli *circa* 1700. In correspondence, Leibniz questioned Bernoulli's method. Of more interest here is the fact that Laplace *circa* 1813 used both the Bernoullian and Bayesian approaches to estimate p and presented slightly discrepant normal approximations without comment. Poisson in 1830 also used both methods but achieved normal approximations which were in agreement. De Morgan in 1837 drew attention to the differences in logical processes used and queried Poisson's results. Todhunter himself believed Poisson to have been correct. Unfortunately, the question of the differences in Bernoullian and Bayesian approaches was confounded with the question of accuracy of normal approximations and was destined to remain obscure for around 100 years. See Todhunter (1865) pp. 57, 73, 554–558, for discussion and references.

At present, the Bayesian school is showing renewed vigor and is increasingly in conflict with what I have called above the Bernoullian school. Within the latter school there are disagreements between the many who generally follow Neyman and the few who prefer R.A. Fisher. The following two statements summarize a previously given (Dempster (1964)) attitude to the Neyman-Fisher differences: (i) Neyman's methods while often available and useful are not fully satisfying, and (ii) Fisher, while extraordinarily inventive and mostly on the right track, was unable to give coherence to his system and in particular failed to perfect his fiducial argument.

I believe that the methods of this paper are close to Fisher's viewpoint. The arguments given here resemble the fiducial argument in that they produce posterior probabilities using the sampling hypothesis and parametric hypotheses but no prior distribution. I believe also that the basic reasoning principle described in Sect. 3 is essentially what Fisher relied on in his fiducial argument.

At the same time the new methods of this paper can be viewed as belonging under a common umbrella with the Bayesian methods. This umbrella is described in a later paper (Dempster (1965)). There the logic underlying upper and lower probability systems is given more generally. Rules are given for combining independent sources of information. The methods of this paper implicitly apply these rules to the combination of information from individual sample observations. If a prior distribution is available, it may be combined with the sample information *according to the same rules*, and the result is the standard Bayesian answer (Dempster (1965)).

2 Construction of the Sampling Model

Throughout the following discussion measure-theoretic details are not supplied, mostly because they are obvious in the range of examples of present interest.

The basic components of the theory are a pair of spaces $\mathcal{A}$ and $\mathcal{X}$. $\mathcal{A}$ represents the *population* being sampled, and each *population individual* $a \varepsilon \mathcal{A}$ has a corresponding *observable characteristic* $x \varepsilon \mathcal{X}$. The mapping $a \rightarrow x$ thus assumed to exist is regarded as unknown but subject to certain restrictions posed below. The statement that a population individual a comes under observation as part of a sample is construed to mean that the x corresponding to a becomes known to the observer. The observer is not allowed, however, to identify a.

A unique probability measure μ over $\mathcal{A}$ is assumed given. This plays the role of the law governing the random sampling operation. A *finite* population of size N is represented by a set of N elements, and the natural measure μ governing random sampling is the measure assigning probability $1/N$ to each of the N elements. The reader may supply the obvious definitions of a random sample $a_1, a_2, \cdots, a_n$ from $\mathcal{A}$, sampling either with replacement or without replacement as desired. When an infinite population is postulated, an appropriate choice of $\mathcal{A}$ and μ is less clear, and, to the extent that various choices may be transformed into one another, the choice is more or less arbitrary. A convenient representation for the infinite population structures used in this paper takes $\mathcal{A}$ to be a simplex and μ to be the uniform distribution over the simplex. A random sample $a_1, a_2, \cdots, a_n$ from an infinite population $\mathcal{A}$ is defined, as one would expect, to be a drawing from the product measure μ^n over the product space $\mathcal{A}^n$.

Besides $\mathcal{A}, \mathcal{X}$ and μ, the user of the theory must specify in each instance (i) a class of contemplated mappings $a \rightarrow x$, and (ii) a family of probability measures over $\mathcal{X}$ whose typical member may be denoted by $\mathcal{F}_\theta$ where θ ranges over a space Θ. The family of measures $\mathcal{F}_\theta$ is used in the theory to define two postulates restricting the class of contemplated mappings $a \rightarrow x$, namely

(P1) the probability measure over $\mathcal{X}$ induced by the measure μ over $\mathcal{A}$ under any contemplated mapping $a \rightarrow x$ must be $\mathcal{F}_\theta$ for some $\theta \varepsilon \Theta$, and

(P2) exactly one mapping in the class of contemplated mappings $a \rightarrow x$ leads to the induced measure $\mathcal{F}_\theta$ over $\mathcal{X}$ for each $\theta \varepsilon \Theta$.

(P1) and (P2) together imply a one-one correspondence between the class of contemplated mappings $a \rightarrow x$ and the family of measures $\mathcal{F}_\theta$. The two postulates are kept separate in the exposition because (P1) is easier to swallow than (P2). A discussion of (P2) will be given shortly.

In any application of the theory, a random sample $a_1, a_2, \cdots, a_n$ is drawn from $\mathcal{A}$ as specified above. The observer identifies the corresponding $x_1, x_2, \cdots, x_n$ under the true mapping $a \rightarrow x$. He is then asked to draw inferences concerning which member of the class of contemplated mappings is the true member or, equivalently, concerning which θ in Θ is the true θ. The suggested mode of inference is given in Sect. 3.

The $N(\theta, 1)$ example of Sect. 1 may be used as a first illustration of the theory. Take $\mathcal{A}$ to be the whole real line and take μ to be the $N(0, 1)$ distribution over $\mathcal{A}$. Take $\mathcal{X}$ to be the whole real line, and take $\mathcal{F}_\theta$ to be the $N(\theta, 1)$

distribution, where the range space Θ of θ is also the whole real line. Finally, define the class of contemplated mappings $a \to x$ to be

$$a \to x = \theta + a, \tag{1}$$

where the dual interpretation of θ as a parameter for the class of mappings and as a parameter for the class of distributions $\mathcal{F}_\theta$ defines the one-one correspondence satisfying (P1) and (P2). The essential feature of this illustration is the preservation of the natural orderings on $\mathcal{A}$ and $\mathcal{X}$ under the whole class of mappings from $\mathcal{A}$ to $\mathcal{X}$. The particular representation of $\mathcal{A}$ and μ is not essential, and any monotone one-one transformation, for example carrying μ on $\mathcal{A}$ into a uniform distribution on (0, 1), could be used to obtain an alternative representation. This example will be termed a *structure of the first kind* in the later discussion of this section. Note that, as remarked in Sect. 1, the only probability law operating is the law of n independent $N(0, 1)$ random variables applied to $a_1, a_2, \cdots, a_n$.

The sampling model proposed above differs from the conventional formulation of mathematical statistics in that the population being sampled is explicitly represented by a mathematical space, namely the space $\mathcal{A}$ of population individuals. The presence of this space makes it possible to ask certain questions within the framework of the model which were only dimly conceivable under the old formulation. Specifically, the old formulation provided a mathematical representation of a population distribution such as $\mathcal{F}_\theta$ for an observable characteristic, but it did not describe how each population individual contributed to the overall distribution. In real life, however, it is legitimate to ask at least what each individual's x might be under a contemplated hypothesis $\mathcal{F}_\theta$. In other words, what mapping or mappings $a \to x$ should be regarded as permissible for a given θ within the limits specified by (P1)?

One answer to this question is to allow *any* set of mappings consistent with (P1). This is tantamount to refusing to be interested in the question. Postulate (P2) goes to the other end of the spectrum and requires that *only one* mapping shall be allowed for each given θ. An underlying motivation for this directive is the general principle that parsimony is a good thing in model-building. Of course, (P2) goes only part way to answering the question, since it does not say which mapping $a \to x$ shall be the only one allowed for a given θ. Two classes of specific answers, hence specific instances of the theory, will shortly be given. (P2) itself provides a guideline, adopted in a speculative spirit by this investigation in order to examine the statistical methodology which follows naturally from it.

Another consequence of explicitly introducing the population space $\mathcal{A}$ is the insertion of the random sampling hypothesis into the model where it naturally belongs. In the conventional formulation, a distinct law based on independent and identically distributed random variables is assumed to govern $x_1, x_2, \cdots, x_n$ for each distinct θ. In the present formulation, the collection of distinct laws is replaced by a single law μ^n which is overtly meant to describe the operation of sampling from $\mathcal{A}$. Note especially that in the new approach the $\mathcal{F}_\theta$ are not regarded as probability laws in the ordinary sense, i.e., a random variable x governed by the law $\mathcal{F}_\theta$ is nowhere postulated. The $\mathcal{F}_\theta$ play the roles not of sampling distributions but rather of deterministic laws describing the contemplated population distributions of x.

The remainder of this section describes two classes of completely specified sampling models of the proposed kind. These will be called the class of *structures of the first kind* and the class of *structures of the second kind*. The first class, which has been illustrated above, assumes $\mathcal{A}$ and $\mathcal{X}$ to be ordered. Unfortunately such an ordering of $\mathcal{X}$ restricts consideration essentially to a univariate characteristic. The second class is designed to remove this restriction so that either multivariate or univariate x may be handled. To keep the discussion simple, $\mathcal{X}$ will be assumed finite of size k where $k \geqq 2$. In other words the observable characteristic is multinomial, assuming values in one of k *categories* which constitute $\mathcal{X}$. In this multinomial context the use of a structure of the first kind presupposes that the k categories possess a natural order, while the use of a structure of the second kind poses no such restriction and treats all k categories symmetrically.

Motivation and definition will now be given for the class of structures of the first kind. When the observable characteristic is assumed to classify the population individuals into k ordered categories, it is not implausible to suppose that the population individuals possess an ordering consistent with the partial ordering induced by the mapping $a \to x$, with the same basic ordering of $\mathcal{A}$ holding whatever mapping $a \to x$ is contemplated. It is then but a short step to suppose that the population individuals are distributed over a real line and a further short step to regard this distribution as being monotonely transformable and thence transformed into a uniform distribution over the interval $(0, 1)$. Such a uniform distribution over $(0, 1)$ induces a given $\mathcal{F}_\theta$ over $\mathcal{X}$ under a mapping $a \to x$ such that a on the intervals $(0, p_1), (p_1, p_1 + p_2), \cdots, (p_1 + p_2 + \cdots + p_{k-1}, 1)$ map respectively into categories $1, 2, \cdots, k$ of $\mathcal{X}$, where p_i defines the probability of category i under $\mathcal{F}_\theta$ for $i = 1, 2, \cdots, k$. This mapping is illustrated in Fig. 1. Except for its indeterminacy at a finite set of points of $\mathcal{A}$, this is the only mapping $a \to x$ which satisfies (P1) for a given $\mathcal{F}_\theta$ and which preserves the ordering on $\mathcal{A}$ and $\mathcal{X}$. Any resolution of the indeterminacy for each $\mathcal{F}_\theta$ yields a class of contemplated mappings in

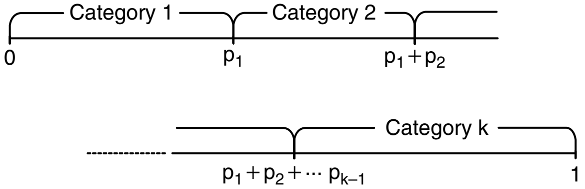


Fig. 1. The interval (0, 1) of population individuals and their corresponding multinomial categories for a given $(p_1, p_2, \cdots, p_k)$ in a structure of the first kind

the desired one-one correspondence with the class of all distributions over $\mathcal{X}$. To complete the definition of a structure of the first kind it remains only to specify a family of distributions $\mathcal{F}_\theta$, and this may be done arbitrarily.

Consider now the class of structures of the second kind. Here, the k multinomial categories are to be treated without regard to order. A natural means to this end is to increase the dimension of the proposed $\mathcal{A}$ so it may have the capability to reflect a multivariate observable characteristic. The following simple scheme is proposed: Suppose that $\mathcal{A}$ consists of the points of a $(k-1)$-dimensional simplex. Using barycentric coordinates, the general point of such a simplex may be represented by a k-tuple of real numbers $(\alpha_1, \alpha_2, \cdots, \alpha_k)$ where

$$\alpha_j \geqq 0 \text{ for } j = 1, 2, \cdots, k, \text{ and } \sum\nolimits_{j=1}^{k} \alpha_j = 1. \tag{2}$$

The vertices $I_1, I_2, \cdots, I_k$ of the simplex are represented by the k-tuples $(1, 0, \cdots, 0), (0, 1, \cdots, 0), \cdots, (0, 0, \cdots, 1)$. Suppose that μ is defined to be the uniform probability measure over the simplex $\mathcal{A}$. Specifying a mapping $a \to x$ is equivalent to specifying a partition of $\mathcal{A}$ into $\pi_1, \pi_2, \cdots, \pi_k$ where $a \varepsilon \pi_i$ maps into category $i \varepsilon \mathcal{X}$, for $i = 1, 2, \cdots, k$. The mapping $a \to x$ corresponding to a given $\mathcal{F}_\theta$ under (P1) and (P2) must have an associated partition satisfying

$$\mu(\pi_i) = p_i, \tag{3}$$

where p_i is the probability of category i under $\mathcal{F}_\theta$, for $i = 1, 2, \cdots, k$. Such a partition is defined by considering the point P in $\mathcal{A}$ with coordinates $(p_1, p_2, \cdots, p_k)$ and defining π_i for $i = 1, 2, \cdots, k$ to be the simplex with vertices P and I_j for $1 \leqq j \leqq k, j \neq i$. (Points on the common boundaries of the π_i may be arbitrarily assigned.) A set of mappings of this type, in one-one correspondence with a specified family of distributions $\mathcal{F}_\theta$, will be said to define a structure of the second kind.

The case $k = 3$ is illustrated in Fig. 2.

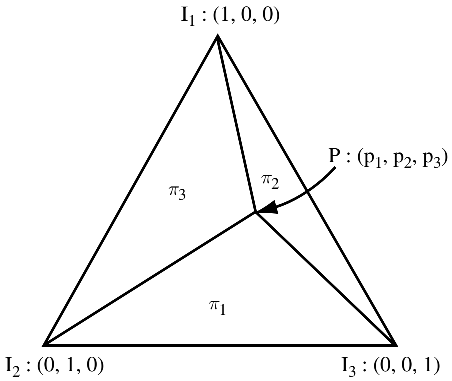


Fig. 2. The triangle of population individuals associated with a structure of the second kind when $k = 3$

There are many other structures satisfying postulates (P1) and (P2). The two special classes of structures proposed above were selected because of their mathematical simplicity. I have been unable to find any others with comparably clean properties. The idea behind the class of structures of the first kind, namely the idea of monotonely transforming the distribution of an observable into a uniform distribution on (0,1), is a familiar one in statistical theory, and some of the resulting inferences resemble those coming from confidence and fiducial arguments. The idea behind the class of structures of the second kind is unfamiliar, but, I think, not drastically different from the idea behind the first class and worth developing so that its potential may be understood.

3 Inference Methods for the Proposed Sampling Model

The first task here is to define inferences about an unknown parameter θ, given an observed sample $x_1, x_2, \cdots, x_n$, when a model of the type defined in Sect. 2 is assumed. Later in the section the discussion will be broadened to include inferences made jointly about θ and a future sample $y_1, y_2, \cdots, y_m$ from the same population.

As conceived here, the aim of inference is to assign a probability distribution to θ. Any probability deduced from such a distribution is intended for interpretation in the usual prospective way as long as θ remains unknown. For example, if the statement Pr $(\theta > 5.1) = .035$ should be made about a real parameter θ, this statement would be intended to convey the same type of information as the statement that the probability is .035 of drawing a white ball from an urn containing 35 white balls and 965 black balls.

It turns out that the reasoning developed here leads in general not to precise probability statements but to bounded probability statements about any event determined by θ and $y_1, y_2, \cdots, y_m$. For example, in place of a statement such as Pr $(\theta > 5.1) = .035$, a statement such as $.010 \leqq$ Pr $(\theta > 5.1) \leqq .063$ might be found. The aim of inference and the interpretation of probability remains as before. The difference is simply that the logical apparatus carried by the statistician is able to produce only bounds for the desired posterior probabilities.

The central idea follows. Throughout this section an infinite population is assumed, so that the sample is represented by a point drawn at random from the space $\mathcal{A}^n$ according to the measure μ^n. That is, *before the sample is drawn*, prospective probability judgments concerning which sample $a_1, a_2, \cdots, a_n$ will appear are governed by the measure μ^n over $\mathcal{A}^n$. *After the sample is drawn*, this law is generally not appropriate for prospective probability judgments because the observations $x_1, x_2, \cdots, x_n$ typically rule out many of the points of $\mathcal{A}^n$ as possible samples. It is proposed here to consider the subspace of $\mathcal{A}^n$ which does represent the range of samples still possible after $x_1, x_2, \cdots, x_n$ become known, to restrict the measure μ^n to this subspace, and to use the restricted measure for prospective probability judgments after $x_1, x_2, \cdots, x_n$ are known.

Accordingly, define R_n to be the subspace of $\mathcal{A}^n$ consisting of points $a_1, a_2, \cdots, a_n$ such that

$$a_1 \to x_1, a_2 \to x_2, \cdots, a_n \to x_n, \tag{4}$$

under some mapping $a \to x$ in the class of contemplated mappings, i.e., R_n consists of the set of samples which could have produced the observed data $x_1, x_2, \cdots, x_n$. Define the measure ν_n over R_n from

$$\nu_n(A) = \mu^n(A)/\mu^n(R_n), \tag{5}$$

for $A \subset R_n$. This is just the familiar device of conditioning by R_n. The restricted measure ν_n over R_n is regarded here as appropriate for prospective probability judgments about $a_1, a_2, \cdots, a_n$ after $x_1, x_2, \cdots, x_n$ are known.

It is assumed in (5) that $\mu^n(R_n) > 0$. This assumption is essentially met by the structures of the first and second kinds as defined in Sect. 2 when $\mathcal{X}$ is finite. For these structures, either $\mu^n(R_n) > 0$ or an observation x_i has fallen in a category of $\mathcal{X}$ assigned zero measure by all $\mathcal{F}_\theta$, and the latter possibility means that the data contradict the model with certainty. The extension of the theory to cover continuous observables is touched on in Sect. 6.

A sample $a_1, a_2, \cdots, a_n$ will be called *consistent with the data* $x_1, x_2, \cdots, x_n$ *and with* θ in Θ if (4) holds for the mapping $a \to x$ corresponding to θ. After the data are fixed, this consistency concept defines a mapping from R_n to Θ. If the mapping should be one-one, then the measure ν_n over R_n induces a measure over Θ which may be used for prospective probability judgments about the unknown θ. In general, however, this mapping from R_n to Θ is one-many, with the consequence that ν_n induces a system of upper and lower probability judgments about θ rather than a single measure.

This system of upper and lower probabilities is defined as follows. Given any event Σ determined by θ, i.e., any subset Σ of Θ belonging to an appropriate class of subsets, define $\bar{R}_n(\Sigma)$ to be the set of points of R_n which are consistent with the data for at least one θ in Θ, and define $\underline{R}_n(\Sigma)$ to be the set of points of R_n which are consistent with the data for no θ not in Σ. Thence define *the upper probability* $\bar{P}(\Sigma)$ of Σ and *the lower probability* $\underline{P}(\Sigma)$ of Σ to be

$$\bar{P}(\Sigma) = \nu_n(\bar{R}_n(\Sigma)) \text{ and } \underline{P}(\Sigma) = \nu_n(\underline{R}_n(\Sigma)). \tag{6}$$

The rationale behind the definitions (6) is that $\bar{P}(\Sigma)$ includes "as much" of the measure ν_n as can be transferred from R_n to Θ under the various one-one mappings consistent with the one-many consistency mapping from R_n to Θ prescribed above. Similarly, $\underline{P}(\Sigma)$ includes "as little" of the measure as can be transferred under the same circumstances. Thus, prospective probability judgments based on ν_n transfer naturally into a system of upper and lower probability judgments applied to events $\Sigma \subset \Theta$.

The calculus of these upper and lower probability judgments is developed more fully in a later paper (Dempster (1965)), but a few obvious properties are included here.

Since $R_n \supset \bar{R}_n(\Sigma) \supset \underline{R}_n(\Sigma)$ it follows that

$$0 \leqq \underline{P}(\Sigma) \leqq \bar{P}(\Sigma) \leqq 1. \tag{7}$$

Also it is easily checked that $\bar{R}_n(\Sigma)$ and $\underline{R}_n(\Theta - \Sigma)$ form a disjoint pair with union R_n so that

$$\bar{P}(\Sigma) = 1 - \underline{P}(\Theta - \Sigma). \tag{8}$$

Finally, since $R_n = \bar{R}_n(\Theta) = \underline{R}_n(\Theta)$, it follows that

$$\bar{P}(\Theta) = \underline{P}(\Theta) = 1. \tag{9}$$

For any real parameter ϕ determined by θ, upper and lower cumulative distribution functions may be defined as

$$\bar{H}(Z) = \bar{P}(\phi) \leqq Z), \text{ and } \underline{H}(Z) = \underline{P}(\phi \leqq Z). \tag{10}$$

Corresponding upper and lower expectations of ϕ may then be defined as

$$\bar{E}(\phi) = \int_{-\infty}^{\infty} Z \; d\underline{H}(Z) \text{ and } \underline{E}(\phi) = \int_{-\infty}^{\infty} Z \; d\bar{H}(Z). \tag{11}$$

The behavior of these operators under linear transformations is governed by

$$\begin{aligned} \bar{E}(a + b\phi) &= a + b\bar{E}(\phi), \qquad \text{if } b > 0, \\ &= a + b\underline{E}(\phi), \qquad \text{if } b < 0, \end{aligned} \tag{12}$$

where a and b are real constants. The expectations (11) are suggested as guides for betting or decision procedures whose loss functions are linear in ϕ.

Inferences about further sample observations $y_1, y_2, \cdots, y_m$ may be defined using ideas very similar to those above. The observed sample $a_1, a_2, \cdots, a_n$ and a future sample $b_1, b_2, \cdots, b_m$ are governed prior to any sampling by the law μ^{n+m} over $\mathcal{A}^{n+m}$.

The observations are $x_1, x_2, \cdots, x_n$ as before, but the unknowns are now $\theta, y_1, y_2, \cdots, y_m$ in the space $\Theta \times \mathcal{X}^m$. The space $R_{n,m}$ of samples possible after observation consists of those $a_1, a_2, \cdots, a_n, b_1, b_2, \cdots, b_m$ satisfying

$$\begin{aligned} &a_1 \rightarrow x_1, a_2 \rightarrow x_2, \cdots, a_n \rightarrow x_n, \\ &\qquad b_1 \rightarrow y_1, b_2 \rightarrow y_2, \cdots, b_m \rightarrow y_m \end{aligned} \tag{13}$$

for some mapping $a \rightarrow x$ in the class of contemplated mappings and for some $\theta, y_1, y_2, \cdots, y_m$. The initial measure μ^{n+m} over $\mathcal{A}^{n+m}$ leads to a measure $\nu_{n,m}$ appropriate for postsample judgments. Given any event Σ^* determined by $\theta, y_1, y_2, \cdots, y_m$ the subsets $\bar{R}_{n,m}(\Sigma^*)$ and $\underline{R}_{n,m}(\Sigma^*)$ of $R_{n,m}$ are defined analogously to $\bar{R}_n(\Sigma)$ and $\underline{R}_n(\Sigma)$ above, i.e., $\bar{R}_{n,m}(\Sigma^*)$ is the set of points in

$R_{n,m}$ which could have given rise to $x_1, x_2, \cdots, x_n$ for some $\theta, y_1, y_2, \cdots, y_m$ in Σ^* and $\underline{R}_{n,m}(\Sigma^*)$ is the set of points in $R_{n,m}$ which could have given rise to $x_1, x_2, \cdots, x_n$ for no $\theta, y_1, y_2, \cdots, y_m$ not in Σ^*. Thence

$$\bar{P}(\Sigma^*) = \nu_{n,m}(\bar{R}_{n,m}(\Sigma^*)) \text{ and } \underline{\mathrm{P}}(\Sigma^*) = \nu_{n,m}(\underline{\mathrm{R}}_{n,m}(\Sigma^*)). \tag{14}$$

Any event determined by θ alone has upper and lower probabilities derivable by (14) or by (6). It is clear, however, that the two sets of inferences concur, as would be desired.

The following two sections are intended to illustrate the foregoing definitions in a pair of non-trivial situations. Section 4 deals with finite $\mathcal{X}$ of size $k = 2$ (binomial sampling), the family $\mathcal{F}_\theta$ consisting of all possible distributions over the two categories of $\mathcal{X}$. A structure of the first kind is assumed, but this is also trivially a structure of the second kind when $k = 2$. The illustration of Sect. 5 assumes a structure of the second kind with general k but finite Θ.

4 Binomial Sampling

Illustrative inferences are worked out here for the structure of the first kind defined by setting $k = 2$ and allowing $\mathcal{F}_\theta$ to range over all distributions on the two categories of $\mathcal{X}$. As is usually done with binomial sampling, the parameter p on $0 \leqq p \leqq 1$ will be used for the distributions over $\mathcal{X}$, where p denotes the probability of category 1 and $1 - p$ denotes the probability of category 2. The population individuals are supposed uniformly distributed on the interval (0, 1) under this structure of the first kind. (The corresponding structure of the second kind would differ only in the nonessential way that the population individuals would be uniformly distributed over the line segment (one-dimensional simplex) joining the points with coordinates (0, 1) and (1,0).) The mapping $a \rightarrow x$ corresponding to a given p is ambiguous at $a = p$. This ambiguity does not affect the resulting inference, but for definiteness $a = p$ will be assumed to map into category 1.

The population individuals $a_1, a_2, \cdots, a_n, b_1, b_2, \cdots b_m$ representing the observed sample of size n and a future sample of size m are supposed drawn at random according to a uniform distribution over $\mathcal{A}^{n+m}$ which is here a unit cube in $n + m$ dimensions. The sample data $x_1, x_2, \cdots, x_n$ marks each individual of the observed sample as belonging to category 1 or category 2. The observation vector $x_1, x_2, \cdots, x_n$ will be replaced here by the single quantity T defined to be the total number of sample observations in category 1. To assume that only T is observed, rather than the actual configuration $x_1, x_2, \cdots, x_n$, has no effect on the resulting inferences because the spaces R_n and $R_{n,m}$ corresponding to each of the $\binom{n}{T}$ configurations with given T are disjoint and isomorphic. Consequently, the only effect on (5) is to multiply both numerator and denominator of the right side by $\binom{n}{T}$. It is a theorem, not proved here, that the inferences based on multinomial data, represented by either a structure of the first kind or a structure of the second kind, are not affected if the individual sample observations $x_1, x_2, \cdots, x_n$ are thrown away and only $T_1, T_2, \cdots, T_k$ retained, where T_i denotes the number of sample observations in category i.

Upper and lower probabilities will be computed for the events

$$\Sigma = \{\alpha \leqq p \leqq \beta\} \tag{15}$$

and

$$\Sigma^* = \{r \leqq S \leqq t\}, \tag{16}$$

where S is the number of category 1 observations in a future sample of size m. These upper and lower probabilities depend of course on the observed T.

Consider first (15). A point $a_1, a_2, \cdots, a_n$ in $\mathcal{A}^n$ is consistent with the observed T and the parameter value p if and only if

$$a_{(T)} \leqq p < a_{(T+1)} \tag{17}$$

where $a_{(1)} \leqq a_{(2)} \leqq \cdots \leqq a_{(n)}$ denote the ordered random variables $a_1, a_2, \cdots, a_n$ and where $a_{(0)} = 0$ and $a_{(n+1)} = 1$. It follows that $\bar{R}_n(\Sigma)$ is the subset of $\mathcal{A}^n$ such that the intersection of the intervals $[a_{(T)}, a_{(T+1)})$ and $[\alpha, \beta]$ is nonempty. R_n is the special case of $\bar{R}_n(\Sigma)$ when $\alpha = 0$ and $\beta = 1$ so that $R_n = \mathcal{A}^n$.

Since the measures $\dot{\nu}_n$ and μ^n coincide here, the definition (6), reduces to

$$\bar{P}(\Sigma) = \mu^n(\bar{R}_n(\Sigma)). \tag{18}$$

To calculate this it is convenient to write

$$\bar{R}_n(\Sigma) = \left\{\alpha < a_{(T)} \leqq \beta\right\} \cup \left\{a_{(T)} \leqq \alpha < a_{(T+1)}\right\} \tag{19}$$

which is a union of disjoint sets. Thus

$$\begin{aligned}\bar{P}(\Sigma) &= T\binom{n}{T}\int_\alpha^\beta p^{T-1}(1-p)^{n-T}dp \\ &\quad + \binom{n}{T}\alpha^T(1-\alpha)^{n-T}, \qquad \text{if } 1 \leqq T \leqq n, \\ &= (1-\alpha)^n, \qquad\qquad\qquad\qquad \text{if } T = 0.\end{aligned} \tag{20}$$

An alternative to (19) is

$$\bar{R}_n(\Sigma) = \left\{\alpha < a_{(T+1)} \leqq \beta\right\} \cup \left\{a_{(T)} \leqq \beta < a_{(T+1)}\right\} \tag{21}$$

which leads to an alternative to (20), namely

$$\begin{aligned}\bar{P}(\Sigma) &= (n-T)\binom{n}{T}\int_{\alpha}^{\beta} p^{T}(1-p)^{n-T-1}dp \\ &\quad+\binom{n}{T}\beta^{T}(1-\beta)^{n-T}, \qquad \text{if } 0 \leqq T \leqq n-1, \\ &= \beta^{n}, \qquad\qquad\qquad\qquad\quad \text{if } T = n.\end{aligned} \tag{22}$$

Another alternative may be found by replacing the beta integrals in (20) or (22) by binomial sums, yielding

$$\bar{P}(\Sigma) = \sum\nolimits_{i=0}^{T}\binom{n}{i}\alpha^{i}(1-\alpha)^{n-i} + \sum\nolimits_{i=T}^{n}\binom{n}{i}\beta^{i}\,(1-\beta)^{n-i} - 1. \tag{23}$$

A similar variety of forms is possible for $\underline{P}(\Sigma)$. $\underline{R}_{n}(\Sigma)$ is the event that the interval $[a_{(T)}, a_{(T+1)})$ is contained in the interval $[\alpha, \beta]$. Writing

$$\underline{R}_{n}(\Sigma) = \left\{\alpha \leqq a_{(T)} \leqq \beta\right\} - \left\{\alpha \leqq a_{(T)} \leqq \beta, a_{(T+1)} > \beta\right\} \tag{24}$$

and

$$\begin{aligned}&\{\alpha \leqq a_{(T)} \leqq \beta, a_{(T+1)} > \beta\} \\ &\qquad\qquad = \left\{a_{(T)} \leqq \beta < a_{(T+1)}\right\} - \left\{a_{(T)} < \alpha, a_{(T+1)} > \beta\right\},\end{aligned} \tag{25}$$

it is seen that

$$\underline{P}(\Sigma) = T\binom{n}{T}\int_{\alpha}^{\beta} p^{T-1}(1-p)^{n-T}dp - \binom{n}{T}[\beta^{T} - \alpha^{T}](1-\beta)^{n-T}, \tag{26}$$

at least if $1 \leqq T \leqq n-1$. Special interpretations are needed if $T = 0$ and $T = n$, and these may be handled by checking directly that

$$\begin{aligned}\underline{P}(\Sigma) &= 0, && \text{if } T = 0,\ \alpha > 0, \\ &= 1-(1-\beta)^{n}, && \text{if } T = 0,\ \alpha = 0, \\ &= 0, && \text{if } T = n,\ \beta < 1, \\ &= 1-\alpha^{n}, && \text{if } T = n,\ \beta = 1.\end{aligned} \tag{27}$$

Just as (20) may be replaced by (22) and (23), (26) may be replaced by

$$\begin{aligned}\underline{P}(\Sigma) &= (n-T)\binom{n}{T}\int_{\alpha}^{\beta} p^{T}(1-p)^{n-T-1}dp \\ &\qquad\qquad - \binom{n}{T}\alpha^{T}[(1-\alpha)^{n-T} - (1-\beta)^{n-T}],\end{aligned} \tag{28}$$

or, replacing the integrals by sums,

$$\begin{aligned}\bar{P}(\Sigma) &= \sum\nolimits_{i=0}^{T-1}\binom{n}{i}\alpha^{i}(1-\alpha)^{n-i} \\ &\qquad + \binom{n}{T}\alpha^{T}(1-\beta)^{n-T} + \sum\nolimits_{i=T+1}^{n}\binom{n}{T}\beta^{i}(1-\beta)^{n-i} - 1.\end{aligned} \tag{29}$$

Note that

$$\bar{P}(\Sigma) - \underline{P}(\Sigma) = \binom{n}{T} [\alpha^T(1-\alpha)^{n-T} + \beta^T(1-\beta)^{n-T} - \alpha^T(1-\beta)^{n-T}]. \quad (30)$$

Also, in the special case $\alpha = \beta$,

$$\bar{P}(p = \alpha) = \binom{n}{T} \alpha^T(1-\alpha)^{n-T} \text{ and } \underline{P}(p = \alpha) = 0. \quad (31)$$

Several general features of the above inferences are worthy of remark. The small upper probability (31) assigned to any particular value $p = \alpha$ is proportional to the conventional likelihood at $p = \alpha$. This likelihood is the probability content of the region $\bar{R}_n(p = \alpha)$ in $\mathcal{A}^n$, and such regions sweep out the region R_n as α ranges over $0 \leqq \alpha \leqq 1$. If the regions $\bar{R}_n(p = \alpha)$ had not overlapped for different α, then all upper and lower probabilities would have coincided and both would have been derivable from a posterior density proportional to likelihood. It will next be shown that the overlapping decreases as n increases in the sense that the upper and lower probabilities tend towards agreement with a distribution whose density is proportional to likelihood.

Large sample behavior may be studied by supposing that $n \to \infty$ and $T \to \infty$ in such a way that $T/n \to \rho$. By considering the limiting normal behavior of binomial distributions, it becomes clear that

$$\bar{P}(\Sigma) - \underline{P}(\Sigma) = O(1/n^{\frac{1}{2}}) \quad (32)$$

uniformly in α and β, and consequently that either $\bar{P}(\Sigma)$ or $\underline{P}(\Sigma)$ may be approximated by

$$\bar{P}(\Sigma) \sim \underline{P}(\Sigma) \sim \Phi(\beta^*) - \Phi(\alpha^*) \quad (33)$$

where Φ denotes the cdf of the $N(0,\, 1)$ distribution,

$$\beta^* = (\beta - T/n)/[n(T/n)(1 - T/n)]^{\frac{1}{2}}, \quad (34)$$

and

$$\alpha^* = (\alpha - T/n)/[n(T/n)(1 - T/n)]^{\frac{1}{2}}. \quad (35)$$

(The symbols $\sim$ in (33) mean that the ratios tend to unity as $n \to \infty$.) The normal approximation (33) extends to show that, if the arguments α^* and β^* tend to constants as $n \to \infty$, the posterior inferences may be computed from a normal density function whose ratio to the likelihood tends to a constant. The limiting posterior inference considered here is also that reached by a Bayesian argument with any well-behaved prior density and the same limiting conditions. That is, in a circumstance where the Bayesian would say that the choice of a prior distribution does not matter, the present theory yields the same answer.

Consider next how to find $\bar{P}(\Sigma^*)$ and $\underline{P}(\Sigma^*)$ where Σ^* was defined in (16). This requires consideration of the sample space $\mathcal{A}^{n+m}$ from which the pair of samples is drawn. For Σ^* to hold it is necessary and sufficient that

$$b_{(r)} \leqq p < b_{(t+1)} \tag{36}$$

where $b_{(1)} \leqq b_{(2)} \leqq \cdots \leqq b_{(m)}$ denote the ordered random variables $b_1, b_2, \cdots, b_m$ with the additional conventions that $b_{(0)} = 0$ and $b_{(m+1)} = 1$. On the other hand, (17) must hold if p is to be consistent with the observation T. Thus $\bar{R}_{n,m}(\Sigma^*)$ is the subset of $\mathcal{A}^{n+m}$ such that the intervals $[b_{(r)}, b_{(t+1)})$ and $[a_{(T)}, a_{(T+1)})$ are not disjoint. Writing

$$\bar{R}_{n,m}(\Sigma^*) = \left\{b_{(r)} \leqq a_{(T)} < b_{(t+1)}\right\} \cup \left\{a_{(T)} < b_{(r)} < a_{(T+1)}\right\}, \tag{37}$$

it is seen that finding $\bar{P}(\Sigma^*)$ is reducible to a combinatorial problem concerning the $\binom{n+m}{n}$ equally likely relative orderings of the samples $a_1, a_2, \cdots, a_n$ and $b_1, b_2, \cdots, b_m$.

For example, the event $\{a_{(T)} \leqq b_{(r)} < a_{(T+1)}\}$ may be expressed as the event that $b_{(r)}$ has rank $r+T$ in the combined samples. Under this event, the first $r+T-1$ members of the combined sample consist of T of the a_i and $r-1$ of the b_j, and the last $m+n-r-T$ members of the combined sample consist of $n-T$ of the a_i and $m-r$ of the b_j. Thus

$$\mu^{n+m}(a_{(T)} \leqq b_{(r)} < a_{(T+1)}) = \binom{r+T-1}{T}\binom{m+n-r-T}{n-T} \Big/ \binom{m+n}{n}. \tag{38}$$

This and subsequent formulas apply generally when $0 \leqq r \leqq t \leqq m$ *and* $0 \leqq T \leqq n$ *provided that* $\binom{x}{y}$ *is regarded as zero when* $x < y$.

By reasoning similar to that producing (38), one finds from (37) that

$$\begin{aligned}\bar{P}(\Sigma^*) = &\sum\nolimits_{i=r}^{t} \binom{i+T-1}{i}\binom{m+n-i-T}{m-i} \Big/ \binom{m+n}{n} \\ &+ \binom{r+T-1}{T}\binom{m+n-r-T}{n-T} \Big/ \binom{m+n}{n}.\end{aligned} \tag{39}$$

Similarly $\underline{R}_{n,m}(\Sigma^*)$ may be expressed as the event that the interval $[a_{(T)}, a_{(T+1)})$ is contained in the interval $[b_{(r)}, b_{(t)})$, so that $\underline{R}_{n,m}(\Sigma^*)$ may be written

$$\left\{b_{(r)} \leqq a_{(T)} < b_{(t+1)}\right\} - \left\{b_{(r)} \leqq a_{(T)} < b_{(t+1)} < a_{(T+1)}\right\} \tag{40}$$

while the second event on the right side of (40) may be written

$$\left\{a_{(T)} < b_{(t+1)} < a_{(T+1)}\right\} - \left\{a_{(T)} < b_{(r)}, b_{(t+1)} < a_{(T+1)}\right\}. \tag{41}$$

From (40) and (41) one has

$$\begin{aligned}\underline{\mathrm{P}}(\Sigma^*) = \sum\nolimits_{i=r}^{t} & \binom{i+T-1}{i}\binom{m+n-i-T}{m-i} \Big/ \binom{m+n}{n} \\ & - \left[\binom{t+T}{T} + \binom{r+T-1}{T}\right]\binom{m+n-T-t-1}{n-T} \Big/ \binom{m+n}{n}.\end{aligned} \tag{42}$$

From (39) and (42) it follows that

$$\begin{aligned}\bar{\mathrm{P}}(\Sigma^*) - \underline{\mathrm{P}}(\Sigma^*) = & \left[\binom{t+T-1}{T}\binom{m+n-r-T}{n-T}\right. \\ & + \binom{t+T}{T}\binom{m+n-T-t-1}{n-T} - \binom{r+T-1}{T} \\ & \left.\binom{m+n-T-t-1}{n-T}\right] \Big/ \binom{m+n}{n}.\end{aligned} \tag{43}$$

There is an obvious analogy between the set of formulas (15), (19), (20), (24), (25), (26), (30) and (16), (37), (39), (40), (41), (42), (43), respectively. This analogy has an important statistical consequence. As $m \to \infty, r \to \infty$ and $t \to \infty$ in such a way that $r/m \to \alpha$ and $t/m \to \beta$, one might conjecture that $\bar{P}(\Sigma^*) \to \bar{P}(\Sigma)$ and $\underline{P}(\Sigma^*) \to \underline{P}(\Sigma)$, i.e., that inferences about p should be the same as inferences about the proportion of category 1 observations in a subsequent infinite sample. The validity of these limiting properties is evident from the fact that $b_{(r)}$ and $b_{(t+1)}$ converge in probability to α and β together with the fact that the events governing $\bar{P}(\Sigma)$ and $\underline{P}(\Sigma)$ depend on the interval (α, β) in precisely the same way that the events governing $\bar{P}(\Sigma^*)$ and $\underline{P}(\Sigma^*)$ depend on the interval $(b_{(r)}, b_{(t+1)})$. Thus $\bar{P}(\Sigma^*)$ and $\underline{P}(\Sigma^*)$ actually cover $\bar{P}(\Sigma)$ and $\underline{P}(\Sigma)$ as limiting cases.

Finally, to present a simple result, suppose that $\bar{P}_1$ and $\underline{P}_1$ denote upper and lower probabilities that the next sample individual will be observed in category 1 given that T of the first n sample individuals were observed in category 1. From (39) and (40) with $m = 1$ and $r = t = 1$,

$$\bar{P}_1 = (T+1)/(n+1) \text{ and } \underline{\mathrm{P}}_1 = T/(n+1). \tag{44}$$

5 Structures of the Second Kind with Finite Θ

Let $\mathcal{X}$ be a set of k observable categories. Let

$$\Theta = \{1, 2, \cdots, q\}. \tag{45}$$

index a set of q specified distributions over $\mathcal{X}$, say $\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_q$. Let Σ be any subset of Θ. The aim here is to develop formulas for $\bar{P}(\Sigma)$ and $\underline{P}(\Sigma)$

based on sample observations $x_1, x_2, \cdots, x_n$ where the sampling model is a structure of the second kind as defined in Sect. 2.

According to these definitions, $\mathcal{A}$ is represented by a $(k-1)$-dimensional simplex with vertices $I_1, I_2, \cdots, I_k$ and μ is the uniform probability measure over $\mathcal{A}$. Each distribution $\mathcal{F}_i$ determines a point

$$P_i = (p_{i1}, p_{i2}, \cdots, p_{ik}) \tag{46}$$

of $\mathcal{A}$ where, for $i = 1, 2, \cdots, q$ and $j = 1, 2, \cdots, k$, the probability of category j under $\mathcal{F}_i$ is denoted by p_{ij}. Each P_i determines a partition of $\mathcal{A}$ into simplexes $\pi_{i1}, \pi_{i2}, \cdots, \pi_{ik}$ where π_{ij} denotes the simplex with the same vertices as $\mathcal{A}$ except that I_j is replaced by P_i. The mapping $a \rightarrow x$ corresponding to $\theta = i$ is the mapping which sends $a\ \varepsilon \pi_{ij}$ into category j (with some rule to make the mapping specific on the boundaries of the π_{ij}). In accordance with the postulate (P1)

$$\mu(\pi_{ij}) = p_{ij}, \tag{47}$$

for $i = 1, 2, \cdots, q$ and $j = 1, 2, \cdots, k$ (c.f., (3)).

Consider first inferences based on a sample of size $n = 1$ when the sample observation x_1 falls in category j of $\mathcal{X}$. The regions $R_1, \bar{R}_1(\Sigma)$ and $\underline{R}_1(\Sigma)$ whose measures determine $\bar{P}(\Sigma)$ and $\underline{P}(\Sigma)$ are given by

$$R_1 = \cup_{i \varepsilon \Theta} \pi_{ij}, \tag{48}$$

$$\bar{R}_1(\Sigma) = \cup_{i \varepsilon \Sigma} \pi_{ij}, \tag{49}$$

and

$$\underline{\mathrm{R}}_1(\Sigma) = R_1 - \bar{R}_1(\Theta - \Sigma). \tag{50}$$

It turns out to be simpler to characterize intersections of the π_{ij} for given j rather than unions. The intersections are also important for understanding the passage from $n = 1$ to general n. The approach therefore will be to express the probabilities of the unions (48) and (49) in terms of the probabilities of intersections.

A simplex with the same vertices as $\mathcal{A}$ except that the vertex I_j of $\mathcal{A}$ is replaced by a general point of $\mathcal{A}$ will be called for short a *simplex of type j*. The vertex which replaces I_j will be called the *free vertex*. For convenience, the simplex of type j with free vertex P will be denoted by $\pi_j(P)$. For example, π_{ij} above may also be denoted by $\pi_j(P_i)$.

Using obvious vector space operations of addition and multiplication by a scalar, a general point Q of the simplex $\pi_j(P)$ may be characterized as

$$Q = r_j P + \sum\nolimits_{l=1, l \neq j}^{k} r_l I_l \tag{51}$$

where $r_l \geqq 0$ for $l = 1, 2, \cdots, k$ and $\sum_1^k r_l = 1$. The following two lemmas will be deduced from (51).

Lemma 1. *If Q lies in $\pi_j(P)$ then $\pi_j(Q) \subset \pi_j(P)$, and conversely.*

Lemma 2. *Suppose that* $P = \sum_1^k p_l I_l$ *and* $Q = \sum_1^k q_l I_l$ *where* $p_l \geqq 0$ *and* $q_l \geqq 0$ *for* $l = 1, 2, \cdots, k$ *and* $\sum_1^k p_l = \sum_1^k q_l = 1$. *Then* Q *lies in* $\pi_j(P)$ *if and only if*

$$q_l / q_j \geqq p_l / p_j \tag{52}$$

for $l = 1, 2, \cdots, k$.

The converse part of Lemma 1 is immediate and the direct part requires only a simple application of the definitions of $\pi_j(Q)$ and $\pi_j(P)$, and so is omitted.

To prove Lemma 2, note that the comparison of (51) with $Q = \sum_1^k q_l I_l$ yields

$$q_j = r_j p_j, \text{ and } q_l = r_l + r_j p_l \text{ for } l \neq j. \tag{53}$$

If Q lies in $\pi_j(P)$ then (53) holds with $r_l \geqq 0$ and $r_j = q_j / p_j$, so that (52) follows. Conversely, starting from (52) and defining $r_1, r_2, \cdots, r_k$ from (53) it follows that $r_l \geqq 0$ for $l = 1, 2, \cdots, k$ and

$$\begin{aligned} \sum\nolimits_{l=1}^k r_l &= r_j + \sum\nolimits_{l=1, l \neq j}^k (q_l - r_j p_l) \\ &= r_j + \sum\nolimits_{l=1}^k (q_l - r_j p_l) \\ &= r_j + 1 - r_j \cdot 1 \\ &= 1, \end{aligned} \tag{54}$$

as required.

The basic result about intersections, which is stated in Theorem 1, asserts that the intersection of a finite set of simplexes of type j is again a simplex of type j.

Theorem 1. *Suppose that* P_i *is defined by* (46) *for* i *in a subset* Σ *of the integers* $1, 2, \cdots, q$. *Suppose that* $Q = \sum_i^k q_l I_l$ *is defined by*

$$q_l = \max\nolimits_{i \varepsilon \Sigma} \{p_{il}/p_{ij}\} / \sum\nolimits_{u=1}^k \max\nolimits_{i \varepsilon \Sigma} \{p_{iu}/p_{ij}\} \tag{55}$$

for $l = 1, 2, \cdots, k$. *Then*

$$\pi_j(Q) = \bigcap i \varepsilon \Sigma \pi_j(P_i). \tag{56}$$

To prove Theorem 1, consider finding a point Q lying in the desired intersection and having maximum coordinate q_j. From (52) it follows that

$$q_l / q_j \geqq p_{il} / p_{ij} \tag{57}$$

for i in Σ and hence that

$$q_l / q_j \geqq \max\nolimits_{i \varepsilon \Sigma} \{p_{il}/p_{ij}\} \tag{58}$$

for $l = 1, 2, \cdots, k$. Summing and using $\sum_1^k q_l = 1$ gives

$$q_j \leqq [\sum_{l=1}^{k} \max_{i\varepsilon\Sigma} \{p_{il}/p_{ij}\}]^{-1}. \tag{59}$$

Moreover, if (59) is changed to an equality, it is easily seen that Q defined by (55) is the only point consistent with (58) and $\sum_1^k q_l = 1$, i.e., Q is the unique point in the desired intersection with maximum coordinate q_j. This explains where (55) came from.

That

$$\pi_j(Q) \subset \bigcap_{i\varepsilon\Sigma} \pi_j(P_i) \tag{60}$$

follows from Lemma 1. That

$$\pi_j(Q) \supset \bigcap_{i\varepsilon\Sigma} \pi_j(P_i) \tag{61}$$

follows by applying Lemma 2 to a general point in the intersection and showing that it satisfies the requirement of the converse application of Lemma 2 to $\pi_j(Q)$. Thus Theorem 1 is proved.

Returning now to *inference* from a sample of size $n = 1$ with x_1 in category j, and reverting to the notation π_{ij} in place $\pi_j(P_i)$, the important consequence of Theorem 1 is that

$$\mu(\bigcap_{i\varepsilon\Sigma} \pi_{ij}) = [\sum_{u=1}^{k} \max_{i\varepsilon\Sigma} \{p_{iu}/p_{ij}\}]^{-1}. \tag{62}$$

This follows from (3), which shows that $\mu(\pi_j(Q)) = q_j$, and from (55) with $l = j$. Note that the numerator of (55) is unity when $l = j$.

Since $\bar{R}_1(\Sigma)$ is a union of simplexes of type j as in (49), $\mu(\bar{R}_1(\Sigma))$ may be expressed in terms of the quantities defined by (62) applied to all subsets of Σ. Specifically, suppose that Σ_{1a} for $a = 1, 2, \cdots$ denote the single element subsets of Σ, that Σ_{2b} for $b = 1, 2, \cdots$ denote the two-element subsets of Σ, that Σ_{3c} for $c = 1, 2, \cdots$ denote the three-element subsets of Σ, and so on. Then

$$\mu(\bar{R}_1(\Sigma)) = \sum_a \mu(\bigcap_{i\varepsilon\Sigma_{1a}} \pi_{ij}) - \sum_b \mu(\bigcap_{i\varepsilon\Sigma_{2b}} \pi_{ij}) + \sum_c \mu(\bigcap_{i\varepsilon\Sigma_{3c}} \pi_{ij}) - \cdots . \tag{63}$$

Formula (63) may also be applied when Σ is replaced successively by Θ and by $\Theta - \Sigma$ to determine $\mu(R_1)$ and $\mu(\bar{R}_1(\Theta - \Sigma)) = \mu(R_1) - \mu(\underline{R}_1(\Sigma))$. These along with $\mu(\bar{R}_1(\Sigma))$ determine $\bar{P}(\Sigma)$ and $\underline{P}(\Sigma)$.

Consideration of the simplest case $q = 2$ may help to illuminate the foregoing. Here Θ consists of two elements and there are two non-trivial subsets namely Σ_1 consisting of $i = 1$ and Σ_2 consisting of $i = 2$. Thus only three numbers are required to determine all upper and lower probabilities, namely

$$\begin{aligned} \mu(\bar{R}_1(\Sigma_1)) &= p_{1j}; \\ \mu(\bar{R}_1(\Sigma_2)) &= p_{2j}; \\ \mu(\bar{R}(\Sigma_1) \cap \bar{R}(\Sigma_2)) &= [\sum\nolimits_{u=1}^{k} \max\{p_{1u}/p_{1j}, p_{2u}/p_{2j}\}]^{-1}. \end{aligned} \tag{64}$$

Denoting $\mu(\bar{R}(\Sigma_1) \cap \bar{R}(\Sigma_2))$ by p_{12j} for short, it follows that

$$\mu(R_1) = p_{1j} + p_{2j} - p_{12j} \tag{65}$$

and hence that

$$\begin{aligned} &\bar{P}(\Sigma_1) = p_{1j}/(p_{1j} + p_{2j} - p_{12j}) \text{ and} \\ &\underline{P}(\Sigma_1) = (p_{1j} - p_{12j})/(p_{1j} + p_{2j} - p_{12j}). \end{aligned} \tag{66}$$

For samples of general size n, consideration must be directed to regions in the product space $\mathcal{A}^n$. If the observations $x_1, x_2, \cdots, x_n$ fall in categories $c_1, c_2, \cdots, c_n$, respectively, and if $\Sigma_i \subset \Theta$ is the subset consisting of i only, then

$$\bar{R}_n(\Sigma_i) = \pi_{ic_1} \times \pi_{ic_2} \times \cdots \times \pi_{ic_n} \tag{67}$$

for $i = 1, 2, \cdots, q$. For general Σ, unions of regions like (67) are needed. As already mentioned it is easier to first find intersections. In fact, for general Σ,

$$\bigcap\nolimits_{i\epsilon\Sigma} \bar{R}_n(\Sigma_i) = (\bigcap\nolimits_{i\epsilon\Sigma} \pi_{i\epsilon_1}) \times (\bigcap\nolimits_{i\epsilon\Sigma} \pi_{i\epsilon_2}) \times \cdots \times (\bigcap\nolimits_{ie\Sigma} \pi_{i\epsilon_n}) \tag{68}$$

and thence

$$\mu^n(\bigcap\nolimits_{i\varepsilon\Sigma} \bar{R}_n(\Sigma_i)) = \Pi_{m=1}^{n} \mu(\bigcap\nolimits_{i\varepsilon\Sigma\ \pi_{ic_m}}). \tag{69}$$

Each term in the product on the right side of (69) is of the form (62) for different j. Formula (63) must be generalized by replacing the terms on the right side by products of n terms as in (69). Then the computation of upper and lower probabilities proceeds as before.

The task of determining inferences for a sample of size n may therefore be summarized as follows. For the mth sample individual with observation x_m in category c_m, compute the vector of $2^q - 1$ quantities (62) with $j = c_m$ and Σ ranging over the $2^q - 1$ non-empty subsets of $\{1, 2, \cdots, q\}$. Having such a vector for each sample individual, combine these n vectors into a single vector by multiplying the corresponding elements as indicated by (69). From this sample vector compute $\mu(\bar{R}_n(\Sigma))$ for any Σ as in the generalization of (63) and thence determine upper and lower probabilities as required.

Again the case $q = 2$ is especially simple because the vector of $2^q - 1$ quantities required for each individual reduces to three quantities as in (64). Thus for each sample individual s there is a triple $(p_{1j(s)}, p_{2j(s)}, p_{12j(s)})$ for $s = 1, 2, \cdots, n$ where $j(s)$ denotes the observational category into which individual s falls. The inferences (66) are modified by replacing $(p_{1j}, p_{2j}, p_{12j})$ with

$$(\Pi_{s=1}^{n} p_{1j}(s), \Pi_{s=1}^{n} p_{2j(s)}, \Pi_{s=1}^{n} p_{12j(s)}). \tag{70}$$

6 Concluding Remarks

The following discussion of qualitative aspects of the proposed inference methods may help the reader to evaluate these methods.

Unlike the fiducial argument which Fisher limited to continuous observables only, the present methods have been developed above in detail only for finite $\mathcal{X}$. Interestingly enough, the extension to continuous observables poses greater difficulty in the case of structures of the first kind than in the case of structures of the second kind.

Pick up again the $N(\theta, 1)$ example of Sect. 2 which illustrates a structure of the first kind extended in the obvious way to cover real x. If the procedures of Sect. 3 are applied to the $N(\theta, 1)$ example, it is found for $n = 1$ that $R_1 = \mathcal{A}$ and that ν_1 is the $N(0, 1)$ distribution over $\mathcal{A}$ which, from (1), induces a $N(x_1, 1)$ distribution for θ. In other words, Fisher's fiducial argument is reproduced in this simple case. For general sample sizes, R_n becomes the line in n-space consisting of samples $a_1, a_2, \cdots, a_n$ satisfying $x_i = a_i + \theta$ for $i = 1, 2, \cdots, n$. Unfortunately $\mu^n(R_n)$ is now zero and (5) cannot be used. This breakdown is not fatal because the observable x may be approximated by a multinomial observable specifying which of a large set of k mutually exclusive and exhaustive intervals contains x. In this way R_n is approximated by a cylinder with $\mu^n(R_n) > 0$. As $k \to \infty$ in an appropriate way, the cross-section of the cylinder shrinks to the vanishing point and the posterior distribution induced on θ approaches the $N(\bar{x}, 1/n)$ distribution. This answer is the same as that given by the fiducial argument, but the reasoning is quite different: the present method conditions by R_n while the fiducial argument uses sufficiency to reduce consideration to $\bar{x}$.

It thus appears that multinomial approximation may be used to extend the reasoning of Sect. 3 to continuous observables. At this point a snag arises in connection with structures of the first kind but not, remarkably enough, in connection with structures of the second kind. The snag is that different multinomial approximations may lead in the limit to different inferences. This does not happen for location parameter situations, such as the $N(\theta, 1)$ example, but a little analysis shows that it does happen for general families $\mathcal{F}_\theta$ with sampling represented by a structure of the first kind. On the other hand, for structures of the second kind, the fundamental quantity (62) does approach a common limit under a wide range of approximating conditions, i.e.,

$$[\sum_{u=1}^{k} \max_{i \varepsilon \Sigma} \{p_{iu}/p_{ij}\}]^{-1} \to [\int \max_{i \varepsilon \Sigma} \{f_i(x)/f_i(x_1)\} \;\; dx]^{-1} \qquad (71)$$

where $(p_{i1}, p_{i2}, \cdots, p_{ik})$ approximates a continuous distribution with density $f_i(x)$. This remarkable property means that the structures of the second kind extend in an unambiguous way to yield inferences for general univariate or multivariate observables. For example, inferences about all the parameters of a multivariate normal distribution from a sample of any size are uniquely defined using a structure of the second kind.

This uniqueness property together with the ability to handle multivariate observables make the inferences based on the structures of the second kind appear very attractive to the author. These inferences have the property that upper and lower probabilities differ even with continuous observables, which is also plausible in small samples.

The new methods pay for the absence of a prior distribution by being able to specify only upper and lower posterior probabilities. If two hypotheses Σ_1 and Σ_2 are "close" in the sense that $\bar{R}_n(\Sigma_1)$ and $\bar{R}_n(\Sigma_2)$ overlap considerably, then it becomes difficult to decide between such hypotheses because both $\bar{P}(\Sigma_1)$ and $\bar{P}(\Sigma_2)$ are close to $\bar{P}(\Sigma_1 \cup \Sigma_2)$ and there is no unambiguous division of posterior probability between them. Consider an extreme case where $\Sigma_1 = \{\theta_1\}, \Sigma_2 = \{\theta_2\}$ and $\mathcal{F}_{\theta_1} \equiv \mathcal{F}_{\theta_2}$. Here the hypotheses might fairly be judged indistinguishable, and the present methods react by finding $\bar{P}(\Sigma_1 \cup \Sigma_2) = \bar{P}(\Sigma_1) = \bar{P}(\Sigma_2)$ and $\underline{P}(\Sigma_1) = \underline{P}(\Sigma_2) = 0$. (The Bayesian would, of course, distinguish between such Σ_1 and Σ_2 on the basis of his prior distribution alone.) As illustrated in Sect. 4, it typically happens that the overlapping of $\bar{R}_n(\Sigma_1)$ and $\bar{R}_n(\Sigma_2)$ becomes less serious as n increases, i.e., large samples have high resolving power.

As in other theories of inference, the concept of likelihood plays a prominent role, but the interpretation of likelihood is radically changed. Here, the standard likelihood function $L(\theta)$ is proportional to $\mu^n(\bar{R}_n(\{\theta\}))$ or to $\nu_n(\bar{R}_n(\{\theta\})) = \bar{P}(\{\theta\})$. While $L(\theta)$ for each θ is the measure of a set, the sets corresponding to different θ overlap in an important way which is not defined by the function $L(\theta)$ itself. Thus, all the relevant information is not contained in $L(\theta)$. In large samples, however, "nearly" all the relevant information resides in $L(\theta)$, as illustrated in Sect. 4.

Noting that the sampling model specifies a measure μ over $\mathcal{A}$ in addition to a family of distributions $\mathcal{F}_\theta$, a reader might jump to the conclusion that μ is playing a role analogous to the prior distribution adopted by a Bayesian. Such an analogy would be specious. The measure μ simply idealizes the assertion that all samples are equally likely. As such it belongs to the category of assumption which is usually regarded as objective, in contrast to the Bayesian prior distribution which is often frankly subjective. It is not the assumption of μ which gives the present methods their distinctiveness, but rather the postulate (P2), or, more precisely, the classes of structures of the first and second kind which translate (P2) into precise models.

References

Dempster, A. P. (1963). On direct probabilities. *J. Roy. Statist. Soc. Ser. B* **20** 102–107.

Dempster, A. P. (1964). On the difficulties inherent in Fisher's fiducial argument. *J. Amer. Statist. Assoc.* **59** 56–66.

Dempster, A. P. (1965). On a class of mathematical structures yielding upper and lower probabilities. Unpublished research report.

Todhunter, I. (1865). *A History of the Mathematical Theory of Probability.* Reprinted (1949) by Chelsea, New York.

3

Upper and Lower Probabilities Induced by a Multivalued Mapping[*]

Arthur P. Dempster

Abstract. A multivalued mapping from a space X to a space S carries a probability measure defined over subsets of X into a system of upper and lower probabilities over subsets of S. Some basic properties of such systems are explored in Sects. 1 and 2. Other approaches to upper and lower probabilities are possible and some of these are related to the present approach in Sect. 3. A distinctive feature of the present approach is a rule for conditioning, or more generally, a rule for combining sources of information, as discussed in Sects. 4 and 5. Finally, the context in statistical inference from which the present theory arose is sketched briefly in Sect. 6.

1 Introduction

Consider a pair of spaces X and S together with a multivalued mapping Γ which assigns a subset $\Gamma x \subset S$ to every $x \,\varepsilon\, X$. Suppose that μ is a probability measure which assigns probabilities to the members of a class $\mathcal{F}$ of subsets of X. If μ is acceptable for probability judgments about an uncertain outcome $x \,\varepsilon\, X$, and if this uncertain outcome x is known to correspond to an uncertain outcome $s \,\varepsilon\, \Gamma x$, what probability judgments may be made about the uncertain outcome $s \,\varepsilon\, S$? The answer to this question would be a familiar one if Γ were single-valued, for under wide conditions a single-valued Γ would carry the measure μ over subsets of X into a unique probability measure over subsets of S. For multivalued Γ, however, one is led to consider upper and lower probabilities defined as follows over subsets of S.

For any $T \subset S$ define

$$T^* = \{x \,\varepsilon\, X, \Gamma x \cap T \neq \varnothing\} \tag{1}$$

and

[*] This research was supported by the Office of Naval Research. Reproduction in whole or in part is permitted by the U. S. Government. Distribution of the document is unlimited.

$$T_* = \{\mathrm{x} \,\varepsilon\, X, \Gamma x \neq \varnothing, \Gamma x \subset T\} . \tag{2}$$

In particular, $S^* = S_*$ is the domain of Γ. Define $\mathcal{E}$ to be the class of subsets T of S such that T^* and T_* belong to $\mathcal{F}$. Suppose that $\mathrm{S} \,\varepsilon\, \mathcal{E}$. Finally, define the *upper probability* of $\mathrm{T}\varepsilon\mathcal{E}$ to be

$$P^*(T) = \mu(T^*)/\mu(S^*) \tag{3}$$

and the *lower probability* of $T\varepsilon\mathcal{E}$ to be

$$P_*(T) = \mu(T_*)/\mu(S^*). \tag{4}$$

$P^*(T)$ and $P_*(T)$ are defined only if $\mu(S^*) \neq 0$.

Since T^* consists of those $x \,\varepsilon\, X$ which can possibly correspond under Γ to an $s \,\varepsilon\, T$, one may naturally regard $\mu(T^*)$ to be the largest possible amount of probability from the measure μ which can be transferred to outcomes $s \,\varepsilon\, T$. Similarly T_* consists of those $x \,\varepsilon\, X$ which must lead to an $s \,\varepsilon\, T$, so that $\mu(T_*)$ represents the minimal amount of probability which can be transferred to outcomes $s \,\varepsilon\, T$. The denominator $\mu(S^*)$ in (3) and (4) is a renormalizing factor necessitated by the fact that the model permits, in general, outcomes in X which do not map into a meaningful subset of S. The offending subset $\{x \,\varepsilon\, X, \Gamma x = \varnothing\}$ must be removed from X and the measure of the remaining set S^* renormalized to unity. It would have been possible to restrict the formulation so that $\mu(S^*) = 1$, but it will be convenient in Sects. 4 and 5 to have the general model.

The case of finite $S = \{s_1, s_2, \cdots, s_m\}$ will now be developed somewhat further. Suppose that $S_{\delta_1 \delta_2 \cdots \delta_m}$ denotes the subset of S which contains s_i if $\delta_i = 1$ and excludes s_i if $\delta_i = 0$, for $i = 1, 2, \cdots, m$. The 2^m subsets of S so defined are the possible Γx, and they determine a partition of X into

$$X = \bigcup\nolimits_{\delta_1 \delta_2 \cdots \delta_m} X_{\delta_1 \delta_2 \cdots \delta_m} \tag{5}$$

where

$$X_{\delta_1 \delta_2 \cdots \delta_m} = \{x \,\varepsilon\, X, \Gamma x = S_{\delta_1 \delta_2 \cdots \delta_m}\} . \tag{6}$$

For any $T \subset S$, the subsets T^* and T_* are unions of subsets of the form $X_{\delta_1 \delta_2 \cdots \delta_m}$ and hence $P^*(T)$ and $P_*(T)$ are uniquely determined by the 2^m quantities

$$p_{\delta_1 \delta_2 \cdots \delta_m} = \mu\left(X_{\delta_1 \delta_2 \cdots \delta_m}\right) . \tag{7}$$

It is assumed, of course, that each $X_{\delta_1 \delta_2 \cdots \delta_m}$ is in $\mathcal{F}$. Note that *any* set of 2^m non-negative numbers $p_{\delta_1 \delta_2 \cdots \delta_m}$ with sum unity determines a possible set of upper and lower probabilities for all $T \subset S = \{s_1, s_2, \cdots, s_m\}$.

Table 1 displays formulas for all possible upper and lower probabilities when $m = 3$. For example, if $T = S_{110} = \{s_1, s_2\}$, then $T^* = X_{100} \cup X_{010} \cup X_{110} \cup X_{101} \cup X_{011} \cup X_{111}$ and $T_* = X_{100} \cup X_{010} \cup X_{110}$, and therefore

$$\mu(T^*) = p_{100} + p_{010} + p_{110} + p_{101} + p_{011} + p_{111} \tag{8}$$

Table 1. Upper and lower probabilities when $S = \{s_1, s_2, s_3\}$

T	$P^*(T)$	$P_*(T)$
$\varnothing$	0	0
$\{s_1\}$	$(p_{100} + p_{110} + p_{101} + p_{111})/(1 - p_{000})$	$p_{100}/(1 - p_{000})$
$\{s_2\}$	$(p_{010} + p_{110} + p_{011} + p_{111})/(1 - p_{000})$	$p_{010}/(1 - p_{000})$
$\{s_3\}$	$(p_{001} + p_{101} + p_{011} + p_{111})/(1 - p_{000})$	$p_{001}/(1 - p_{000})$
$\{s_1, s_2\}$	$(p_{100} + p_{010} + p_{110} + p_{101} + p_{011} + p_{111})/(1 - p_{000})$	$(p_{100} + p_{010} + p_{110})/(1 - p_{000})$
$\{s_1, s_3\}$	$(p_{100} + p_{001} + p_{110} + p_{101} + p_{011} + p_{111})/(1 - p_{000})$	$(p_{100} + p_{001} + p_{101})/(1 - p_{000})$
$\{s_2, s_3\}$	$(p_{010} + p_{001} + p_{110} + p_{101} + p_{011} + p_{111})/(1 - p_{000})$	$(p_{010} + p_{001} + p_{011})/(1 - p_{000})$
S	1	1

and

$$\mu(T_*) = p_{100} + p_{010} + p_{110}. \tag{9}$$

These need only be divided by

$$\mu(S^*) = 1 - p_{000} \tag{10}$$

to become upper and lower probabilities as defined in (3) and (4). Similar arguments yield the rest of Table 1.

This section closes with several more definitions. The term *variate* will be used for a real-valued function defined over S. Subject to measurability requirements, any variate V has an *upper distribution function* $F^*(v)$ and a *lower distribution function* $F_*(v)$ defined by

$$\begin{aligned} F^*(v) &= P^*(V \leqq v), \\ F^*(v) &= P_*(V \leqq v), \end{aligned} \tag{11}$$

for $-\infty < v < \infty$. The corresponding upper and lower expected values $E^*(V)$ and $E_*(V)$ are defined by

$$\begin{aligned} E^*(V) &= \int_{-\infty}^{\infty} v \; dF_*(v) \\ E_*(V) &= \int_{-\infty}^{\infty} v \; dF^*(v). \end{aligned} \tag{12}$$

(The interchange of upper and lower stars is necessary here in order to have both $F_*(v) \leqq F^*(v)$ and $E_*(V) \leqq E^*(V)$.)

The concepts of upper expected value and lower expected value generalize the concepts of upper probability and lower probability, respectively. For, if the variate Z is defined to be the indicator function of $T \subset S$, i.e., if

$$\begin{aligned} Z(s) &= 1 \qquad \text{for } s \varepsilon T, \\ &= 0 \qquad \text{otherwise,} \end{aligned} \tag{13}$$

then it follows from (12) that

$$\begin{aligned} E^*(Z) &= P^*(T) \\ E_*(Z) &= P_*(T). \end{aligned} \tag{14}$$

2 The Class of Compatible Measures Over S

Given a system of upper and lower probabilities for the subsets $\mathcal{E}$ of S determined as above from $(X, \mathcal{F}, \mu)$ and Γ, it is natural to ask for the class $\mathcal{C}$ of probability measures P such that

$$P_*(T) \leqq P(T) \leqq P^*(T) \tag{15}$$

for all $T \, \varepsilon \, \mathcal{E}$. Clearly $\mathcal{C}$ is the same as the class of probability measures P such that

$$E_*(V) \leqq E(V) \leqq E^*(V) \tag{16}$$

for all variates V for which $E_*(V)$ and $E^*(V)$ are defined and finite, and where $E(\cdots)$ refers to expectation with respect to P. The class $\mathcal{C}_1$ will be called the class of measures *compatible* with the given system of upper and lower probabilities.

It is convenient to begin with a constructive definition of a class $\mathcal{C}_1$ of measures P and to prove ultimately that $\mathcal{C} = \mathcal{C}_1$. A general member of the class $\mathcal{C}_1$ is defined by specifying a probability measure $\gamma_{\Gamma x}$ over each possible $\Gamma x \subset S$ and taking $P(T) = \int \gamma_{\Gamma x}(T \cap \Gamma x) \ d_\mu(x)$. To avoid topological complexities only the case of finite S will be considered in detail. Consider, therefore, the following method of constructing a probability measure P over the finite sample space $S = \{s_1, s_2, \cdots, s_m\}$ given the 2^m quantities $p_{\delta_1 \delta_2 \cdots \delta_m}$ defined in (7).

Suppose that each $p_{\delta_1 \delta_2 \cdots \delta_m}$ other than $p_{00 \cdots 0}$ is partitioned into a sum of m non-negative pieces

$$p_{\delta_1 \delta_2 \cdots \delta_m} = \sum\nolimits_{i=1}^{m} p^{(i)}_{\delta_1 \delta_2 \ldots \delta_m} \tag{17}$$

where $p^{(i)}_{\delta_1 \delta_2 \cdots \delta_m} = 0$ unless $\delta_i = 1$. Define the measure P from

$$P\{s_i\} = \sum\nolimits_{\delta_1 \delta_2 \cdots \delta_m} p^{(i)}_{\delta_1 \delta_2 \cdots \delta_m} / (1 - p_{00 \ldots 0}) \tag{18}$$

for $i = 1, 2, \cdots, m$. The motivation behind this definition of P is that in the logic of the situation $p_{\delta_1 \delta_2 \cdots \delta_m}$ is a piece of probability that may attach to any s_i for which $\delta_i = 1$. The partition (17) specifies the subpieces to be attached to each eligible s_i and (18) collects the appropriate subpieces from all $p_{\delta_1 \delta_2 \cdots \delta_m}$.

For example, when $m = 3$, one needs the decompositions

$$\begin{aligned} p_{110} &= p^{(1)}_{110} + p^{(2)}_{110}; \\ p_{101} &= p^{(1)}_{101} + p^{(3)}_{101}; \\ p_{011} &= p^{(2)}_{011} + p^{(3)}_{011}; \\ p_{111} &= p^{(1)}_{111} + p^{(2)}_{111} + p^{(3)}_{111}; \end{aligned} \tag{19}$$

and the corresponding measure P is defined from

$$\begin{aligned} P_{\{s_1\}} &= \left(p_{100} + p^{(1)}_{110} + p^{(1)}_{101} + p^{(1)}_{111}/(1 - p_{000})\right), \\ P_{\{s_2\}} &= \left(p_{010} + p^{(2)}_{110} + p^{(2)}_{011} + p^{(2)}_{111}/(1 - p_{000})\right), \\ P_{\{s_3\}} &= \left(p_{001} + p^{(3)}_{101} + p^{(3)}_{011} + p^{(3)}_{111}/(1 - p_{000})\right). \end{aligned} \tag{20}$$

The class of all measures P determined by such partition schemes will be denoted by $\mathcal{C}_1$. These measures are compatible in the sense of (15); indeed,

$$\begin{aligned} P_*(T) &= \min_{P \varepsilon \mathcal{C}_1} P(T), \\ P^*(T) &= \max_{P \varepsilon \mathcal{C}_1} P(T) \end{aligned} \tag{21}$$

for each $T \subset S$. More generally,

$$\begin{aligned} E_*(V) &= \min_{P \varepsilon \mathcal{C}_1} E(V), \\ E^*(V) &= \max_{P \varepsilon \mathcal{C}_1} E(V) \end{aligned} \tag{22}$$

for any variate V.

Before proving (22), it is convenient to introduce a finite subclass of $\mathcal{C}_1$ with several important properties, including the property that the extremes in (21) and (22) are all attained within this finite subclass. Suppose that $\pi(1), \pi(2), \cdots, \pi(m)$ is a permutation of $1, 2, \cdots, m$. The partition (17) may be determined in such a way that $p_{\delta_1 \delta_2 \cdots \delta_m} = p^{(i)}_{\delta_1 \delta_2 \cdots \delta_m}$ for that i which appears first in the permutation $\pi(1), \pi(2), \cdots, \pi(m)$ subject, of course, to the restriction $\delta_i = 1$. Determining this partition for each $\delta_1, \delta_2, \cdots, \delta_m$ determines a specific member of $\mathcal{C}_1$ associated with the permutation $\pi(1), \pi(2), \cdots,$ $\pi(m)$. The $m!$ members of $\mathcal{C}_1$ which are determined in this way are not necessarily distinct. They will be called the *extremal* members of $\mathcal{C}_1$ for reasons to become evident.

Given any variate V there is at least one permutation $\pi(1), \pi(2), \cdots, \pi(m)$ such that

$$V(s_{\pi(1)}) \leqq V(s_{\pi(2)}) \leqq \cdots \leqq V(s_{\pi(m)}). \tag{23}$$

It will now be shown that $\min_{P \varepsilon \mathcal{C}_1} \ E_*(V)$ is achieved when P is the extremal measure associated with any such $\pi(1), \pi(2), \cdots, \pi(m)$. Note first that for any measure P and any permutation satisfying (23)

$$E(V) = V(s_{\pi(1)}) + \sum\nolimits_{j=2}^{m} [V(s_{\pi(j)}) - V(s_{\pi(j-1)})] \cdot P\{s_{\pi(j)}, s_{\pi(j+1)}, \cdots, s_{\pi(m)}\}. \quad (24)$$

Second, it is claimed that the $(m-1)$ terms in the sum on the right side of (24) are simultaneously minimized by choosing P to be the extremal measure associated with any permutation satisfying (23). Indeed, $P\{s_{\pi(j)}, s_{\pi(j+1)}, \cdots, s_{\pi(m)}\}$ is minimized by requiring that the partition (17) concentrate as much as possible on $p^{(i)}_{\delta_1\delta_2\cdots\delta_m}$ with $i = \pi(1), \pi(2), \cdots, \pi(j-1)$. The partition defining the extremal measure corresponding to $\pi(1), \pi(2), \cdots, \pi(m)$ is clearly one means of assuring such a concentration. Furthermore, the definition of lower probability implies that this minimum of $P\{s_{\pi(j)}, s_{\pi(j+1)}, \cdots, s_{\pi(m)}\}$ is $P_*\{s_{\pi(j)}, s_{\pi(j+1)}, \cdots, s_{\pi(m)}\}$. The first half of (22) is thus proved; the other half follows similarly using the reverse permutation $\pi(m), \pi(m-1), \cdots, \pi(1)$.

Defining $\mathcal{C}_2$ to be the class of measures P formed by taking mixtures of the extremal measures, it is clear from their definitions that each of the classes $\mathcal{C}, \mathcal{C}_1$, and $\mathcal{C}_2$ are closed under the operation of mixing. It is also clear from the relations proved above that $\mathcal{C}_2 \subset \mathcal{C}_1 \subset \mathcal{C}$. This section concludes by showing that $\mathcal{C} = \mathcal{C}_1 = \mathcal{C}_2$, i.e., that these three possible definitions of compatibility are equivalent.

Any measure P determines a point

$$P = (p^{(1)}, p^{(2)}, \cdots, p^{(m)}) \quad (25)$$

in the $(m-1)$-dimensional simplex with the m vertices $(1, 0, \cdots, 0)$, $(0, 1, \cdots, 0), \cdots, (0, 0, \cdots, 1)$ where

$$p^{(i)} = P\{s_i\} \quad (26)$$

for $i = 1, 2, \cdots, m$. Any class of measures P defines a subset of the simplex and a class closed under mixing defines a convex subset. Thus $\mathcal{C}, \mathcal{C}_1$, and $\mathcal{C}_2$ may be identified with convex subsets of the simplex. Any convex set in $(m-1)$-dimensional space is uniquely determined by the pairs of planes of support determined by all families of parallel planes of dimension $m-2$. To show that $\mathcal{C} = \mathcal{C}_2$ one need only check that they have all the same planes of support.

In the present formulation, the intersection of the simplex with any plane of dimension $m-2$ consists of all those measures P for which a variate V has the same expectation. For example, the plane of points $\mathbf{P}$ such that

$$a_1 p^{(1)} + a_2 p^{(2)} + \cdots + a_m p^{(m)} = c \quad (27)$$

contains all measures P such that $E(V) = c$ where V is defined by

$$V(s_i) = a_i \quad (28)$$

for $i = 1, 2, \cdots, m$. Of course, V is unique only up to a linear transformation of the form $a + bV$ and the family of planes parallel to (27) shares the family of variates $a + bV$. It follows that the planes of support of a closed convex subset of the simplex in the family of planes parallel to (27) are those which maximize and minimize $E(V)$ over choices of P in the closed convex subset. From (16) and (22), and because the extrema in (22) occur in $\mathcal{C}_2$, it follows that the closed convex subsets $\mathcal{C}$ and $\mathcal{C}_2$ have the same pairs of planes of support, as was required to prove.

From all this, it is seen that the class of compatible measures is a closed convex polygon in the simplex, having at most $m!$ vertices, namely, the extremal measures P. There may be as few as m distinct vertices; for example, the class of compatible measures may be the whole simplex in the "informationless" model where $p_{11\cdots 1} = 1$ and all other $p_{\delta_1\delta_2\cdots\delta_m} = 0$.

3 Other Approaches

The approach to upper and lower probabilities introduced above may be placed in a clearer perspective by considering a hierarchy of approaches, suggested to the author by L.J. Savage. Again consider for simplicity the case of finite S.

Any class $\mathcal{C}$ of probability measures P over the subsets $T \subset S$ defines upper and lower probabilities

$$\begin{aligned} P^*(T) &= \sup\nolimits_{P\varepsilon\mathcal{C}} P(T); \\ P_*(T) &= \inf\nolimits_{P\varepsilon\mathcal{C}} P(T). \end{aligned} \tag{29}$$

Since the same upper and lower probabilities are yielded by the convex closure of $\mathcal{C}$ as by $\mathcal{C}$ itself, one might as well restrict $\mathcal{C}$ to be a closed convex set of measures.

Define Ω to be the class of all closed convex subsets of the simplex, i.e., all sets of probability measures over the subsets of S which are closed under mixing. Define $\Omega_1 \subset \Omega$ to consist of those closed convex sets of measures defined solely by inequalities on probabilities of events. Finally, define Ω_2 to consist of sets of compatible measures as defined in Sect. 2, where the definition (15) assures that $\Omega_2 \subset \Omega_1$. It is clear, see for example Fig. 1, that Ω_1 is properly contained in Ω. It will next be shown that Ω_2 is properly contained in Ω_1.

For any member of Ω_2, define

$$p'_{\delta_1\delta_2\cdots\delta_m} = p_{\delta_1\delta_2\cdots\delta_m}/(1 - p_{00\cdots 0}) \tag{30}$$

if at least one $\delta_i = 1$, and define $p'_{00\cdots 0} = 0$. The set of $p'_{\delta_1\delta_2\cdots\delta_m}$ determine the same $\mathcal{C}$ as do the original $p_{\delta_1\delta_2\cdots\delta_m}$ with the simplification that the normalizing factor $1 - p_{00\cdots 0}$ may be ignored. Thus, for example when $m = 3$, all lower probabilities (and hence upper probabilities from (34)) may be formed from

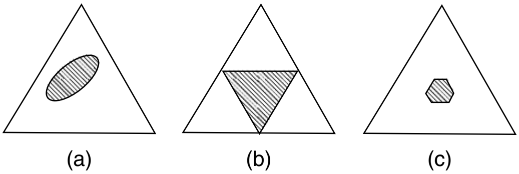


Fig. 1. Three types of convex subsets of the triangle: case (a) a general convex subset, case (b) a subset in Ω_1 but not in Ω_2 as described in the text, and case (c) a subset in Ω_2 with $p_{100} = p_{010} = p_{001} = \frac{1}{4}$ and $p_{110} = p_{101} = p_{011} = p_{111} = \frac{1}{16}$

$$\begin{aligned} P_* \{s_1\} &= p'_{100}; \\ P_* \{s_2\} &= p'_{010}; \\ P_* \{s_3\} &= p'_{001}; \\ P_* \{s_1,\ s_2\} &= p'_{100} + p'_{010} + p'_{110}; \\ P_* \{s_1,\ s_3\} &= p'_{100} + p'_{001} + p'_{101}; \\ P_* \{s_2,\ s_3\} &= p'_{010} + p'_{001} + p'_{011}. \end{aligned} \tag{31}$$

These relations may be solved to yield

$$\begin{aligned} p'_{100} &= P_* \{s_1\}, && \text{and similarly for } p'_{010} \text{ and } p'_{001}; \\ p'_{110} &= P_* \{s_1, s_2\} - P_* \{s_1\} - P_* \{s_2\} && \text{and similarly for } p'_{101} \text{ and } p'_{011}; \\ p'_{111} &= 1 (= P_* \{s_1,\ s_2,\ s_3\}) - P_* \{s_1,\ s_3\} - P_* \{s_1, s_3\} - P_* \{s_2, s_3\} \\ &\quad + P_* \{s_1\} + P_* \{s_2\} + P_* \{s_3\}. \end{aligned} \tag{32}$$

The obvious extension of (32) to general m is easily proved by induction, and is omitted here.

The relations (32) may be applied to any member of Ω_1 using on the right side the bounding planes of support for that member. The result is a set of $p'_{\delta_1 \delta_2 \cdots \delta_m}$ which may be used as in (31) to determine the bounds of probabilities and hence give back the member of Ω_1. It also follows from (32) that the $p'_{\delta_1 \delta_2 \cdots \delta_m}$ sum to unity, but a difference between Ω_1 and Ω_2 arises because the $p'_{\delta_1 \delta_2 \cdots \delta_m}$ need not all be non-negative in Ω_1. A simple example of the latter when $m = 3$ is pictured in case (b) of Fig. 1. For this example, $P_*\{s_1\} = P_*\{s_2\} = P_*\{s_3\} = 0$ while $P_*\{s_1, s_2\} = P_*\{s_1, s_3\} = P_*\{s_2, s_3\} = \frac{1}{2}$ and (32) yields $p'_{100} = p'_{010} = p'_{001} = 0$, $p'_{110} = p'_{101} = p'_{011} = \frac{1}{2}$ and $p'_{111} = -\frac{1}{2}$. Thus there are closed convex subsets in Ω_1 which are not in Ω_2, where Ω_2 is the class of primary interest in this paper.

Many of the basic relationships of ordinary probability theory have analogues for systems of upper and lower probabilities. For example, in Ω one has

$$P_*(\varnothing) = P^*(\varnothing) = 0, \qquad P_*(S) = P^*(S) = 1; \tag{33}$$

if the complement of T is denoted by $\bar{T}$, then

$$P_*(T) + P^*(\bar{T}) = 1; \tag{34}$$

if T and R are mutually exclusive, then

$$\begin{aligned} P_*(T) + P_*(R) \leqq P_*(T \cup R) &\leqq P_*(T) + P^*(R) \\ \leqq P^*(T \cup R) &\leqq P^*(T) + P^*(R); \end{aligned} \tag{35}$$

if $E_*(V) = \inf_{P\varepsilon\mathcal{C}} E(V)$ and $E^*(V) = \sup_{P\varepsilon\mathcal{C}} E(V)$ are used to define upper and lower expectations, then

$$E_*(V) = -E^*(-V) \tag{36}$$

or more generally

$$\begin{aligned} E_*(a + bV) &= a + bE_*(V) \text{ if } b \geqq 0 \\ &= a + bE^*(V) \text{ if } b \geqq 0 \end{aligned} \tag{37}$$

together with a similar formula for $E^*(a + bV)$; for any pair of variates V and W,

$$\begin{aligned} E_*(V) + E_*(W) \leqq E_*(V + W) &\leqq E_*(V) + E^*(W) \\ &\leqq E^*(V + W) \leqq E^*(V) + E^*(W). \end{aligned} \tag{38}$$

Note that (38) and (36) generalize (35) and (34), respectively. To prove (38), for example, note that there exists a measure in $\mathcal{C}$ such that $E_*(V + W) = E(V + W) = E(V) + E(W) \geqq E_*(V) + E_*(W)$. The remaining parts of (38) follow from the first part together with (36).

It is interesting to note that the definitions (12) and (22) do not coincide in Ω as they do in Ω_2, i.e., for general convex sets it can happen that

$$\textstyle\int_{-\infty}^{\infty} v \; dF^*(v) < \inf_{P\varepsilon\mathcal{C}} E(V). \tag{39}$$

This comes about because there is in general no measure P which simultaneously minimizes each of the terms in (24).

Another relation which holds in Ω_2 but not in general in Ω is

$$\begin{aligned} P_*(T) + P_*(R) \leqq P_*(T \cup R) + P_*(T \cap R) &\leqq P_*(T) + P^*(R) \\ \leqq P^*(T \cup R) + P^*(T \cap R) &\leqq P^*(T) + P^*(R) \end{aligned} \tag{40}$$

for any $T, R \subset S$. Simple counterexamples may be found in Ω even for $m = 3$. To prove (40) in Ω_2 define $T_1 = T \cap R, T_2 = T - T_1, T_3 = R - T_1$, and $T_4 = S - (T_1 \cup T_2 \cup T_3)$. Then, analogous to (32) define

$$\begin{aligned} t_{1000} &= P_*(T_1), \text{ etc.}, \\ t_{1100} &= P_*(T_1 \cup T_2) - P_*(T_1) - P_*(T_2), \text{ etc.}, \\ t_{1110} &= P_*(T_1 \cup T_2 \cup T_3) - P_* \, (T_1 \cup T_2) - P_*(T_1 \cup T_3) - P_*(T_2 \cup T_3) \\ &\quad + P_*(T_1) + P_*(T_2) + P_*(T_3). \text{ etc.}, \\ t_{1111} &= 1 - P_*(T_1 \cup T_2 \cup T_3) - \cdots + P_*(T_1 \cup T_2) + \cdots - P_*(T_1) \\ &\quad - \cdots - P_*(T_4). \end{aligned} \tag{41}$$

By a simple argument of inclusion and exclusion, these $2^4 - 1$ quantities are non-negative and sum to unity. Like (32), the relations (41) may be solved to yield lower probabilities and thence upper probabilities in terms of $t_{\delta_1\delta_2\delta_3\delta_4}$ for every event determined by T_1, T_2, T_3, and T_4 or equivalently by T and R. The relations (40) follow simply by replacing each quantity with its expression in terms of the $t_{\delta_1\delta_2\delta_3\delta_4}$ and using the fact that each $t_{\delta_1\delta_2\delta_3\delta_4} \geqq 0$.

I do not know whether (39) can happen or whether (40) can fail in Ω_1. The literature on upper and lower probabilities is to my knowledge quite small. Good (1962) has presented an axiomatic approach which he believes simplifies but does not necessarily agree with an earlier axiomatic approach of Koopman (1940a), (1940b). I have not attempted to produce a compact set of probability axioms sufficient to characterize Ω, Ω_1 or Ω_2, as the case may be. Nor does Good appear to discuss models with the mathematical concreteness of the families Ω, Ω_1 or Ω_2. Smith (1961), (1965) has also discussed upper and lower probabilities, largely in the context of upper and lower betting odds. Since upper and lower odds for any bet are equivalent to a pair of planes of support for a convex set $\mathcal{C}$ of measures P, it appears that Smith is considering the family Ω. Fishburn (1964) considers upper and lower probabilities and their corresponding expectations apparently in the framework Ω_1. There appears to be no hint of the family Ω_2 in any of the work referred to.

From the viewpoint of a reader to whom probabilities are essentially determinants of bets or rational decisions, it may seem undesirable to restrict the class of convex subsets $\mathcal{C}$ to Ω_1 or even less to Ω_2, since any member of Ω would seem to be a defensible position for a rational consistent man. On the other hand, when upper and lower probabilities can be traced back to a single measure μ, a more stringent kind of logic can be introduced in the area of conditioning. This concept of conditioning and its generalization to the concept of combining independent sources of information are the crux of this paper and, I believe, the most attractive feature of restriction to Ω_2.

4 Upper and Lower Conditional Probabilities

Given a system of upper and lower probabilities defined over subsets $T \subset S$ by $(X, \mathcal{F}, \mu)$ and Γ, what are the appropriate upper and lower conditional probabilities of T given R, i.e., probabilities appropriate when $S - R$ is ruled impossible? The obvious answer is to use the same $(X, \mathcal{F}, \mu)$ and Γ except restricting Γ to subsets of R, or more precisely, using the multivalued mapping Γ' from X to R defined by

$$\Gamma' x = \Gamma x \cap R. \tag{42}$$

The upper and lower conditional probabilities defined by Γ' may be expressed simply in terms of the unconditional upper and lower probabilities defined by Γ, i.e.,

$$\begin{aligned}&P^*(T|R) = P^*(T \cap R)/P^*(R);\\&P_*(T|R) = 1 - P^*(\bar{T}|R) = 1 - P^*(\bar{T} \cap R)/P^*(R).\end{aligned} \tag{43}$$

The first line of (43) is an application of (3) with Γ' in place of Γ, and the second line of (43) follows from (34) and the first line of (43). Note that upper and lower conditional probabilities given R are undefined unless $P^*(R) > 0$, i.e., unless the range of Γ' includes more than $\varnothing$.

The following lemma is a consequence of the above definitions.

Lemma 1. *If T_1 and T_2 are mutually exclusive subsets of R, then*

$$\begin{aligned}P_*(T_1)/P^*(T_2) \leqq P_*(T_1|R)/P^*(T_2|R) &\leqq P^*(T_1|R)/P_*(T_2|R)\\ &\leqq P^*(T_1)/P_*(T_2).\end{aligned} \tag{44}$$

Only the first inequality need be proved, since the second is obvious and the third follows from the first. To prove the first write

$$\begin{aligned}P_*(T_1|R)/P^*(T_2|R) &= (1 - P^*(R - T_1|R))/P^*(T_2|R)\\ &= (P^*(R) - P^*(R - T_1))/P^*(T_2)\\ &\geqq P_*(T_1)/P^*(T_2),\end{aligned} \tag{45}$$

where the inequality between the last two numerators follows from (35). Relations (44) assert that the elimination of possibilities extraneous to a given bet serves to tighten the upper and lower betting odds appropriate to that bet. Note that these upper and lower betting odds do not in general come together, even when $R = T_1 \cup T_2$.

The definition of upper and lower conditional probabilities given above relies for its motivation on the structure of Ω_2. In Ω or Ω_1 one could use (43) but it would no longer appear natural; instead, one might regard

$$P^{**}(T|R) = \sup_{P\varepsilon\mathcal{C}} P(T|R), \quad P_{**}(T|R) = \inf_{\mathrm{P}\varepsilon\mathcal{C}} P(T|R) \tag{46}$$

as the natural definitions of upper and lower conditional probabilities. The relationship between (43) and (46) as alternatives in Ω_2 may be clarified as follows:

Define $T_1 = T \cap R, T_2 = R - T$, and $T_3 = S - R$. Analogous to (41), define

$$\begin{aligned}t_{100} &= P_*(T_1), \text{etc.},\\ t_{110} &= P_*(T_1 \cup T_2) - P_*(T_1) - P_*(T_2), \text{ etc.},\\ t_{111} &= 1 - P_*(T_1 \cup T_2) - P_*(T_1 \cup T_3) - P_*(T_2 \cup T_3) + P_*(T_1)\\ &\quad + P_*(T_2) + P_*(T_3),\end{aligned} \tag{47}$$

which in Ω_2 are seven non-negative quantities summing to unity. From (43) and (47) it follows that

$$P^*(T|R) = (t_{100} + t_{110} + t_{101} + t_{111})/$$
$$(t_{100} + t_{010} + t_{110} + t_{101} + t_{011} + t_{111}),$$
$$P_*(T|R) = t_{100}/(t_{100} + t_{010} + t_{110} + t_{101} + t_{011} + t_{111}). \tag{48}$$

On the other hand, the maximum and minimum of $P(T|R) = P(T_1)/(P(T_1)+P(T_2))$ are found by distributing the pieces (47) appropriately among T_1, T_2 and T_3 where t_{100} must go to T_1, while t_{110} may go to T_1 or T_2, and so on. Thus

$$P^{**}(T|R) = (t_{100} + t_{110} + t_{101} + t_{111})/(t_{100} + t_{110} + t_{101} + t_{111} + t_{010}),$$
$$P_{**}(T|R) = t_{100}/(t_{100} + t_{010} + t_{110} + t_{011} + t_{111}). \tag{49}$$

From (48) and (49)

$$P^{**}(T|R) \geqq P^*(T|R) \geqq P_*(T|R) \geqq P_{**}(T|R). \tag{50}$$

Thus the additional structure used in the definitions (43) serves to pull the upper and lower probabilities inward relative to the less structured definitions (46).

In Sect. 5 the definitions (43) will be seen as a very special case of a method of assimilating new information into a system of upper and lower probabilities.

5 Combination of Independent Sources of Information

A probability measure may be regarded as defining degrees of belief which quantify a state of partial knowledge. Any such measure arises in some way from a limited range of human experience which will be called a source of information. A mechanism for combining such sources of information is a virtual necessity for a theory of probability oriented to statistical inference. The mechanism adopted here assumes *independence* of the sources, a concept whose real world meaning is not so easily described as its mathematical definition. Opinions of different people based on overlapping experiences could not be regarded as independent sources. Different measurements by different observers on different equipment would often be regarded as independent, but so would different measurements by one observer on one piece of equipment: here the question concerns independence of errors. In the application referred to in Sect. 6, the independent sources are taken to be non-overlapping random samples from a population, together with prior information which may be regarded as a distillation of previous samples or experiences.

The sources considered here are mathematically defined by their basic probability spaces $(X_i, \mathcal{F}_i, \mu_i)$ and multivalued mappings Γ_i, where i indexes the source. The space S into which Γ_i maps is the same for each i, i.e., the different sources are giving information about the same uncertain outcome in S. If the n sources $i = 1, 2, \cdots, n$ are assumed independent, then the combined source $(X, \mathcal{F}, \mu)$ and Γ is defined from

$$\begin{aligned} X &= X_1 \times X_2 \times \cdots \times X_n, \\ \mathcal{F} &= \mathcal{F}_1 \times \mathcal{F}_2 \times \cdots \times \mathcal{F}_n, \\ \mu &= \mu_1 \times \mu_2 \times \cdots \times \mu_n, \\ \Gamma_x &= \Gamma_1 x \cap \Gamma_2 x \cap \cdots \cap \Gamma_n x \end{aligned} \tag{51}$$

for all $x \,\varepsilon\, X$. The product measure space $(X, \mathcal{F}, \mu)$ is motivated by the usual definition of statistical independence. The definition of Γ reflects the idea that $x_i \,\varepsilon\, X_i$ is consistent with a particular $s \,\varepsilon\, S$ if and only if $s \,\varepsilon\, \Gamma_i x_i$, for $i = 1, 2, \cdots, n$, and consequently $x = (x_1, x_2, \cdots, x_n) \;\varepsilon X$ is consistent with that s if and only if s belongs to all of the $\Gamma_i x_i$ simultaneously.

It is a characteristic of the above combination rule that neither upper probabilities, nor lower probabilities nor probabilities of the type $p_{\delta_1 \delta_2 \cdots \delta_m}$ have a simple product rule of combination. A set of probabilities $q_i(T)$ which do obey a simple product rule is defined as follows: For the systems $(X, \mathcal{F}, \mu)$ and Γ defined by (51) from the systems $(X_i, \mathcal{F}_i, \mu_i)$ and Γ_i, and for any $T \subset S$, set

$$\tilde{T} = \{x \,\varepsilon\, X, \Gamma x \supset T\} \text{ and } \tilde{T}_i = \{x_i \,\varepsilon\, X_i, \Gamma_i x_i \supset T\} \tag{52}$$

and set

$$q(T) = \mu(\tilde{T}) \text{ and } q_i(T) = \mu_i(\tilde{T}_i). \tag{53}$$

It follows immediately that

$$\tilde{T} = \tilde{T}_1 \times \tilde{T}_2 \times \cdots \times \tilde{T}_n \tag{54}$$

and hence that

$$q(T) = q_1(T) \times q_2(T) \times \cdots \times q_n(T). \tag{55}$$

It will be seen shortly that, at least for finite S, the probabilities $q(T)$ are sufficient to determine all upper and lower probabilities defined by a system $(X, \mathcal{F}, \mu)$ and Γ and hence from (55) they provide a convenient form, ready for further combination, for storing the information in a given source. Note also that, if T consists of a single element $s \,\varepsilon\, S$, then $\tilde{T} = T^*$ so that $q(T) = \mu(T^*)$ whence from (3) the $q\{s\}$ as s ranges over S are proportional to $P^*\{s\}$.

The foregoing ideas will now be concretely illustrated using a finite S, beginning with $m = 3$. A source is characterized here by $p_{000}, p_{100}, p_{010}, p_{001}$, $p_{110}, p_{101}, p_{011}$ and p_{111}. If the $q(T)$ corresponding to the $T = \varnothing, \{s_1\}, \{s_2\}$, $\{s_3\}, \{s_1, s_2\}, \cdots, \{s_1, s_2, s_3\}$ are denoted by $q_{000}, q_{100}, q_{010}, q_{001}, q_{110}, \cdots$, q_{111}, it follows directly that

$$\begin{aligned} &q_{000} = 1 = p_{100} + p_{010} + p_{001} + p_{110} + p_{101} + p_{011} + p_{111}, \\ &q_{100} = p_{100} + p_{110} + p_{101} + p_{111}, \qquad \text{and similarly for } q_{010} \text{ and } q_{001}, \\ &q_{110} = p_{110} + p_{111}, \qquad \text{and similarly for } q_{101} \text{ and } q_{011}, \text{ and} \\ &q_{111} = p_{111}. \end{aligned} \tag{56}$$

The relations (56) may be solved to yield

$$\begin{aligned}
p_{000} &= 1 - q_{100} - q_{010} - q_{001} + q_{110} + q_{101} + q_{011} - q_{111}, \\
p_{100} &= q_{100} - q_{110} - q_{101} + q_{111}, \qquad \text{and similarly for } p_{010} \text{ and } p_{001}, \\
p_{110} &= q_{110} - q_{111}, \qquad \text{and similarly for } q_{101} \text{ and } p_{011}, \text{ and} \\
p_{111} &= q_{111}, \qquad (57)
\end{aligned}$$

thus showing that the set of a $q(T)$ determine the $p_{\delta_1\delta_2\delta_3}$. Note also that the extensions of (56) and (57) from $m = 3$ to general m are evident and easily proved.

A pair of sources $i = 1, 2$ may be characterized by their $p^{[i]}_{\delta_1\delta_2\delta_3}$ or by their $q^{[i]}_{\delta_1\delta_2\delta_3}$. The relations $q_{\delta_1\delta_2\delta_3} = q^{[1]}_{\delta_1\delta_2\delta_3} q^{[2]}_{\delta_1\delta_2\delta_3}$ from (55) together with the relations (56) and (57) applied to the two sources and their combination yield

$$\begin{aligned}
p_{000} &= p^{[1]}_{100}p^{[2]}_{010} + p^{[1]}_{100}p^{[2]}_{001} + p^{[1]}_{100}p^{[2]}_{011} + p^{[1]}_{010}p^{[2]}_{100} + p^{[1]}_{010}p^{[2]}_{001} + p^{[1]}_{010}p^{[2]}_{101} \\
&\quad + p^{[1]}_{001}p^{[2]}_{100} + p^{[1]}_{001}p^{[2]}_{010} + +p^{[1]}_{001}p^{[2]}_{110} + p^{[1]}_{110}p^{[2]}_{100} + p^{[1]}_{101}p^{[2]}_{010} + p^{[1]}_{011}p^{[2]}_{100}, \\
p_{100} &= p^{[1]}_{100}p^{[2]}_{100} + p^{[1]}_{100}p^{[2]}_{110} + p^{[1]}_{100}p^{[2]}_{101} + p^{[1]}_{100}p^{[2]}_{111} \\
&\quad + p^{[1]}_{110}p^{[2]}_{100} + p^{[1]}_{110}p^{[2]}_{101} + p^{[1]}_{101}p^{[2]}_{100} + p^{[1]}_{101}p^{[2]}_{110} + p^{[1]}_{111}p^{[2]}_{100}, \qquad (58) \\
&\qquad \text{and similarly for } p_{010} \text{ and } p_{001}, \\
p_{110} &= p^{[1]}_{110}p^{[2]}_{110} + p^{[1]}_{110}p^{[2]}_{111} + p^{[1]}_{111}p^{[2]}_{110}, \\
&\qquad \text{and similarly for } p_{101} \text{ and } p_{011}, \text{and} \\
p_{111} &= p^{[1]}_{111}p^{[2]}_{111}.
\end{aligned}$$

The general rule here, extended to any m, is that

$$p_{\delta_1\delta_2\cdots\delta_m} = \sum p^{[1]}_{\delta_1{}'\delta_2{}'\cdots\delta_m{}'} p^{[2]}_{\delta_1{}'\delta_2{}''\cdots\delta_m{}''} \qquad (59)$$

with summation over all $(\delta_1{}', \delta_2{}', \cdots, \delta_m{}', \delta_1{}'', \delta_2{}'', \cdots, \delta_m{}'')$ such that $\delta_i = \delta_i{}'\delta_i{}''$ for $i = 1, 2, \cdots, m$. It is clear, however, that combining sources directly in terms of $p_{\delta_1\delta_2\cdots\delta_m}$ is awkward, and by referring from (58) back to Table 1 one sees that the situation is no better in terms of upper and lower probabilities.

This section concludes with two important properties of the combination rule. To introduce the first of these, note that, if a source $(X_1, \mathcal{F}_1, \mu_1)$ and Γ_1 is combined with an informationless source, then the result is again the original source $(X_1, \mathcal{F}_1, \mu_1)$ and Γ_1. By an informationless source is meant an $(X_2, \mathcal{F}_2, \mu_2)$ and Γ_2 such that $\Gamma_2 x_2 = S$ for all $x_2 \varepsilon X_2$, i.e., a source for which $P^*(T) = 1$ and $P_*(T) = 0$ for every T other than $\varnothing$ and S. The more general version of the first property asserts that, if a source $(X_1, \mathcal{F}_1, \mu_1)$ and Γ_1 is combined with a source $(X_2, \mathcal{F}_2, \mu_2)$ and Γ_2 where $\Gamma_2 x_2 = R \subset S$ for all $x_2 \varepsilon X_2$, then

$$P^*(T) = {P_1}^*(T|R), \qquad P_*(T) = P_{1*}(T|R) \tag{60}$$

for $T \subset S$, where ${P_1}^*(T|R)$ and $P_{1*}(T|R)$ are upper and lower conditional probabilities for the system $(X_1, \mathcal{F}_1, \mu_1)$ and Γ_1 according to the definitions (43). In other words, the rule of this section is sufficiently general to include the definition of conditioning as a special case. The relations (60) are immediate consequences of the definitions adopted.

The second property concerns sharp sources. A source will be called *sharp* if it is sharp with respect to T for all $T\varepsilon\mathcal{E}$, and will be called *sharp with respect to T if* $P^*(T) = P_*(T)$. Thus a sharp source is an ordinary probability measure over the events $T\varepsilon\mathcal{E}$. Assuming finite S, it will be shown that *a source which is sharp with respect to a given T remains sharp with respect to T after combination with any other source.* A similar property therefore holds for sharpness with respect to all T. Thus, if sharpness is once achieved by a user of this theory, it remains a characteristic of all subsequent states of knowledge of the user.

The demonstration depends on a simple lemma:

Lemma 2. *A source is sharp with respect to* $T\varepsilon\mathcal{E}$ *if and only if* $q(R) = 0$ *for every* R *such that* $R \cap T \neq \varnothing$ *and* $R \cap \bar{T} \neq \varnothing$. Clearly, $P^*(T) - P_*(T) \geqq q(R)$ for any R such that $R \cap T \neq \varnothing$ and $R \cap \bar{T} \neq \varnothing$, so that $P^*(T) = P_*(T)$ implies $q(R) = 0$ for all such R. Conversely, if $q(R) = 0$ for all such R, then no Γx which intersects both T and $\bar{T}$ may have positive probability, which implies that $P^*(T) = P_*(T)$. In view of the simple combination property of the $q(R)$ function, the sharpness property of the preceding paragraph follows immediately from the above lemma.

A new kind of limit theorem becomes possible in discussions of upper and lower probability, namely results about convergence to sharpness. For example, in view of (51) one would expect the combination of n sources to be sharper than a typical member of the sources combined. Thus rates of convergence to sharpness deserve definition and study. An illustration of this may be found in equation (32) of Dempster (1966).

6 An Application

The theory proposed in this paper has been implicitly applied to statistical inference in an earlier paper (Dempster (1966)). The nature of the application will be sketched briefly. Individual sample observations may be regarded as sources whose information may be combined according to the rule of Sect. 5. In such an individual source, the role of X is played by a space representing the possible sample individuals, and the role of S is represented by a parameter space or more generally by the product of a parameter space and a space of future observations. Before a particular sample observation is recorded, the source defined by that sample individual is informationless, but after conditioning by the sample observation one generally gets a non-trivial system of upper and lower probabilities referring to the parameters or to the parameters and future observations jointly.

The combination of many sample individuals appears to lead to sharp inferences which agree with standard asymptotic inferences given either by Bayesian methods or by confidence methods. It is also suggested as valid to treat a prior distribution as a source of information independent of sample data. If source 1 is taken to be combined sample information, and source 2 is taken to be prior information, and if this prior information is sharp and has a density, then combination of the two sources reduces in this special case to the familiar product of likelihood and prior leading to a Bayesian posterior density.

Further concrete examples of these applications to inference will be forthcoming soon.

Acknowledgment

Many helpful comments were provided by L.J. Savage who read an earlier version of this paper.

References

Dempster, A. P (1966). New approaches for reasoning towards posterior distributions based on sample data. *Ann. Math. Statist.* **37** 355–374.

Fishburn, Peter C. (1964). *Decision and Value Theory.* Wiley, New York.

Good, I. J. (1962). The measure of a non-measurable set. *Logic, Methodology and Philosophy of Science* (edited by Ernest Nagel, Patrick Suppes, and Alfred Tarski). Stanford Univ. Press. 319–329.

Koopman, B. O. (1940a). The axioms and algebra of intuitive probability. *Ann. Math.* **41** 269–292.

Koopman, B. O. (1940b). The bases of probability. *Bull. Amer. Math. Soc.* **46** 763–774.

Smith, C. A. B. (1961). Consistency in statistical inference and decision, (with discussion). *J. Roy. Statist. Soc. Ser. B* **23** 1–25.

Smith, C. A. B. (1965). Personal probability and statistical analysis, (with discussion). *J. Roy. Statist. Soc. Ser. A* **128** 469–499.

4

A Generalization of Bayesian Inference*

Arthur P. Dempster

Abstract. Procedures of statistical inference are described which generalize Bayesian inference in specific ways. Probability is used in such a way that in general only bounds may be placed on the probabilities of given events, and probability systems of this kind are suggested both for sample information and for prior information. These systems are then combined using a specified rule. Illustrations are given for inferences about trinomial probabilities, and for inferences about a monotone sequence of binomial p_i. Finally, some comments are made on the general class of models which produce upper and lower probabilities, and on the specific models which underlie the suggested inference procedures.

1 Introduction

REDUCED to its mathematical essentials, Bayesian inference means starting with a global probability distribution for all relevant variables, observing the values of some of these variables, and quoting the conditional distribution of the remaining variables given the observations. In the generalization of this paper, something less than a global probability distribution is required, while the basic device of conditioning on observed data is retained. Actually, the generalization is more specific. The term *Bayesian* commonly implies a global probability law given in two parts, first the marginal distribution of a set of parameters, and second a family of conditional distributions of a set of observable variables given potential sets of parameter values. The first part, or *prior distribution*, summarizes a set of beliefs or state of knowledge in hand before any observations are taken. The second part, or *likelihood function*, characterizes the information carried by the observations. Specific generalizations are suggested in this paper for both parts of the common Bayesian model, and also for the method of combining the two parts. The

* Read at a RESEARCH METHODS MEETING of the Society, February 14th, 1968, Professor R.L. PLACKETT in the Chair.

components of these generalizations are built up gradually in Sect. 2 where they are illustrated on a model for trinomial sampling.

Inferences will be expressed as *probabilities* of events defined by unknown values, usually unknown parameter values, but sometimes the values of observables not yet observed. It is not possible here to go far into the much-embroiled questions of whether probabilities are or are not objective, are or are not degrees of belief, are or are not frequencies, and so on. But a few remarks may help to set the stage. I feel that the proponents of different specific views of probability generally share more attitudes rooted in the common sense of the subject than they outwardly profess, and that careful analysis renders many of the basic ideas more complementary than contradictory. Definitions in terms of frequencies or equally likely cases do illustrate clearly how reasonably objective probabilities arise in practice, but they fail in themselves to say what probabilities mean or to explain the pervasiveness of the concept of probability in human affairs. Another class of definitions stresses concepts like degree of confidence or degree of belief or degree of knowledge, sometimes in relation to betting rules and sometimes not. These convey the flavour and motivation of the science of probability, but they tend to hide the realities which make it both possible and important for cognizant people to agree when assigning probabilities to uncertain outcomes. The possibility of agreement arises basically from common perceptions of symmetries, such as symmetries among cases counted to provide frequencies, or symmetries which underlie assumptions of exchangeability or of equally likely cases. The importance of agreement may be illustrated by the statistician who expresses his inferences about an unknown parameter value in terms of a set of betting odds. If this statistician accepts any bet proposed at his stated odds, and if he wagers with colleagues who consistently have more information, perhaps in the form of larger samples, then he is sure to suffer disaster in the long run. The moral is that probabilities can scarcely be "fair" for business deals unless both parties have approximately the same probability assessments, presumably based on similar knowledge or information. Likewise, probability inferences can contribute little to public science unless they are as objective as the web of generally accepted fact on which they are based. While knowledge may certainly be personal, the communication of knowledge is one of the most fundamental of human endeavours. Statistical inference can be viewed as the science whose formulations make it possible to communicate partial knowledge in the form of probabilities.

Generalized Bayesian inference seeks to permit improvement on classical Bayesian inference through a complex trade-off of advantages and disadvantages. On the credit side, the requirement of a global probability law is dropped and it becomes possible to work with only those probability assumptions which are based on readily apparent symmetry conditions and are therefore reasonably objective. For example, in a wide class of sampling models, including the trinomial sampling model analysed in Sect. 2, no probabilities are assumed except the familiar and non-controversial representation of a sample as n independent and identically distributed random elements from a population. Beyond this, further assumptions like specific parametric forms or prior distributions for parameters need be put in only to the extent that they appear to command a fair degree of assent.

The new inference procedures do not in general yield exact probabilities for desired inferences, but only bounds for such probabilities. While it may count as a debit item that inferences are less precise than one might have hoped, it is a credit item that greater flexibility is allowed in the representation of a state of knowledge. For example, a state of total ignorance about an uncertain event T is naturally represented by an upper probability $P^*(T) = 1$ and a lower probability $P_*(T) = 0$. The new flexibility thus permits a simple resolution of the old controversy about how to represent total ignorance via a probability distribution. In real life, ignorance is rarely so total that (0, 1) bounds are justified, but ignorance is likely to be such that a precise numerical probability is difficult to justify. I believe that experience and familiarity will show that the general range of bounds $0 \leq P_*(T) \leq P^*(T) \leq 1$ provides a useful tool for representing degrees of knowledge.

Upper and lower probabilities apparently originated with Boole (1854) and have reappeared after a largely dormant period in Good (1962) and Smith (1961, 1965). In this paper upper and lower probabilities are generated by a specific mathematical device whereby a well-defined probability measure over one sample space becomes diffused in its application to directly interesting events. In order to illustrate the idea simply, consider a map showing regions of land and water. Suppose that $0{\cdot}80$ of the area of the map is visible and that the visible area divides in the proportions $0{\cdot}30$ to $0{\cdot}70$ of water area to land area. What is the probability that a point drawn at random from the *whole* map falls in a region of water? Since the visible water area is $0{\cdot}24$ of the total area of the map, while the unobserved $0{\cdot}20$ of the total area could be water or land, it can be asserted only that the desired probability lies between $0{\cdot}24$ and $0{\cdot}44$. The model supposes a well-defined uniform distribution over the whole map. Of the total measure of unity, the fraction $0{\cdot}24$ is associated with water, the fraction $0{\cdot}56$ is associated with land, and the remaining fraction $0{\cdot}20$ is ambiguously associated with water or land. Note the implication of total ignorance of the unobserved area. There would be no objection to introducing other sources of information about the unobserved area. Indeed, if such information were appropriately expressed in terms of an upper and lower probability model, it could be combined with the above information using a rule of combination defined within the mathematical system. A correct analogy can be drawn with prior knowledge of parameter values, which can likewise be formally incorporated into inferences based on sample data, using the same rule of combination. The general mathematical system, as given originally in Dempster (1967a), will be unfolded in Sect. 2 and will be further commented upon in Sect. 4.

If the inference procedures suggested in this paper are somewhat speculative in nature, the reason lies, I believe, not in a lack of objectivity in the probability assumptions, nor in the upper and lower probability feature. Rather, the source of the speculative quality is to be found in the logical relationships between population members and their observable characteristics which are postulated in each model set up to represent sampling from a population. These logical relationships are conceptual devices, which are not regarded as empirically checkable even in principle, and they are somewhat arbitrary. Their acceptability will be analysed in Sect. 5 where it will be argued that the arbitrariness may correspond to something real in the nature of an uncertainty principle.

A degree of arbitrariness does not in itself rule out a method of statistical inference. For example, confidence statements are widely used in practice despite the fact that many confidence procedures are often available within the same model and for the same question, and there is no well-established theory for automatic choice among available confidence procedures. In part, therefore, the usefulness of generalized Bayesian inference procedures will require that practitioners experiment with them and come to feel comfortable with them. Relatively few procedures are as yet analytically tractable, but two examples are included, namely, the trinomial sampling inference procedures of Sect. 2, and a procedure for distinguishing between monotone upward and monotone downward sequences of binomial p_i as given in Sect. 3. Another model is worked through in detail in Dempster (1967b).

Finally, an acknowledgement is due to R.A. Fisher who announced with characteristic intellectual boldness, nearly four decades ago, that probability inferences were indeed possible outside of the Bayesian formulation. Fisher compiled a list of examples and guide-lines which seemed to him to lead to acceptable inferences in terms of probabilities which he called *fiducial probabilities*. The mathematical formulation of this paper is broad enough to include the fiducial argument in addition to standard Bayesian methods. But the specific models which Fisher advocated, depending on ingenious but often controversial *pivotal quantities*, are replaced here by models which start further back at the concept of a population explicitly represented by a mathematical space. Fisher did not consider models which lead to separated upper and lower probabilities, and indeed went to some lengths, using sufficiency and ancillarity, and arranging that the spaces of pivotal quantities and of parameters be of the same dimension, in order to ensure that ambiguity did not appear. This paper is largely an exploration of fiducial-like arguments in a more relaxed mathematical framework. But, since Bayesian methods are more in the main stream of development, and since I do explicitly provide for the incorporation of prior information, I now prefer to describe my methods as extensions of Bayesian methods rather than alternative fiducial methods. I believe that Fisher too regarded fiducial inference as being very close to Bayesian inference in spirit, differing primarily in that fiducial inference did not make use of prior information.

2 Upper and Lower Probability Inferences Illustrated on a Model for Trinomial Sampling

A pair of sample spaces X and S underlie the general form of mathematical model appearing throughout this work. The first space X carries an ordinary probability measure μ, but interest centres on events which are identified with subsets of S. A bridge is provided from X to S by a logical relationship which asserts that, if x is the realized sample point in X, then the realized sample point s in S must belong to a subset Γx of S. Thus a basic component of the model is a mathematical transformation which associates a subset Γx of S with each point x of X. Since the Γx determined by a specific x contains in general many points (or *branches* or *values*), the transformation $x \rightarrow \Gamma x$ may be called a *multivalued mapping.* Apart from measurability considerations, which are ignored in this paper, the general model is defined by the elements introduced above and will be labelled (X, S, μ, Γ) for convenient reference. Given (X, S, μ, Γ), upper and lower probabilities $P^*(T)$ and $P_*(T)$ are determined for each subset T of S.

In the cartographical example of Sect. 1, X is defined by the points of the map, S is defined by two points labelled "water" and "land", μ is the uniform distribution of probability over the map, and Γ is the mapping which associates the single point "water" or "land" in S with the appropriate points of the visible part of X and associates both points of S with the points of the unseen part of X. For set-theoretic consistency, Γx should be regarded as a single point subset of S, rather than a single point itself, over the visible part of X, but the meaning is the same either way.

The general definitions of $P^*(T)$ and $P_*(T)$ as given in Dempster (1967a) are repeated below in more verbal form. For any subset T of S, define T^* to be the set of points x in X for which Γx has a non-empty intersection with T, and define T_* to be the set of points x in X for which Γx is contained in T but is not empty. In particular, the sets S^* and S_* coincide. The complement $X - S^*$ of S^* consists of those x for which Γx is the empty set. Now define the *upper probability* of T to be

$$P^*(T) = \mu(T^*)/\mu(S^*) \tag{1}$$

and the *lower probability* of T to be

$$P_*(T) = \mu(T_*)/\mu(S^*). \tag{2}$$

Note that, since $T_* \subset T^* \subset S^*$, one has

$$0 \leq P_*(T) \leq P^*(T) \leq 1. \tag{3}$$

Also, if $\bar{T}$ is the complement of T in S, then $\bar{T}_*$ and $\bar{T}^*$ are respectively the complements of T^* and T_* in S^*, so that

$$P_*(\bar{T}) = 1 - P^*(T) \text{ and } P^*(\bar{T}) = 1 - P_*(T). \tag{4}$$

Other formal consequences of the above definitions are explored in Dempster (1967a).

The heuristic conception which motivates (1) and (2) is the idea of carrying probability elements $d\mu$ from X to S along the branches of the mapping Γx. The ambiguity in the consequent probability measure over S occurs because the probability element $d\mu(x)$ associated with x in X may be carried along any branch of Γx or, more generally, may be distributed over the different branches of Γx for each x. Part of the μ measure, namely the measure of the set $X - S^*$ consisting of points x such that Γx is empty, cannot be moved from X at all. Since there is an implicit assumption that some s in S is actually realized, it is appropriate to condition by S^* when defining relevant probabilities. This explains the divisor $\mu(S^*)$ appearing in (1) and (2). Among all the ways of transferring the relevant probability $\mu(S^*)$ from X to S along branches of Γx, the largest fraction which can possibly follow branches into T is $P^*(T)$, while the smallest possible fraction is $P_*(T)$. Thus conservative probability judgements may be rendered by asserting only that the probability of T lies between the indicated upper and lower bounds.

It may also be illuminating to view Γx as a random *set* in S generated by the random *point* x in X, subject to the condition that Γx is not empty. After conditioning on $S^*, P^*(T)$ is the probability that the random set Γx intersects the fixed set T, while $P_*(T)$ is the probability that the random set Γx is contained in the fixed set T.

A probability model like (X, S, μ, Γ) may be modified into other probability models of the same general type by conditioning on subsets of S. Such conditioning on observed data defines the generalized Bayesian inferences of this paper. Beyond and generalizing the concept of conditioning, there is a natural rule for combining or multiplying several independent models of the type (X, S, μ, Γ) to obtain a single product model of the same type. For example, the models for n independent sample observations may be put together by the product rule to yield a single model for a sample of size n, and the model defining prior information may be combined with the model carrying sample information by the same rule. The rules for conditioning and multiplying will be transcribed below from Dempster (1967a) and will be illustrated on a model for trinomial sampling. First, however, the elements of the trinomial sampling model will be introduced for a sample of size one.

Each member of a large population, shortly to be idealized as an infinite population, is supposed known to belong to one of three identifiable categories c_1, c_2 and c_3, where the integer subscripts do not indicate a natural ordering of the categories. Thus the individuals of the population could be balls in an urn, identical in appearance apart from their colours which are red (c_1) or white (c_2) or blue (c_3). A model will be defined which will ultimately lead to procedures for drawing inferences about unknown population proportions of c_1, c_2 and c_3, given the categories of a random sample of size n from the population. Following Dempster (1966), the individuals of the population will be explicitly represented by the points of a space U, and the randomness associated with a sample individual drawn from U will be characterized by a probability measure over U. Thus, a finite population of size N could be represented by any finite space U with N elements, with random sampling represented by the uniform distribution of probability over the N elements of U. Such a finite population model is analysed in detail in Dempster (1967b). Here, however, the population is treated as infinite, and, for reasons tied up with the trinomial observable, the space U is identified with a triangle. Convenient barycentric coordinates for a general point of U are

$$\mathbf{u} = (u_1, u_2, u_3), \tag{5}$$

where $0 \leq u_1, 0 \leq u_2, 0 \leq u_3$ and $u_1 + u_2 + u_3 = 1$. See Fig. 1. It is further supposed that a random sample of size one means an individual $\mathbf{u}$ drawn according to the uniform distribution ρ over the triangle U. In the model (X, S, μ, Γ) representing a random sample of size one from a trinomial population the roles of X and μ will be played by U and ρ.

Two further spaces enter naturally into the model for a single trinomial observation. The first is the three-element space $C = \{c_1, c_2, c_3\}$ whose general member c represents the observable category of the sample individual. The second is the space Π whose general point is

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3), \tag{6}$$

with $0 \leq \pi_1, 0 \leq \pi_2, 0 \leq \pi_3$ and $\pi_1 + \pi_2 + \pi_3 = 1$, where π_i is to be interpreted for $i = 1, 2, 3$ as the proportion of the population falling in category c_i. Note that Π is a mathematical copy of U, but its applied meaning is distinct from that of U. The role of S in the general model (X, S, μ, Γ) will be played by the product space $C \times \Pi$ which represents jointly the observation on a single random individual together with the population proportions of c_1, c_2 and c_3.

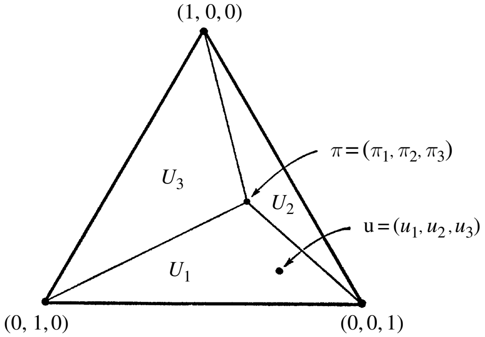


Fig. 1. A triangle representing the space U, showing the barycentric coordinates of the three vertices of U together with a general point $\mathbf{u} = (u_1, u_2, u_3)$. The three closed sub-triangles labelled U_1, U_2 and U_3 with a common vertex at $\boldsymbol{\pi}$ represent the subsets of U consisting of points $\mathbf{u}$ such that B$\mathbf{u}$ contains $(c_1, \boldsymbol{\pi}), (c_2, \boldsymbol{\pi})$ and $(c_3, \boldsymbol{\pi})$, respectively

Finally, the role of Γ is played by B where, for any $\mathbf{u}$ in U, the set B$\mathbf{u}$ in $C \times \Pi$ consists of the points $(c_i, \boldsymbol{\pi})$ such that

$$\frac{\pi_i}{u_i} = \max\left(\frac{\pi_1}{u_1}, \frac{\pi_2}{u_2}, \frac{\pi_3}{u_3}\right), \tag{7}$$

for $i = 1, 2, 3$, To understand the definition of B, but not yet the motivation for the definition, it is helpful to visualize $C \times \Pi$ as a stack of three triangles as in Fig. 2 where the three levels correspond to the three points of C. The contributions to B$\mathbf{u}$ from the three levels of $C \times \Pi$ are shown as shaded areas in Fig. 2. It is important also to understand the inverse mapping B^{-1} which carries points of $C \times \Pi$ to subsets of U, where

$$U_i = B^{-1}(c_i, \boldsymbol{\pi}) \tag{8}$$

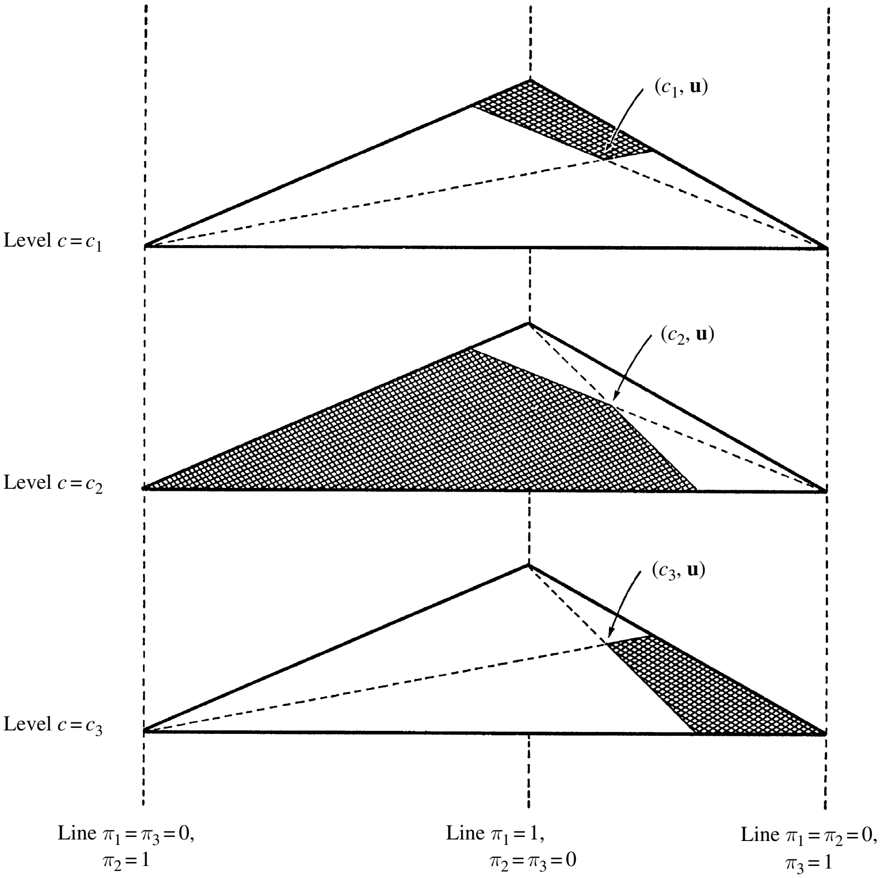

$(c_1, \mathbf{u})$
Level $c = c_1$
$(c_2, \mathbf{u})$
Level $c = c_2$
$(c_3, \mathbf{u})$
Level $c = c_3$
Line $\pi_1 = \pi_3 = 0$, $\pi_2 = 1$
Line $\pi_1 = 1$, $\pi_2 = \pi_3 = 0$
Line $\pi_1 = \pi_2 = 0$, $\pi_3 = 1$

Fig. 2. The space $C \times \Pi$ represented as triangles on three levels. The three closed shaded regions together make up the subset $B\mathbf{u}$ determined from a given $\mathbf{u}$

is defined to be the subset of U consisting of points $\mathbf{u}$ for which $\mathbf{Bu}$ contains $(c_i, \boldsymbol{\pi})$. The subsets U_1, U_2, U_3 defined by a given $\boldsymbol{\pi}$ in Π are illustrated in Fig. 1.

It is easily checked with the help of Fig. 1 that

$$\rho(U_i) = \pi_i \text{ and } \rho(U_i \cap U_j) = 0 \tag{9}$$

for $i, j = 1, 2, 3$ and $i \neq j$. It will be shown later that the property (9) is a basic requirement for the mapping B defined in (7). Other choices of U and B could be made which would also satisfy (9). Some of these choices amount to little more than adopting different coordinate systems for U, but other possible choices differ in a more fundamental way. Thus an element of arbitrariness enters the model for trinomial sampling at the point of choosing U and B. The present model was introduced in Dempster (1966) under the name *structure of the second kind.* Other possibilities will be mentioned in Sect. 5.

All of the pieces of the model $(U, C \times \Pi, \rho, B)$ are now in place, so that upper and lower probabilities may be computed for subsets T of $C \times \Pi$. It turns out, however, that $P^*(T) = 1$ and $P_*(T) = 0$ for interesting choices of T, and that interesting illustrations of upper and lower probabilities are apparent only after conditioning. For example, take T to be the event that category c_1 will be observed in a single drawing from the population, i.e. $T = C_1 \times \Pi$, where C_1 is the subset of C consisting of c_1 only. To check that $P^*(T) = 1$ and $P_*(T) = 0$, note (i) that $T^* = U$ because every $\mathbf{u}$ in U lies in U_1 of Fig. 1 for some $(c_1, \boldsymbol{\pi})$ in $C_1 \times \Pi$, and (ii) that T_* is empty because no $\mathbf{u}$ in U lies in U_1 for all $(c_1, \boldsymbol{\pi})$ in $C_1 \times \Pi$. In general, any non-trivial event governed by C alone or by Π alone will have upper probability unity and lower probability zero. Such a result is sensible, for if no information about $\boldsymbol{\pi}$ is put into the system no information about a sample observation should be available, while if no sample observation is in hand there should be no available information about $\boldsymbol{\pi}$. (Recall the interpretation suggested in Sect. 1 that $P^*(T) = 1$ and $P_*(T) = 0$ should convey a state of complete ignorance about whether or not the real world outcome s will prove to lie in T.)

Turning now to the concept of upper and lower *conditional* probabilities, the definition which fits naturally with the general model (X, S, μ, Γ) arises as follows. If information is received to the effect that sample points in $S - T$ are ruled out of consideration, then the logical assertion "x in X must correspond to s in $\Gamma x \subset S$" is effectively altered to read "x in X must correspond to s in $\Gamma x \cap T \subset S$". Thus the original model (X, S, μ, Γ) is *conditioned on* T by altering (X, S, μ, Γ) to $(X, S, \mu, \tilde{\Gamma})$, where the multivalued mapping $\tilde{\Gamma}$ is defined by

$$\tilde{\Gamma} x = \Gamma x \cap T. \tag{10}$$

Under the conditioned model, an outcome in $S - T$ is regarded as impossible, and indeed the set $S - T$ has upper and lower conditional probabilities both zero. It is sufficient for practical purposes, therefore, to take the conditional model to be $(X, T, \mu, \tilde{\Gamma})$ and to consider upper and lower conditional probabilities only for subsets of T.

Although samples of size one are of little practical interest, the model for a single trinomial observation provides two good illustrations of the definition of a conditioned model. First, it will be shown that conditioning on a fixed value of $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ results in π_i being both the upper and lower conditional probability of an observation c_i, for $i = 1, 2, 3$. This result is equivalent to (9) and explains the importance of (9), since any reasonable model should require that the population proportions be the same as the probabilities of the different possible outcomes in a single random drawing *when the population proportions are known.* Second, it will be shown that non-trivial inferences about $\boldsymbol{\pi}$ may be obtained by conditioning on the observed category c of a single individual randomly drawn from U.

In precise mathematical terms, to condition the trinomial sampling model $(U, C \times \Pi, \rho, B)$ on a fixed $\boldsymbol{\pi}$ is to condition on $T = C \times \tilde{\Pi}$, where $\tilde{\Pi}$ is the subset of Π consisting of the single point $\boldsymbol{\pi}$. T itself consists of the three points $(c_1, \boldsymbol{\pi})$, $(c_2, \boldsymbol{\pi})$ and $(c_3, \boldsymbol{\pi})$ which in turn define single point subsets T_1, T_2 and T_3 of T. The conditioned model may be written $(U, T, \rho, \tilde{B})$ where $\tilde{B}\mathbf{u} = B\mathbf{u} \cap T$ for all $\mathbf{u}$. By referring back to the definition of B as illustrated in Figs. 1 and 2, it is easily checked that the set of $\mathbf{u}$ in U such that $\tilde{B}\mathbf{u}$ intersects T_i is the closed triangle U_i appearing in Fig. 1, while the set of $\mathbf{u}$ in U such that $\tilde{B}\mathbf{u}$ is contained in T_i is the open triangle U_i, for $i = 1, 2, 3$. Whether open or closed, the triangle U_i has measure π_i, and it follows easily from (9) that the upper and lower conditional probabilities of T_i given T are

$$P^*(T_i|T) = P_*(T_i|T) = \pi_i, \tag{11}$$

for $i = 1, 2, 3$, Note that $\tilde{B}\mathbf{u}$ is not empty for any $\mathbf{u}$ in U, so that the denominators in (1) and (2) are both unity in the application (11).

Consider next the details of conditioning the trinomial model on a fixed observation c_1. The cases where a single drawing produces c_2 or c_3 may be handled by permuting indices. Observing c_1 is formally represented by conditioning on $\tilde{T} = C_1 \times \Pi$ where C_1 as above is the subset of C consisting of c_1 alone. In the conditional model $(U, \tilde{T}, \rho, \tilde{B})$, the space $\tilde{T}$ is represented by the first level in Fig. 2 while $\tilde{B}\mathbf{u}$ is represented by the closed shaded region in that first level. Since $\tilde{B}\mathbf{u}$ is non-empty for all $\mathbf{u}$ in U, the ρ measure may be used directly without renormalization to compute upper and lower conditional probabilities given $\tilde{T}$. An event R defined as a subset of Π is equivalently represented by the subset $C_1 \times R$ of $\tilde{T}$. The upper conditional probability of $C_1 \times R$ given $\tilde{T}$ is the probability that the random region $\tilde{B}\mathbf{u}$ intersects $C_1 \times R$ where $(c_1, \mathbf{u})$ is uniformly distributed over $C_1 \times \Pi$. See Fig. 3. Similarly, the lower conditional probability of $C_1 \times R$ given $\tilde{T}$ is the probability that the random region $\tilde{B}\mathbf{u}$ is contained in $C_1 \times R$. For example, if R is the lower portion of the triangle where $0 \leq \pi_1 \leq \pi_1''$, then

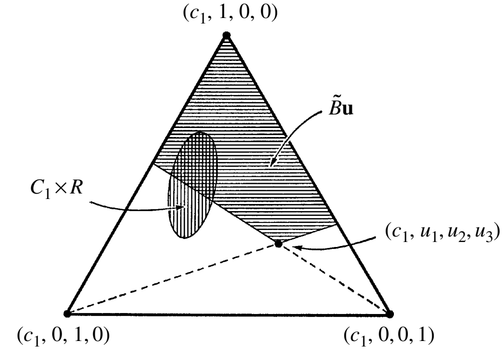


Fig. 3. The triangle $\tilde{T} = C_1 \times \Pi$ for the model conditioned on the observation c_1. Horizontal shading covers the region $\tilde{B}\mathbf{u}$, while vertical shading covers a general fixed region $C_1 \times R$

$$P^*(C_1 \times R|\tilde{T}) = 1 - (1 - \pi_1'')^2 = \pi_1''(2 - \pi_1'') \text{ and } P_*(C_1 \times R|\tilde{T}) = 0.$$

Or, in more colloquial notation,

$$P^*(0 \leq \pi_1 \leq \pi_1''|c = c_1) = \pi_1''(2 - \pi_1'') \text{ and } P_*(0 \leq \pi_1 \leq \pi_1''|c = c_1) = 0.$$

More generally, it can easily be checked that

$$P^*(\pi_1' \leq \pi_1 \leq \pi_1''|c = c_1) = \pi_1''(2 - \pi_1''), \tag{12}$$

while

$$\left.\begin{aligned} P_*(\pi_1' \leq \pi_1 \leq \pi_1''|c = c_1) &= 0 && \text{if } \pi_1'' < 1 \\ &= (1 - \pi_1')^2 && \text{if } \pi_1'' < 1, \end{aligned}\right\} \tag{13}$$

for any fixed π_1' and π_1'' satisfying $0 \leq \pi_1' \leq \pi_1'' \leq 1$. Likewise,

$$P^*(\pi_2' \leq \pi_2 \leq \pi_2''|c = c_1) = 1 - \pi_2', \tag{14}$$

while

$$\left.\begin{aligned} P_*(\pi_2' \leq \pi_2 \leq \pi_2''|c = c_1) &= 0 && \text{if } \pi_2' > 0 \\ &= \pi_2'' && \text{if } \pi_2' = 0, \end{aligned}\right\} \tag{15}$$

for any fixed π_2' and π_2'' satisfying $0 < \pi_2' < \pi_2'' < 1$. Relations (14) and (15) also hold when subscripts 2 and 3 are interchanged. Formulae (12) to (15) are the first instances of generalized Bayesian inferences reached in this paper, where, as will shortly be explained, prior knowledge of $\boldsymbol{\pi}$ is tacitly assumed to have the null form such that all upper probabilities are unity and all lower probabilities are zero. For example, the model asserts that, if a single random individual is observed to belong in category c_1, and no prior knowledge of $\boldsymbol{\pi}$ is assumed, it may be inferred that at least half the population belongs in c_1 with probability between $\frac{1}{4}$ and 1.

A collection of n models $(X^{(i)}, S, \mu^{(i)}, \Gamma^{(i)})$ for $i = 1, 2, \ldots, n$ may be *combined* or *multiplied* to obtain a *product model* (X, S, μ, Γ). The formal definition of (X, S, μ, Γ) is given by

$$\left.\begin{aligned} X &= X^{(1)} \times X^{(2)} \times \ldots \times X^{(n)}, \\ \mu &= \mu^{(1)} \times \mu^{(2)} \times \ldots \times \mu^{(n)} \\ \text{and} \qquad & \\ \Gamma\mathbf{x} &= \Gamma^{(1)}x^{(1)} \cap \Gamma^{(2)}x^{(2)} \cap \ldots \cap \Gamma^{(n)}x^{(n)}, \end{aligned}\right\} \tag{16}$$

where $\mathbf{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})$ denotes a general point of the product space X. The product model is appropriate where the realized values $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ are regarded as independently random according to the probability measures $\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(n)}$, while the logical relationships implied by $\Gamma^{(1)}, \Gamma^{(2)}, \ldots, \Gamma^{(n)}$ are postulated to apply simultaneously to a common realized outcome s in S. It may be helpful to view the models $(X^{(i)}, S, \mu^{(i)}, \Gamma^{(i)})$ as separate sources of information about the unknown s in S. In such a view, if the n sources are genuinely independent, then the product rule (16) represents the legitimate way to pool their information.

The concept of a product model actually includes the concept of a conditioned model which was introduced earlier. Proceeding formally, the information that T occurs with certainty may be represented by a degenerate model (Y, S, ν, Δ), where Y consists of a single point y, while $\Delta y = T$ and y carries ν measure unity. Multiplying a general model (X, S, μ, Γ) by (Y, S, ν, Δ) produces essentially the same result as conditioning the general model (X, S, μ, Γ) on T. For $X \times Y$ and $\mu \times \nu$ are isomorphic in an obvious way to X and μ, while $\Gamma x \cap \Delta y = \Gamma x \cap T = \tilde{\Gamma} x$ as in (10). Thus the objective of taking account of information in the special form of an assertion that T must occur may be reached either through the rule of conditioning or through the rule of multiplication, with identical results. In particular, when $T = S$ the degenerate model (Y, S, ν, Δ) conveys no information about the uncertain outcome s in S, both in the heuristic sense that upper and lower probabilities of non-trivial events are unity and zero, and in the formal sense that combining such a (Y, S, ν, Δ) with any information source (X, S, μ, Γ) leaves the latter model essentially unaltered.

Product models are widely used in mathematical statistics to represent random samples of size n from infinite populations, and they apply directly to provide the general sample size extension of the trinomial sampling model $(U, C \times \Pi, \rho, B)$. A random sample of size n from the population U is represented by $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}$ independently drawn from U according to the same uniform probability measure ρ. More precisely, the sample $(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)})$ is represented by a single random point drawn from the product space

$$U^n = U^{(1)} \times U^{(2)} \times \ldots \times U^{(n)} \tag{17}$$

according to the product measure

$$\rho^n = \rho^{(1)} \times \rho^{(2)} \times \ldots \times \rho^{(n)}, \tag{18}$$

where the pairs $(U^{(1)}, \rho^{(1)}), (U^{(2)}, \rho^{(2)}), \ldots, (U^{(n)}, \rho^{(n)})$ are n identical mathematical copies of the original pair (U, ρ). In a similar way, the observable categories of the n sample individuals are represented by a point in the product space

$$C^n = C^{(1)} \times C^{(2)} \times \ldots \times C^{(n)}, \tag{19}$$

where $C^{(i)}$ is the three-element space from which the observable category $c^{(i)}$ of the sample individual $\mathbf{u}^{(i)}$ is taken. The interesting unknowns before sampling are $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ and $\boldsymbol{\pi}$, which define a point in the space $C^n \times \Pi$. Accordingly, the model which represents a random sample of size n from a trinomial population is of the form $(U^n, C^n \times \Pi, \rho^n, B^n)$, where it remains only to define B^n. In words, B^n is the logical relationship which requires that (7) shall hold for each $\mathbf{u}^{(i)}$. In symbols,

$$B^n \left(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}\right) = B^{(1)}\mathbf{u}^{(1)} \cap B^{(2)}\mathbf{u}^{(2)} \cap \ldots \cap B^{(n)}\mathbf{u}^{(n)}, \tag{20}$$

where $B^{(i)}\mathbf{u}^{(i)}$ consists of those points $(c^{(1)}, c^{(2)}, \ldots, c^{(n)}, \boldsymbol{\pi})$ in $C^n \times \Pi$ such that

$$\pi_k/u_k^{(i)} = \max\left\{\left(\pi_1/u_1^{(i)}\right), (\pi_2/u_2^{(i)}), (\pi_3/u_3^{(i)})\right\} \tag{21}$$

for $k = 1, 2, 3$.

The model $(U^n, C^n \times \Pi, \rho^n, B^n)$ now completely defined provides in itself an illustration of the product rule. For (17), (18) and (20) are instances of the three lines of (16), and hence show that $(U^n, C^n \times \Pi, \rho^n, B^n)$ is the product of the n models $(U^{(i)}, C^n \times \Pi, \rho^{(i)}, B^{(i)})$ for $i = 1, 2, \ldots, n$, each representing an individual sample member.

As in the special case $n = 1$, the model $(U^n, C^n \times \Pi, \rho^n, B^n)$ does not in itself provide interesting upper and lower probabilities. Again, conditioning may be illustrated either by fixing $\boldsymbol{\pi}$ and asking for probability judgments about $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ or conversely by fixing $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ and asking for probability judgments (i.e. generalized Bayesian inferences) about $\boldsymbol{\pi}$. Conditioning on fixed $\boldsymbol{\pi}$ leads easily to the expected generalization of (11). Specifically, if T is the event that $\boldsymbol{\pi}$ has a specified value, while $\tilde{T}$ is the event that $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ are fixed, with n_i observations in category c_i for $i = 1, 2, 3$, then

$$P^*(\tilde{T}|T) = P_*(\tilde{T}|T) = \pi_1^{n_1}\pi_2^{n_2}\pi_3^{n_3}. \tag{22}$$

The converse approach of conditioning on $\tilde{T}$ leads to more difficult mathematics.

Before $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ are observed, the relevant sample space $C^n \times \Pi$ consists of 3^n triangles, each a copy of Π. Conditioning on a set of recorded observations $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ reduces the relevant sample space to the single triangle associated with those observations. Although this triangle is actually a subset of $C^n \times \Pi$, it is essentially the same as Π and will be formally identified with Π for the remainder of this discussion. Conditioning the model $(U^n, C^n \times \Pi, \mu^n, B^n)$ on $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ leads therefore to the model $(U^n, \Pi, \mu^n, \tilde{B}^n)$

where $\tilde{B}^n$ is defined by restricting B^n to the appropriate copy of Π. The important random subset $\tilde{B}^n(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)})$ of Π defined by the random sample $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}$ will be denoted by V for short. V determines the desired inferences, that is, the upper and lower probabilities of a fixed subset R of Π are respectively the probability that V intersects R and the probability that V is contained in R, both conditional on V being non-empty.

V is the intersection of the n random regions $B^{(i)}\mathbf{u}^{(i)}$ for $i = 1, 2, \ldots, n$ where each $B^{(i)}\mathbf{u}^{(i)}$ is one of the three types illustrated on the three levels of Fig. 2, the type and level depending on whether the observation $c^{(i)}$ is c_1, c_2 or c_3. Figure 4 illustrates one such region for $n = 4$. It is easily discovered by experimenting with pictures like Fig. 4 that the shaded region V may have 3,4,5 or 6 sides, but most often is empty. It is shown in Appendix A that V is non-empty with probability $n_1!n_2!n_3!/n!$ under independent uniformly distributed $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}$. Moreover, conditional on non-empty V, six random vertices of V are shown in Appendix A to have Dirichlet distributions. Specifically, define $\mathbf{W}^{(i)}$ for $i = 1, 2, 3$ to be the point $\boldsymbol{\pi}$ in V with maximum coordinate π_i and define $\mathbf{Z}^{(i)}$ for $i = 1, 2, 3$ to be the point $\boldsymbol{\pi}$ in V with minimum coordinate π_i. These six vertices of V need not be distinct, but are distinct with positive probability and so have different distributions. Their distributions are

$$\left.\begin{array}{ll}
\mathbf{W}^{(1)}: & D(n_1+1, n_2, n_3), \\
\mathbf{W}^{(2)}: & D(n_1, n_2+1, n_3), \\
\mathbf{W}^{(3)}: & D(n_1, n_2, n_3+1), \\
\mathbf{Z}^{(1)}: & D(n_1, n_2+1, n_3+1), \\
\mathbf{Z}^{(2)}: & D(n_1+1, n_2, n_3+1), \\
\mathbf{Z}^{(3)}: & D(n_1+1, n_2+1, n_3),
\end{array}\right\} \tag{23}$$

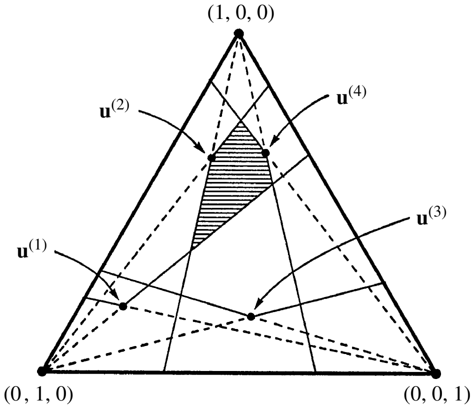


Fig. 4. The triangle Π representing the sample space of unknowns after $n = 4$ observations $c^{(1)} = c_1, c^{(2)} = c_3, c^{(3)} = c_1, c^{(4)} = c_2$ have been taken. The shaded region is the realization of V determined by the illustrated realization of $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{u}^{(3)}$ and $\mathbf{u}^{(4)}$

where $D(r_1, r_2, r_3)$ denotes the Dirichlet distribution over the triangle Π whose probability density function is proportional to

$$\pi_1^{r_1-1}\pi_2^{r_2-1}\pi_3^{r_3-1}.$$

The Dirichlet distribution is defined as a continuous distribution over Π if $r_i > 0$ for $i = 1, 2, 3$. Various conventions, not listed here, are required to cover the distributions of the six vertices when some of the n_i are zero.

Many interesting upper and lower probabilities follow from the distributions (23). For example, the upper probability that π_1 exceeds π_1' is the probability that V intersects the region where $\pi_1 \geq \pi_1'$ which is, in turn, the probability that the first coordinate of $\mathbf{W}^{(1)}$ exceeds π_1'. In symbols,

$$\begin{aligned} P^*(\pi_1 \geq \pi_1' | n_1, n_2, n_3) &= \int_{\pi_1'}^1 \int_0^1 \frac{n!}{n_1!(n_2-1)!(n_3-1)!} \pi_1^{n_1}\pi_2^{n_2-1}\pi_3^{n_3-1} d\pi_1 d\pi_2 \\ &= \int_{n_1{}'}^1 \frac{n!}{n_1!(n_2+n_3-1)!} \pi_1^{n_1}(1-\pi_1)^{n_2+n_3-1} d\pi_1 \end{aligned} \tag{24}$$

if $n_2 > 0$ and $n_3 > 0$. Similarly, $P_*(\pi_1 \geq \pi_1' | n_1, n_2, n_3)$ is the probability that the first coordinate of $\mathbf{Z}^{(1)}$ exceeds π_1', that is,

$$P_*(\pi_1 \geq \pi_1' | n_1, n_2, n_3) = \int_{\pi_1{}'}^1 \frac{(n+1)!}{(n_1-1)!(n_2+n_3+1)!} \pi_1^{n_1-1}(1-\pi_1)^{n_2+n_3+1} d\pi_1, \tag{25}$$

again assuming no prior information about $\boldsymbol{\pi}$. Two further analogues of the pair (24) and (25) may be obtained by permuting the indices so that the role of 1 is played successively by 2 and 3. In a hypothetical numerical example with $n_1 = 2, n_2 = 1, n_3 = 1$ as used in Fig. 4, it is inferred that the probability of at least half the population belonging in c_1 lies between $\frac{3}{16}$ and $\frac{11}{16}$. In passing, note that the upper and lower probabilities (24) and (25) are formally identical with Bayes posterior probabilities corresponding to the pseudo-prior distributions $D(1, 0, 0)$ and $D(0, 1, 1)$, respectively. This appears to be a mathematical accident with a limited range of applicability, much like the relations between fiducial and Bayesian results pointed out by Lindley (1958). In the present situation, it could be shown that the relations no longer hold for events of the form $(\pi_1' \leq \pi_1 \leq \pi_1'')$.

The model $(U^n, C^n \times \Pi, \rho^n, B^n)$ has the illuminating feature of remaining a product model *after* conditioning on the sample observations. Recall that the original model $(U^n, C^n \times \Pi, \rho^n, B^n)$ is expressible as the product of the n models $(U^{(i)}, C^n \times \Pi, \rho^{(i)}, B^{(i)})$ for $i = 1, 2, \ldots, n$. Conditioning the original model on the observations yields $(U^n, \tilde{T}, \rho^n, \tilde{B}^n)$ where, as above, $\tilde{T}$ is the subset of $C^n \times \Pi$ with $c^{(1)}, c^{(2)}, \ldots, c^{(n)}$ fixed at their observed values and

$$\tilde{B}^n(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}) = B^n(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}) \cap \tilde{T}. \tag{26}$$

Conditioning the ith component model on the ith sample observation yields $(U^{(i)}, \tilde{T}^{(i)}, \rho^{(i)}, \tilde{B}^{(i)})$, where $\tilde{T}^{(i)}$ is the subset of $C^n \times \Pi$ with $c^{(i)}$ fixed at its observed value, and

$$\tilde{B}^{(i)}\mathbf{u}^{(i)} = B^{(i)}\mathbf{u}^{(i)} \cap \tilde{T}^{(i)}, \tag{27}$$

for $i = 1, 2, \ldots, n$. It is clear that

$$\tilde{T} = \tilde{T}^{(i)} \cap \tilde{T}^{(2)} \cap \ldots \cap \tilde{T}^{(n)}, \tag{28}$$

and from (20), (26), (27) and (28) it follows that

$$\tilde{B}^n(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}) = \tilde{B}^{(1)}\mathbf{u}^{(1)} \cap \tilde{B}^{(2)}\mathbf{u}^{(2)} \cap \ldots \cap \tilde{B}^{(n)}\mathbf{u}^{(n)}. \tag{29}$$

From (28) and (29) it is immediate that the model $(U^n, \tilde{T}, \rho^n, \tilde{B}^n)$ is the product of the n models $(U^{(i)}, \tilde{T}^{(i)}, \rho^{(i)}, \tilde{B}^{(i)})$ for $i = 1, 2, \ldots, n$. The meaning of this result is that inferences about π may be calculated by traversing two equivalent routes. First, as above, one may multiply the original n models and condition the product on $\tilde{T}$. Alternatively, one may condition the original n models on their associated $\tilde{T}^{(i)}$ and then multiply the conditioned models. The availability of the second route is conceptually interesting, because it shows that the information from the ith sample observation $c^{(i)}$ may be isolated and stored in the form $(U^{(i)}, \tilde{T}^{(i)}, \rho^{(i)}, \tilde{B}^{(i)})$, and when the time comes to assemble all the information one need only pick up the pieces and multiply them. This basic result clearly holds for a wide class of choices of U and B, not just the particular trinomial sampling model illustrated here.

The separability of sample information suggests that prior information about π should also be, stored as a model of the general type (X, Π, μ, Γ) and should be combined with sample information according to the product rule. Such prior information could be regarded as the distillation of previous empirical data. This proposal brings out the full dimensions of the generalized Bayesian inference scheme. Not only does the product rule show how to combine individual pieces of sample information: it handles the incorporation of prior information as well. Moreover, the sample information and the prior information are handled symmetrically by the product rule, thus banishing the asymmetric appearance of standard Bayesian inference. At the same time, if the prior information is given in the standard form of an ordinary probability distribution, the methods of generalized Bayesian inference reproduce exactly the standard Bayesian inferences.

A proof of the last assertion will now be sketched in the context of trinomial sampling. An ordinary prior distribution for an unknown π is represented by a model of the form (X, Π, μ, Γ) where Γ is single-valued and hence no ambiguity is allowed in the computed probabilities. Without loss of generality, the model (X, Π, μ, Γ) may be specialized to (Π, Π, μ, I), where I is the identity mapping and μ is the ordinary prior distribution over Π. For simplicity, assume that μ is a discrete distribution with probabilities $p_1, p_2, \ldots, p_d$ assigned to points $\pi_1, \pi_2, \ldots, \pi_d$ in Π. From (16) it follows that the mapping associated with a product of models is single-valued if the mapping associated with any component model is single-valued. If a component model not only has a single-valued mapping, but has a discrete measure μ as well, then the product model is easily seen to reduce to another discrete distribution over the same carriers $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_d$. Indeed the second line of (16) shows that the product model assigns probabilities $P(\boldsymbol{\pi}_i)$ to $\boldsymbol{\pi}_i$ which are proportional to $p_i l_i$, where l_i is the probability that the random region V includes the point $\boldsymbol{\pi}_i$. Setting $\boldsymbol{\pi_i} = (\boldsymbol{\pi_{i1}, \pi_{i2}, \pi_{i3}})$, it follows from the properties of the random region V that

$$l_i = \pi_{i1}^{n_1} \pi_{i2}^{n_2} \pi_{i3}^{n_3}, \tag{30}$$

which is just the probability that all of the independent random regions whose intersection is V include $\boldsymbol{\pi}_i$. Normalizing the product model as indicated in (1) or (2) leads finally to

$$P(\boldsymbol{\pi}_i) = \frac{p_i l_i}{p_1 l_1 + p_2 l_2 + \ldots + p_d l_d} \tag{31}$$

for $i = 1, 2, \ldots, d$, which is the standard form of Bayes's theorem. This result holds for any choices of U and B satisfying (9). Note that l_i is identical with the likelihood of $\boldsymbol{\pi}_i$.

Generalized Bayesian inference permits the use of sample information alone, which is mathematically equivalent to adopting the informationless prior model in which all upper probabilities are unity and all lower probabilities are zero. At another extreme, it permits the incorporation of a familiar Bayesian prior distribution (if it is a genuine distribution) and then yields the familiar Bayesian inferences. Between these extremes a wide range of flexibility exists. For example, a prior distribution could be introduced for the coordinate π_1 alone, while making no prior judgment about the ratio π_2/π_3. Alternatively, one could specify prior information to be the same as that contained in a sample of size m which produced m_i observations in category c_i for $i = 1, 2, 3$. In the analysis of quite small samples, it would be reasonable to attempt to find some characterization of prior information which could reflect tolerably well public notions about $\boldsymbol{\pi}$. In large samples, the inferences clearly resemble Bayesian inferences and are insensitive to prior information over a wide range.

3 A Second Illustration

Consider a sequence of independent Bernoulli trials represented by z_i with

$$P(z_i = 1|p_i) = p_i \text{ and } P(z_i = 0|p_i) = 1 - p_i, \text{ for } i = 1, 2, \cdots, n, \tag{32}$$

where it is suspected that the sequence p_i is subject to a monotone upward drift. In this situation, the common approach to a sequence of observations

z_i is to apply a test of the null hypothesis $\{p_1 = p_2 = \ldots = p_n\}$ designed to be sensitive against the alternative hypothesis $\{p_1 \leq p_2 \leq \ldots \leq p_n\}$. The unorthodox approach suggested here is to compute upper and lower probability inferences for the pair of symmetric hypotheses $\{p_1 \geq p_2 \geq \ldots \geq p_n\}$ and $\{p_1 \leq p_2 \leq \ldots \leq p_n\}$ under the overall prior assumption that the sequence p_i is monotone, either increasing or decreasing, with probability one. A small upper probability for either of these hypotheses would be evidence for drift in the direction contrary to that indicated by the hypothesis. Upper and lower probabilities may also be computed for the null hypothesis $\{p_1 = p_2 = \ldots = p_n\}$, but the upper probability will usually be vanishingly small in sample sequences of moderate length however little trend is apparent, while the lower probability is always zero.

The model described could apply in simple bioassays or learning situations. A wider range of applications could be achieved in several ways, for example by allowing several observations at each p_i or postulating Markov-type dependence in the z_i sequence. But the aim here is to focus attention as simply as possible on one feature of the new methods, namely their ability to handle the problem of many nuisance parameters which plagues the more traditional forms of statistical inference. Plausible inferences may be obtained despite the presence of as many continuous parameters as there are dichotomous observables.

Under the binomial analogue of the trinomial model treated in Sect. 2, a single binomial observable z is represented before observation by the model $(U, Z \times P, \rho, B)$ where

$$U = \{u : 0 \leq u \leq 1\}\,, \tag{33}$$

$$Z = \{z : z = 0 \text{ or } z = 1\}\,, \tag{34}$$

$$P = \{p : 0 \leq p \leq 1\}\,, \tag{35}$$

ρ is the uniform distribution over U, and

$$\begin{aligned} Bu = \{(z,p) : z = 0 \text{ and } u \leq p \leq 1, \text{or}\} \\ z = 1 \text{ and } 0 \leq p \leq u\}. \end{aligned} \tag{36}$$

After conditioning on z, this model becomes effectively (U, P, ρ, B_z), where

$$\begin{aligned} B_z u &= \{p : u \leq p \leq 1\} \text{ if } z = 0, \\ &= \{p : 0 \leq p \leq u\} \text{ if } z = 1. \end{aligned} \tag{37}$$

A conditioned model of this kind may be constructed for each of n independent observations z_i and associated parameters p_i. Combining these n sources of information about $p_1, p_2, \ldots, p_n$ yields a single model $(U^n, P^n, \rho^n, B_{(z_1, z_2, \ldots, z_n)})$, where

$$U^n = \{(u_1, u_2, \ldots, u_n) : 0 \leq u_i \leq 1 \text{ for } i = 1, 2, \ldots, n\}\,, \tag{38}$$

$$P^n = \{(p_1, p_2, \ldots, p_n) : 0 \leq p_i \leq 1 \text{ for } i = 1, 2, \ldots, n\}\,, \tag{39}$$

ρ^n is the uniform distribution over the cube U, and

$$B_{(Z_1,z_2,\ldots,z_n)}(u_1,u_2,\ldots,u_n) = \{(p_1,p_2,\ldots,p_n) : p_i \in B_{z_i}u_i \text{ for } i = 1,2,\ldots,n\}\,. \tag{40}$$

The combined model would be appropriate for unrestricted inferences about an unknown $(p_1,p_2,\ldots,p_n)$ based on observations $(z_1,z_2,\ldots,z_n)$. However, when consideration is restricted to the subset S of P^n in which $p_1,p_2,\ldots,p_n$ is a monotone sequence, the sharpness of the inferences is much improved.

Define T_1 and T_2 to be the subsets of S for which $p_1 \le p_2 \le \ldots \le p_n$ and $p_1 \ge p_2 \ge \ldots \ge p_n$, respectively. Define $T_{12} = T_1 \cap T_2$ to be the subset of S for which $p_1 = p_2 = \ldots = p_n$. An immediate objective is to characterize T_1^*, T_2^* and T_{12}^*, from whose ρ^n measure the desired inferences will follow. For example, T_1^* consists of all points $(u_1,u_2,\ldots,u_n)$ for which there exists some $(p_1,p_2,\ldots,p_n)$ satisfying $p_1 \le p_2 \le \ldots \le p_n$ and such that p_i lies in $B_{z_i}u_i$, for $i = 1,2,\ldots,n$. With the help of Fig. 5 it is easily checked that

$$T_1^* = \{(u_1,u_2,\ldots,u_n) : u_i \le u_j, \text{ whenever } z_i = 1, z_j = 0, i < j\}\,. \tag{41}$$

By symmetry,

$$T_2^* = \{(u_1,u_2,\ldots,u_n) : u_i \ge u_j, \text{ whenever } z_i = 0, z_j = 1, i < j\}\,. \tag{42}$$

Finally,

$$T_{12}^* = \{(u_1,u_2,\ldots,u_n) : u_i \le u_j, \text{ whenever } z_i = 1, z_j = 0\}\,. \tag{43}$$

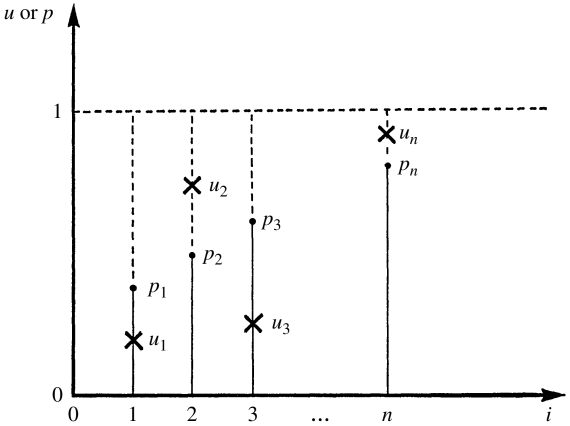


Fig. 5. The plotted values $p_1,p_2,\ldots,p_n$ determine a point P^n for which $p_1 \le p_2 \le \ldots \le p_n$ The plotted values $u_1,u_2,\ldots,u_n$ determine a point of U^n for which p_1 lies in $B_1\ z_1, p_2$ lies in $B_0 z_2, p_3$ lies in $B_1\ z_3,\ldots,p_n$ lies in $B_0 z_n$. The interpretation is that $(u_1,u_2,\ldots,u_n)$ lies in the region T_1^* determined by the observation $z_1 = 1, z_2 = 0, z_3 = 1,\ldots,z_n = 0$

It is clear that $T_{12}^* = T_1^* \cap T_2^*$ and that $T_{12}^*, T_1^* - T_{12}^*$ and $T_2^* - T_{12}^*$ are disjoint sets whose union is S^*.

U^n may be decomposed into $n!$ geometrically similar simplexes, each characterized by a particular ordering of the values of the coordinates $(u_1, u_2, \ldots, u_n)$. These simplexes are in one-to-one correspondence with the permutations

$$(1, 2, \cdots, n) \to (1^*, 2^*, \ldots, n^*),$$

where for every $(u_1, u_2, \ldots, u_n)$ in a given simplex the corresponding permutation obeys $u_{1^*} \leq u_{2^*} \leq \ldots \leq u_{n^*}$. Since the characterizations (41), (42) and (43) involve only order relations among coordinates u_i, each of the simplexes is either included or excluded as a unit from T_1^* or T_2^* or T_{12}^*. And since each of the $n!$ simplexes has ρ^n measure $1/n!$, the ρ^n measures of T_1^* or T_2^* or T_{12}^* may be found by counting the appropriate number of simplexes and dividing by $n!$. Or, instead of counting simplexes, one may count the permutations to which they correspond. The permutation

$$(1, 2, \ldots, n) \to (1^*, 2^*, \cdots, n^*)$$

carries the observed sequence $(z_1, z_2, \ldots, z_n)$ of zeros and ones into another sequence $(z_{1^*}, z_{2^*}, \ldots, z_{n^*})$ of zeros and ones. According to the definition of T_1^*, a simplex is contained in T_1^* if and only if its corresponding permutation has the property that $i^* < j^*$ for all $i < j$ such that $z_i = 1$ and $z_j = 0$, i.e. any pair ordered (1,0) extracted from $(z_1, z_2, \ldots, z_n)$ must retain the same order in the permuted sequence $(z_{1^*}, z_{2^*}, \ldots, z_{n^*})$. Similarly, to satisfy T_2^* any pair ordered (0,1) extracted from $(z_1, z_2, \ldots, z_n)$ must have its order reversed in the permuted sequence, while to satisfy $T_{12}^* = T_1^* \cap T_2^*$ the sequence $(z_{1^*}, z_{2^*}, \ldots, z_{n^*})$ must consist of all ones followed by all zeros.

If $(z_1, z_2, \ldots, z_n)$ contains n_1 ones and n_2 zeros, then a simple counting of permutations yields

$$\rho(T_{12}^*) = \frac{n_1! n_2!}{n!} \tag{44}$$

A simple iterative procedure for computing $\rho^n(T_1^*)$ or $\rho^n(T_2^*)$ is derived in Appendix B by Herbert Weisberg. The result is quoted below and illustrated on a numerical example.

For a given sequence of observations $z_1, z_2, \ldots$ of indefinite length define $N(n)$ to be the number of permutations of the restricted type counted in T_1^*. $N(n)$ may be decomposed into

$$N(n) = \sum_{k=0}^{r} N(k, n), \tag{45}$$

where $N(k, n)$ counts the subset of permutations such that $(z_{1^*}, z_{2^*}, \ldots, z_{n^*})$ has k zeros preceding the rightmost one. Since no zero which follows the rightmost one in the original sequence $(z_1, z_2, \ldots, z_n)$ can be permuted to the left of any one under any allowable permutation, the upper limit r in (45) may be taken as the number of zeros preceding the rightmost one in the original sequence $(z_1, z_2, \ldots, z_n)$. In the special case of a sequence consisting entirely of zeros, all of the zeros will be assumed to follow the rightmost one so that $N(k,n) = 0$ for $k > 0$ and indeed $N(n) = N(0,n) = n!$. Weisberg's iterative formula is

$$\begin{aligned} N(k,n+1) &= \sum_{j=0}^{k-1} N(j,n) + (n_1 + 1 + k)N(k,n) \quad \text{if } z_{n+1} = 1 \\ &= (n_2 + 1 - k)N(k,n). \quad \text{if } z_{n+1} = 0, \end{aligned} \tag{46}$$

where n_1 and n_2 denote as above the numbers of ones and zeros, respectively, in $(z_1, z_2, \ldots, z_n)$.

Formula (46) has the pleasant feature that the counts for the sequences $(z_1), (z_1, z_2), (z_1, z_2, z_3), \ldots$ may be built up successively, and further observations may be easily incorporated as they arrive. Consider, for example, the hypothetical observations

$$(z_1, z_2, \ldots, z_7) = (0, 0, 1, 1, 0, 1, 1).$$

Table 1 shows

$$z_n, N(0,n), \ldots, N(r,n)$$

on line n, for $n = 1, 2, \ldots, 7$, from which $N(7) = 1680$. The number of permutations consistent with T_2^* is found by applying the same iterative process to the sequence (1,1,0,0,1,0,0) with zeros and ones interchanged. This yields Table 2 from which $N(7) = 176$. The number of permutations common to T_1^* and T_2^* is 3! 4! = 144. Thus $\rho^n(T_1^*) = 1680/7!, \rho^n(T_2^*) = 176/7!, \rho^n(T_{12}^*) = 144/7!$, and $\rho^n(S^*) = (1680 + 176 - 144)/7! = 1712/7!$. Consequently, the upper and lower probabilities of T_1, T_2 and T_{12} conditional on S and $(z_1, z_2, \ldots, z_7) = (0, 0, 1, 1, 0, 1, 1)$ are

$$P^*(T_1) = \frac{1680}{1712},\ p_*(T_1) = \frac{1536}{1712},\ P^*(T_2) = \frac{176}{1712},\ P_*(T_2) = \frac{32}{1712},$$
$$P^*(T_{12}) = \frac{144}{1712},\ p_*(T_{12}) = 0.$$

Table 1.

n	z_n	$N(0,n)$	$N(1,n)$	$N(2,n)$	$N(3,n)$
1	0	1			
2	0	2			
3	1	2	2	2	
4	1	4	8	12	
5	0	12	16	12	
6	1	36	76	84	40
7	1	144	416	640	480

Table 2.

n	z_n	$N(0,n)$	$N(1,n)$	$N(2,n)$
1	1	1		
2	1	2		
3	0	2		
4	0	4		
5	1	12	4	4
6	0	36	8	4
7	0	144	24	8

Since more than 10% of the measure could apply to a monotone non-increasing sequence, the evidence for an increasing sequence is not compelling.

For the extended sequence of observations 0,0,1,1,0,1,1,0,1,1,1,. . ., the lower and upper probabilities of a monotone downward sequence after n observations are exhibited in Table 3.

Table 3.

n	$P_*(T_2)$	$P^*(T_2)$
1	0	1
2	0	1
3	0	0.333
4	0	0.167
5	0.167	0.417
6	0.048	0.190
7	0.019	0.103
8	0.188	0.319
9	0.065	0.148
10	0.028	0.080
11	0.014	0.047

4 Comments on the Method of Generating Upper and Lower Probabilities

Although often notationally convenient, it is unnecessary to use models (X, S, μ, Γ) outside of the subclass where the inverse of Γ is single-valued. For the model $(X, \tilde{S}, \mu, \tilde{\Gamma})$ with

$$\tilde{\mathrm{S}} = X \times S \tag{47}$$

and

$$\tilde{\Gamma}x = \{x\} \times \Gamma x \tag{48}$$

does belong to the stated subclass, and yields

$$(P^*(T), P_*(T)) = (\tilde{P}^*(X \times T)\tilde{P}_*(X \times T)) \tag{49}$$

for any $T \subset S$, where the left side of (49) refers to any original model (X, S, μ, Γ) and the right side refers to the corresponding model $(X, \tilde{S}, \mu, \tilde{\Gamma})$. Moreover, the model $(X, \tilde{S}, \mu, \tilde{\Gamma})$ provides upper and lower probabilities for all subsets of $X \times S$, not just those of the form $X \times T$. On the other hand, it was assumed in applying the original form (X, S, μ, Γ) that the outcome x in X is conceptually unobservable, so that no operational loss is incurred by the restriction to subsets of the form $X \times T \subset \tilde{S}$.

Underlying the formalism of (X, S, μ, Γ) or its equivalent $(X, \tilde{S}, \mu, \tilde{\Gamma})$ is the idea of a probability model which assigns a distribution only over a partition of a complete sample space, specifically the distribution μ over the partition of $\tilde{S} = X \times S$ defined by X. Thus the global probability law of an ordinary probability measure space is replaced by a marginal distribution or what might be called a *partial* probability law. The aim therefore is to establish a useful probability calculus on marginal or partial assumptions.

I believe that the most serious challenges to applications of the new calculus will come not from criticism of the logic but from the strong form of ignorance which is necessarily built into less-than-global probability laws. To illustrate, consider a simple example where w_1 denotes a measured weight, w_2 denotes a true weight, and $x = w_1 - w_2$ denotes a measurement error. Assume that ample relevant experience is available to justify assigning a specific error distribution μ over the space X of possible values of x. The situation may be represented by the model (X, W, μ, Γ) with X and μ as defined, with $W = \{(w_1, w_2); w_1 \geq 0, w_2 \geq 0\}$, and Γ defined by the relation $x = w_1 - w_2$. Conditioning the model on an observed w_1 leaves one with the same measure μ applied to $w_1 - w_2$, except for renormalization which restricts the measure to $w_1 \geq 0$. The result is very much in the spirit of the fiducial argument (although there is some doubt about Fisher's attitude to renormalization). I am unable to fault the logic of this fiducial-like argument. Rather, some discomfort is produced by distrust of the initial model, in particular by its implication that every uncertain event governed by the true weight w_2 has initial upper and lower probabilities one and zero. It would be hard to escape a feeling in most real situations that a good bit of information about a parameter is available, even if difficult to formalize objectively, and that such information should clearly alter the fiducial-like inference if it could be incorporated. One way to treat this weakness is openly to eschew the use of prior information, while not necessarily denying its existence, that is, to assert that the statistician should summarize only that information which relies on the observation w_2 and the objectively based error distribution μ. Because of the conservatism implicit in the definition of upper and lower probabilities, the approach of rejecting soft information seems likely to provide conservative inferences on an average, but I have not proved theorems to this effect. The difficulty is that the rejection of all soft information, including even information about parametric forms, may lead to unrealistically weak inferences. The alternative approach is to promote vague information into as precise a model as one dares and combine it in the usual way with sample information.

Some comments on the mathematics of upper and lower probabilities are appropriate. A very general scheme for assigning upper and lower probabilities to the subsets of a sample space S is to define a family $\mathcal{C}$ of measures P over S and to set

$$P^*(T) = \sup_{\mathcal{C}} P(T), \quad P_*(T) = \inf_{\mathcal{C}} P(T). \tag{50}$$

Within the class of systems of upper and lower probabilities achieved in this way for different $\mathcal{C}$, there is a hierarchical scheme of shrinking subclasses ending with the class of systems defined by models like (X, S, μ, Γ). (See Dempster, 1967a.). The family $\mathcal{C}$ corresponding to a given (X, S, μ, Γ) consists of all measures P which for each x distribute the probability element $d\mu(x)$ in some way over Γx. Some readers may feel that all systems should be allowed, not just the subclass of this paper. In doing so, however, one loses the conception of a source of information as being a single probability measure. For, in the unrestricted formulation of (50), the class $\mathcal{C}$ consists of conceptually distinct measures such as might be adopted by a corresponding class of personalist statisticians, and the conservatism in the bounds of (50) amounts to an attempt to please both extremes in the class of personalist statisticians. I believe that the symmetry arguments underlying probability assignments do not often suggest hypothetical families $\mathcal{C}$ demanding simultaneous satisfaction. Also, the rules of conditioning and, more generally, of combination of independent sources of information do not extend to the unrestricted system (50), and without these rules the spirit of the present approach is lost.

The aim of this short section has been to suggest that upper and lower probabilities generated by multivalued mappings provide a flexible means of characterizing limited amounts of information. They do not solve the difficult problems of what information should be used, and of what model appropriately represents that information. They do not provide the only way to discuss meaningful upper and lower probabilities. But they do provide an approach with a well-rounded logical structure which applies naturally in the statistical context of drawing inferences from samples to populations.

5 Comments on the Models Used for Inference

The models used here for the representation of sampling from a population take as their point of departure a space whose elements correspond to the members of the population. In addition to the complex of observable characteristics usually postulated in mathematical statistics, each population member is given an individual identity. In conventional mathematical statistics the term *hypothesis* is often used for an unknown population distribution of observable characteristics, but the presence of the population space in the model leads directly to the more fundamental question of how each hypothesized population distribution applies to the elements of the population space, that is, under a given hypothesis what are the observable characteristics of each population member? In the trinomial sampling model of Sect. 2, the question is answered by the multivalued mapping B defined in (7). As illustrated in Fig. 1, B asserts that for each hypothesis $\boldsymbol{\pi}$ the population space U partitions into three regions U_1, U_2, U_3 corresponding to the observable characteristics c_1, c_2, c_3. More generally, the observable characteristics may be multinomial with k categories $c_1, c_2, \ldots, c_k$ and the population space U may be any space with an associated random sampling measure ρ. For a given hypothesis $\boldsymbol{\pi} = (\boldsymbol{\pi_1, \pi_2, \ldots, \pi_k})$ the question is answered by determining subsets $U_1, U_2, \ldots, U_k$ of U which specify that a population member in U_i is permitted to have characteristic c_i under $\boldsymbol{\pi}$, for $i = 1, 2, \ldots, k$. Having reached this point in building the model, it seems reasonable to pose the restriction which generalizes (9), namely,

$$\rho(U_i) = \pi_i \text{ and } \rho(U_i \cap U_j) = 0 \tag{51}$$

for $i, j = 1, 2, \ldots, k$ and $i \neq j$. The reason for (51) as with (9) is simply to have π_i represent both upper and lower probabilities of c_i for a single drawing with a given $\boldsymbol{\pi}$.

Now it is evident that the above information by no means uniquely determines a model for multinomial sampling. Indeed, one may start from any continuous space U with measure ρ, and for each $\boldsymbol{\pi}$ specify a partition $U_1, U_2, \ldots, U_k$ satisfying (51) but otherwise completely arbitrary. In other words, there is a huge bundle of available models. In Dempster (1966), two choices were offered which I called *models of the first kind* and *models of the second kind.* The former assumes that the multinomial categories $c_1, c_2, \ldots, c_k$ have a meaningful order, and is uniquely determined by the assumption that the population members have an order consistent with the order of their observable characteristics under any hypothesis $\boldsymbol{\pi}$. (See Dempster, 1967b.) The restriction to ordered categories implies essentially a univariate characteristic, and because that restriction is so severe the following discussion is mostly aimed at a general multinomial situation with no mathematical structure assumed on the space of k categories. The general model of the second kind is defined by extending (5), (6) and (7) in the obvious way from $k = 3$ to general k. This model treats the k categories with complete symmetry, but it is not the only model to do so, for one can define $\mathbf{B}^{-1}$ arbitrarily for $\boldsymbol{\pi}$ such that $\pi_1 \leq \pi_2 \leq \ldots \leq \pi_k$, and define $\mathbf{B}^{-1}$ for other $\boldsymbol{\pi}$ by symmetry. But the general model of the second kind is strikingly simple, and **I** recommend it because I can find no competitor with comparable aesthetic appeal.

The status of generalized Bayesian inference resembles that of Bayesian inference in the time of Bayes, by which I mean that Bayes must have adopted a uniform prior distribution because no aesthetically acceptable competitor came to mind. The analogy should be carried further, for even the principles by which competitors should be judged were not formulated by Bayes, nor have the required principles been well formulated for the models discussed here. I believe that the principles required by the two situations are not at all analogous, for the nature and meaning of a prior distribution has become quite clear over the last two centuries and the concept may be carried more or less whole over to generalized Bayesian inference. The choice of a model satisfying (51), on the other hand, has no obvious connection with prior information as the term is commonly applied relative to information about postulated unknowns. In the case of generalized Bayesian inference, I believe the principles for choosing a model to be closely involved with an *uncertainty principle* which can be stated loosely as: *The more information which one extracts from each sample individual in the form of observable characteristics, the less information about any given aspect of the population distribution may be obtained from a random sample of fixed size.* For example, a random sample of size $n = 1000$ from a binomial population yields quite precise and nearly objective inferences about the single binomial parameter p involved. On the other hand, if a questionnaire given to a sample of $n = 1000$ has been sufficient to identify each individual with one of 1,000,000 categories, then it may be foolhardy to put much stock in the sample information about a binomial p chosen arbitrarily from among the $2^{1,000,000} - 2$ non-trivial available possibilities. Conceptually, at least, most real binomial situations are of the latter kind, for a single binomial categorization can be achieved only at the expense of suppressing a large amount of observable information about each sample individual. The uncertainty principle is therefore a specific instance of the general scientific truism that an investigator must carefully delimit and specify his area of investigation if he is to learn anything precise.

Generalized Bayesian inference makes possible precise formulations of the uncertainty principle. For example, the model of the second kind with $k = 2$ and $n = 1000$ yields inferences which most statisticians would find nearly acceptable for binomial sampling. On the other hand, it is a plausible conjecture that the model of the second kind with $k = 1,000,000$ and $n = 1000$ would yield widely separated upper and lower probabilities for most events. The high degree of uncertainty in each inference compensates for the presence of a large number of nuisance parameters, and protects the user against selection effects which would produce many spurious inferences. Use of the model of the first kind with $k = 1,000,000$ and $n = 1000$ would very likely lead to closer bounds than the model of the second kind for binomial inferences relating to population splits in accord with the given order of population members. And it is heuristically clear that models could be constructed which for each π would place each point of U in each of $U_1, U_2, \ldots, U_k$ as π^* varies over an arbitrarily small neighbourhood about π. Such a model would present an extreme of uncertainty, for all upper and lower probability inferences would turn out to be one and zero, respectively. It is suggested here that the choice of a model can only be made with some understanding of the specific reflections of the uncertainty principle which it provides. For the time being, I judge that the important task is to learn more about the inferences yielded by the aesthetically pleasing models of the second kind. Eventually, intuition and experience may suggest a broader range of plausible models.

Models of the second kind were introduced above for sampling from a general multinomial population with k categories and unknown $1 \times k$ parameter vector $\boldsymbol{\pi}$. But the range of application of these models is much wider. First, one may restrict $\boldsymbol{\pi}$ to parametric hypotheses of the general form $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta}, \boldsymbol{\phi}, \ldots)$. More important, the multinomial may be allowed to have an infinite number of categories, as explained in Dempster (1966), so that general spaces of discrete and continuous observable characteristics are permissible. It is possible therefore to handle the standard parametric hypotheses of mathematical statistics. Very few of these have as yet proved analytically tractable.

At present, mainly qualitative insights are available into the overview of statistical inference which the sampling models of generalized Bayesian inference make possible. Some of these insights have been mentioned above, such as the symmetric handling of prior and sample information, and the uncertainty principle by which upper and lower probabilities reflect the degree of confusion produced by small samples from complex situations. It is interesting to note also that parametric hypotheses and prior distributions, which are viewed as quite different in conventional statistical theory, play indistinguishable roles in the logical machinery of generalized Bayesian inference. For a parametric hypothesis such as $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta}, \boldsymbol{\phi}, \ldots)$ may be represented by a model of the general type (X, S, μ, Γ), which assigns all of its probability ambiguously over the subset of $\boldsymbol{\pi}$ allowed by $\boldsymbol{\pi}(\boldsymbol{\theta}, \boldsymbol{\phi}, \ldots)$ as $\theta, \phi, \ldots$ range over their permitted values, and this model combines naturally with sample information using the rule of combination defined in Sect. 2 and suggested there to be appropriate for the introduction of prior information.

Concepts which appear in standard theories of inference may reappear with altered roles in generalized Bayesian inference. *Likelihood* is a prime example. The ordinary likelihood function $L(\boldsymbol{\pi})$ based on a sample from a general multinomial population is proportional to the upper probability of the hypothesis $\boldsymbol{\pi}$. This may be verified in the trinomial example of Sect. 2 by checking that the random region illustrated in Fig. 4 covers the point $\boldsymbol{\pi}$ with probability $\pi_1^{n_1}\pi_2^{n_2}\pi_3^{n_3}$. The general result is hardly more difficult to prove. Now the upper probability of $\boldsymbol{\pi}$ for all $\boldsymbol{\pi}$ does not contain all the sample information under generalized Bayesian inference. Thus the likelihood principle fails in general, and the usual sets of sufficient statistics under exponential families of parametric hypotheses no longer contain all of the sample information. The exception occurs in the special case of ordinary Bayesian inference with an ordinary prior distribution, as illustrated in (31). Thus the failure of the likelihood principle is associated with the uncertainty which enters when upper and lower probabilities differ. In passing, note that marginal likelihoods are defined in the general system, that is, the upper probabilities of specific values of θ from a set of parameters $\theta, \phi, \ldots$ are well defined and yield a function $L(\theta)$ which may be called the marginal likelihood of θ alone. If the prior information consists of an ordinary prior distribution of θ alone, with no prior information about the nuisance parameters, then $L(\theta)$ contains all of the sample information about θ.

Unlike frequency methods, which relate to sequences of trials rather than to specific questions, the generalized Bayesian inference framework permits direct answers to specific questions in the form of probability inferences. I find that significance tests are inherently awkward and unsatisfying for questions like that posed in the example of Sect. 4, and the main reason that Bayesian inference has not replaced most frequency procedures has been the stringent requirement of a precise prior distribution. I hope that I have helped to reduce the stringency of that requirement.

References

Boole, G. (1854). *An Investigation of the Laws of Thought.* New York: Reprinted by Dover (1958).

Dempster, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Ann. Math. Statist.*, **37**, 355–74.

—— (1967a). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.*, **38**, 325–39.

—— (1967b). Upper and lower probability inferences based on a sample from a finite univariate population. *Biometrika*, **54**, 515–528.

Good, I. J. (1962). The measure of a non-measurable set. *Logic, Methodology and Philosophy of Science* (edited by Ernest Nagel, Patrick Suppes and Alfred Tarski), pp. 319–329. Stanford University Press.

Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. *J. R. Statist. Soc.* B, **20**, 102–107.

Smith, C. A. B. (1961). Consistency in statistical inference and decision (with discussion). *J. R. Statist. Soc.* B, **23**, 1–25.

—— (1965). Personal probability and statistical analysis (with discussion). *J. R. Statist. Soc.* A, **128**, 469–499.

Appendix A

A derivation is sketched here for the distributions (23) relating to specific vertices of the random region R defined by (20). R is the intersection of n regions $B^{(i)}\mathbf{u}^{(i)}$, for $i = 1, 2, \ldots, n$, as illustrated in Fig. 4. The region $B^{(i)}\mathbf{u}^{(i)}$ corresponding to $\mathbf{u}^{(i)}$, which gives rise to an observation c_1, consists of points $\mathbf{u}$ such that $u_3/u_1 \leq u_3^{(i)}/u_1^{(i)}$ and $u_2/u_1 \leq u_2^{(i)}/u_1^{(i)}$. The intersection of the n_1 regions corresponding to the n_1 observations c_1 is a region R_1 consisting of points $\mathbf{u}$ such that

$$u_3/u_1 \leq c_{13} \text{ and } u_2/u_1 \leq c_{12}, \tag{A.1}$$

where $c_{13} = \min(u_3^{(i)}/u_1^{(i)})$ and $c_{12} = \min(u_2^{(i)}/u_1^{(i)})$, the minimization being over the subset of i corresponding to observations c_1. Note that R_1 together with the n_1 regions which define it are all of the type pictured on level 1 of Fig. 2. By permuting subscripts, define the analogous regions R_2 with coordinates c_{23}, c_{21} and R_3 with coordinates c_{31}, c_{32}, where R_2 and R_3 are of the types pictured on levels 2 and 3 of Fig. 2, respectively. One is led thus to the representation

$$R = R_1 \cap R_2 \cap R_3. \tag{A.2}$$

Any particular instance of the region R which contains at least one point is a closed polygon whose sides are characterized by fixed ratios of pairs of coordinates u_i, u_j. Thus R may be described by a set of six coordinates

$$b_{ij} = \max_{\mathbf{u} \in R} (u_j / u_i) \tag{A.3}$$

for $i \neq j$. From (A.1), (A.2), and (A.3) it follows that

$$b_{ij} \leq c_{ij} \tag{A.4}$$

for $i \neq j$. Moreover, equality holds if the corresponding side of R_i is also a side of R, while the inequality is strict if the side of R_i misses R entirely. The reader may wish to satisfy himself that R may have 3, 4, 5 or 6 sides in which case the strict inequality in (A.4) holds for 3, 4, 5 or 6 pairs i, j (with probability one).

If R is considered a random region, while R^0 is a fixed region of the same type with coordinates b_{ij}^0, then

$$\begin{aligned} \mathrm{P}(R \supset R^0) &= \mathrm{P}(b_{ij} \geq b_{ij}^0) \text{ for all } i \neq j \\ &= (1 + b_{12}^0 + b_{13}^0)^{-n_1} (1 + b_{21}^0 + b_{23}^0)^{-n_2} (1 + b_{31}^0 + b_{32}^0)^{-n_3}. \end{aligned} \tag{A.5}$$

To prove (A.5) note first that the three events

$$\left\{ b_{12} \geq b_{12}^0, b_{13} \geq b_{13}^0 \right\}, \quad \left\{ b_{21} \geq b_{21}^0, b_{23} \geq b_{23}^0 \right\}, \quad \left\{ b_{31} \geq b_{31}^0, b_{32} \geq b_{32}^0 \right\}$$

are equivalent respectively to the three events

$$\left\{ c_{12} \leq b_{12}^0, c_{13} \geq b_{13}^0 \right\}, \quad \left\{ c_{21} \geq b_{21}^0, c_{23} \geq b_{23}^0 \right\}, \quad \left\{ c_{31} \geq b_{31}^0, c_{32} \geq b_{32}^0 \right\}.$$

In the latter form, the three events are clearly independent, for they depend on disjoint sets of independent $\mathbf{u}^{(i)}$, and their three probabilities are the three factors in (A.5). For example, the first event says that the n_1 points $\mathbf{u}^{(i)}$ corresponding to observations c_1 fall in the subtriangle $u_2/u_1 \geq b_{12}^0$ and $u_3/u_1 \geq b_{13}^0$ whose area is the fraction $(1 + b_{12}^0 + b_{13}^0)^{-1}$ of the area of the whole triangle U.

It will be convenient to denote the right side of (A.5) by $F(b_{12}^0, b_{13}^0, b_{21}^0, b_{23}^0, b_{31}^0, b_{32}^0)$ which defines, as the b_{ij}^0 vary, a form of the joint cumulative distribution function of the b_{ij}. This c.d.f. should be handled with care. First, it is defined only over the subset of the positive orthant in six dimensions such that the b_{ij}^0 define a non-empty R^0. Many points in the orthant are ruled out by relations like $b^0_{12} \le b^0_{13}b^0_{32}$ which are implicit in (A.3). Second, the distribution of the b_{ij} is not absolutely continuous over its six-dimensional domain, but assigns finite probability to various boundary curved surfaces of dimensions 5, 4 and 3, corresponding to random R with 5, 4 and 3 sides. Nevertheless it is not difficult to deduce (23) from (A.5).

Suppose that u^* denotes the vertex of R with maximum first coordinate. This vertex lies, with probability one, at the intersection of two of the six sides of R_1, R_2 and R_3. By looking at the vertices defined by all possible pairs of sides it is easily checked that exactly three possibilities exist for $\mathbf{u}^*$, namely,

$$\left.\begin{array}{lll} \text{(i)} & u_1^*/u_2^* = c_{21} & \text{and } u_1^*/u_3^* = c_{31}, \\ \text{(ii)} & u_3^*/u_2^* = c_{23} & \text{and } u_1^*/u_3^* = c_{31}, \text{ or} \\ \text{(iii)} & u_1^*/u_2^* = c_{21} & \text{and } u_2^*/u_3^* = c_{32}, \end{array}\right\} \tag{A.6}$$

The probability density function of u^* may be formed by summing the contributions from the three possibilities (i),(ii),(iii). The contribution from case (i) will be expressed first in terms of c_{21}, c_{31} and then transformed to u_1^*, u_2^*. Consider the event E that *both* $\{b^0_{21} < c_{21} < b^0_{21} + \delta, b^0_{31} < c_{31} < b^0_{31} + \varepsilon\}$ *and* that the lines c_{21} and c_{31} intersect in a point which maximizes the first coordinate. The latter condition may be written

$$\{c_{12} \ge v_2/v_1, c_{13} \ge v_3/v_1, c_{23} \ge v_3/v_2, c_{32} \ge v_2/v_3,\} \tag{A.7}$$

where $\mathbf{v} = (v_1, v_2, v_3)$ is the point at which the lines c_{21} and c_{31} intersect, or

$$\{c_{12} \ge c_{21}^{-1}, c_{13} \ge c_{31}^{-1}, c_{23} \ge c_{21}c_{31}^{-1}, c_{32} \ge c_{21}^{-1}c_{31}.\} \tag{A.8}$$

Thus, apart from terms of second order and higher in δ and ε,

$$\begin{aligned} \Pr(E) =& F\Big\{\left(b^0_{21}+\varepsilon\right)^{-1}, \left(b^0_{31}+\delta\right)^{-1}, b^0_{21}+\varepsilon, (b^0_{21}+\varepsilon)(b^0_{31}+\delta)^{-1}, \\ & \quad b^0_{31}+\delta, (b^0_{21}+\varepsilon)^{-1}(b^0_{31}+\delta)\} \\ & - F\Big\{\left(b^0_{21}+\varepsilon\right)^{-1}, \left(b^0_{31}\right)^{-1}, b^0_{21}+\varepsilon, (b^0_{21}+\varepsilon)(b^0_{31})^{-1}, \\ & \quad b^0_{31}, (b^0_{21}+\varepsilon)^{-1}b^0_{31}\} \\ & - F\Big\{\left(b^0_{21}\right)^{-1}, \left(b^0_{31}+\delta\right)^{-1}, b^0_{21}, b^0_{21}\left(b^0_{31}+\delta\right), b^0_{31}+\delta, \\ & \quad (b^0_{21})^{-1}\left(b^0_{31}+\delta\right)\} \\ & + F\Big\{\left(b^0_{21}\right)^{-1}, \left(b^0_{31}\right)^{-1}, b^0_{21}, b^0_{21}(b^0_{31})^{-1}, b^0_{31}, \left(b^0_{21}\right)^{-1}b^0_{31}\Big\}. \end{aligned} \tag{A.9}$$

That is, the required case (i) contribution is found in terms of c_{21}, c_{31} represented by b^0_{21}, b^0_{31} by differentiating F with respect to its third and fifth arguments and then substituting $(b^0_{21})^{-1}, (b^0_{31})^{-1}, (b^0_{21}(b^0_{31})^{-1}, (b^0_{21})^{-1}b^0_{31}$ in order for the other four arguments. Expressing the result in terms of the coordinates $\mathbf{u} = (u_1, u_2, u_3)$ at which the lines b^0_{21} and b^0_{31} intersect, one finds

$$n_2 n_3 u_1^{n_1} u_2^{n_2+1} u_3^{n_3+1}$$

which, after multiplying by

$$\partial(u_1, u_2)/\partial\left(b^0_{21}, b^0_{31}\right) = u_1 u_2^{-2} u_3^{-2}$$

gives the density contribution

$$n_2 n_3 u_1^{n_1+1} u_2^{n_2-1} u_3^{n_3-1} \tag{A.10}$$

expressed in terms of u_1, u_2 and of course $u_3 = 1 - u_1 - u_2$. The contributions from cases (ii) and (iii) may be found similarly to be

$$n_2 n_3 u_1^{n_1} u_2^{n_2-1} u_3^{n_3} \text{ and } n_2 n_3 u_1^{n_1} u_2^{n_2} u_3^{n_3-1}. \tag{A.11}$$

Since

$$u_1 + u_2 + u_3 = 1,$$

the sum of the three parts is

$$n_2 n_3 u_1^{n_1} u_2^{n_2-1} u_3^{n_3-1},$$

or

$$\frac{n_1! n_2! n_3!}{n!} \left\{ \frac{n!}{n_1!\,(n_2-1)!(n_3-1)} u_1^{n_1} u_2^{n_2-1} u_3^{n_2-1} \right\}, \tag{A.12}$$

where the first term is the probability that $\mathbf{u}^*$ is anywhere, i.e. that R is not empty, while the second is the Dirichlet density given in (23).

The density of the point with minimum first coordinate may be found by a similar argument. The analogue of (A.6) is

$$\left.\begin{array}{ll} \text{(i)} & u_2^*/u_1^* = c_{21} \text{ and } u_3^*/u_1^* = c_{13}, \\ \text{(ii)} & u_2^*/u_3^* = c_{32} \text{ and } u_3^*/u_1^* = c_{13}, \text{ or} \\ \text{(iii)} & u_2^*/u_1^* = c_{12} \text{ and } u_3^*/u_2^* = c_{23}, \end{array}\right\} \tag{A.13}$$

and the corresponding three components of density turn out to be

$$n_1(n_1+1) u_1^{n_1-1} u_2^{n_2} u_3^{n_2}, \; n_1(n_3) u_1^{n_1-1} u_2^{n_2} u_3^{n_2}, \text{ and } n_1 n_2 u_1^{n_1-1} u_2^{n_2} u_3^{n_3} \tag{A.14}$$

which sum to

$$\frac{n_1! n_2! n_3!}{n!} \left\{ \frac{(n+1)!}{(n_1-1) n_2! n_3! - 1} u_1^{n_1-1} u_2^{n_2} u_3^{n_2} \right\}, \tag{A.15}$$

which, like (A.12) is the product of the probability that R is not empty and the Dirichlet density specified in (23).

The remaining four lines of (23) follow by symmetry. The probability that R is not empty may be obtained directly by an argument whose gist is that, for any set of n points in U, there is exactly one way to assign them to three cells of sizes n_1, n_2, n_3 corresponding to observations c_1, c_2, c_3 in such a way that R is not empty. This latter assertion will not be proved here.

5

On Random Sets and Belief Functions*

Hung T. Nguyen

1 Introduction

The mathematical theory of evidence, as developed by Shafer [1, 2], is based, in the main, upon the notion of lower-probability measures in the work of Dempster on statistical inference (e.g., [3]). Such set-functions have been employed in many different fields such as theory of capacities (Choquet, [4]), stochastic geometry (Kendall [5], Matheron [6]), random fields (Spitzer [7]), and set-valued Markov processes (Harris [8]).

This paper deals with a closer relationship between Dempster's scheme of multivalued mappings and Shafer's belief functions. The basic probability assignment is regarded as the probability distribution of a random set, the notion of condensability is expressed in terms of a multivalued mapping and is related to a general notion of regularity of probability measures. These points of view are useful for applying the notion of belief to fuzzy analysis where multivalued mappings are replaced by fuzzy mappings, and propositions are of the form "X is A," where A is the label of some fuzzy set [9] of a universe of discourse, possibly a continuum.

2 Measurability of Multivalued Mappings

Let $(X, \mathcal{A}), (S, \mathcal{B}), (\mathcal{P}(S), \mathcal{B})$ be three measurable spaces, where $\mathcal{P}(S)$ denotes the collection of all subsets of the set S.

Consider a multivalued mapping:

$$\varGamma : X \to \mathcal{P}(S).$$

We shall formulate two notions of measurability for $\varGamma$: the first one is needed for defining the lower (and upper) probability measure, the second

* Research sponsored in part by the National Science Foundation under Grant MCS-76-06693 and the Army Research Office Grant under DAHC04-75-G0056.

one for considering random sets. Note that these notions of measurability have been investigated, for example, by Debreu [10] in a topological setting.

First, consider two inverses of T:

(a) Lower-inverse:

$$\begin{gathered} \Gamma_* : \mathcal{P}(S) \to \mathcal{P}(X), \\ T \in \mathcal{P}(S), \qquad \Gamma_*(T) = T_* = \{x \in X : \Gamma_x \neq \phi, \Gamma_x \subset T\}. \end{gathered}$$

(b) Upper-inverse:

$$\begin{gathered} \Gamma^* : \mathcal{P}(S) \to \mathcal{P}(X), \\ T \in \mathcal{P}(S), \qquad \Gamma^*(T) = T^* = \{x \in X : \Gamma_x \cap T \neq \phi\}. \end{gathered}$$

Remark 1. The names of these inverse of Γ are given in the way that is related to lower and upper probability measures. The lower-inverse [resp. upper-inverse] is called upper-inverse [resp. lower-inverse] by Berge [11], and strong inverse [resp. weak inverse] by Debreu [10].

Definition 1. *The multivalued mapping Γ is said to be strongly measurable, with respect to $\mathcal{A}$ and $\mathcal{B}$, iff:*

$$\forall B \in \mathcal{B}, \quad \Gamma^*(B) \in \mathcal{A}.$$

Example 1. Let X be a topological space and $\mathcal{A}$ its Borel σ-field; S is a finite set with its discrete topology. If Γ is lower-semicontinuous on X (i.e., for each $x_0 \in X$, for any V open in S such that $V \cap \Gamma_{x_0} \neq \phi$, there exists a neighborhood U of X_0 such that: $x \in U \Rightarrow V \cap \Gamma_x \neq \phi$), then Γ is strongly measurable, with respect to $\mathcal{A}$ and $\mathcal{P}(S)$, since $\forall A \subset S \Gamma^*(A)$ is open in X.

Now consider Γ as a point-to-point mapping from X to $\mathcal{P}(S)$, where "points" in $\mathcal{P}(S)$ are in fact subsets of S. The collection of all subsets of $\mathcal{P}(S)$ is denoted by $\mathcal{P}\mathcal{P}(S)$.

Let Γ^{-1} be the inverse mapping of Γ, i.e.,

$$\begin{gathered} \Gamma^{-1} : \mathcal{P}\mathcal{P}(S) \to \mathcal{P}(X), \\ \hat{T} \in \mathcal{P}\mathcal{P}(S), \qquad \Gamma^{-1}(\hat{T}) = \left\{x \in X : \Gamma_x \in \hat{T}\right\}. \end{gathered}$$

If $\hat{\mathcal{B}}$ is a σ-field on $\mathcal{P}(S)$, then as usual, Γ is said to be measurable, with respect to $\mathcal{A}$ and $\hat{\mathcal{B}}$, iff:

$$\forall \hat{T} \in \hat{\mathcal{B}}, \qquad \Gamma^{-1}(\hat{T}) \in \mathcal{A}.$$

Remark 2. Let $\mathcal{J}$ be the class of all finite subsets of the set S. For $I \in \mathcal{J}$, let π_I be the projection from $\mathcal{P}(S)$ to $\mathcal{P}(I)$, i.e.,

$$A \in \mathcal{P}(S), \quad \pi_I(A) = A \cap I.$$

A finite-dimensional cylinder set in $\mathcal{P}(S)$ is a subset $\hat{A}$ of $\mathcal{P}(S)$ of the form:

$$\hat{A} = \pi_I^{-1}(A), \quad \text{where} \quad I \in \mathcal{J}, \quad \text{and} \quad A \subset \mathcal{P}(I).$$

In particular, if $A = \{I_1\}, I_1 \subset I$, then:

$$\hat{A} = \{B \subset S : B \supset I_1, B' \supset I - I_1\}.$$

Note that if $I_1, I_2 \in \mathcal{J}$ and $I_1 \cap I_2 = \phi$, then:

$$\pi_{I_1 \cup I_2}^{-1}(I_1) = \{B \subset S : B \supset I_1,\ B' \supset I_2\}.$$

Let $\mathcal{C}$ denote the class of all finite dimensional cylinder sets in $\mathcal{P}(S)$, and $\mathcal{F} = \sigma(\mathcal{C})$, the σ-field of $\mathcal{P}(S)$, generated by $\mathcal{C}$. It is clear that if Γ is strongly measurable (with respect to $\mathcal{A}$ and $\mathcal{B}$) and if $\mathcal{J} \subset \mathcal{B}$, then Γ is measurable (with respect to $\mathcal{A}$ and $\mathcal{F}$).

3 Lower-Probability Measure and Belief Functions

Definition 2. *A source is a probability space $(X, \mathcal{A}, \mathbf{P})$ and a multivalued mapping $\Gamma : X \to \mathcal{P}(S)$. For simplicity, we assume that $S^* \in \mathcal{A}$ and $\mathbf{P}(S^*) = 1$. Let $\mathcal{B}$ be a σ-field on S, we assume that Γ is strongly measurable (with respect to $\mathcal{A}$ and $\mathcal{B}$), and in addition:*

If $T \in \mathcal{B}$, then

$$\tilde{T} = \{x \in X : \Gamma x \supset T\} \in \mathcal{A}.$$

The lower and upper probability measures $\mathbf{P}_*, \mathbf{P}^*$ are defined respectively by:

$$\mathbf{P}_*(B) = \mathbf{P}(B_*),$$
$$\mathbf{P}^*(B) = \mathbf{P}(B^*).$$

Note that $\mathbf{P}^*(B) = 1 - \mathbf{P}_*(B')$.

Dempster [3] considered also the set-function:

$$Q(B) = \mathbf{P}(\tilde{B}).$$

Remark 3. In the study of random fields [7] and set-valued Markov processes [8], the set-functions Q and $\mathbf{P}^*$, in the case where Γ is regarded as a random set, are called the correlation function and incidence function, respectively.

Let f be a set-function: $\mathcal{B} \to \mathbb{R}$. Two types of successive differences of $f(B), B \in \mathcal{B}$, with respect to parameters $B_i \in \mathcal{B}, i = 1, \ldots, n+1$, are defined as follows:

(i) $\nabla_1(B; B_1)_f = f(B) - f(B \cup B_1), \nabla_{n+1}(B; B_1, \ldots, B_{n+1})_f = \nabla_n(B; B_1, \ldots, B_n)_f - \nabla_n(B \cup B_{n+1}; B_1, \ldots, B_v)_f$,

(ii) $\mathit{\Delta}_1(B;B_1)_f = f(B) - f(B\mathrm{D}B_1), \mathit{\Delta}_{n+1}(B;B_1,\ldots,B_{n+1})_f = \mathit{\Delta}_n(B;B_1,\ldots,B_n)_f - \mathit{\Delta}_n(B\cap B_{n+1};B_1,\ldots,B_n)_f$.

Following Choquet [4], we say that:

(a) f is alternating of infinite order if $\nabla_n \leqslant 0$ for all n,
(b) f is monotone of infinite order if $\mathit{\Delta}_n \geqslant 0$ for all n.

Properties of $\mathbf{P}_*$ and $\mathbf{P}^*$ can be summarized as follows:

Proposition 1. *(i)* $\mathbf{P}_*(\phi) = 0, \mathbf{P}_*(S) = 1$,
(ii) $\mathbf{P}_*$ *is monotone of infinite order.*
(iii) If $B_n \in \mathcal{B}$ *is a decreasing sequence, then:*

$$\mathbf{P}_*(B_n) \downarrow \mathbf{P}_*\left(\bigcap_n A_n\right).$$

In a dual way:

Proposition 2. *(i)* $\mathbf{P}^*(\phi) = 0, \mathbf{P}^*(S) = 1$,
(ii) $\mathbf{P}^*$ *is alternating of infinite order.*
(iii) If $B_n \in \mathcal{B}$ *is an increasing sequence, then:*

$$\mathbf{P}^*(B_n) \uparrow \mathbf{P}^*\left(\bigcup_n B_n\right).$$

These facts can be seen from the definition of $\mathbf{P}_*$ *and* $\mathbf{P}^*$ *in terms of* $\mathbf{P}$*, and the fact that:*

$$\mathit{\Gamma}_*\left(\bigcap_i B_i\right) = \bigcup_i T_*(B_i),$$
$$\mathit{\Gamma}^*\left(\bigcup_i B_i\right) = \bigcup_i \mathit{\Gamma}^*(B_i).$$

Remark 4. (a) We have only $\mathit{\Gamma}_*(\bigcup_t B_i) \supseteq \bigcup_i \mathit{\Gamma}_*(B_i)$.

(b) In particular, the lower-probability measure $\mathbf{P}_*$ [resp. $\mathbf{P}^*$] is strongly superadditive [resp. strongly subadditive].

(c) For the time being, no topological notions are considered. For further application to fuzzy analysis, where $S = [0,1]$ or some compact set of the real line, the topology will play an important role. Let us point out a result in [6] (Choquet's theorem) concerning a functional associated with a random closed set [this functional plays the role of probability distribution function of a real random variable]: If S is a locally compact space, the space F of closed subsets of S is topologized in some suitable way, σ_F denotes its Borel σ-field, and T is a set-function defined on the space $\mathcal{K}$ of compact sets of S, then the following are equivalent:

(i) T is an alternating Choquet Capacity of infinite order such that T takes values in [0, 1] and $T(\phi) = 0$.

(ii) There exists a unique probability measure $\hat{\mathbf{P}}$ on σ_F such that

$$T(K) = \hat{\mathbf{P}}\left[\{A \in F : A \cap K \neq \phi\}\right], \quad \forall K \in \mathcal{K}.$$

Definition 3. *We recall here the notion of belief function on a finite set S. A belief function Bel on S is a set-function from $\mathcal{P}(S)$ to [0, 1] such that:*

(i) $\mathrm{Bel}(\phi) = 0$,
(ii) $\mathrm{Bel}(S) = 1$,
(iii) For any k,

$$\mathrm{Bel}\left(\bigcup_{i=1}^{k} A_i\right) \geqslant \sum_{\substack{I \neq \phi \\ I \subset \{1,\dots,k\}}} (-1)^{|I|+1}\, \mathrm{Bel}\left(\bigcap_{i \in I} A_i\right),$$

where $|I|$ denotes the number of elements in I.

Note that a belief function Bel is increasing and there exists a set-function:

$$m : \ \mathcal{P}(S) \to [0,1]$$

such that:

(a) $m(\phi) = 0$,
(b) $\Sigma_{A \in \mathcal{P}(S)}(A) = 1$,
(c) $\mathrm{Bel}(A) = \Sigma_{B \subset A} m(B)$.

m is called the basic probability assignment *[2], and*

$$m(A) = \sum_{B \subseteq A} (-1)^{|A-B|}\, \mathrm{Bel}(B).$$

Note also that (iii) is equivalent to the nonnegativity of m.

Remark 5. The representation problem of belief functions in terms of measure algebra and allocation of probability has been fully discussed in [1].

4 Random Sets and Belief Functions

Consider a source $(X, \mathcal{A}, \mathbf{P}), \Gamma : X \to \mathcal{P}(S)$.

Let $\mathcal{B}$ be a σ-field on S. We assume that Γ is strongly measurable (with respect to $\mathcal{A}$ and $\mathcal{B}$).

Proposition 3. *The lower-probability measure $\mathbf{P}_*$ on $\mathcal{B}$ is deduced from the probability distribution of Γ considered as a random set.*

Proof. Let $\hat{\mathcal{B}}$ be the σ-field on $\mathcal{P}(S)$ defined by:

$$\widehat{T} \in \widehat{\mathcal{B}} \Leftrightarrow \Gamma^{-1}(\widehat{T}) \in \mathcal{A}.$$

Thus, with respect to $\mathcal{A}$ and $\hat{\mathcal{B}}, \Gamma$ is a measurable mapping. We say that Γ is a random set by specifying its probability distribution $\hat{\mathbf{P}}$ on $\mathcal{B}$:

$$\widehat{T} \in \mathcal{B}, \qquad \widehat{\mathbf{P}}(\widehat{T}) = \mathbf{P}[\Gamma^{-1}(\widehat{T})].$$

If $A \in \mathcal{P}(S)$, denote by $I(A)$ the principal ideal generated by A, i.e., $I(A) = \{B \subset S : B \subset A\}$, then $\forall B \in \mathcal{B}, I(B) \in \mathcal{B}$. Indeed: $\Gamma^{-1}(I(B)) = B_* \in \mathcal{A}$ by strong measurability of Γ. It follows that: $\hat{\mathbf{P}}[I(B)] = \mathbf{P}_*(B), \forall B \in \mathcal{B}$.

Proposition 4. *In the finite case, the probability distribution of the random set Γ is precisely the basic probability assignment.*

Proof. Since S is finite, and we assume that $\tilde{A} \in \mathcal{A}$ for all $A \subseteq S$, it is clear that $\hat{\mathcal{B}} = \mathcal{P}\mathcal{P}(S)$.

On the other hand, since:

$$\begin{aligned} m(A) &= \sum_{B \subseteq A} (-1)^{|A|-|B|} \mathbf{P}_*(B) \\ \Rightarrow \mathbf{P}_*(B) &= \sum_{B \subset A} m(B) = \sum_{B \in I(A)} m(B) = \hat{m}[I(A)], \end{aligned}$$

where $\hat{m}$ is the probability measure on $\mathcal{P}\mathcal{P}(S)$ with density m. But

$$\begin{aligned} \mathbf{P}[I(A)] = \mathbf{P}_*(A) \Rightarrow \widehat{\mathbf{P}}\left(\{A\}\right) &= \sum_{B \subset A} (-1)^{|A|-|B|} \mathbf{P}_*(B) \\ &= \hat{m}\left(\{A\}\right). \end{aligned}$$

Remark 6. (i) For $A \subset S$, let $F(A)$ be the principal filter generated by A; then: $\forall A \subset S$ (or more generally, $A \in \mathcal{B}$, in the infinite case) $\tilde{A} \in \mathcal{A} \Leftrightarrow F(A) \in \mathcal{B}$.

(ii) Let X be a topological space, and $\mathcal{A}$ its Borel σ-field. Let S be a Hausdorff, locally compact space, and $\mathcal{B}$ its Borel σ-field.

$F, \mathcal{G}, \mathcal{K}$ denote respectively the collection of all closed, open, compact subsets of S. As a topological space, where the topology is generated by $\{F^K, K \in \mathcal{K}\}$ and $\{F_G, G \in \mathcal{G}\}$, with

$$\begin{aligned} F^K &= \{A \in F : A \cap K = \phi\}, \\ F_G &= \{A \in F : A \cap G \neq \phi\}, \end{aligned}$$

the space F is a Hausdorff, compact space. If $\Gamma : X \to \mathcal{F}$ is continuous, then:

$$\forall B \in \mathcal{B}, \qquad \{x \in X : \mathit{\Gamma}_x = B\} \in \mathcal{A}.$$

Note that if A_* and $\tilde{A} \in \mathcal{A}$ then:

$$A_* \cap \tilde{A} = \{n : \mathit{\Gamma}_x = A\} \in \mathcal{A}.$$

(iii) In this finite case, the existence of the biunivocal correspondence between belief functions on S and probability distributions of random sets is established by using the fact that to construct $\hat{\mathbf{P}}$, it is sufficient to construct its density on $\mathcal{P}(S)$, on one hand; and on the other hand, given a set function υ (belief function) on $\mathcal{P}(S)$, we define $\hat{\mathbf{P}}[\mathbf{I}(\mathbf{A})] = \boldsymbol{\upsilon}(\mathbf{A})$, and we are in conditions of application of the Mobius inversion theorem [12] to obtain $\hat{\mathbf{P}}(\{\mathbf{A}\})$ via the Mobius function:

$$\mu\,(A, B) = -\,(-1)^{|A|-|B|}\,, \qquad A \subset B.$$

(iv) If $\mathcal{R}(\mathit{\Gamma})$ denotes the range of $\mathit{\Gamma}$, it is sufficient to consider $I(A) = \{B \in \mathcal{R}(\mathit{\Gamma}) : B \subset A\}$.

Example 2. Let $E = \{A_t, t \in [0,1]\}$ be a family of subsets of S such that:

(a) $A_0 = S$,
(b) $A_1 = \phi$,
(c) $s \leq t \Leftrightarrow A_s \supseteq A_t$.

Let $\mathcal{E}$ be the σ-field on E defined as follows:

$$\hat{T} \in \mathcal{E} \Leftrightarrow \hat{T} = \{A_t\}_{t \in T}\,,$$

where $T \in \mathcal{B}_1$ the Borel σ-field of the unit interval [0, 1].

Let $\mathit{\Gamma}$ be a random set taking values in $(E, \mathcal{E})$ with probability distribution $\hat{\mathbf{P}}$:

$$\hat{\mathbf{P}}\left[\mathit{\Gamma} \in \widehat{T}\right] = \mathit{\Lambda}(T),$$

where $\mathit{\Lambda}$ is the Lebesgue measure on [0, 1].

Let

$$I(A_1) = \{A_s : A_s \subseteq A_t\}\,.$$

Then

$$I(A_t) \in \mathcal{E} \quad \text{for all } t \in [0,1],$$

since

$$I(A_t) = \{A_s\}_{s \in [t,1]}.$$

Define a belief function v on E by:

$$v(A_t) = \hat{\mathbf{P}}[\mathit{\Gamma} \in I(A_t)] = 1 - t.$$

5 Regularity and Condensability

In this paragraph, given a scheme $(X, \mathcal{A}, \mathbf{P}), \mathit{\Gamma} : X \to \mathcal{P}(S)$, we assume that $\mathit{\Gamma}$ is strongly measurable with respect to $\mathcal{A}$ and $\mathcal{P}(S)$. Thus, the belief function $\mathbf{P}_*$ and the upper probability measure $\mathbf{P}^*$ are defined on $\mathcal{P}(S)$. Following Shafer [1], we say that the upper probability measure $\mathbf{P}^*$ is condensable iff $\mathbf{P}^*$ has the following approximation property:

$$\forall A \in \mathcal{P}(S), \qquad \mathbf{P}^*(A) = \sup_{B \in \mathcal{J} \cap \mathcal{P}(A)} \mathbf{P}^*(B) \tag{1}$$

where $\mathcal{J}$ denotes the collection of all finite subsets of S. Recall that, if A_n is an increasing sequence in $\mathcal{P}(S)$, then:

$$\mathbf{P}^* \left(\bigcup_n A_n \right) = \sup_n \mathbf{P}^*(A_n).$$

The condensability of $\mathbf{P}^*$ is stronger than this sequential increasing continuity. In fact [1], $\mathbf{P}^*$ is condensable if and only if for any upward net A_i in $\mathcal{P}(S), i \in I$, we have:

$$\mathbf{P}^* \left(\bigcup_I A_i \right) = \sup_I \mathbf{P}^*(A_i). \tag{2}$$

The fact that (2) implies (1) can be seen as follows: Let $A \in \mathcal{P}(S)$, and $T = \mathcal{J} \cap \mathcal{P}(A)$. It is obvious that T is an upward net in $\mathcal{P}(S)$, and $A = \bigcup_{I \in T}$ thus:

$$\mathbf{P}^*(A) = \mathbf{P}^* \left(\bigcup_{I \in T} I \right) = \sup_{I \in \mathcal{J} \cap \mathcal{P}(A)} \mathbf{P}^*(I).$$

Recall also that the upper-inverse $\mathit{\Gamma}^*$ of $\mathit{\Gamma}$ maps $\mathcal{P}(S)$ into $\mathcal{A}$, since $\mathit{\Gamma}$ is strongly measurable, and:

(i) $\mathit{\Gamma}^*$ is increasing,
(ii) $\mathit{\Gamma}^*(\bigcup_I A_i) = \bigcup_I \mathit{\Gamma}^*(A_i)$.

As a consequence, if A_i is an upward net in $\mathcal{P}(S)$, then $\mathit{\Gamma}^*(A_i)$ is an upward net in $\mathcal{A}$.

We now proceed to give a first characterization of condensability of $\mathbf{P}^*$ in terms of $\mathit{\Gamma}$.

Let $\hat{\mathcal{A}}(\mathbf{P})$ be the subset of $\mathcal{P}(\mathcal{A})$ defined by:

$$\begin{aligned} \hat{A} \in \hat{\mathcal{A}}(\mathbf{P}) &\Leftrightarrow \bigcup_{A \in \widehat{A}} A \in \mathcal{A} \\ &\Leftrightarrow \mathbf{P} \left(\bigcup_{A \in \widehat{A}} A \right) = \sup_{A \in \widehat{A}} \mathbf{P}(A). \end{aligned}$$

Let $U[\mathcal{P}(S)]$ be the set of all upward nets in $\mathcal{P}(S)$. Define the mapping $\hat{\Gamma}$, from $\mathcal{PP}(S)$ into $\mathcal{PP}(X)$ [in fact into $\mathcal{PP}(\mathcal{A})$], induced by $\hat{\Gamma}^*$, as follows:

$$\hat{\Gamma}(\hat{A}) = \left\{\Gamma^*(A), A \in \hat{A}\right\}.$$

Proposition 1. *A necessary and sufficient condition for the condensability of* $\mathbf{P}^*$ *is that* $\hat{\Gamma}$ *maps* $U[\mathcal{P}(S)]$ *into* $\mathcal{A}(\mathbf{P})$.

Proof. Suppose that $\mathbf{P}^*$ is condensable. Let $A_i, i \in I$, be an upward net in $\mathcal{P}(S)$. By strong measurability of $\Gamma, \Gamma(\bigcup_I A_i) \in \mathcal{A}$, thus $\bigcup_I \Gamma^*(A_i) \in \mathcal{A}$. We have:

$$\mathbf{P}^*\left(\bigcup_I A_i\right) = \mathbf{P}\left[\left(\bigcup_I A_i\right)^*\right] = \mathbf{P}\left[\bigcup_I A_i^*\right] = \sup_I \mathbf{p}^*(A_i) = \sup_I \mathbf{P}(A_i^*).$$

Thus $\{A_i, i \in I\} \in \hat{\mathcal{A}}(\mathbf{P})$.

The sufficiency follows immediately from the definition of $\mathcal{A}(\mathbf{P})$.

There is another way to study the condensability of the upperprobability measure $\mathbf{P}^*$, associated with the scheme $(X, \mathcal{A}, \mathrm{E}), \Gamma : X \to \mathcal{P}(S)$, uniquely in terms of the probability space $(X, \mathcal{A}, \mathbf{P})$ and Γ. As before, the upper-inverse Γ^* will play an important role. For this purpose, we shall first introduce a general notion of regularity for probability measures (or generally, for measures); using this notion, we shall express the condensability of $\mathbf{P}^*$ in terms of Γ^* as a criterion and study some consequences.

Notion of ρ-regularity. Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space. Let $(E, \leqslant)$ be a partially ordered set, and $F \subset E$. Finally, let ρ be a mapping from E to $\mathcal{A}$.

Definition 4. *We say that the probability measure* $\mathbf{P}$ *is regular with respect to the system* (E, F, ρ) *(or simply ρ-regular, if E and F are fixed) iff:*

$$\forall x \in E, \qquad \mathbf{P}[p(x)] = \sup_{A \in \hat{\rho}(x)} \mathbf{P}(A),$$

where

$$\hat{\rho}(x) = \{\rho(y) : y \in F, y \leqslant x\}.$$

Remark 7. (i) Let, E, F be subclasses of $\mathcal{A} : F \subset E \subset \mathcal{A}$; and $\rho : E \to \mathcal{A}$ the canonical injection. Then the ρ-regularity of $\mathbf{P}$ is the usual one, i.e.,

$$\forall A \in E, \qquad \mathbf{P}(A) = \sup_{B \in F \cap \mathcal{P}(A)} \mathbf{P}(B).$$

Here, $\hat{\rho}(A) = F \cap \mathcal{P}(A)$.

(ii) If $\mathbf{P}$ is ρ-regular, and ρ increasing, then:

$$\mathbf{P}[\rho(x)] = \sup_{A \in \rho(F) \cap \mathcal{P}[\rho(x)]} \mathbf{P}(A).$$

Consider again the scheme $(X, \mathcal{A}, \mathbf{P}), \Gamma : X \to \mathcal{P}(S)$, with Γ strongly measurable. Denote by $\mathcal{J}$ the collection of all finite subsets of S. Put $\mathcal{J}^* = \Gamma^*(\mathcal{J})$ and $\mathcal{A}^* = \Gamma^*[\mathcal{P}(S)]$. Consider the system $(\mathcal{P}(S), \mathcal{J}, \Gamma^*)$.

We say that $\mathbf{P}$ is Γ^*-regular if $\mathbf{P}$ is regular with respect to the system $(\mathcal{P}(S), \mathcal{J}, \Gamma^*)$. Then it is straightforward that:

Proposition 2. *The following are equivalent:*

(i) $\mathbf{P}^*$ *is condensalbe,*
(ii) $\mathbf{P}$ *is* Γ^**-regular.*

Proposition 3. *If* $\mathbf{P}$ *is* Γ^**-regular, then:*

$$\forall A \in \mathcal{A}^*, \qquad \mathbf{P}(A) = \sup_{T \in \mathcal{J}^* \cap \mathcal{P}(A)} \mathbf{P}(T).$$

Proof. Let $B \in \mathcal{P}(S)$ such that $A = \Gamma^*(B)$. We have:

$$\mathbf{P}(A) = \mathbf{P}[\Gamma^*(B)] = \sup_{T \in \hat{\Gamma}^*(B)} \mathbf{P}(T),$$

where

$$\hat{\Gamma}^*(B) = \{\Gamma^*(I), I \in \mathcal{J}, I \subset B\}$$

Thus:

$$\begin{aligned} \mathbf{P}(A) &= \sup_{I \in \mathcal{J}, I \subset B} \mathbf{P}[\Gamma^*(I)] \\ &\leqslant \sup_{I \in \mathcal{J}, \Gamma^*(I) \subset \Gamma^*(B)} \mathbf{P}[\Gamma^*(I)] \quad \text{since } \Gamma^* \text{ is increasing.} \end{aligned}$$

We obtain, in fact, equality since $\mathbf{P}$ is increasing.

Remark 8. If $(\Omega, \mathcal{A}, \mathbf{P})$ is a probability space and $\hat{T} \subset \mathcal{B} \subset \mathcal{A}$, we say that $\mathbf{P}$ is (inner) regular on $\mathcal{B}$ if:

$$\forall B \in \mathcal{B}, \quad \mathbf{P}(B) = \sup_{T \in \mathcal{C} \cap \mathcal{P}(B)} \mathbf{P}(T).$$

More generally, let ψ be a mapping from $\mathcal{B}$ into $\mathcal{P}(\mathcal{C})$ such that $\psi(B) \subset \mathcal{C} \cap \mathcal{P}(B)$ for all $B \in \mathcal{B}$. We can say that $\mathbf{P}$ is regular with respect to $(\mathcal{C}, \mathcal{B}, \psi)$ iff: $\forall B \in \mathcal{B}, \mathbf{P}(B) = \sup_{T \in \psi(B)} \mathbf{P}(T)$. If the upper-inverse Γ^* is injective [11], i.e.,

$$A \neq B \Rightarrow \Gamma^*(A) \cap \Gamma^*(B) = \phi,$$

then $\mathbf{P}^*$ is condensable if and only if $\mathbf{P}$ is regular with respect to $(\mathcal{J}^*, \mathcal{A}^*, \psi)$ where:

$$A \in \mathcal{A}^*, \qquad A \neq \phi,$$

$\psi(A) = \{\Gamma^*(I), I \in \mathcal{J}, I \subset B\}$, where B is the unique element of $\mathcal{P}(S)$ such that $A = \Gamma^*(B)$.

Proposition 4. *If for each $B \in \mathcal{P}(S)$, there exists a sequence $\{I_n\}_{n \in N}$ elements of $\mathcal{J}$ such that:*

$$\Gamma^*(B) \bigcup_n \Gamma^*(J_n),$$

then $\mathbf{P}$ *is regular on* $\mathcal{A}^*$ *with respect to* $\mathcal{J}^*$*.*

Proof. Let $A \in \mathcal{A}^*, A = \Gamma^*(B)$ for some $B \in \mathcal{P}(S)$.

Since $\mathcal{J}$ is closed under finite union, and Γ^* preserves (arbitrary) unions, we can assume that the sequence $\{\Gamma^*(I_n)\}_{n \in N}$ is increasing.

By monotone continuity of $\mathbf{P}$, we have:

$$\mathbf{P}(A) = \mathbf{P}[\Gamma^*(B)] = \sup_n \mathbf{P}[\Gamma^*(I_n)] \leqslant \sup_{I \in \mathcal{J}, \Gamma^*(I) \subset A} \mathbf{P}(\Gamma^*(I)).$$

We then get equality since $\mathbf{P}$ is increasing.

Proposition 5. *If S is countable, then* $\mathbf{P}$ *is Γ^*-regular.*

Proof. Each $B \in \mathcal{P}(S)$ can be written as:

$$B = \bigcup_n I_n \qquad \text{with} \qquad I_n \in \mathcal{J}, \text{ and } I_n \text{ increasing,}$$

$$\Gamma^*(B) = \bigcup_n \Gamma^*(I_n) \qquad \text{with} \qquad \{\Gamma^*(I_n)\}_n \text{ incresing.}$$

Thus:

$$\begin{aligned} \mathbf{P}[\Gamma^*(B)] &= \sup_n \mathbf{P}[\Gamma^*(I_n)] \\ &\leqslant \sup_{I \in \mathcal{J}, I \subset B} \mathbf{P}[\Gamma^*(I)] \\ &\leqslant \sup_{I \in \mathcal{J}, \Gamma^*(I) \subset \Gamma^*(B)} \mathbf{P}[\Gamma^*(I)] \leq \mathbf{P}[\Gamma^*(B)] \end{aligned}$$

References

1. G. Shafer, "Allocations of Probability: A Theory of Partial Belief," Ph.D. Thesis, Univ. Microfilms, Ann Arbor, Mich., 1974.
2. G. Shafer, "A Mathematical Theory of Evidence," Princeton Univ. Press, Princeton, N.J., 1976.
3. A. Dempster, Upper and lower probabilities induced by multivalued mapping, *Ann. Math. Statist.* **38** (1967), 325–339.
4. G. Choquet, Theory of capacities, *Ann. Inst. Fourier, Univ. (Grenoble)* **5** (1953–1954), 131–296.
5. D. G. Kendall, "Foundations of a Theory of Random Sets" in Stochastic Geometry, pp. 322–376, Wiley, New York, 1974.

6. G. Matheron, "Random Sets and Integral Geometry," Wiley, New York, 1975.
7. F. Spitzer, "Random Fields and Interacting Particle Systems," *Math. Assos. Amer.*, Washington, D.C., 1971.
8. T. E. Harris, On a class of set-valued Markov processes, *Ann. Probability* **4** (1976), 175–194.
9. L. A. Zadeh, Fuzzy sets, *Inform. Contr.* **8** (1965), 338–353.
10. G. Debreu, Integration of correspondences, *in* "Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (1967)," Vol. II, Part 1, pp. 351–372.
11. Cl. Berge, "Espaces Topologiques, Fonctions Multivoques," Dunod, Paris, 1959.
12. G. C. Rota, Theory of Mobius functions, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **2** (1964), 340–368.

6

Non-Additive Probabilities in the Work of Bernoulli and Lambert*

Glenn Shafer

1 Introduction and Summary

Jacob Bernoulli, a 17th century pioneer in the mathematical theory of chance, and Johann Heinrich Lambert, a broad-ranging 18th century scholar, both studied non-additive probabilities. The purpose of this essay is to understand their work, both in its historical context and in relation to the modern theory of non-additive epistemic probability presented in my monograph *A Mathematical Theory of Evidence* (1976).

The starting point of our effort to understand Bernoulli and Lambert is the modern distinction between aleatory probability (or "objective" probability), which we think of as a feature of the world, and epistemic probability (or "subjective" probability), which is more strictly a feature of our knowledge or belief. An aleatory probability is the probability of a chance event. Because of their relation to frequencies, aleatory probabilities must be additive. The epistemic probability of a proposition, on the other hand, is simply a measure of how certain we are of the proposition; epistemic probabilities bear no necessary relation to frequencies and, in my opinion, need not be additive.

Until the late seventeenth century there was a similar distinction between chance, or randomness, and probability, which was an attribute of opinion. As Ian Hacking has stressed in *The Emergence of Probability* 1975 [36], these two concepts were, until about 1660, remarkably unrelated.[1] Considerable progress had been made in the theory of games of chance before 1660; Cardano had written on the subject around 1560 [62] and Galileo around 1620, the correspondence between Pascal and Fermat occurred in 1654, and Huygens published his *De ratiociniis in aleae ludo* in 1657 [42]. But in none of

* From *Archive for History of Exact Sciences*, Vol. 19, Number 4, 1978, 309–370, © Springer-Verlag 1978. The author has updated some of the references and footnotes, added some new footnotes, and corrected a few typographical errors.

[1] This is also discussed by M.G. Kendall, in Sect. 33 of "The beginnings of a probability calculus," 1956 [44].

these writings do we find the word *probability*. And the philosophers who discussed probability before 1660 seem, similarly, to have seldom perceived any connection between their subject and chance. For medieval and Renaissance thinkers, probability (Latin *probabilitas*) belonged to the realm of opinion and argument, where the random was quite out of place.

The connection between probability and chance seems to have first been made in a discussion of the philosophical concept of probability rather than in a discussion of the mathematical theory of chance. In his textbook on logic *l'Art de penser*, published in 1662 [2], Antoine Arnauld perceived the relevance to his subject of the emerging theory of chance and suggested that the principles of this new theory be used when considering the "probabilities" of gain and loss in everyday life. For Arnauld, the chances that the new theory had learned to calculate were probabilities. He could take this view because of the epistemic aspect of these chances; they were known a priori and hence were unequivocally a feature of one's knowledge.

Jacob Bernoulli was the first substantial contributor to the theory of games of chance to grapple with its connection with probability. Bernoulli began his work on chance and probability in the 1680's, some twenty years after Arnauld had written his textbook. In view of the fame of that textbook, and in view of the similar ideas that we find in the writings of Leibniz, we need not be surprised that Bernoulli made the connection between probability and chance. But how exactly was this connection to be understood? How could one mathematize the concept of probability and use the methods of the theory of games of chance to calculate probabilities while preserving the generality of probability and its role in the assessment of argument? Bernoulli addressed this question with subtlety and penetration.

In his *Ars Conjectandi*, published posthumously in 1713 [6], Bernoulli explains that probability is a degree of subjective certainty—a measure of our knowledge. Probabilities are calculated from arguments, and their properties depend on the nature of the arguments. Most importantly, the probabilities from different arguments can be combined to produce what we might today call probabilities "based on the total evidence." In calculating probabilities, Bernoulli uses the methods of the theory of games of chance: he uses Huygens' rule for calculating "expectations." Yet not all Bernoulli's probabilities have the special features of those in games of chance; in some cases, we notice, the probability of a thing and the probability of its opposite do not add to one. It is to this fact that I refer when I write of Bernoulli's "non-additive probabilities."

Bernoulli's subtle understanding of the connection between probability and chance did not endure. We can discern two important reasons for its failure. First, his theory of combining arguments was not completely satisfactory in its own terms and could not be compared as a mathematical theory with the already well-developed theory of games of chance. Secondly, his understanding was a bit too subtle; it was natural for his successors to simplify it by identifying his "probability" with "ease of happening" as it was understood in games of chance. In the long run, this simplification was encouraged by Bernoulli's own *law of large numbers.* This theorem tells us that in those cases where the ease of happening of a thing is not known a priori, it may be learned a posteriori, from the observation of frequencies. Bernoulli thought one could use such frequencies to find the ease of happening of various cases in individual arguments; the probabilities of these individual arguments could then be calculated and combined according to general rules. Bernoulli's successors tended to leave aside his struggle with the combination of arguments and to think of every probability as an "ease of happening," to be found directly from frequencies.

Bernoulli sought to develop a truly general and hence essentially epistemic theory of probability. But the main effect of his effort, ironically, was to attach the name *probability* to an increasingly aleatory theory of chance. It is clear to us today that Bernoulli's law of large numbers is a theorem about aleatory probabilities rather than a theorem about epistemic probabilities. And it is precisely in this respect that it is a departure from the earlier achievements of the theory of games of chance. Arnauld had found it possible to relate the earlier theory to probability because of the a priori nature of the chances it calculated. Granted that there was always an aleatory aspect to these chances—granted that ease of happening always seemed to be a fact about the world—still, these "chances" were known a priori; they were a feature of our knowledge. Though he did not realize it, Bernoulli destroyed this epistemic aspect of chance with his law of large numbers. A chance which can only be known a posteriori is not initially a feature of our knowledge.

After Bernoulli, and especially after the work of his successors Montmort and De Moivre, we already have our modern perplexity. We have a mathematical theory that is basically about chance but bears the name probability, a general belief that the theory must indeed apply to the epistemic ideas that this name unavoidably evokes, and bewilderment as to how it can do so.

Johann Heinrich Lambert stands out as the single 18th century scholar who was able to break away from the identification of probability with the additive theory of chance. In his *Neues Organon*, published in 1764 [51], Lambert took up and extended Bernoulli's theory of argument. He explicitly recognized and sought to explain the possible non-additivity of the probabilities of propositions, and he corrected and extended Bernoulli's rules for combining probabilities based on different arguments. Lambert's rule of combination turns out to be a special case of Dempster's rule of combination, a rule that plays a central role in *A Mathematical Theory of Evidence.*

In Sect. 4 below, I examine in detail the ideas on probability that we find in Bernoulli's mathematical diary[2] and in *Ars Conjectandi.* In Sects. 3 and 5, I examine the historical context of Bernoulli's work, with special attention to

[2] Bernoulli's diary, or *Meditationes*, has not been translated into English. In 1975, B.L. van der Waerden published it in the original Latin with commentary in German, in Vol. 3 of *Die Werke von Jakob Bernoulli* [7]. This volume also includes

the relative lack of connection between probability and chance before Bernoulli and the near fusion of the two concepts after him. In Sect. 6, I discuss Lambert's treatment of probability in *Neues Organon.* And in Sects. 2 and 7, I relate the work of Bernoulli and Lambert to modern ideas. In Sect. 2, I set forth as precisely as possible the ideas that inform my examination of the historical record, and in Sect. 7, I argue for the revival of Bernoulli and Lambert's conceptions of non-additivity and combination in a modern theory of epistemic probability.

2 Aleatory vs. Epistemic Probability

The word *probability* is used today in a great variety of ways, and modern scholars have distinguished many different "kinds" of probability. But the broad distinction I have already mentioned, the distinction between *aleatory* and *epistemic* probability, is the most common distinction and, I believe, the most fundamental.

Aleatory probabilities find their role in the paradigm of chance; they are the numbers assigned to the various possible outcomes of a chance event. The aleatory probability (or *chance*) of each outcome is thought to measure its propensity to occur, and this is a feature of the objective world, for it is approximated by the frequency with which the outcome does occur when a large number of "trials" of the event are observed. An epistemic probability, on the other hand, describes our knowledge. It is a number that represents, albeit with a usually ludicrous affectation of precision, the degree to which we are certain of something, or alternatively, the degree to which we believe it or the degree to which our evidence supports it.

There are, of course, nuances in the way we understand aleatory probability. In its purest form, the idea of aleatory probability is openly opposed to determinism; it presents us with an image of "random phenomena," phenomena that are truly governed by chance laws and by no deeper laws, so that the chances given by these laws are fundamental facts about nature. This is the image that informs the branch of pure mathematics now called the *mathematical theory of probability*, and it is also the image to which the Copenhagen interpretation of quantum mechanics appeals. Yet it is common, especially among statisticians and others attuned to the skeptical epistemology of our age, to treat this image as a more tool, a tool by which we "model reality."

But these nuances must not be allowed to obscure the fundamental fact that aleatory probabilities are not, in the first instance at least, features of our knowledge in the way that epistemic probabilities are. When we posit that a certain phenomenon is random—i.e., governed by a chance law, we hardly ever, in this cautious age, posit that we know the chances. There are occasions,

facsimiles of Johann de Witt's treatise on annuities and Nicolaus Bernoulli's dissertation, as well as several scholarly studies cited in this paper.

in sciences such as physics and genetics, where a theory tells us the chances, but even here we are cautious: we must test the theory by observing frequencies and thus empirically verifying the chances. And commonly, as when we practice statistical inference, we stress our initial ignorance of the chances. We link the different possible values of these chances with different possibilities as to other facts, and our intent is to gather unprejudiced empirical evidence about the former and to count it as evidence about the latter.

Sometimes we do come to know the aleatory probabilities for the various outcomes of a chance event, at least to a tolerable degree of precision. These aleatory probabilities are then part of our knowledge—part of our evidence, if your will, as to which outcome will result from a given trial, past or future. If these aleatory probabilities are our only evidence, then they would seem to warrant identical numbers as epistemic probabilities. *Quod facile est in re*, *id probabile est in mente*, as Leibniz put it: “That which is easy in fact is probable in the mind.”[3] But it cannot be too strongly stressed that we usually do have other evidence. If the trial is in the past, we may have evidence from memory, from testimony, or from knowledge of its consequences. And even if it is in the future, we may in fact have evidence relating to it as a special case, contrary to the “useful model” which would make it a mere trial of a chance event. To know an aleatory probability is to know a great deal, and it is rare that we know so much and nothing more.

Our modern distinction between aleatory and epistemic probability cannot be traced to the distinction between chance and probability that preceded Bernoulli. For this older distinction disappeared in the 18th century, in a process of synthesis that culminated in the work of Bayes and Laplace. But the modern distinction can be said to be over a century old; it has historical roots in the dissolution of the Laplacean synthesis, which began in the mid-19th century.

Poisson, in his *Recherches sur la probabilité des jugements*... (1837 [64], pp. 30–31), formulated a distinction close to the modern one, and proposed to mark it with the words *chance and probability.* The *chance* of an event, he suggested, is a property of the event in itself, independent of our knowledge; *probability*, in contrast, is relative to our knowledge. The distinction must have been widely understood, but the terminology never gained wide acceptance. Today, at least, this terminology is likely to strike the scientist as impudent, for it would revoke the license of several centuries' standing by which the theory of chance bears the grander title *probability.* But I believe the terminology is thoroughly justified by history and etymology, by the most basic facts of the European languages, and by the clarity it would bring into discussions of our subject. I am prone to adopt it whenever I believe I can do so without confusing the reader.

Cournot, in his *Exposition de la théorie des chances et des probabilités* (1843 [14], p. v), proposed to mark the distinction with the terms *objective*

[3] Quoted by Hacking, 1975 [36], p. 128.

probability and *subjective probability.* This is a natural terminology. It has been quite popular during the past several decades, and is still more current than the relatively recent *aleatory vs. epistemic* terminology. But I fear it has lost its power to mark the broad distinction at which it aims. The term *objective probability* now suggests a strong commitment to the objective reality of chance, and is therefore wont to inspire debates somewhat to one side of the distinction sought. And the term *subjective probability*, having become identified with a view that emphasizes the personal and arbitrary aspects of belief, has lost the broader connotations of *epistemic probability.*

Ernest Nagel, in his *Principles of the Theory of Probability* (1939), described the various kinds of probability as different ways of interpreting the "probability calculus"—i.e., the mathematical theory of probability. This has become a quite common view of the matter. The name *probability*, in this view, belongs most properly to the branch of pure mathematics that now bears that name and to the set functions it studies—functions P that are defined on an algebra of subsets of a set Θ and satisfy three axioms: (1) $P(\phi) = 0$; (2) $P(\Theta) = 1$; and (3) $P(A \cup B) = P(A) + P(B)$ whenever $A \cap B = \phi$. Every kind of numerical probability, it is held, obeys these rules and is hence an interpretation of this mathematical theory.

I reject this view. I believe that epistemic probabilities need not obey the rule of additivity, axiom (3) above. And thus, in my view, aleatory and epistemic probability are distinguished by their mathematics as well as by their meaning.

The additivity of aleatory probability is compelled by the fact that these probabilities are mirrored by frequencies. If Θ is the set of all possible outcomes of a chance event and A and B are disjoint subsets of Θ, then the frequency with which the event's outcome falls in $A \cup B$ is necessarily the sum of the frequency with which it falls in A and the frequency with which it falls in B. And since these frequencies, when based on a sufficiently large number of trials, approximate $P(A \cup B)$, $P(A)$, and $P(B)$, respectively, the relation $P(A \cup B) = P(A) + P(B)$ follows.

The view that epistemic probabilities should also be additive is very venerable and is thoroughly engrained in current thought. Among the unreflective, the view has undoubtedly derived strength from a failure to distinguish between aleatory and epistemic probability, and from a misunderstanding of the significance of the pure mathematics of additive probability. More thoughtful scholars have advanced a variety of explicit arguments for requiring epistemic probabilities to be additive. And perhaps most importantly, the Bayesian theory of statistical inference has exerted a powerful influence in favor of such additivity.

A basic task of statistical inference, at least as I understand it, is to compute, from the evidence provided by a limited number of trials of a chance event, epistemic probabilities for assertions as to values of the event's aleatory probabilities. The Bayesian theory provides a method for computing such epistemic probabilities, a method based in an essential way on the assumption they are additive. Since at least the mid-19th century, this method has been the principal argument for the usefulness of numerical epistemic probability. And as such, it has been an argument for the additivity of epistemic probability.

(The Bayesian theory is named after Thomas Bayes, the English clergyman who first formulated it; the enigmatic essay in which he did so was published in 1764 [3], shortly after his death. Laplace appears to have rediscovered the Bayesian method; in any case, his perspicuous explanation of it, first published in 1774 [52], was historically responsible for its wide acceptance (Stigler 1975 [76]). It should be noted that Bayes was entirely oblivious to any distinction between aleatory and epistemic probability, and that Laplace avoided making any such distinction explicit. It should also be noted that the early thinking about probability and chance studied in this essay antedates the Bayesian theory.)

The theory of epistemic probability presented in my monograph *A Mathematical Theory of Evidence* (1976 [70]) is related in many ways to the Bayesian tradition; it is inspired, in large part, by criticisms of that tradition in general and its approach to statistical inference in particular. But in this new theory, the axiom of additivity is replaced by a more general inequality. This permits greater freedom of expression, especially when the evidence is scant. And, as we see in this essay, it permits the revival of important aspects of Bernoulli's and Lambert's ideas.

3 Probability and Chance before Bernoulli

In order to understand the meaning and significance of Bernoulli's work, we must understand the concepts of chance and probability that preceded him and trace the development of these concepts during the 17th century. To this end, this section reviews the early concept of probability, the early (pre-1660) theory of games of chance, and the application of this theory to mortality statistics in the years from 1660 to 1700. We also review Arnauld's linking of probability and chance and Leibniz's views on probability. I rely heavily on the work of Hacking [36].

3.1 Probability and Chance before 1660

J. van Brakel (1976 [11], p. 124) has suggested that the ancients had, roughly speaking, three epistemological categories: "(i) that of which certain knowledge is possible, (ii) that of which probable knowledge is possible, and (iii) that of which no knowledge is possible." The first two categories correspond to Plato's distinction between knowledge (*episteme*, translated into Latin by *scientia*) and opinion (*doxa*, translated into Latin by *opinio*). The third category was, by definition, the realm of randomness. These categories seem to have endured through the middle ages: probability was an attribute of opinion, and randomness was understood in such a way as to exclude, so it would seem, the recognition of statistical regularities or the development of a mathematics of chance.

These categories were transformed during the Renaissance, though we do not thoroughly understand why and how. Hacking (1975 [36]) traces the transformation to the notion of *sign*, as it had been understood in the low sciences of the middle ages. In the medieval understanding, opinion was based on testimony: a probable opinion was one approved by some authority or by the testimony of respected judges.[4] Our modern, broader concept of evidence was lacking. But the low sciences extended the notion of testimony by including signs—the testimony of nature. By the end of the Renaissance, sign was transformed into a new concept of evidence; Arnauld in his *l'Art de penser* (1662 [2]) could then distinguish two kinds of evidence: *external* evidence, or the evidence of testimony, and *internal* evidence, or the evidence of things.

It is clear that the emergence of the concept of internal evidence changed the meaning of the word *probability* (Latin *probabilitas*, French *probabilité*). Hacking goes further; he suggests that the origin of the new concept of internal evidence in the older concept of sign was reflected in a tendency of philosophers to relate issues of evidence and probability to wagers in games of chance. For signs are only sometimes accurate, and one must take into account the proportion of the time they are accurate. Hacking cites Thomas Hobbes, who wrote around 1640 as follows:

> ...though a man have always seen the day and night to follow one another hitherto, yet can he not hence conclude that they shall do so, or that they have done so eternally: *experience concludeth nothing universally.* If the signs hit twenty times for one missing, a man may lay a wager of twenty to one of the event; but may not conclude it for a truth.

Perhaps, Hacking hints, the new concept of internal evidence was even responsible in part for enabling the 16th and 17th centuries to develop a theory of games of chance.[5]

Be this as it may, the originators of the theory of games of chance were not concerned with general issues of evidence and probability, and they did not use the concept of probability when they dealt with games of chance. Cardano and Galileo sought simply to compute the relative ease with which different outcomes can happen in a game of chance. And Pascal, Fermat and Huygens took as their starting point a problem of equity, the "problem of points."

The problem of points was the problem of dividing the stakes when a game of chance is left unfinished—e.g., if a point is awarded for each play and

[4] For a discussion of Thomas Aquinas' views on what makes an opinion probable (Latin *probabilis*), see Byrne (1968 [12]).

[5] The author reviewed Hacking's book in 1976 [71]. Hacking's views on the evolution of the notion of evidence before Pascal were subsequently challenged by several authors. See *The Science of Conjecture: Evidence and Probability before Pascal*, by James Franklin, Johns Hopkins University Press, 2001, especially p. 373 ff.

a player must win three points to win the game, then how are the stakes to be divided if play is halted when one player has one point and the other has two? Fermat's approach to this question was to count the *hasards* favorable to the one player and to the other, for the stakes were to be divided in the same ratio [30]. Pascal and Huygens preferred when possible to use a more conceptual approach, one based on computations of value. Pascal sought to find the value to a player of each point won, and Huygens reasoned in terms of the value of a player's chance (Dutch *kans*, French *chance*), meaning, broadly, his prospects. In NeoLatin this value was called the player's *expectatio or spei*; thus our present usage: *expectation* in English and *espérance* in French.

The aleatory features of this early theory cannot be denied. It was, after all, about chance. Its vocabulary included Latin words (*aleae, sors*, etc.) used since antiquity to evoke randomness. The correspondence between Pascal and Fermat tends to be narrowly addressed to the problem of points, but in the more general discussions of Cardano, Galileo and Huygens, we find modes of expression that must be given an aleatory rather than an epistemic interpretation.[6] Galileo, for example, remarks that some numbers in a dice-game are made more easily and more frequently (*piu facilmente e piu frequentemente*) than others. And ease of occurrence seems to be a basic concept for Huygens as well. These early writers did not, of course, indulge in hints of indeterminism of the sort typical of contemporary explications of the idea of chance. (See Sect. 2 above.) They knew of the indeterminism that some of the ancients had associated with the random, but Christianity supplied them with a determined determinism, and we can only speculate as to whether they felt any tension between this determinism and the notion of ease of occurrence.

Yet the theory also had epistemic features, stemming, as I have already mentioned, from its a priori nature. Even though Cardano and Galileo seem to have thought of chance as a feature of the world, they betrayed no discomfort with the thought that chances can be known a priori. Notice, moreover, that the problem of equity addressed by Pascal, Fermat and Huygens is an a priori problem. It is implicit in the problem that the game ought to be fair, and that the stakes ought to be divided on the assumption that it is fair. Hence it is natural to use a vocabulary that takes knowledge of its fairness for granted.

3.2 The Port-Royal *Logic*

The connection between probability and chance seems to have first been made in print in 1662, the year of Pascal's death, by his friend Antoine Arnauld, who was the leader of the Port-Royal Jansenists and in whose defense Pascal had written his *Provinciales.* Arnauld's discussion of probability occurs in the final chapter of *la Logique, ou l' art de penser*, which was first published in that year and later became known simply as the Port-Royal *Logic.* (Arnauld wrote the *Logic* in collaboration with Pierre Nicole, but he is generally considered the senior author of the book and the most probable author of the final chapter.)

[6] See Hacking [36], pp. 49–56.

From a modern perspective, Arnauld's discussion seems rather banal. But because of its historical significance, and especially its significance for Bernoulli, I will quote it at length:

> . . . [Many people] consider only the magnitude and the importance of the benefit they hope for or the detriment they fear, without considering at all the verisimilitude and the probability that that benefit or that detriment will or will not materialize.
>
> . . . in order to judge what one should do to obtain a good or avoid an evil, it is necessary to consider not only the good or the evil itself, but also the probability that it will or will not materialize, and to consider geometrically the proportion that these things have together. This can be clarified by this example:
>
> There are games in which ten persons each put up a crown, one of them wins everything, and all the others lose. Here each person risks only the loss of one crown, and stands to gain nine. If one considered only the gain and loss in themselves, it would seem that each person is at an advantage. But it must further be recognized that while each person can win nine crowns and is risking only the loss of one, it is also nine times more probable in the case of each that he will lose his crown and not win the nine. Hence each person has nine crowns to hope for, one crown to lose, nine degrees of probability of losing one crown, and a single one of winning the nine crowns: which leaves the matter in a perfect equality.
>
> All games of this kind are equitable, to the extent that games can be equitable, and those that do not meet this condition are obviously unfair. And thereby one can show that there is an obvious unfairness in those sorts of games that are called lotteries, for since the operator of the lottery usually takes a tenth of everything as his share, the whole body of players is cheated in the same way as a man who bets ten pistoles against nine in an equal game—i.e., in a game where there is as much verisimilitude of gain as of loss. This being disadvantageous for the whole body of players, it is also disadvantageous for each player, because it follows that the probability of loss exceeds the probability of gain more than the profit one hopes for exceeds the injury to which one exposes oneself, which is to lose what one has put up.
>
> There is sometimes so little verisimilitude in the success of something that no matter how profitable it is and no matter how little one must risk in order to obtain it, it is best not to risk it. Accordingly, it would be foolish to risk twenty sous against ten million pounds, or against a kingdom, on the condition that one could not win it unless a child should suddenly compose the first twenty verses of Virgil's *Aeneid* in the course of arranging at random the letters of a printing press. Indeed, there is not a moment in one's life but that, without

> thinking, one risks it more than a prince would be risking his kingdom in betting on this condition.
>
> These reflections seem trifling, and indeed they are, if one leaves the matter at that. But we can use them in more important matters, and the main use we should make of them is to make us more reasonable in our hopes and fears. There are, for example, many people who are excessively frightened when they hear thunder. If the thunder makes them think of God and of death, so much to the good—one cannot think too much on these things. But if it is merely the danger of dying from the thunder that causes their extraordinary dread, then it is easy to make them see that it is not reasonable. For from among two million people, one such death would be a lot; and one can even say that there is hardly any violent death that is less common. Hence since the fear of an evil ought to be proportional not only to the magnitude of the evil but also to the probability of the event, and since there is hardly any kind of death less common than death from thunder, there are also hardly any that should cause us less fear - especially since this fear does nothing to help us avoid death.[7]

Here, suddenly, probability has been connected with chance and has become numerical: "nine times more probable," "the probability of loss exceeds the probability of gain," "the fear of an evil ought to be proportional not only to the magnitude of the evil but also to the probability of the event." Most fundamentally, the word *probability* has been used to draw an analogy between games of chance and everyday life. Notice, though, that Arnauld stops just short of our conception of numerical probability: he does not anticipate Bernoulli by singling out the ratio of the number of favorable cases to the total number of cases and assigning the name *probability* to this ratio.

Looking backwards, it seems remarkable that Pascal himself did not appropriate the word *probability* in his discussion of chance. (One recent student of the mathematical theory of probability, Alfred Renyi [67], found it so incredible that he composed a letter for Pascal to send to Fermat, rectifying the omission.) After all, Pascal likened one's decision whether to accept the Catholic religion to a wager in a game of chance. And whether the Catholic religion is true might be considered an opinion whose probability can be discussed. But there is no evidence that Pascal ever used *probability* in connection with chance. Perhaps he avoided the word because of his opposition to probabilism, a doctrine of the Jesuits which gave Christians exceptional latitude in choosing among probable opinions in difficult matters of conscience. Or perhaps, in spite of his famous wager, he was simply too old-fashioned to speak of the "probability of an event."

[7] This passage is from the last chapter of the book: Chapter XV of Part IV in the first edition, Chap. XVI in later editions.

It is difficult to assess the originality of Arnauld's writing, and hence its influence. Was his use of the word probability as novel as it appears to be from the written record? Or had the vernacular already appropriated the word to describe games of chance? Did his linking of probability with randomness seem familiar to his readers, or bizarre?

This much can be said: Though the Port-Royal *Logic* was the most widely read and successful logic text of its time, it did not convince all its readers that games of chance provide a fundamental model for epistemic probability. It remained quite possible after 1662 to formulate the notion of probability in a purely epistemic and non-numerical way. The English philosopher John Locke may be advanced as a case in point; when he undertook to write his own treatise on human reasoning in 1671, he produced this account of probability:

> Probability then is a likeliness to be true. The very notation of the word signifying as much, and from its derivation may be thus defined: "Probabile est quod probari potest," i.e., a proposition for which there be arguments or proofs to make it pass or be received for true.[8]

Probability admits of degrees; Locke preceded the statement just quoted with the remark that there are "degrees of probability from the very neighborhood of certainty and evidence quite down to improbability and unlikeliness even to the confines of impossibility." And Locke further makes it clear that the degree of probability depends on the strength and number of the arguments pro and con. But this goes no farther than Thomas Granger's maxim of 1620 "Many probabilities concurring prevail much."[9] There is no hint that probability can be numerical, and no connection with chance. This same purely epistemic treatment persisted in the final version of Locke's *Essay Concerning Human Understanding*, published in 1690 [59].

3.3 The Demographers

The tendency of the late 17th century to associate epistemic ideas with chance seems to have been furthered, or at least reflected, by the activities of the practical statisticians of the age—the demographers. Stimulated by the new theory of games of chance, as well as by political curiosity and by the fashion of selling annuities, Graunt, Petty, and Halley in England and de Witt, Hudde, and the brothers Huygens in Holland began during this period to compile life tables and calculate life expectancies. Hacking has surveyed the work of these scholars, from the publication of Graunt's pioneering *Natural and Political Observations* in 1662 to the publication of Halley's articles on degrees of mortality in 1693.[10]

[8] See p. 56 of Aaron and Gibb's edition of Locke's draft [60].

[9] *Divine Logike*, p. 80 [35].

[10] For another survey, see Kohli and van der Waerden [45]. This essay is especially valuable for its overview of early practices regarding annuities and for its precis of de Witt's tract of 1671 on annuities.

One is struck by the frank dependence of this work on the vocabulary and methods of the theory of games of chance. Graunt was mainly concerned with numerical facts, yet he writes of "hazards" and betting. The later authors are more explicitly concerned with the theory of chance. de Witt and Hudde consulted with Huygens himself, who had already corresponded with his brother Lodewijk about statistics of mortality.

Yet these authors are soon using a more epistemic vocabulary than had been seen in the early theory of games of chance. In 1669, Huygens described what we would now call a man's median life expectancy with the French phrase *age auquel il y a égale apparence qu'il parviendra ou ne parviendra pas* (literally, "age such that there is equal appearance of his reaching or not reaching it").[11] The undeniably epistemic *apparence* echoes Arnauld and was shortly echoed in demographic contexts by de Witt's Dutch *apparenz* (1671) and Petty's English *likelihood* (1674). Such epistemic terms are, or course, entirely in order. One does not examine mortality statistics because of an interest in randomness. Rather, one is interested in what verisimilitude is given to various possible future happenings by our knowledge of men's mortality in general. Still, none of these authors followed Arnauld in using the weightier epistemic term *probability.*

It might be thought that the need to gather statistics about mortality should have weakened the *a priori* nature of the exercise and thus its epistemic flavor, but this seems not to be the case. As Hacking shows, there was considerable confusion about the roles of statistics and a priori knowledge in the construction of mortality curves. And the statistics were soon at hand in any case. It was only long after Bernoulli's statement of his law of large numbers that problems began to arise from the habit of giving epistemic status to still unknown facts.

3.4 Leibniz

A survey of the thinking about probability that preceded and influenced Jacob Bernoulli can hardly omit reference to Gottfried Wilhelm Leibniz (1646–1716), the great German philosopher and mathematician whose version of the calculus inspired the mathematics of both Jacob and Johann Bernoulli, and who displayed an unbounded optimism concerning the possibilities for mathematizing thought. Unfortunately, it is difficult to summarize Leibniz's opinions or to assess his influence; his was an undisciplined and prolific genius, and most of his work remained unpublished long after his death.[12]

Leibniz first discussed probability in *De conditionibus*, the dissertation he submitted for his bachelor's degree in law at Leipzig in 1665. *De conditionibus*

[11] See p. 537 of Vol. VI of Huygens' complete works [43]. In the passage from Arnauld above, the French *apparence* was translated by *versisimilitude.*

[12] For a guide to Leibniz' scattered thoughts on probability, see Couturat, 1901 [15], pp. 239–282, and Hacking [36], 1975, pp. 85–91, 122–142. Both authors, in my judgment, occasionally credit Leibniz with too much.

deals with the idea of conditional rights (Latin *jus conditionale*). Towards the end of this dissertation, Leibniz suggests that rights be represented as numbers; an absolute right (*jus purum*) is unity, a non-existent right (*jus nullum*) is zero, and a conditional right is a fraction. Whether a right is non-existent, conditional, or absolute depends, he explains, on whether the condition on which it is based is impossible, contingent, or necessary. And the magnitude of a conditional right depends on the probability of the existence of the condition: *Quanto major probabilitas est existentiae Conditionis, tanto majoris jus Conditionale*, "The greater the probability of the existence of the condition, the greater the conditional right."[13]

De conditionibus was published in Leipzig in 1665, revised in 1667, and republished as part of *Specimina juris* in 1669. But it was not, apparently, influential or even widely available in later years.[14] While granting its relative lack of influence, Hacking (pp. 85–91) emphasizes its intellectual significance: Leibniz appears to have been moving towards a numerical conception of probability neither inspired by nor connected with the theory of games of chance.[15]

Leibniz became familiar with the contemporary work on games of chance and annuities during his stay in Paris from 1672 to 1676 [40]. It was during these years that he became a serious mathematician and did his most important mathematical work, and Huygens and Arnauld were among the mathematicians with whom he rubbed shoulders. Within a few years of his return to Germany he had written two essays on chance: *De incerti aestimatione*, an unpolished memorandum dated 1678,[16] and *Essai de quelques raisonnements sur la vie humaine et sur le nombre des hommes*, written around 1682.[17] Mathematically, *De incerti aestimatione* is a confused and on the whole unsuccessful attempt to master Huygens' method; philosophically, it can be read as an attempt to justify that method by principles of jurisprudence.[18] The *Essai de quelques raisonnements* is Leibniz' contribution to the problem of annuities. Neither essay achieves anything new mathematically, but they are

[13] The idea of basing property settlements on epistemic probabilities is not so foreign to modern practice. Out-of-court settlements between large American corporations and the Internal Revenue Service are often explicitly based on the lawyers' judgment of the probability of the I.R.S. winning in court. Notice, incidentally, that additive probabilities are most natural in this context.

[14] See Ravier, 1937 [66], pp. 3–4, and Couturat, 1901 [15], p. 240.

[15] The case is not quite proven. It is plausible that Leibniz had not heard of the new theory of games of chance during his student years at Leipzig and Altdorf, nearly certain that he did not understand the new theory's mathematics. But would he have not yet seen *l'Art de penser* in 1667? He certainly had seen it by 1671, when he praised it in a letter to Arnauld. (See Lewis, 1952 [58], p. 4.)

[16] First published by Biermann and Faak, 1957 [9]. A translation into French, with commentary, appears in *l'estime des apparences*, by G.W. Leibniz, edited by Marc Parmentier, Vrin, Paris, 1995.

[17] See Couturat 1902 [15], p. 274.

[18] Hacking's report (p. 145) that this memorandum mentions the notion of degree of certainty is in error. And I cannot agree with Hacking's discernment (pp. 125–128) of distinct aleatory and epistemic approaches within the memorandum.

notable for their recognition of the role of the concept of probability in the theory of games of chance. Because of his understanding of the role of probability in jurisprudence, Leibniz took it for granted that Huygens' problem of equity was a problem about probabilities. And he readily used *probablement* in discussions of mortality statistics, where Huygens had contented himself with *apparence.*

To what extent did Leibniz's ideas on probability influence Jacob Bernoulli? We have to be impressed by Leibniz's strong, almost instinctive conviction that *probability* could be made numerical and by his unhesitating classification of the contemporary work on chance and annuities under the general heading of probability. His many correspondents of the 1680s and 1690s must have also been impressed, and hence it is plausible that his ideas indirectly influenced Bernoulli. But the direct influence seems to have been scant. The correspondence between the two indicates that Leibniz learned Bernoulli was working on probability in 1703; by this time Bernoulli had been pondering the subject for many years and was about to begin writing Part IV of *Ars Conjectandi.*[19] In April of that year Leibniz inquired about Bernoulli's work, and in letters written in 1703 and 1704, Bernoulli explained and defended his law of large numbers. On the whole, the new ideas in this correspondence seem to be coming from Bernoulli. And as late as February, 1705, 6 months before his death, Bernoulli remarks that he has only two of Leibniz's works: *De Arte Combinatoria and Hypothesis Physica nova.* Apparently he never saw *De conditionibus*, *De incerti aestimatione*, or the *Essai de quelques raisonnements.*

There is one important point on which Leibniz anticipated Bernoulli and may have influenced him: the ambiguous use of the phrase *aeque possibiles*, "equally possible." This phrase occurs in *De incerti aestimatione*; there Leibniz requires that *eventus sint aeque faciles seu aeque possibiles*, "the events be equally easy or equally possible." The use of *aeque faciles* in this context is familiar, but the use of *aeque possibiles* seems to be new with Leibniz. It is an intriguingly ambiguous phrase, for a thing can be possible either relative to our knowledge or relative to nature. And this ambiguity can be related to basic issues in Leibniz's philosophy: that which is most possible in the sense of internal consistency has the greatest power to come into existence.[20] Bernoulli apparently never saw *De incerti aestimatione* and certainly did not have a full view of Leibniz's philosophy, but at one point in their correspondence,[21] Leibniz casually mentions equipossibility; he reports that de Witt has used equally possible cases. Bernoulli may have been consciously following Leibniz when he himself wrote about "equally possible cases."

[19] Their correspondence is in Gerhardt's collection of Leibniz' mathematical works, [57], Vol. III, pp. 3–110. The passages commented on here are on pp. 71, 77–78, 83–84, 87–89, and 95. For a commentary on the correspondence, see Kohli 1975 [48].

[20] See Hacking [36], p. 138.

[21] See [57], Vol. III, p. 84.

4 Jacob Bernoulli's Treatment of Probability

In the late 17th century the Bernoullis were prominent merchants in the city of Basel, where they had settled after having been driven from Antwerp by Catholic persecution a century earlier. Jacob, born in 1654, was the first of the family to become a mathematician, and he did so only after the conclusion of his studies in philosophy at the University of Basel. He established extensive contacts with European mathematicians during his travels from 1676 to 1682; on his return to Basel he devoted himself to physics and mathematics, finally accepting a chair in mathematics at the University of Basel in 1687. He became well known during his lifetime for his work on the infinitesimal calculus; he and his younger brother Johann were the first to master Leibniz abbreviated presentation of the differential calculus, and they eventually became his champions in the dispute with Newton. Jacob died while still holding his chair at Basel, on August 16, 1705.

Bernoulli's great treatise on probability, *Ars Conjectandi*, was not published until 1713, 8 years after his death. But we now know, from B.L. van der Waerden's study of Bernoulli's *Meditationes* (his mathematical diary) that he had already worked out the treatise's most basic results, including the law of large numbers and its proof, in three periods of study between 1684 and 1689.[22] It appears that he began, in 1684 or 1685, with a study of the five problems that Huygens had posed at the end of *De ratiociniis in aleae ludo*. In a second period, during the fall or winter of 1685/86, he went beyond the realm of games of chance to think about probabilities in connection with aspects of daily life such as marriage contracts, weather, and testimony. Finally, sometime during the period from 1687 to 1689, after a period of attention to other matters (including Leibniz infinitesimal calculus), he returned to probability and stated and proved his law of large numbers.

It is the second period, in the fall or winter of 1685/86, that is of greatest interest to us. For it was here that Bernoulli made the transition from chance to probability, and in the course of trying to compute probabilities in practical questions, found the motivation for his law of large numbers. The crucial passage in his *Meditationes* is his study of a problem concerning a marriage contract.[23] Titius is marrying Caja, and an agreement is to be made concerning the division of the estate between him and the children in case she (Caja) dies before him. But the size of the estate will vary according to whether one or both of their fathers are still living. Titius proposes two possible agreements: one specifies that 2/3 of the estate will fall to him in any case; the other varies the proportion according to which fathers are still living. Which agreement is most advantageous to Titius? The answer depends, obviously, on the chances Caja has of outliving one or both of the fathers. Bernoulli makes

[22] Van der Waerden's discussion of the dates for the *Meditationes* is in [7], Vol. 3, pp. 385–389.

[23] This passage is in [7], Vol. 3, pp. 42–48; van der Waerden's commentary is on pp. 364–369.

various assumptions from which these chances can be calculated, and proceeds to calculate them. In the course of these calculations, we suddenly find him writing about parts of certainty and probabilities; from one calculation, for example, he obtains "1/5 of certainty," or "one probability where five make the whole certainty." This kind of language reappears in *Ars Conjectandi*, but it is in contrast with the language of Huygens, which Bernoulli had tended to use in earlier passages in his *Meditationes*.

The practicality of Titius' problem not only led Bernoulli to the vocabulary of probability; it also led him to realize that we cannot choose a priori among various assumptions about different individuals' chances of death. Instead, we must observe actual patterns of mortality for similar individuals. And, Bernoulli adds, such a need for observations often arises in civil and ethical matters: "The safest way to estimate probabilities in these matters is not a priori or from first principles but *a posteriori* or from the outcomes observed in many similar examples." One must agree with van der Waerden's conclusion that this passage marks the inspiration for the law of large numbers. We see, moreover, that the example of mortality statistics was at the heart of this inspiration. Indeed, Bernoulli published a note on Titius' problem in 1686, in which he remarked, with apparent reference to Graunt's work, that the probabilities could be estimated from data like that collected in London and Paris.[24]

When Bernoulli sat down to write *Ars Conjectandi*, he once again began with Huygens' theory and Huygens' vocabulary. The first three parts of the treatise deal with the theory of games of chance and with combinatorics; the word *probabilitas* does not appear. But in Part IV he presents his numerical conception of probability and undertakes to show how the methods of the theory of games of chance can be used to calculate probabilities in practical life—*in civilibus, moralibus, et oeconomicis*.

In the following pages I quote Part IV of the *Ars Conjectandi* extensively and examine its approach to probability in detail. As I show in Sects. 4.1 and 4.2 below, the approach is based, in the first instance, on the notion of argument. In Sects. 4.3 and 4.4, I examine Bernoulli's rules for calculating probabilities and for combining the probabilities obtained from different arguments; it turns out that the probabilities obtained from these rules are, in general, non-additive. His rules of combination are cast in terms of the enumeration of cases from which he calculates the probabilities for each argument, but I recast them simply in terms of the probabilities, and compare them in this form with the more general and more satisfactory rule later proposed by Lambert. (See Sect. 6.4 below.) In Sect. 4.5, I note, but do not quote in full, Bernoulli's discussion of the problem of judging when his rules of combination are appropriate.

[24] For the note, see p. 283 of Vol. 1 of [7]. For van der Waerden's comments, see pp. 367–368 of Vol. 3.

The idea of calculating probabilities from arguments is, on the face of it, completely epistemic and completely divorced from chance or aleatory probability. But Bernoulli proposes to make such calculations using Huygens' method: he analyzes an argument by distinguishing cases that "are equally possible, or can happen with equal ease," and then calculates the probabilities afforded by the argument using the same formula that Huygens used to calculate expectations. And he introduces his law of large numbers as a tool for determining the ease with which different cases happen.

The law of large numbers is, quite justly, the most celebrated of Bernoulli's contributions in *Ars Conjectandi* and probably his most important contribution to mathematics. But it did not turn out to be primarily a contribution to the theory of epistemic probability. Instead, it quickly became, in the eighteenth century, the linchpin of an essentially aleatory theory; and since the mid-nineteenth century it has been transmuted into the "frequency interpretation of probability," whereby that aleatory theory is sharply divorced from epistemic conceptions. I do not examine the law of large numbers in this paper, for my purpose is rather to study Bernoulli's epistemic ideas, which have been all but forgotten in the shadow of that theorem's fame. In Sects. 4.5 and 4.6 below I do, however, consider the implications of Bernoulli's presentation of the theorem for his successors' conception of probability.

Ars Conjectandi ends with the proof of Bernoulli's law of large numbers. We are told that Bernoulli had not finished the treatise, but we are not told what he intended to add (Kohli 1975 [46]). It appears that he intended to clarify and illustrate the application of this theorem. Indeed, his repeated requests of Leibniz to help him find de Witt's tract tend to confirm the suspicion that he hoped to do so with mortality statistics (Kohli 1975 [48]).

4.1 Bernoulli's Conception of Probability

The conception of probability that Bernoulli presents in Part IV of *Ars Conjectandi* is decidedly and forthrightly epistemic. It encompasses probabilities in games of chance, but here, as in general, probability is a measure of our knowledge; a game of chance is merely an example where we lack certainty of what is or will be, and hence have only probability.

Bernoulli begins with a discussion of the notion of certainty. There are, he tells us, two kinds of certainty, *objective* certainty and *subjective* certainty. Every truth is completely certain "objectively and in itself." But we ourselves may be completely certain of something or only partly certain. This is certainty considered "subjectively and in relation to us"; it admits of degree because it is a "measure of our knowledge."

> The certainty of things considered in relation to us is not the same for all things, but varies manifoldly, being sometimes greater, sometimes less. Those things concerning which (by revelation, reasoning, perception, trial, self-knowledge, or otherwise) there is such certainty

> that we can in no way doubt their being or future being—these things enjoy absolute and utmost certainty. All other things hold an imperfect measure of certainty in our minds, greater or less according as there are more or fewer probabilities arguing that the thing is, will be, or has been.
>
> For probability is a degree of certainty and differs from it as a part from a whole. Indeed, if whole and absolute certainty, which we denote by the letter a or by the unit 1, is supposed for the sake of discussion to consist of five probabilities or parts, three of which stand for the being or future being of some event, the rest against, then that event is said to have $\frac{3}{5}a$, or $\frac{3}{5}$ of certainty. [*Ars Conjectandi*, p. 211]

Notice that a numerical probability is not merely a measure of intensity; it is a part of a whole. The idea that probability is a part of certainty is one whose time seems to have come at the end of the seventeenth century. Leibniz announced it in 1687, in a letter to V. Placcius ([56], Vol. VI, I, p. 36):

> In imitation of the mathematicians, I shall think of certainty or truth as a whole and of probabilities as parts, so that probabilities relate to truth as acute angles relate to a right angle.

And, as we shall see in Sect. 5.2 below, an anonymous Englishman offered a similar formulation to the English public in 1699 [1]. It would seem that for Bernoulli's generation it was necessary to represent probability as a part of a whole in order to treat it as a number. If it was to be a fraction, it had to be a fraction of something; and since probability was epistemic the something had to be subjective certainty.

Bernoulli's conception of probability is closely related to his conception of contingency. Contingency, like subjective certainty, is a feature of our knowledge. The outcome of a throw of a die, for example, is contingent not because it is inherently undetermined but because it lies beyond our ken.

> ... It is most certain, given the position, speed, and distance of a die from the gambling table at the moment when it leaves the hand of the thrower, that the die cannot fall otherwise than as it actually does fall. Likewise, given the present constitution of the atmosphere and given the strength, position, motion, direction and velocity of the winds, vapors and clouds, and the laws of operation by which all these things act on one another, tomorrow's weather cannot be other than what it in fact will be. Indeed, these effects follow from their own proximate causes with no less necessity than the phenomena of eclipses follow from the motion of the heavenly bodies. Nevertheless, the usage is observed that only the eclipses are reckoned as necessary, while the falls of the die and the future weather are reckoned as contingent. The only reason for this is that those things which are supposed to be given in order to determine these subsequent effects, and are indeed given in nature, are nevertheless not sufficiently known to us. And if they

> were, the study of geometry and physics is not sufficiently cultivated to enable these effects to be calculated from them in the manner in which eclipses can be computed and predicted from known principles of Astronomy. Before Astronomy had been brought to this point of perfection, eclipses themselves had to be regarded, no less than these other two and for the same reason, as future contingencies. Hence it follows that what can be seen as contingent by one person at one time is necessary to another person (or even the same person) at another time, after its causes have become known. Thus contingency especially depends on our knowledge... [*Ars Conjectandi*, pp. 212–213]

4.2 The Classification of Arguments

Probabilities are calculated from arguments:

> Probabilities are appraised from the *number* together with the *weight* of the arguments which in any way prove or indicate that a thing is, will be, or has been. By *weight*, moreover, I mean the force of proof. [*Ars Conjectandi*, p. 214]

The arguments which give rise to probabilities are not, of course, demonstrative or necessary. Rather, they are contingent arguments. And in order to calculate the probabilities they warrant, we must understand the nature of this contingency. As we shall see, some contingent arguments yield additive probabilities, while others yield non-additive probabilities.

In this Sect. I examine Bernoulli's classification of contingent arguments, and in the next Sect. I examine his methods for calculating probabilities from the various types of contingent arguments. Finally, in Sect. 4.4 below, I show how Bernoulli combines the probabilities obtained from the various arguments bearing on a question to obtain probabilities based on the total evidence.

Bernoulli distinguishes two ways in which an argument may be contingent. It may be contingent whether the argument arises or "exists" (Latin *existere*, to come into existence). And it may be contingent whether the argument "proves" (Latin *indicare,*[25] to indicate). Some arguments are contingent in one of these ways but not the other; some are contingent in both ways:

> One who examines the various arguments upon which an opinion or conjecture is based should take notice of a threefold distinction among them. Indeed, certain of them *exist necessarily and prove contingently*; others *exist contingently and prove necessarily*; finally, others *both exist and prove contingently.* I explain the distinction by

[25] The verb *indicare* is most often translated by "to indicate," but "to prove" seems preferable in this context. Bernoulli thinks of a contingent argument as one which proves in some cases and does not prove in others. In the cases where it does prove it definitely establishes the thing to be proved. It would not convey Bernoulli's meaning to say that it merely "indicates" in these cases.

> examples: My brother has not written me for a long time; I am not sure whether his indolence or his business is to blame; also I fear he might in fact have died. Here there are three arguments concerning the interrupted writing: indolence, death, and business. The first of these exists necessarily (by a hypothetical necessity, since I know and assume my brother to be lazy), but proves contingently, for it might have happened that this indolence did not keep him from writing. The second exists contingently (for my brother may still be among the living), but proves necessarily, since a dead man cannot write. The third both exists contingently and proves contingently, for he might or might not have business, and if he has any it may not be so great as to keep him from writing. Another example: I consider a gambler who, by the rules of a game, would win a prize if he threw a seven with two dice, and I wish to conjecture what hope he has of so winning. Here the argument for his winning is a throw of the seven, which proves it necessarily (by a necessity from the agreement entered into by the players) but exists only contingently, since other numbers of points can occur besides seven. [*Ars Conjectandi*, pp. 217–218]

It is difficult for us to grasp exactly what Bernoulli means by the existence of an argument, in part because his notion of an argument (Latin *argumentum*) is so broad; he applies the name to any testimony by a witness or an authority, as well as to any sign or circumstance "that seems to have a sort of bond with the thing to be proven." If, however, we think of an argument as consisting of premises and conclusion, then we surely come close to Bernoulli's meaning if we say that the argument exists contingently when the premises do not necessarily hold, and that the argument proves contingently when the premises do not necessarily entail the conclusion.

After distinguishing the different ways in which an argument can be contingent, Bernoulli distinguishes between *pure* arguments and *mixed* arguments. This distinction is of particular interest to us, because pure arguments give rise to non-additive probabilities.

> ...I call those arguments *pure* which prove a thing in certain cases in such a way that they prove nothing positively in other cases; I call those *mixed* which prove the thing in some cases in such a way that they prove the contrary in the remaining cases. Here is an example: A certain man has been stabbed with a sword in the midst of a rowdy mob, and it is established by the testimony of trustworthy men who were standing at a distance that the crime was committed by a man in a black cloak. If it is found that Gracchus and three others in the crowd were wearing tunics of that color, this tunic is something of an argument that the murder was committed by Gracchus, but it is mixed; for in one case it proves his guilt, in three cases his innocence, according to whether the murder was perpetrated by himself or by one of the remaining three; for it is not possible that one of these

> perpetrated it without Gracchus being thereby supposed innocent. But if indeed in a subsequent hearing Gracchus paled, this pallor of face is a pure argument; for it proves Gracchus' guilt if it arises from a guilty conscience, but it does not, on the other hand, prove his innocence if it arises otherwise; for it could be that Gracchus pales from a different cause yet is still the murderer. [*Ars Conjectandi*, p. 218]

In other words, the evidence of a mixed argument points to both sides of the question it addresses, whereas the evidence of a pure argument points only to the positive side.

4.3 The Calculation of Probabilities

According to my monograph *A Mathematical Theory of Evidence*, the probability p of a proposition and the probability q of its negation should obey $0 \leqq p \leqq 1$, $0 \leqq q \leqq 1$ and $p + q \leqq 1$. The case $p > 0$ and $q = 0$ corresponds to the presence of evidence in favor of the proposition and the absence of evidence against it, whereas the case $p > 0$ and $q > 0$ corresponds to the presence of evidence on both sides of the question. The case $p > 0$, $q > 0$ and $p + q = 1$ (additivity) occurs only when there is very strong evidence on both sides of the question.

From the perspective of this modern theory, it is natural to expect Bernoulli's analysis of argument to lead him to non-additive probabilities. A pure argument in favor of a proposition, for example, should give rise to probabilities $p > 0$ and $q = 0$; there is evidence for the proposition but none against it. And an argument which exists contingently but proves necessarily ought also to give rise to probabilities of this form; for the possibility that an argument for a proposition may fail does not in itself give any positive support for the proposition's negation. An argument which exists necessarily and is mixed should, no doubt, yield additive probabilities; but an argument which exists contingently and is mixed when it does exist should perhaps yield probabilities p and q with $p > 0$, $q > 0$ and $p + q < 1$.

The passage quoted in the last section is followed by a list of rules for calculating probabilities. Though Bernoulli does not say so explicitly, these rules do indeed sometimes yield non-additive probabilities.

> It is clear from what has been said thus far that the force of proof by which any given argument avails depends on the large number of cases whereby it can exist or not exist, prove or not prove, or even prove the contrary of the thing. Indeed, the degree of certainty or probability which the argument generates can be computed from these cases by the doctrine of the first Part [i.e., Part I of the book], just as the fates of gamblers in games of chance are usually investigated. In order to show this, we assume b is the number of cases where a

> given argument exists,[26] c is the number where it does not exist, and $a = b + c$ is the number of both together. Similarly, we assume β is the number of cases where it proves, γ is the number where it does not prove or else proves the contrary of the thing and $\alpha = \beta + \gamma$ is the number of both together. Moreover, I suppose that all the cases are equally possible, or can happen with equal ease. Otherwise, discretion must be applied and in the place of any case that happens more easily than the others one must count as many cases as it happens more easily. For example, in place of a case that happens three times more easily than the others I count three cases which can happen equally as easily as the others.
>
> 1. So first let the argument *exist contingently and prove necessarily.* By what has just been said, there will be b cases where the argument exists and thus proves the thing (or 1), and c cases where it does not exist and thus proves nothing. By Corollary I of Proposition III of Part I,[27] this is worth
>
> $$\frac{b \cdot 1 + c \cdot 0}{a} = \frac{b}{a},$$
>
> so that such an argument establishes $\frac{b}{a}$ of the thing, or of the certainty of the thing. [*Ars Conjectandi*, pp. 218–219]

What proportion of the certainty of the contrary is established by such an argument? Bernoulli does not pause to answer this question, but the only sensible answer is *zero*; the argument cannot provide support for the contrary merely by not existing. Hence we have probabilities $p = \frac{b}{a}$ and $q = 0$ for the thing, and $p + q < 1$ unless $p = 1$.

Bernoulli continues:

> 2. Next let the argument *exist necessarily and prove contingently.* By hypothesis, there will be β cases where it proves the thing, and γ cases where it does not prove or proves the contrary; this now gives a force of argument for proving the thing of
>
> $$\frac{\beta \cdot 1 + \gamma \cdot 0}{\alpha} = \frac{\beta}{\alpha}$$

[26] A literal translation yields "the number of cases in which it can happen that a given argument exists." But this English phrase is ambiguous. Bernoulli does not mean merely that the argument might or might not exist in these cases. He means that it *would* exist in each of these cases.

[27] Proposition III of Part I is Huygens' rule for computing an expectation: if there are p cases where one obtains a prize a and q cases where one obtains a prize b, and all $p + q$ cases are equally easy, then one's expectation is $(pa + qb/(p + q)$. Bernoulli's Corollary I to this proposition is merely the special case where $b = 0$. Bernoulli is applying this corollary to the case where the prize one obtains is certainty or lack of it.

> Therefore an argument of this kind establishes $\frac{\beta}{\alpha}$ of the certainty of the thing; and moreover, if it is mixed it establishes (as is clear in the same way)
>
> $$\frac{\gamma \cdot 1 + \beta \cdot 0}{\alpha} = \frac{\gamma}{\alpha}$$
>
> of the certainty of the contrary. [*Ars Conjectandi*, p. 219]

So we have additive probabilities in the mixed case: $\frac{\beta}{\alpha} + \frac{\gamma}{\alpha} = 1$. But if the argument is pure then it establishes none of the certainty of the contrary, and it would seem that $\frac{\gamma}{\alpha}$ must be replaced by zero.

> 3. If some argument *exists contingently and proves contingently*, I suppose first that it exists, in which case it is judged in the manner just shown to prove $\frac{\beta}{\alpha}$ of the thing and moreover, if it is mixed, $\frac{\gamma}{\alpha}$ of the contrary. Hence, since there are b cases where it exists and c cases where it does not exist and hence cannot prove anything, this argument is worth
>
> $$\frac{b \cdot \frac{\beta}{\alpha} + c \cdot 0}{a} = \frac{b\beta}{a\alpha}$$
>
> for proving the thing, and, if it is mixed, is worth
>
> $$\frac{b \cdot \frac{\gamma}{\alpha} + c \cdot 0}{a} = \frac{b\gamma}{a\alpha}$$
>
> for proving the contrary. [*Ars Conjectandi*, p. 219]

So in the case of a mixed argument which exists only contingently, we do indeed have a positive probability p for the thing and a positive probability p for its contrary such that $p + q < 1$. For

$$\frac{b\beta}{a\alpha} + \frac{b\gamma}{a\alpha} = \frac{b}{a},$$

which is less than one if the argument really exists only contingently.

It is clear that Bernoulli is calculating non-additive probabilities in the passages we have just quoted, and that he is not in the least embarrassed to do so. It does not even occur to him to remark on their non-additivity. The contemporary understanding of probability is so dominated by the model of chance that non-additive probabilities are apt to strike us as bizarre, artificial, or simply impossible. But Bernoulli was striving for an understanding genuinely grounded in the analysis of argument, and from this perspective the pure arguments, with their one-sided probabilities, seem exceedingly natural; it is the mixed arguments, with their chance-like probabilities, that seem relatively artificial.

It should be added, however, that Bernoulli did not consistently recognize the implications of possible non-additivity. At one point, for example, he identifies zero probability with impossibility:

> That is possible which has even a small part of certainty, impossible which has none or infinitely little. Thus that is possible which has $\frac{1}{20}$ or $\frac{1}{30}$ of certainty. [*Ars Conjectandi*, p. 211]

But, as the example of the pure argument shows, a small or zero degree of certainty for a thing need not imply great or complete certainty for its contrary.

4.4 The Combination of Arguments

The touchstone of a truly epistemic theory of probability is its emphasis on combination. There can be only one true aleatory probability for a chance event, and hence there is no question of combining different aleatory probabilities. But there is usually more than one argument for or against a proposition, and the epistemic probabilities from each argument must be combined to obtain a probability based on the total evidence.

Bernoulli turns to the problem of combination immediately after showing us how to calculate probabilities from a single argument. He does not give a single general rule of combination as Lambert was to do later. Instead, he gives several separate rules: one for combining two or more pure arguments, one for combining two or more mixed arguments, and one for combining pure with mixed arguments. As Lambert pointed out, the first two of these rules are sensible, while the third is not. Lambert's own rule, as we shall see, includes the first two of Bernoulli's rules as special cases and corrects the third; it also deals with more complex possibilities that Bernoulli addressed awkwardly or not at all.

When he turns to the problem of combination, Bernoulli suppresses consideration of the two distinct ways in which an argument may be contingent, and thus simplifies his notation. In his previous notation he had considered $a\alpha = (b+c)(\beta+\gamma)$ cases, $b\beta$ in which the argument proves "the thing" and $b\gamma + c\beta + c\gamma$ in which it proves the contrary or else nothing at all. Now he denotes the total number of cases for the first argument he considers simply by a, the number of these in which the argument proves by b, and the remainder by c. And he uses a similar notation for additional arguments.

He first adduces his rule for combining pure arguments:

> 4. If, further, more arguments are assembled for the proof of the same thing, and denoted

Arguments	1st	2nd	3rd	4th	5th	etc.
number of cases						
Total	a	d	g	p	s	etc.
Proving	b	e	h	q	t	etc.
Non-proving or proving the contrary.....	c	f	i	r	u	etc.

> then the force of proof resulting from the assemblage of all the arguments is computed as follows. First let all the arguments be *pure.*

> Then, as we have seen, the weight of the first argument considered alone will be $\frac{b}{a} = \frac{a-c}{a}$. (This stands for $\frac{\beta}{\alpha}$ if the argument proves contingently, or for $\frac{b\beta}{a\alpha}$ if it also exists contingently.) Now consider another argument which in c or $d-f$ cases proves the thing (or 1), and in f cases proves nothing, so that the weight of the first argument alone, which has been shown to be $\frac{a-c}{a}$, remains effective; the weight from both arguments together will be worth
>
> $$\frac{(d-f)1 + i(\frac{a-c}{a})}{d} = \frac{ad-cf}{ad} = 1 - \frac{cf}{ad}$$
>
> of the thing. Let a third argument be added; there will be h or $g-i$ cases that prove the thing, and i cases in which the argument is null and the two earlier proofs retain their power of proof by themselves, $\frac{ad-cf}{ad}$; whence the force of all three is judged to be
>
> $$\frac{(g-i)1 + i(\frac{ad-cf}{ad})}{g} = \frac{adg-cfi}{adg} = 1 - \frac{cfi}{adg}.$$
>
> And so on successively if there be further arguments at hand. From this it is clear that all the arguments taken together induce a probability which falls short of absolute certainty of the thing, or unity, by that part of unity that is obtained by dividing the product of the non-proving cases by the product of all the cases in all the arguments. [*Ars Conjectandi*, p. 220]

Dispensing with the analysis into cases, we can express this rule abstractly: if there are n pure arguments for a proposition and the i^{th} gives it probability p_i then the n together give it probability

$$1 - [(1-p_1)(1-p_2)\cdots(1-p_n)]. \tag{1}$$

The student of aleatory probability will recognize this as the chance of the occurrence of at least one of n independent events which have chances $p_1, p_2, \ldots, p_n$, respectively.

Bernoulli next adduces his rule for combining mixed arguments.

> 5. Next let all the arguments be mixed. Since the number of proving cases in the first argument is b, in the second e, in the third h, etc., and the number proving the contrary, c, f, i, etc., the probability of the thing to the probability of the contrary is as b is to c on the strength of the first argument alone, as e is to f on the strength of the second alone, and as h is to i on the strength of the third alone, etc. Hence it is evident enough that the total force of proof resulting from the assemblage of all the arguments should be composed of the forces of all the arguments taken singly, i.e., that the probability of the thing to the probability of its contrary should be in the ratio of $beh\cdots$ to

> $cfi \cdots$. Hence the absolute probability of the thing is $\frac{beh}{beh+cfi}$, and the absolute probability of the contrary is $\frac{cfi}{beh+cfi}$. [*Ars Conjectandi*, pp. 220–221]

(Notice the implicit assumption of additivity. The probability for the thing and the probability for its contrary add to one in the case of each mixed argument alone; it is taken for granted that they still do so when the arguments are combined.) As with the rule for combining pure arguments, we can express this rule abstractly: if there are n mixed arguments for a proposition and the i^{th} gives it probability p_i and hence gives its negation probability $q_i = 1 - p_i$ then the n together give the proposition probability

$$\frac{p_1 p_2 \cdots p_n}{p_1 p_2 \cdots p_n + q_1 q_2 \cdots q_n} \tag{2}$$

and give its negation probability

$$\frac{q_1 q_2 \cdots q_n}{p_1 p_2 \cdots p_n + q_1 q_2 \cdots q_n}. \tag{3}$$

After Laplace it became possible to contrive a Bayesian argument to justify (2) and (3), at least in the case where the n "arguments" are the testimonies of n different witnesses. (See Sect. 5.2 below.) But it would be a gross anachronism to attribute any Bayesian argument to Bernoulli.

We now come to the rule which was refuted by Lambert, the rule for combining pure with mixed arguments.

> 6. On the other hand, let some of the arguments be *pure* (say the first three) and some mixed (say the two others). Consider first the pure ones alone, which by Sect. 4 prove $\frac{adg-cfi}{adg}$ of the certainty of the thing, falling short of unity by $\frac{cfi}{adg}$. Hence there are $adg - cfi$ cases, as it were, where these three arguments together prove the thing, or unity, and cfi cases in which they prove nothing and consequently give the mixed arguments alone an opportunity to prove something. But by Sect. 5 above these two arguments prove $\frac{qt}{qt+ru}$ of the thing and $\frac{ru}{qt+ru}$ of the contrary. So the probability of the thing resulting from all the arguments is
>
> $$\begin{aligned}\frac{(adg - cfi)1 - (cfi)(\frac{qt}{qt+ru})}{adg} &= \frac{adgqt + adgru - cfiru}{adgqt + adgru} \\ &= 1 - \frac{cfiru}{adg(qt+ru)},\end{aligned}$$
>
> which falls short of complete certainty or unity by the product of $\frac{cfi}{adg}$ (the deficit from unity of the probability of the thing resulting from the pure arguments alone according to Sect. 4) by $\frac{ru}{qt+ru}$, the absolute probability of the contrary computed from the mixed arguments by Sect. 5 above. [*Ars Conjectandi*, p. 221]

For simplicity in re-expressing this rule, suppose we wish to combine one pure argument with one mixed argument, that the pure argument gives our proposition probability p_1 (and its negation probability 0 of course), while the mixed argument gives it probability p_2 (and hence its negation probability $1 - p_2$). Then Bernoulli's rule yields a probability of

$$p_1 + (1 - p_1)p_2 = 1 - (1 - p_1)(1 - p_2) \tag{4}$$

for the proposition. We shall examine this rule more closely shortly; it is the rule with which Lambert found fault.

Finally, Bernoulli awkwardly addresses the case where there are pure arguments on both sides of an issue.

> 7. Now if besides the arguments that tend to prove a thing, other pure arguments urging the contrary arise, then both categories of arguments must be weighed separately according to the preceding rules so as to establish the ratio that holds between the probability of the thing and the probability of the contrary. Here it should be noted that if the arguments adduced on each side are strong enough, it may happen that the absolute probability of each side significantly exceeds half of certainty, i.e., that both of the contraries are rendered probable, though relatively speaking one is less probable than the other. So it is possible that one thing should have $\frac{2}{3}$ of certainty while its contrary will have $\frac{3}{4}$; in this way both contraries will be probable, yet the first less probable than its contrary, in the ratio $\frac{2}{3}$ to $\frac{3}{4}$, or 8 to 9. [*Ars Conjectandi*, p. 221]

Not only does Bernoulli allow the probability of a thing and its contrary to add to less than one, he also allows them to add to more than one! Notice that he refuses, in the example he presents, to “ renormalize” so as to adjust the absolute probabilities to $\frac{8}{17}$ and $\frac{9}{17}$ This refusal is surely not accidental, for while $\frac{8}{17}$ and $\frac{9}{17}$ might seem reasonable final values in this example, a method which thus introduces additivity in the case of conflicting pure arguments would not be satisfactory in general. Suppose, indeed, that one pure argument proves $\frac{1}{10}$ of the certainty of a thing and another pure argument proves $\frac{1}{100}$ of the certainty of the opposite. Then it would seem that the two together, like each singly, fail to prove very much of anything; but a method that renormalizes to force additivity would have it that they together prove $\frac{10}{11}$ of the certainty of the thing.

It is easy to understand Lambert's criticism (Sect. 6.4 below) of Bernoulli's rule for combining a pure with a mixed argument. Indeed, the rule seems to give an unreasonable priority to the pure argument. For no matter how much the mixed argument disfavors the proposition the rule insists on awarding the proposition at least as much probability as the pure argument did. This is evident from (4): the result is always greater than or equal to p_1 Even when the mixed argument claims certainty in the negation (i.e., $p_2 = 0$), (4) still gives the proposition probability p_1 It is possible to avoid this absurdity; one might argue that Bernoulli meant the rule to apply only when the mixed argument really was an argument *for* the proposition—i.e., when $p_2 > \frac{1}{2}$. (It may be unfair, in any case, to dwell on the extreme case where $p_2 = 0$. In his definition of a mixed argument, Bernoulli had said that it should prove the thing *in casibus nonnullus*—in some cases. And the Latin *nonnullus* does literally mean non-zero.) But on the whole, Lambert's criticism seems pertinent. I believe that Bernoulli himself would have concurred with it had he seen Lambert's more elegant general rule.

Anticipating the presentation of Lambert's ideas in Sect. 6.4 below, I state his general rule in the language of (1)–(4): if one argument gives a proposition probability $p_1$and gives its negation probability q_1 ($p_1+q_1 \leqq 1$), and if another argument gives a proposition probability p_2 and its negation probability q_2 ($p_2+q_2 \leqq 1$), then the two arguments together give the proposition probability

$$\frac{p_1 + p_2 - p_1p_2 - p_1q_2 - p_2q_1}{1 - p_1q_2 - p_2q_1} \tag{5}$$

and give its negation probability

$$\frac{q_1 + q_2 - q_1q_2 - p_1q_2 - p_2q_1}{1 - p_1q_2 - p_2q_1}. \tag{6}$$

In the case where both arguments are pure ($q_1 = q_2 = 0$), (5) reduces to $p_1+p_2-p_1p_2 = 1-(1-p_1)(1-p_2)$, in agreement with (1), and (6) reduces to zero. In the case where both arguments are mixed ($q_1 = 1-p_1$ and $q_2 = 1-p_2$), (5) reduces to

$$\frac{p_1p_2}{p_1p_2 + q_1q_2}$$

and (6) reduces to

$$\frac{q_1q_2}{p_1p_2 + q_1q_2},$$

in agreement with (2) and (3). When the first argument is pure and the second is mixed ($q_1 = 0$ and $q_2 = 1 - p_2$), (5) reduces to

$$\frac{p_2}{1 - p_1(1 - p_2)};$$

unlike (4), this reduces to zero when $p_2 = 0$. Finally, consider two pure but opposed arguments; say $p_1 = \frac{2}{3}$, $q_1 = 0$, $p_2 = 0$, and $q_2 = \frac{3}{4}$, as in Bernoulli's last example. Then (5) becomes $\frac{1}{3}$ and (6) becomes $\frac{1}{2}$; the proposition is less probable than its negation, but the probabilities still are not additive.

4.5 Two Fundamental Questions

Bernoulli's account of epistemic probability raises two questions which are far more fundamental and important than the technical imperfections of his rules of combination. First, when are the rules of combination appropriate? Surely they are inappropriate unless the arguments being combined are somehow totally distinct or “independent.” But what exactly does this mean? Secondly, and more fundamentally, how do we judge the probability due a thing on the basis of a single argument? We count the cases, Bernoulli tells us. But how do we discern cases which happen with equal case? Bernoulli grappled with both these questions, and his efforts to answer them should be of interest to us today, not least because the same questions remain fundamental in any theory of epistemic probability based on a rule of combination.

When is it appropriate to apply the rules of combination? Bernoulli addresses this question immediately after presenting his rules:

> I cannot conceal the fact that in the specific application of these rules I foresee that many things will happen which can cause one to err frequently and shamefully unless one proceeds cautiously in discerning arguments. For sometimes arguments can seem distinct which in fact are one and the same argument. Or, vice versa, those which are distinct can seem identical. Sometimes such things are assumed in one argument as to demolish clearly a contrary argument... [*Ars Conjectandi*, pp. 221–222]

As one illustration of the possible problems, Bernoulli considers again the example of Gracchus. If the murderer in that example is also known to have red hair, and Gracchus is one of three men in the mob who has red hair, then his red hair is another mixed argument for his guilt. But if he is the only one in the mob with both red hair and a black tunic, then he stands convicted, and the red hair and black tunic should not be treated as two distinct arguments to be combined by the rule in Bernoulli’s paragraph Sect. 5.

Bernoulli’s discussion of combination suffers from the fact that he can only combine probabilities for a single proposition and its negation. Greater insight is possible in the case of Dempster’s rule of combination, which can operate on collections of probabilities for larger algebras of propositions. (See Sect. 8.2 of *A Mathematical Theory of Evidence.*) But we are unlikely to improve on Bernoulli’s practical advice: Proceed cautiously in discerning arguments.

How do we appraise probabilities on the basis of a single argument or a single item of evidence? This is surely the most fundamental question facing a theory of epistemic probability. The best answer, perhaps, is that this appraisal is ultimately and simply an act of judgment. (See Sect. 1.6 of *A Mathematical Theory of Evidence.*) But Bernoulli, as we have seen, directs our minds to an enumeration of “equally possible” cases, or cases that “happen with equal ease,” and suggests that we reckon from this enumeration just as we reckon concerning games of chance. And thus for him the question is how to identify these equally possible cases—or, alternatively, how to determine how much more easily a thing happens in one way than in another. As his answer to this question, he offers the famous theorem that Poisson later called the *law of large numbers.*

The problem, as Bernoulli points out, is that "equally possible cases" cannot be identified a priori in matters of practical life as they can be in games of chance:

> It has been shown in the preceding chapter how, from the number of cases in which the arguments for things of any sort can exist or not exist, prove or not prove, or even prove the contrary, their force of proof and the proportionate probabilities of the things can be deduced and reckoned by calculation. And hence it turns out that nothing else is needed in order to form correctly conjectures on any topic whatever but that first the number of these cases be accurately determined and then that it be determined how much more easily some can occur than others. But here, finally, we are in deep water. For this can scarcely be done in the smallest matters. And it hardly succeeds anywhere except in games of chance; the first inventors took pains to ensure fairness by arranging these so that the numbers of cases in which gain or loss must follow are fixed and known, and so that all these cases can happen with equal ease. In most other matters, whether they depend on the operation of nature or on the decisions of men, this is by no means the situation.... [*Ars Conjectandi*, pp. 223]

But even when the relative facilities of different cases cannot be found a priori, it may still be possible to find them a posteriori, by observing the frequencies with which the different cases actually appear in repeated trials:

> ...that which it is not possible to find *a priori* may at least be brought out *a posteriori*, by observing the outcome of many similar examples. For it should be presumed that a particular thing can henceforth happen and not happen in as many cases as it has been found to happen and not happen in similar circumstances in the past. If, for example, you once made a study of 300 men of Titius' present age and constitution, and you observed that 200 of them had died before the end of a decade and that the others had prolonged their lives further, then you will be able safely enough to conclude that there are twice as many cases whereby Titius would have to pay his debt to nature within the next decade as cases whereby this bound could be passed. And if anyone should have watched the weather and noted how many times it was calm or rainy for many years past, or if anyone should have very often watched two players and seen how many times one or the other emerged as victor in their game, he would have thereby discovered the probable ratio between the numbers of cases according to which the same events are able to happen or not happen in similar circumstances, in the past or in the future. [*Ars Conjectandi*, pp. 224–225]

Bernoulli continues this discussion at length and makes his meaning clear: there is a "true ratio," and this ratio can be estimated by the relative frequencies of actual outcomes.

Bernoulli acknowledges that others had already thought of using observations in similar ways, and cites the Port-Royal *Logic* as an example. (We should not be misled by his modesty. The idea that the frequency of an event in the future will be similar to its frequency in the past may have been a familiar one. The idea that both are approximations to and hence can be used to estimate a hidden "true ratio"—this is clearly original.) He claims this originality: he has proven a theorem (his law of large numbers) to the effect that a sufficiently large number of observations allows one to estimate the true ratio to a given accuracy with as great a *probability* as desired. He can even calculate how many observations are needed.

The law of large numbers caught the imagination of Bernoulli's successors and became the basis for a great mathematical theory. But it has been clear for several centuries that as a solution to the general problem of calculating epistemic probabilities, it seldom works. It only works, in fact, where one is dealing with a situation quite analogous to a game of chance; it must be possible to make repeated trials involving circumstances identical or nearly identical with those at hand. There are situations where such repeated trials are possible, and where the circumstances or individuals involved in these trials are sufficiently like the circumstances or individuals in question that probabilities calculated from the repeated trials are relevant and useful; the probabilities calculated from mortality statistics provide, for us as for Bernoulli, the most obvious examples. But more often we contemplate circumstances sufficiently unusual or detailed that repeated trials or observations involving similar circumstances are simply impossible.

4.6 The Ambiguity of Bernoulli's Legacy

Bernoulli had begun Part IV with a concept of probability which was truly general and hence epistemic. But in the end, as we have seen, he based his epistemic probabilities on the enumeration of cases that "can happen with equal ease." And by offering the law of large numbers as a device for measuring "ease of happening," he ensured an ultimately aleatory interpretation for this notion. Bernoulli's immediate successors resolved the conflict between the epistemic and aleatory aspects of his work in a way that largely favored the aleatory. They left aside his analysis of argument and his theory of combination, and they completely lost sight of the non-additive and hence non-aleatory probabilities that he had associated with pure arguments; they seized instead on his mathematical success, the law of large numbers, and they elaborated its role in the theory of chance. But they did not, and perhaps could not, undo the knot by which Bernoulli had bound chance with probability. Theirs was basically a theory of chance, but it had the name probability, and all the pretensions implied by that name.

From our twentieth-century standpoint, the notion of "ease of happening" is decidedly aleatory rather than epistemic; it refers to a property of the world rather than to an aspect of our knowledge. Bernoulli did not, of course, draw the strict distinction between aleatory and epistemic ideas that informs our standpoint; he considered games of chance a special case in his general art of conjecture. Nevertheless, "ease of happening" must have referred primarily, even in Bernoulli's time, to the thing-in-itself rather than to concepts of knowledge; certainly, such terminology had long been associated with chance rather than with epistemology. (Galileo and Huygens, in their discussions of chance, had used Latin terms such as *aeque facile* and *aeque in proclivi*; see Hacking [36], pp. 124–125.) And Bernoulli's explication of the ease of happening of an event by means of his law of large numbers forces one to a non-epistemic understanding: this ease of happening must be a feature of the objective world, for it is measured by the frequency with which the event actually happens.

It was clearly Bernoulli's intention that one should use his law of large numbers in conjunction with his rules for combining arguments. Observations of frequencies are to be used to determine equally possible cases for each argument, and then the arguments are to be combined. But if one's imagination is captured by the law of large numbers, then one is likely to ignore Bernoulli's theory of argument and envision a cruder approach: regard all the evidence as a single "argument" and obtain one's probability by measuring the frequency with which "similar arguments" are correct. The history of probability theory since Bernoulli has been characterized by this crude simplification; we have interpreted his proposal focusing the law of large numbers as a suggestion that probabilities in practical life are reflected in frequencies in the same way as probabilities in games of chance. And the defects of this crude approach have long bedeviled us; when we include all the evidence (all relevant facts and circumstances), the hopelessness of finding similar cases (let alone finding the truth in these other cases) becomes painfully obvious.

It comes close to the mark to say that the mathematical theory that emerged with the proof of the law of large numbers was basically about chance. But Bernoulli's immediate successors could not divorce this theory of chance from epistemic probability. For though the theory could not tell how to calculate epistemic probabilities in practical life, it required epistemic probabilities for its own purposes. Indeed, the law of large numbers tells us that the frequency of a random event will *probably* approximate its true aleatory probability. The "probably" demanded, for at least 150 years after Bernoulli, an epistemic interpretation. And with such an interpretation, it grew into the theory of statistical inference.

Through its role in statistical inference, Bernoulli's law of large numbers can indeed help us calculate epistemic probabilities, but the way it does so is less direct than Bernoulli had envisaged. And the debate over the theorem's role has not led to a general understanding of epistemic probability. Our present need is to understand both the nature of statistical inference and the nature of epistemic probability, and to this end we may do well to attend less to Bernoulli's law of large numbers and more to his theory of argument.

5 Probability and Chance after Bernoulli

After Bernoulli's death, the word *probability* continued to have its broad epistemic meaning in the European languages. But the connection with games of chance quickly came to dominate the thinking of those who sought to understand epistemic probability numerically. With the exception of Lambert, scholars of the 18th century seem to have taken it for granted that all numerical probabilities—including all probabilities of propositions—should obey the rules obeyed by probabilities in games of chance. In particular, probability became additive: the probability of a proposition and the probability of its negation had to add to one. Bernoulli's broader conception, which stressed the analysis of argument and consequently allowed non-additivity, was lost.

The disappearance of Bernoulli's broad conception was not, in my opinion, inevitable. Bernoulli did not invent the idea that probabilities are based on arguments and that their determination requires the combination of arguments. Rather, this idea was already a basic feature of the philosophical concept of probability, and he was compelled to take it into account in his attempt to make the concept numerical. Had Bernoulli been immediately followed by a scholar of Lambert's breadth—by a mathematician who shared Bernoulli's understanding of and interest in the philosophical concept of probability—the theory of argument might have survived. But in fact, Bernoulli was followed by two mathematicians, Montmort and De Moivre, who were relatively narrowly concerned with the theory of games of chance and saw in Bernoulli's struggle with *probability* only a license to give that name to their subject. In Sect. 5.1 below, I examine the relation of Montmort and De Moivre to Bernoulli, and their influence on the 18th century's understanding of probability.

Though Bernoulli's explicit recognition of non-additive probabilities reappeared in the 18th century only in Lambert, some aspects of his ideas on the combination of arguments survived for nearly a century in the form of rules for calculating the "credibility of testimony." These rules first appear in 1699, in an anonymous paper in the *Philosophical Transactions*, and various authors reproduce them in the course of the 18th century, until they are finally replaced by Laplace's Bayesian approach to testimony. These authors do not appear to think of the probabilities calculated by these rules as non-additive, but, as I argue in Sect. 5.2 below, the rules make good sense only from such a perspective.

5.1 Probability becomes Additive

The disappearance of Bernoulli's broad conception of numerical probability may have been partly due to the delay in the publication of *Ars*

Conjectandi.[28] The existence of the treatise became well known during the eight years between Bernoulli's death and its publication. But in the absence of the actual text, it was hardly possible for scholars to form a conception of the subtlety of its ideas. They were left instead to form their own conception as to how the theory of chance could be applied to practical affairs.

The learned public first learned of *Ars Conjectandi* from two eulogies published in 1706, one in the *Journal des Sçavans* [68],[29] and one, by Fontenelle, in the memoirs of the French Academy of Sciences [32].[30] In their report on *Ars Conjectandi*, these eulogies give the general impression that Bernoulli's idea was to study the role of randomness in practical affairs. In the eulogy in the *Journal des Sçavans*, for example, we read as follows:

> The title of the work is supposed to be *de Arte Conjectandi*, "On the Art of Conjecturing." In it the author in effect determines, and reduces to calculation, the various degrees of certitude or of verisimilitude of the conjectures that one can frame about things which depend on chance; and he even extends this to civil life and to practical affairs.

In Pierre Rémond de Montmort's *Essay d'analyse sur les jeux de hazard*, published in 1708, this narrowing of Bernoulli's conception becomes more definite: the notion of numerical probability is essentially narrowed to the paradigm of games of chance, and the attempt to calculate probabilities in situations other than games of chance is taken as an attempt to assimilate those situations completely to this paradigm.

The body of Montmort's book deals strictly with games of chance.[31] But in its preface, he discusses Bernoulli's broader ambitions. He had not seen *Ars Conjectandi*, but he had read the eulogies, and he explains that Bernoulli "had undertaken to give rules for judging the probability of future events of which knowledge is hidden from us, both in games and in other things in life where chance alone plays a part." How can chance be given a large role in a deterministic world? Montmort's explanation echoes the opening paragraphs of Part IV of *Ars Conjectandi*:

> Strictly speaking, nothing depends on chance. When one studies nature, one is quickly convinced that its Author acts in a general and uniform way, characterized by infinite wisdom and foresight. So in order to give chance[32] a meaning that conforms to the true Philosophy,

[28] The delay was due to apprehension on the part of Jacob's widow and son that an editor, even Jacob's brother Johann or nephew Nicolaus, might plagiarize his manuscripts. Eventually the editorship of *Ars Conjectandi* was entrusted to Nicolaus. See Kohli 1975 [46].

[29] According to Montmort (1708 [61], p. iv), the author of this eulogy was Joseph Saurin.

[30] See Kohli 1975 [46].

[31] For assessments of Montmort's contribution to the theory of chance, see David 1962 [17] and Henny 1975 [39].

[32] French *hazard*.

> we must say that all things are governed by laws which are certain but whose ordering we most often do not know, and that those things depend on chance whose natural causes are hidden from us. After making this definition, we may say that the life of man is a game governed by chance.

So life is a game of chance. Ironically, a stubborn determinism has extended the domain of chance: since chance must be understood as an aspect of our knowledge rather than as a feature of reality, one may as well use the paradigm of chance to describe all our uncertainties.

If we could count the cases in practical life as in games of chance, says Montmort, we could proceed as Arnauld had recommended:

> . . . generally, with regard to all things in life about which we have to make decisions, our deliberations should come down, as in the case of betting in games, to a comparison of the number of cases in which a certain event will happen to the number of cases in which it will not happen; or to speak as a geometer, to an examination of whether what we hope for, multiplied by the degree of probability that we will obtain it, equals or exceeds our stake, i.e., the advances that we shall have to make, whether they be effort, money, credit, or whatever.
>
> It follows that the same rules of analysis that we have used in games to determine the players' bets or the way in which they should play should also be used to determine the correct measure of our expectations in our many undertakings and to teach us the way we should behave in order to obtain the greatest possible advantage.

But Montmort has the same problem as Bernoulli did; he does not know how to count the cases. As he puts it, he is unable to "formulate hypotheses which, being based on established facts, could guide and support me. . . " And thus, as he confesses, he is unable to emulate Bernoulli's application of the theory of games of chance to practical life.

Montmort's treatment of games of chance is, for the most part, in Huygens' vocabulary—i.e., he counts cases and computes expectations. On a few occasions early in the book he uses the word *probability*, but only in the vague way that it was used by Arnauld. Notice, though, that the quotation above goes one step beyond Arnauld in a technical sense. Since Montmort would have us compare the value of our stakes with the product of the value of the prize by the degree of probability of obtaining it, he must be thinking of this probability as a number between zero and one—it must mean the ratio of the number of favorable cases to the total number of cases.

Montmort's contribution to the theory of games of chance was soon followed by that of Abraham De Moivre (1667–1754). De Moivre was already a middle-aged and powerful mathematician when Montmort's book aroused his interest, and he quickly broke new ground; he became, by all accounts, the most important contributor to the theory between Bernoulli and Laplace.[33] I shall not attempt here a general assessment of De Moivre's mathematical contributions, but I do wish to note his influence in fixing the word *probability* as a technical term within the theory of games of chance. It is in De Moivre's *Doctrine of Chances* that we first find *probability* and the ratio to which it refers playing a basic role in the vocabulary and methods of this theory.

De Moivre's *De Mensura Sortis* [18], his first essay on the theory of games of chance, was published in 1711, still two years before the appearance of *Ars Conjectandi*. It was inspired by Montmort's essay, and it opens with an explanation of probability and expectation that would sound familiar to a reader of Montmort:

> If p is the number of cases by which some event may happen, and q is the number of cases by which it may not happen, then both the happening and the not happening of the event have their degrees of probability. And if all the cases by which the event can happen or not happen are equally easy, then the probability of happening will be to the probability of not happening as p is to q.
>
> If two players A and B contend about the event so that A wins in the p cases and B wins in the q cases, and a is the amount of the stakes, then the prospect[34] or expectation of A himself will be $\frac{qa}{p+q}$. . . .

The essay solves a long list of problems. But as in the case of Montmort's book, these problems are mostly posed in the language of expectation; *probability* does not play a large role in their statement or solution.

After the appearance of *Ars Conjectandi*, De Moivre developed a new vocabulary. He adopted Bernoulli's idea that *probabilities* are numbers between zero and one, and he took as axiomatic the rule that the probability of an event is the ratio of the favorable number to the total number of cases. His *Doctrine of Chances* [19], published in 1718, opens as follows:

> The Probability of an Event is greater, or less, according to the number of Chances by which it may Happen, compar'd with the number of all the Chances, by which it may either Happen or Fail.
>
> Thus, If an Event has 3 Chances to Happen, and 2 to Fail, the Probability of its Happening may be estimated to be $\frac{3}{5}$, and the Probability of its Failing $\frac{2}{5}$.
>
> Therefore, if the Probability of Happening and Failing are added together, the Sum will always be equal to Unity.

And the word *probability* plays a fundamental role in the statement and solution of the problems that comprise the body of the book. Instead of asking for expectations or the ratios of expectations, these problems demand to know the probabilities of various events.

[33] See David 1962 [17], Schneider 1968 [69], and Kohli 1975 [47].

[34] Latin *sors*, meaning *chance*, *lot*, or *fortune*.

De Moivre's definition of probability was repeated by Laplace and others, and has come to be called the *classical* definition. There is a good deal of justice in the claim that the definition originated with De Moivre. Cardano and Fermat had both calculated the same ratio on occasion, but they had not, of course, called it a probability. Bernoulli did sometimes call this ratio a probability (see Sect. 4.3 above), but only in the case of an argument which exists necessarily.

Notice the rule of additivity that De Moivre announces in the last paragraph of the quotation above. This appears to be the first statement of a rule of additivity for probabilities; the more general rule (that the probability of the disjoint union of two events should be the sum of their probabilities) was first stated by Bayes, about half a century later.

One could argue that De Moivre used *probability* as a technical term within a mathematical theory of chance, and that it was quite proper for him to set aside the question of whether his mathematical definition always fit the accepted epistemic meaning of the word. But his work, by virtue of its mathematical success, quickly became the last word on the nature of numerical probability. With the single exception of Lambert, the mathematicians and philosophers who followed De Moivre took it for granted that probabilities were additive.

Today we understand "pure mathematics" as an exercise unto itself; a mathematician may use whatever words he pleases, and the theorems he proves about these words need not have any significance for other people's use of the words. But the 18th century gave mathematics a greater authority: if the mathematicians had demonstrated that probability could be measured and that numerical probabilities were additive, then this had to be true.

5.2 The Credibility of Testimony

In 1699 a brief anonymous paper entitled "A Calculation of the Credibility of Human Testimony" appeared in the *Philosophical Transactions* of the Royal Society. This paper echoed Bernoulli's conception of probability as a degree of certainty, and set forth two rules for combining the "credibilities" of witnesses. The first of these rules concerns successive testimony (or chains of testimony), and the second concerns concurrent testimony. Both can be construed as applications of Bernoulli's methods, and both were repeated by various authors during the 18th century. As I explain below, these rules make good sense only if one thinks of the probabilities or credibilities they yield as non-additive or "one-sided." Towards the end of the century, when the work of Bayes and Laplace made scholars more accustomed to the implications of additivity for the probabilities of propositions, these rules began to seem wrong-headed and were severely criticized. The 19th century replaced them with a pair of rules that were more sensible in terms of the additive theory.[35]

[35] Work on this topic during the 1980s includes Glenn Shafer's "The combination of evidence," in *International Journal of Intelligent Systems*, 1:155–179, 1986,

The author of “A Calculation of the Credibility of Human Testimony” has a wondrously varied vocabulary. He begins by writing of degrees of certitude and degrees of confidence:

> *Moral Certitude Absolute*, is that in which the Mind of Man entirely acquiesces, requiring no further Assurance: As if one in whom I absolutely confide, shall bring me word of 1200 *l* accruing to me by Gift, or a Ships Arrival; and for which therefore I would not give the least valuable Consideration to be Ensur’d.
>
> *Moral Certitude Incompleat*, has its several Degrees to be estimated by the Proportion it bears to the *Absolute*. As if one in whom I have that degree of Confidence, as that I would not give above One in Six to be ensur’d of the Truth of what he says, shall inform me, as above, concerning 1200 *l*: I may then reckon that I have as good as the Absolute Certainty of a 1000 *l*, or five sixths of Absolute Certainty for the whole Summ.

Later he writes of degrees of certainty and degrees of credibility. All these terms are synonyms; they all refer to a fraction between 0 and 1. At one point he even casually uses the word *probability*: “So if the Probability or Proportion of Certitude transmitted by each Reporter, be $\frac{100}{106}\dots$”; this appears to be the first time that the word was explicitly used in print to denote a fraction. The paper as a whole is shallow in comparison with *Ars Conjectandi*, but it is remarkable as a sign of the extent to which Bernoulli’s ideas were in the air at the end of the 17th century.

The two rules are quite simple. The rule for successive testimony says that if a report has been relayed to us through a chain of n witnesses, each witness having a degree of credibility p, then the credibility of the report is p^n. And the rule for concurrent testimony says that if a fact is testified to simultaneously by n witnesses each with credibility p, then the credibility of their common report is $1-(1-p)^n$. (Here $0 \leqq p \leqq 1$.) Thus the credibility of a report is weakened by transmission through many witnesses but strengthened by the simultaneous concurrence of many witnesses.

There are many grounds on which to criticize these rules. The rule for successive testimony paradoxically assumes certain knowledge as to the length and nature of the chain. And the rule for concurrent testimony can only be defended if one assumes an unlikely independence in the evidence and motives of the different witnesses. (See Sect. 8 of Shafer, 1976 [70].) But we wish only to note here that the rules make no sense at all unless the credibilities are taken to be one-sided: one’s confidence p in the truth of a report is accompanied by confidence 0 (not $1-p$) in its falsehood.

Stephen M. Stigler’s “John Craig and the probability of history: From the death of Christ to the birth of Laplace,” *Journal of the American Statistical Association*, 81:879–887, 1986, and Sandy L. Zabell’s “The probabilistic analysis of testimony,” *Journal of Statistical Planning and Inference*, 20:327–354, 1988.

Consider the rule for successive testimony. It is sensible, and in the same spirit as the multiplication made by Bernoulli in the case of an argument that both exists and proves contingently. Each successive transmission ought to diminish our confidence in the report. But does this signify an increasing confidence that the report is false? Surely not. Rather, we should say that our possibly quite small confidence p^n in the report's truth is still accompanied by a confidence 0 in its falsehood.

The rule concerning concurrent testimony is merely Bernoulli's rule for combining pure arguments, and must be interpreted in the same one-sided way. Notice, for example, that $1-(1-p)^n$ tends to 1 with increasing n even if p is quite small; the report acquires high credibility even if each witness has a credibility of only $p=\frac{1}{10}$, say. This would be absurd if the credibility $p=\frac{1}{10}$ for each witness were thought to entail a confidence of $1-p=\frac{9}{10}$ in the falsehood of his report.

Who wrote "A Calculation of the Credibility of Human Testimony"?[36] Todhunter (1865 [78], p. 55) notes a suggestion that it may have been written by the Scottish mathematician John Craig. Craig, a friend of Newton and a mathematician of some note, did publish a treatise entitled *Theologiae Christianae Principia Mathematica* in 1699 [16], in which he professed to show, *inter alia*, how the probability of a history diminishes with time, distance, and transmission. The suggestion that he was also the author of "A Calculation..." has been repeated many times, and credited to varying degrees. But it is rendered unlikely, I think, by the internal evidence.

The ideas to which Craig's treatise and "A Calculation..." attempt to give mathematical form do indeed overlap. Both argue that the credibility of a narrative diminishes by transmission, that the diminution is slower for a written than for an oral tradition, and that it can be retarded by concurrent chains of transmission. But the mathematical treatments in the two works are quite different. The rules in "A Calculation..." are obviously inspired by the theory of games of chance. Craig's treatise, in contrast, does not betray any hint of a connection with that theory; it takes the initial probability of a history to be an arbitrary positive constant and assumes that it decreases linearly with the length of the chain of witnesses. (There are additional deductions for distance in space and time; these are inversely proportional to the square of the distance!) Internal evidence shows that Craig wrote his treatise in 1696; perhaps "A Calculation..." reflects an alternative approach that occurred to him later. More likely it is someone else's alternative approach.

Craig attracted widespread attention and censure by pretending to calculate the rate of diminution of the Christian faith and deducing therefrom the

[36] Historians of probability now know, as some historians of theology had known all along, that the article was written by George Hooper (1640–1727), Bishop of Bath and Wells. It was reprinted in two collections of Hooper's works published by Oxford, one in 1757 and one in 1855. Brown Grier, of Northern Illinois University, called this to the attention of historians of probability in 1981.

data of the second coming. He also repeated the argument for Pascal's wager and attempted to mathematize pleasure. The author of "A Calculation..." was more prudent and remains anonymous.

The rules for concurrent and successive testimony given by our anonymous author were fairly popular during the 18th century. The author of the article *Probabilité* in Diderot's *Encyclopédie* [28][37] adopted them. According to Todhunter (p. 441), C.-F. Bicquilley adopted them in his *du Calcul des probabilités*, published in 1783. And according to Prevost and Lhulier, writing in 1797 [65] (p. 122), they were advocated in "many other memoirs and courses." We may cite two other 18th century authors who produced rules along similar lines: J.H. Lambert, whose work we examine in more detail in Sect. 6 below, and Nicolaus Bernoulli, Jacob's nephew.

In 1709, Nicolaus defended and published a dissertation [8] in which he sought to apply his uncle's *Ars Conjectandi* to legal questions.[38] Among other topics, he considered how the credit we can give to the innocence of an accused person diminishes from unity as evidence accumulates against him.[39] Nicolaus argued, in effect, that if each item of evidence had sufficient force to reduce the plausibility[40] of his innocence to $\frac{2}{3}$, then n items should reduce it to $\left(\frac{2}{3}\right)^n$. Notice that Nicolaus' rule measures the plausibility of innocence rather than the probability of the arguments for guilt. If, however, we follow Shafer (1976 [70], p. 144) in identifying the plausibility of innocence with the deficit from unity of the probability of guilt, then this rule reduces to Jacob's rule for combining pure arguments and thus agrees with the rule for concurrent testimony: each item of evidence against the accused has probability $\frac{1}{3}$, and all together have probability $1 - (1 - \frac{1}{3})^n = 1 - \left(\frac{2}{3}\right)^n$.

[37] The article on probability in the *Encyclopédie* has sometimes been attibuted to Diderot himself; see, for example, Jean Mayer's "Diderot et le calcul des probabilités dans l'Encyclopédie," in *Revue d'histoire des sciences*, Vol. XLIV, pp. 375–391. Work by Jean-Daniel Candaux (*Recherches sur Diderot et sur l'Encyclopédie*, number 15, October 1993, pp. 71–96) and Thierry Martin ("La logique probabiliste de Gabriel Cramer," *Mathématiques et sciences humaines*," number 176, 2006:4) has shown that it was written by Charles Benjamin de Lubières, who drew its ideas from an unpublished paper by Gabrielle Cramer. See also Martin's "La logique probabiliste de Gabriel Cramer," *Electronic Journal for History of Probability and Statistics* (www.jehps.net), 2(1), November, 2006.

[38] The dissertation is reprinted in Vol. 3 of *Die Werke von Jakob Bernoulli* and discussed in detail by Karl Kohli on pp. 541–556 of that volume. As Kohli's discussion shows, most of Nicolaus' examples can be readily understood from the viewpoint of an additive conception of probability.

[39] See pp. 54–55 of the dissertation, pp. 69–170 of the excerpts published in *Actorum Eruditorum*, or p. 196 of Todhunter [78].

[40] Nicolaus does not use "plausibility" or any other such term. He says merely that the accused's innocence would be worth such and such: *ejus innocentia valeret* $\left(\frac{2}{3}\right)^{10}$, etc.

In Chap. XI of his famous *Théorie analytique des probabilités*, Laplace discussed the probability of testimony and gave Bayesian rules for successive and concurrent testimony.[41] Laplace's rules were the ones commonly received in the 19th century. They are more complicated than the 18th century rules, but they make more sense if one insists on the additivity of probability.

Consider first the case of successive testimony, where one witness reports the report of a second. If we claim, in the case of each witness, that there is a probability p that the witness tells the truth and a probability $1-p$ that he lies, then there is a probability p^2 that both tell the truth and a probability $(1-p)^2$ that they both lie. But the first lies by telling the opposite of the truth and the second lies by telling the opposite of the report of the first, so when both lie the second reports the truth. Hence the total credibility of the report of the second is $p^2+(1-p)^2$ In the case of a chain of three witnesses, similar principles give a credibility of $p^3+3p(1-p)^2$ to the final report; in the case of n witnesses, a credibility of

$$\sum_{0\leqq k\leqq \frac{n}{2}} \binom{n}{2k} p^{n-2k}\,(1-p)^{2k} = \frac{1}{2}+\frac{1}{2}(2p-1)^n .$$

This is Laplace's rule for successive testimony, in the special case where all the witnesses have the same credibility p and must always choose between the same two reports, each of which has prior probability $\frac{1}{2}$.[42]

The derivation of Laplace's rule for concurrent testimony is also straightforward. We suppose again that the witnesses act independently and that each tells the truth with probability p and lies with probability $(1-p)$. Then with probability p^n all will tell the truth and with probability $(1-p)^n$ all will lie. When we find that all n agree, we know that they either all told the truth or all lied; so if we had previously given prior probability $\frac{1}{2}$ to both the truth and falsehood of what has been reported, then we have a posterior probability of

$$\frac{p^n}{p^n+(1-p)^n}$$

that the report is true, and a posterior probability of

$$\frac{(1-p)^n}{p^n+(1-p)^n}$$

that it is false. Notice that though the derivation is based on Bayesian principles, the result is merely Bernoulli's rule for combining mixed arguments.

Though the development of these two Bayesian rules definitely followed and depended on Laplace's 1774 Bayesian paper, they were not, apparently,

[41] Chapter XI first appears in the second edition, published in 1814 [53].

[42] In Laplace's exposition, the witnesses are reporting on which ball was drawn from an urn, so that these are prior probabilities derived from the number of balls in the urn.

first adduced by Laplace himself. The rule for concurrent testimony can be discerned in Condorcet's work; see p. 10 of his 1785 *Essai* [13] and pp. 357 and 400 of Todhunter. And according to Todhunter (p. 463), it was also adduced by Matthew Young in 1798. Both rules were given by Prevost and Lhulier in 1797 [65] (see Sect. 6.6 below), and Prevost claimed priority in that paper for the rule for successive testimony. Both Young and Prevost and Lhulier explicitly noted and rejected the 18th century rules.

6 Lambert's Treatment of Probability

Johann Heinrich Lambert was born in 1728 in Mulhouse in the Alsace. Mulhouse was then part of Switzerland; his family had settled there after fleeing Catholic persecution in Lorraine in 1635. As a youth he did not enjoy the advantages of Bernoulli and Leibniz; his family was poor and his formal education was limited. But his energy and talent enabled him to advance quickly in the world; he became a tutor for a wealthy Swiss family in 1748, and during his 10 years in their employment he developed into a creative and broad-ranging scholar. In 1765 he obtained an academic post in the Berlin academy, where he remained until his death, from a neglected pulmonary infection, in 1777. His literary education was never outstanding; he wrote awkwardly in Latin and French. But during his relatively short career he distinguished himself as a mathematician, natural scientist, and philosopher.

Lambert's contribution to the theory of games of chance, considered as a branch of pure mathematics, was slight. But he frequently applied the theory in his scientific work, and some of these applications were path-breaking.[43] Of particular interest are his discussions of the theory of errors; on one occasion (§§271–306 of his *Photometria*, published in 1760), he formulated what we now call the method of maximum likelihood.[44] Our present interest is in his philosophical treatment of probability, and particularly in the discussion of the probability of propositions in his *Neues Organon*. It was here that he developed Bernoulli's theory of non-additive probability and corrected Bernoulli's rules of combination.

Organon (Greek for "tool") was the name given to Aristotle's treatise on logic. Francis Bacon (1561–1626) had entitled his own treatise on inductive logic *Novum Organon*, in the belief that it would replace Aristotle's work. Lambert wrote his *Neues Organon* while he was still a tutor[45] and published it in 1764. It was his first venture into philosophy and also his best known,

[43] Sheynin [73] is valuable for its wealth of references.

[44] A more careful exposition of this idea was published by Daniel Bernoulli in 1777 [5]. The idea did not survive as an independent approach to statistical inference, for it was absorbed into Laplace's Bayesian synthesis; see Laplace's memoir of 1774.

[45] See Eisenring, 1942 [29], p. 8.

though some philosophers consider his later *Anlage zur Architektonik* (1771) more important.

Neues Organon is divided into four parts, each with a title derived from Greek: 1) the *Dianoiologie*, which studies the laws of thought, 2) the *Alethiologie*, which studies the nature of truth, 3) the *Semiotik*, which studies semantics, and 4) the *Phänomenologie*, which studies how to distinguish appearance from truth. All the passages that I translate below are from Chap. 5 of the *Phänomenologie*, which is entitled *Von dem Wahrscheinlichen* ("On the probable"). The paragraphs of the *Phanomenologie* are numbered, and these numbers are reproduced in the translations.

In Sects. 6.1 and 6.2, I examine Lambert's attitude toward the different "kinds" of probability and quote his derivation of two rules for calculating probabilities of propositions. I then proceed, in Sects. 6.3 and 6.4, to the matters of principal interest; in Sect. 6.3, I show how Lambert's treatment of the syllogism led him to the recognition of the possible non-additivity of probability, and in Sect. 6.4, I present the passages in which he corrects and generalizes Bernoulli's rules of combination. In Sect. 6.5, I record the scant notice that posterity gave to Lambert's ideas.

6.1 Lambert's Conception of Probability

In contrast to most of the mathematicians who studied the doctrine of chances after Bernoulli, Lambert shared Bernoulli's interest in epistemic probability. But there is a great difference between their standpoints. Bernoulli was still trying to perfect the synthesis of the epistemic concept with chance. Lambert, 80 years later, was already in our modern situation: the synthesis had become so well accepted, and aleatory ideas had become so dominant within it, that he could do justice to epistemic ideas only by distinguishing among different "kinds" of probability. (Writing in German, Lambert used the word *Wahrscheinlichkeit*, but then as now, this was clearly understood to be the equivalent of Bernoulli's Latin *probabilitas*.)

He began by distinguishing three kinds of probability related to physical events. Two of these can be classified as aleatory: the first kind, which consists of probabilities that can be known a priori in games of chance and similar setups, and the second kind, which consists of probabilities that resemble those in games of chance but which can only be found a posteriori. The third kind is clearly epistemic; it consists of probabilities that are given to events by virtue of inference from effects or from circumstances.

After dealing with the probabilities of physical events. Lambert turns to his main subject: logic and the probability of propositions. Here, as we might expect, his probabilities are more consistently epistemic.

Lambert's distinction between a priori and a posteriori probabilities is taken, of course, directly from Bernoulli. We need not pause over his account of a priori probabilities, but his account of a posteriori probabilities is of some interest.

> §153. Games of chance have the special feature that the number of possible cases and their individual degrees of possibility can be determined from the structure of the game. In this way, the probability of each *case* can be calculated *a priori*. But it is evident from what has already been said that this could also be done *a posteriori*, were the game repeated for a long time or infinitely many times. For this reason, people have also begun to apply the theory of probability to other situations. Not only in games of chance, but also in countless other matters, nature operates according to very complex laws, and such that only the net results of all these laws can be known from experience. Thus these results have been tabulated in order to find the extent of each law and the probability of the case where it dominates. This is the second general kind of probability, and we will elucidate it in greater detail.
>
> §154. One considers a proposition arising from experience but from which experience sometimes deviates, without it being possible to discuss the circumstances under which one or the other happens. One records both kinds of cases, as they occur and without selection, so as to determine from the totals of each the ratio of the correct cases to the failures. *This ratio determines the natural*[46] *degree of probability of the proposition.* It tells not only that some A are B, but more precisely, how many are and how many are not. Thus, for example, the annual death rate in large cities has been determined from annual records of the numbers living and dying.

He emphasizes that all cases of A must be counted without selection, "as they occur," in order for the calculation to be valid. And he also requires that the tabulation be made under constant conditions, or else that the observations continue long enough for the irregularities to balance out and the ratio to become constant or only imperceptibly variant.

If no such balancing out occurs, then the a posteriori calculation is impossible, and our attention turns to the third, more epistemic kind of probability. This third kind is based either on incomplete induction from effects to causes, or else on the appraisal of circumstances:

> §161. Furthermore, the actual tabulation of cases can only be undertaken and used where there is something constant and definite in the variation in successive causes. Otherwise, if new causes were always appearing, or if some were disappearing without being replaced by equivalent ones, or if they lasted for a short time (as in many human activities), one would be unable to find any constant or definite ratio between different kinds of cases. So in situations such as this, in order to determine with probability or certainty what will

[46] *der Natur gemaß*

> actually happen one must use entirely different principles and take individual circumstances into account.
>
> §162. This leads us to the third general kind of probability. Indeed, we can assure ourselves immediately of the occurrence of an event if we have either seen it or done it ourselves, or if we see effects of it which necessarily imply its occurrence. But if the effects we find are insufficient to establish its occurrence then here again we attain only a certain degree of probability. We get no further when forced to decide whether the event has happened or will happen from circumstances, causes, or motives—especially when the event is subject to hindrance...

Lambert's understanding of induction is based on the idea that no two causes can have exactly the same effects, so that a cause is established once each of its "direct" effects is observed. If only some of these effects are observed, then the cause is only probable. The above passage continues as follows:

> ...Effects are unreliable if they could have derived, collectively or individually, from other causes. But the more numerous and varied such other causes must have been, the less probable it is that they should have all coincided to produce effects collectively derivable from a single event. Here particulars in the effects are especially helpful in inferring a cause. Lacking these, one takes the cause as *a hypothesis* and derives the effect from it. But this sort of reasoning is only an induction, and it must therefore be complete if it is to serve as a proof. For, certainly, when every effect that a cause must entail in given circumstances is definitely observed, one may validly conclude that these effects could not have derived from anything else (Dianoiolog. §569, 595; Alethiol. §176). Incidentally, there is also a distinction concerning the effects themselves, as to whether they are direct or indirect. It is intrinsically sufficient for induction that the direct effects be complete, since the indirect effects derive from them. And it is often necessary to assemble many indirect effects to make up for the lack of one direct one. We make this remark with a view to calculating probabilities from such inductions. All the direct effects together constitute certainty, which is taken in calculations of probability to be a unit, of which the degrees of probability are fractions (Alethiol. §76). Each direct effect yields such a fraction, and determining this fraction obviously comes down to determining what part the effect is of the total....

How do we judge what part of the total effects is constituted by the effects we observe? Lambert does not tell us. Neither has anyone since.

Lambert's treatment of induction in the case of propositions is quite similar to his treatment in the case of physical events:

> §165. We now advance from *physical consequences*,[47] which are really effects and modifications, to *logical* ones. These are more general and include the physical ones as a special class, insofar as they enter into inference. Thus we here consider not *events themselves*, but the *ideas* and *propositions* they afford us. Once again there are different kinds of probabilities to investigate as to composition and departure from certainty. The first we derive from the question of *the extent to which a proposition can be deduced from its consequences.* We established that this is possible in principle for any proposition when we showed, in §175 of the Alethiologie, that *a proposition is necessarily true as long as nothing contradictory can be deduced from it.* Thus if every conclusion that can be validly deduced from a proposition with the help of other true propositions is found to be true, the original proposition is also true.. . .

Suppose, for example, that we wish to establish the proposition that A is B. Then we identify as many as possible of the features of B—i.e., of the predicates that are true of B—and call these C, D, E, F, etc. If we know from experience or from first principles that each of these predicates is also true of A, then we are naturally inclined to conclude that A must be B. If the predicates C, D, E, F, etc., include all possible predicates of B, or more realistically, if they include predicates unique to B, then the conclusion is valid. In some other cases, the proposition A is B will be rendered probable:

> §168. . . . The only remark we wish to add with respect to probability is this: If, in this kind of inference, one amasses a very large number of predicates C, D, E, F, etc., without being able to make any selection, then the presumption increases that one of them, or several of them taken together, is unique to B and thus that the proposition A is B is thereby proven. In fact it increases in proportion to the diversity of the predicates C, D, E, F, etc., and to the extent that none of these predicates appear to follow from others.. . .

6.2 Two Rules from the Doctrine of Chances

Lambert's discussion of the probability of propositions soon leads him to the problem of combining arguments. His first step in dealing with this problem is to adduce two rules from the doctrine of chances. Formally, they are similar to the rules given in the English article of 1699 we noticed in Sect. 5.2 above.

First Lambert addresses the problem of combining several uncertain arguments for the same conclusion, any one of which would fully establish the conclusion if it were certain.

> §169. . . . Here we shall content ourselves with reducing this calculation to the theory of games of chance. Let us imagine a heap of

[47] *Folgen*, hitherto translated as "effects."

> tickets for each argument. In each heap let the ratio of the number of valid or marked tickets to the number of unmarked ones be the same as the ratio of the number of cases where the argument is valid to the number where it is not valid. If we then suppose that Cajus takes a ticket blindly from each heap, the question is how probable it is that there will be no valid tickets among the ones drawn. It will be this probable, or this improbable, that all the arguments one has found on behalf of the proposition do not prove it. The theory of games of chance specifies the following rule for this calculation. *Multiply together the numbers of tickets in the different heaps, and likewise multiply together the numbers of invalid or unmarked tickets in the different heaps. Then the latter product, divided by the former, will give the degree of probability that the arguments do not prove. And if this degree, which is necessarily a fraction, is subtracted from one, then the remainder is the degree of probability that the arguments prove.*

This is, of course, Bernoulli's rule for combining pure arguments.

The context in which Lambert introduces this rule is problematic. He is dealing with the situation where one seeks to prove "*A* is *B*" by observing *A* to have features known to belong to *B*. Each such feature is an argument for the proposition "*A* is *B*"; its strength is measured by the proportion of cases where the feature belongs to *B* rather than to some other subject. Lambert is treating these arguments, in effect, as pure arguments, but one might argue that they are mixed.

Lambert's second rule is a rule for combining the probabilities of uncertain arguments where all these arguments are needed for the conclusion.

> §184. ... We remark that the calculation of degree of probability given above (§169) is actually only applicable where every argument is independent of the others. For each one contributes in itself to the lessening of the improbability, so that if one of them is certain, or if one knows one of them to be correct in a given case, the rest thereby become superfluous. Thus the question as to whether Cajus will draw at least one valid ticket from the several heaps of tickets is directly decided if even only one of these heaps consists of purely valid tickets. ...
>
> §185. On the other hand it is entirely different when the probability of the conclusion of a syllogism is to be determined from the probability of the premises. For then the premises cannot be viewed as separate, mutually independent arguments, because the conclusion necessarily depends on both of them together; the conclusion only follows when all the premises are correct. With this understood, the calculation of the probability of a conclusion even from a whole chain of reasoning can also be reduced to the theory of games of chance. For this purpose we again consider the heaps of tickets—in fact, we consider as many as there are premises in the chain of reasoning. In

> each heap, let the ratio of the number of valid tickets to the number of invalid be the same as the ratio of the cases in which the premise is correct to those in which it is not. Now let Cajus take a ticket blindly from each heap. Then the question is how probable it is that no invalid tickets should be found among the tickets drawn—i.e., that all should be valid. The conclusion will obtain this degree of probability from the given chain of reasoning. The theory of games of chance gives the following rule for this calculation: *Multiply together the numbers of tickets in the different heaps and likewise multiply together the numbers of valid tickets in the different heaps. Then the latter product, divided by the first, will give the degree of probability of the conclusion.* The first product represents the total of all the possible cases, while the latter represents the number of cases in which the conclusion follows, or, what amounts to the same thing, the premises are all correct.

6.3 The Syllogism

Lambert acknowledges quite explicitly that the probabilities for a proposition and its negation may add to less than one. In §212, for example, he tells us that "the degrees of probability obtained for the affirmation and for the denial of the conclusion of a syllogism do not always together form a whole. For often a considerable portion remains undetermined... One must certainly take this indeterminate portion into account if one wants to infer the degree of improbability from the degree of probability." In this Sect. I study Lambert's treatment of probability in the syllogism and show how this treatment led to his cognizance of non-additive probabilities. The essence of the matter, as we shall see, is that a merely probable minor premise in a syllogism of the first figure leads to a non-additively probable conclusion.

As Lambert sees it, the proposition "A is B" can be numerically qualified in three ways. (1) We may qualify the subject, writing

$$\frac{3}{4}A \text{ and } B$$

to mean that $\frac{3}{4}$ of the individuals that are A have the predicate B. (2) We may qualify the predicate, writing

$$C \text{ is } \frac{2}{3}B$$

to mean that C is known to have $\frac{2}{3}$ of the attributes in the concept B. Of course, in order to determine the number $\frac{2}{3}$, a judgment must be made as to the weight of various attributes. As Lambert explains in §191, if M, N, P and Q form a comprehensive (*die seinem Umfang ausfüllen*) list of the attributes in B, none of these are known to be unique to B, and C is only known to have attributes M, N, and P, then

> the conclusion
>
> $$C \text{ is } B$$
>
> can only be drawn with probability, for it remains undecided whether C has the predicate Q. Here the degree of probability depends on the ratio of the magnitude and quantity of the predicates MNP, which have already been found in C, to the magnitude and quantity of those that have yet to be found.

(3) Finally, we may qualify the proposition as a whole. Lambert expresses this by qualifying the connective "is"; he writes, for example,

$$C\ \frac{1}{2} \text{ is } B$$

to indicate that the proposition "C is B" has probability $\frac{1}{2}$.

Notice that the statements "$\frac{3}{4}A$ are B" and "C is $\frac{2}{3}A$" have the form of statements of fact, not of probability. They can give rise, however, to statements of probability. If $\frac{3}{4}A$ are B and we know of a particular individual only that it is an A, then we may say with probability $\frac{3}{4}$ that it is B; similarly, if C is $\frac{2}{3}A$, then we may say with probability $\frac{2}{3}$ that C has any particular attribute that forms part of the predicate A. Lambert represents these deductions by quantifying the first figure of the traditional syllogism. He begins with the syllogism Barbara

$$\begin{aligned} \text{all } A \text{ are } B \\ C \text{ is } A \\ \text{therefore } C \text{ is } B \end{aligned}$$

as his basic example of a syllogism of the first figure, and modifies it to

$$\begin{aligned} \frac{3}{4}A \text{ are } B \\ C \text{ is } A \\ \text{therefore } C\ \frac{3}{4} \text{ is } B \end{aligned}$$

or to

$$\begin{aligned} \text{all } A \text{ are } B \\ C \text{ is } \frac{2}{3}A \\ \text{therefore } C\ \frac{2}{3} \text{ is } B. \end{aligned}$$

After exhibiting these examples, Lambert proceeds to an example where both premises of the syllogism are qualified. In place of the major premise "all A are B" he puts "$\frac{3}{4}A$ are B" and "$\frac{1}{4}A$ are not B," and in place of the minor premise "C is A" he puts "C is $\frac{2}{3}A$." (Notice that we do *not* add that "C is $\frac{1}{3}$ not A." For "C is $\frac{2}{3}A$" means that we have verified that C has $\frac{2}{3}$ of the attributes which make up A; if it fails to have one of the remaining $\frac{1}{3}$, then the conclusion is not "C is $\frac{1}{3}$ not A" but simply "C is not A.") With these premises he obtains a probability of $\frac{1}{2}$ for the conclusion "C is B," and a probability of $\frac{1}{6}$ for its negation:

> §192. . . . It comes out as follows:
>
> $$\frac{3}{4}A \text{ are } B,\ C \text{ is } \frac{2}{3}A, \text{ therefore } C\ \frac{1}{2} \text{ is } B.$$
>
> For here the probability of the one premise is diminished in proportion to the probability of the others. So we can take only $\frac{2}{3}$ of the $\frac{3}{4}$ for the conclusion—i.e., only $\frac{2}{4}$ or $\frac{1}{2}$.
>
> §193. If we make the major premise of this syllogism negative, the result is
>
> $$\frac{1}{4}A \text{ are not } B,\ C \text{ is } \frac{2}{3}A, \text{ therefore } C\ \frac{1}{6} \text{ is not } B.$$
>
> Thus the probability that the syllogism's conclusion is negative is $\frac{1}{6}$, whereas the probability that it is affirmative is $\frac{1}{2}$. Both probabilities together yield $\frac{1}{6}+\frac{1}{2} = \frac{2}{3}$, which is the probability of the minor premise. And nothing more can be obtained from the calculation or from the syllogism itself. For one cannot make the minor premise negative in the first figure of the syllogism; if it is negative, the form of the conclusion remains undetermined. And thus in cases where the minor premise influences the probability of the conclusion, we find only that part of the probability of the conclusion that can be determined according to the form of the syllogism and its rules. That is to say, if one takes a number of cases where the premises are similar in kind and degree to the two presented here, the conclusion will be affirmed in half these cases and denied in $\frac{1}{6}$ of them; in the $\frac{1}{3}$ that remain it is entirely undetermined whether it will be affirmed or denied, entirely or in part.

It should be noted that two sources of indeterminacy are confounded in this example. The first is our ignorance as to whether C possesses the remaining attributes of A. The second lies in the fact that denial of the minor premise leads not to denial of the conclusion but to no conclusion at all. This second source of indeterminacy is the more fundamental insofar as the syllogism is concerned, for it causes indeterminacy in the conclusion even when there is none in the premises. To illustrate this suppose that the premises themselves (rather than their "middle term" A) are only probable, but additively so—e.g., begin with

$$\text{all } A\ \frac{3}{4} \text{ are } B, \qquad \text{all } A\ \frac{1}{4}\text{are not } B$$

and

$$C\,\frac{2}{3}\text{ is }A,\qquad C\,\frac{1}{3}\text{ is not }A.$$

Then, as the general rule that Lambert later (§216) propounds makes clear, Lambert would still obtain a probability of $\frac{1}{2}$ for "C is B" and a probability of $\frac{1}{6}$ for "C is not B." The probability of $\frac{1}{3}$ attached to "C is not A" contributes nothing to the syllogism.

Having perceived the essential role played by indeterminacy, Lambert introduces a more compact and perspicuous notation and generalizes his example by allowing indeterminacy in the major premise as well as the minor:

> §194. Call the affirmative a, the negative e, and the indeterminate u. Then the most complex case with regard to the middle term will be this:
>
> $$\left(\frac{2}{3}a+\frac{1}{4}e+\frac{1}{12}u\right)A\text{ are }B,\ C\text{ is }\left(\frac{3}{5}a+\frac{2}{5}u\right)A,\ \text{therefore}$$
> $$C\left(\frac{2}{5}a+\frac{3}{20}e+\frac{9}{20}u\right)\text{ is }B.$$
>
> In order to explain and prove this formula, which is only a special case so far as the numbers are concerned, we observe the following:
>
> 1. In the major premise, $(\frac{2}{3}a+\frac{1}{4}e+\frac{1}{12}u)$ means that of all individuals that are A, or of all A, there are $\frac{2}{3}$ for which B certainly holds, $\frac{1}{4}$ for which B does not hold, and $\frac{1}{12}$ for which it remains undetermined whether B holds or not. In this way three major premises are combined, as it were, and since $\frac{2}{3}+\frac{1}{4}+\frac{1}{12}=1$, we see that all A will be taken into account in this syllogism.
>
> 2. In the minor premise, $\left(\frac{3}{5}a+\frac{2}{5}u\right)$ represents the sum of the attributes of A. One knows that $\frac{3}{5}$ of these hold for C; the matter is still undetermined for the remaining $\frac{2}{5}$. The part that would be e cannot occur here. For if even a single attribute were in A that one knew was not in C, the minor premise would be denied for certain, and consequently the conclusion would be thoroughly indeterminate.
>
> 3. Now $\left(\frac{2}{3}a+\frac{1}{4}e+\frac{1}{12}u\right)$ is multiplied by $\left(\frac{3}{5}a+\frac{2}{5}u\right)$, and the product is divided into three classes:
>
> $$(\frac{2}{5}aa)+(\frac{3}{20}ae)+(\frac{3}{60}au+\frac{4}{15}au+\frac{2}{20}eu+\frac{2}{60}uu)=\frac{2}{5}a\quad+\frac{3}{20}e\quad+\frac{9}{20}u$$
>
> That is to say, everything involved with u belongs in one class, or to the indeterminate part of the conclusion, the ae belongs in the second class, or to the negative part, and the aa in the third, or to the affirmative part.
>
> 4. So the conclusion says that out of 20 cases where inferences of this kind and degree occur without selection, eight are affirmed,

> three are denied, and nine remain undetermined. Or alternatively, in a particular case one would have eight reasons to affirm the conclusion, three reasons to deny it, and nine reasons to leave it undecided—i.e., not draw a conclusion.

Lambert continues his discussion of the syllogism at great length. After presenting the example we have just quoted, he proceeds to consider examples where fractions are attached to the other parts of the premises—to the terms B and C or to the connectives; and in §216 he finally states a general rule that allows all these possibilities simultaneously. He also extends his methods to longer chains of reasoning and to the other figures of the syllogism. And he does not neglect to caution the reader against applying arguments from probability when other arguments can yield certainty; he is particularly concerned that every effort should be made to complete an incomplete induction before settling for mere probability. (See §§166–183.) But the excerpts we have quoted here surely suffice to convey the gist and spirit of his methods.

6.4 The Combination of Testimony

Lambert's discussion of Bernoulli's rules of combination occurs towards the end of his chapter on probability in the context of a discussion of the credibility of testimony. In this Sect. I examine the passages where Lambert criticizes Bernoulli's rules and proposes a new, more general rule.

Lambert stresses that the assessment of testimony should take into account a common-sense examination of circumstances, motives, and other particulars. But he also discusses how one might assess the credibility of a witness solely on the basis of the witness's general intelligence, character and knowledgeability. Such an assessment may produce a very middling degree of probability for the witness's report, but this can be increased by multiplying the number of witnesses:

> §236. ...one seeks to increase such apparently trifling degrees of probability by amassing separate witnesses. And if indeed this amassment is done without selection, then it certainly cannot be denied that each independent witness can be treated as a separate and independent argument, provided of course that their testimonies agree, and to the extent that they agree...

Notice that he again insists that all the evidence be considered, "without selection."

Lambert uses this idea of combining testimony to introduce his general rule for combining probabilities.

> §237. Consider two witnesses who give the same testimony. Let the credibility of the first be such that for every ten truths he tells three untruths and one lie—viz., if one wants to hit on the truth, one must believe him in ten cases, not believe him in three cases, and believe

> the opposite in one case. We express this by $10a + 3u + 1e$. Similarly, let the credibility of the other be $12a + 5u + 2e$.When these cases are multiplied together the product is
>
> $$120aa + 86au + 15uu + 11eu + 2ee + 32ae.$$
>
> The $32ae$ will be omitted from this product, for it is impossible to believe the testimony on account of one witness and simultaneously believe the opposite on account of the other. Moreover, the $120aa + 86au$ will be consolidated to form $206a$. For though the one witness is not believed in 86 cases, the other still is. The $2ee + 11eu$ are similarly consolidated to form $13e$. For in case of the $11eu$, belief falls to the opposite of the testimony. Thus we have
>
> $$206a + 15u + 13e$$
>
> for the credibility of a single witness who counts for as much as these two together. ... Here is the general formula:
>
> | | |
> |---|---|
> | Witness 1 | $Ma + Nu + Pe$ |
> | Witness 2 | $ma + nu + pe$ |
> | Both | $(Mm + Mn + mN)a + Nnu + (Pp + Pn + pN)e.$ |
>
> If one witness is fully credible, say $n = p = 0$, then all the u and e terms drop out of the product; and since all the remaining cases are a, this shows that the remaining witnesses neither increase nor decrease his credibility. On the other hand, when no witness is fully credible, u and e still enter into the total sum, and consequently the testimony merely has probability.

We can apply the same rule to the case where one witness testifies to the opposite, provided we interchange the a and the e in his credibility.

> §238. ... To retain the previous example, if the second witness testifies to the opposite, then $12a + 5u + 2e$ becomes $2a + 5u + 12e$. Taking the first witness$10a + 3u + e$ into account, the total credibility is $76a + 15u + 53e$, which differs markedly from the previous result. If the witness who says the opposite is fully credible, then $M = N = 0$ in the general formula, and only the e cases remain in the product; his credibility, as a result, is unimpaired by that of the other witnesses. Moreover, in the case of two witnesses who both have full credibility, it is intrinsically impossible that one should say the opposite of the other's testimony. If one posits this case, then $M = N = n = p = 0$ in the formula, and thus all the terms in the product are equal to zero. This says that no such case occurs.

Lambert now remarks that his rule of combination can be used to combine probabilities arising from any independent arguments. And as a general rule of combination, it is an improvement over Bernoulli's rules, both in generality and in soundness.

§239. We might remark in passing that the formula given above can also be used with arguments, when these are independent of one another and make the same proposition probable. In the case of such an argument,

$$12a + 5u + 2e$$

means that in 12 cases the argument proves the proposition, in 5 cases proves nothing (i.e., leaves the proposition undecided), and in 2 cases overturns it (i.e., makes it negative or proves the opposite). If a probable syllogism (§194) produces a proposition of the form

$$\text{all } A \left(\frac{12}{19}a + \frac{5}{19}u + \frac{2}{19}e\right) \text{ are } B$$

then the fractions with which the connective is encumbered represent the credibility of the proposition and thus the weight of the argument. The method of calculation given here differs markedly, by the way, from the methods that appear on p. 220ff of Bernoulli's *Ars conjectandi.* There Mr. Bernoulli accepts two kinds of arguments: namely, those that partly prove and partly do not prove, and then those that partly prove and partly prove the opposite. He calls the first pure arguments and the others mixed arguments. Along with these he would include yet a third kind—namely, those that partly do not prove and partly prove the opposite; but he only announces this kind, and does not bring it into the calculations. Here we have combined these three kinds of arguments into a single general kind. For from the formula $Ma + Nu + Pe$, one can obtain: 1). $Ma + Nu$; 2). $Ma + Pe$; and 3). $Nu + Pe$ by setting P or N or M equal to zero. In this respect the method of calculation presented here is more general than Bernoulli's, for it exhibits all his special cases at once. But it also gives a different result, and this should not be, if both were correct. Rather than expound Bernoulli's method, we will merely remark on his formula (p. 221):

$$1 - \frac{cfi}{adg} \cdot \frac{ru}{qt + ru}.$$

If we take one of the arguments that partly prove nothing[48] and partly prove the opposite to be complete (i.e., if we assume it completely proves the opposite), and if we accordingly set q or t equal to zero, then this formula becomes

$$1 - \frac{cfi}{adg}.$$

[48] This seems to be a slip. The numbers q and t, which Lambert wants to manipulate, refer to mixed arguments, and hence Lambert should say. "If we suppose one of the mixed arguments completely proves the opposite..."

> But it should be zero, because in this case all the affirmative arguments represented in the formula are completely refuted. The reason why this formula says otherwise will be found on p. 221; to wit, Mr. Bernoulli considers all the cases where the pure argument proves to be valid, whether or not the cases from the mixed argument with which they are combined prove the opposite. But in §237 we completely omitted the cases *ae* because they were impossible, and this makes the result in the calculation presented here different from Bernoulli's result.

We conclude by restating Lambert's rule of combination in terms of the probabilities. Consider two arguments bearing on a proposition, and suppose the first provides a probability p_1 for the proposition and a probability q_1 for its negation, while the second provides a probability p_2 for the proposition and a probability q_2 for the negation. In order to apply the rule as Lambert states it in his paragraph §237, quoted above, we must choose numbers M, N, P such that

$$p_1 = \frac{M}{M+N+P} \quad \text{and} \quad q_1 = \frac{P}{M+N+P},$$

and numbers m, n, p such that

$$p_2 = \frac{m}{m+n+p} \quad \text{and} \quad q_2 = \frac{p}{m+n+p}.$$

The rule then yields a probability for the proposition, on the basis of both arguments together, of

$$\begin{aligned}\frac{Mm+Mn+mN}{Mm+Mn+mN+Nn+Pp+Pn+pN} &= \frac{M(m+n+p)+m(M+N+P)-Mm-Mp-mP}{(M+N+P)(m+n+p)-Mp-mP} \\ &= \frac{p_1+p_2-p_1p_2-p_1q_2-p_2q_1}{1-p_1q_2-p_2q_1}\end{aligned} \tag{7}$$

and a probability for its negation of

$$\begin{aligned}\frac{Pp+Pn+pN}{Mm+Mn+mN+Nn+Pp+Pn+pN} &= \frac{P(m+n+p)+p(M+N+P)-Pp-Mp-mP}{(M+N+P)(m+n+p)-Mp-mP} \\ &= \frac{q_1+q_2-q_1q_2-p_1q_2-p_2q_1}{1-p_1q_2-p_2q_1}.\end{aligned} \tag{8}$$

Thus formulae 5 and 6, which we used to state Lambert's rule in Sect. 4.4 above, are indeed accurate.

Lambert's rule of combination is obviously a special case of Dempster's rule for combining belief functions; see pp. 374–376 of Shafer 1976 [72]. Lambert's quantification of the probable syllogism (§194) can also be construed as a special case of Dempster's rule.

6.5 The Unintelligibility of Non-Additivity

The influence of Lambert's ideas is easily summed up: hardly anyone noticed, and no one understood. At the time he wrote he was already in conflict with received opinion; scholars had known since De Moivre that all probabilities are additive. And within a few years of the publication of *Neues Organon*, scholars began to learn, from Bayes, Condorcet, and Laplace, just how this additivity worked in the case of propositions; then Lambert's non-additive probabilities became simply unintelligible.

There is one exception to this general neglect of Lambert's thought. Pierre Prevost and Simon Lhulier,[49] noted and praised Lambert's emendation of Bernoulli and Lambert's general rule of combination in their memoir of 1797. And they seem to have understood the non-additivity implicit in this rule. But after having praised Lambert's ideas, they proceeded to specialize his rule to the case of additivity. This specialization yields the 18th-century (or Bayesian) rule for concurrent testimony (see Sect. 5.2 above), and, in fact, Prevost and Lhulier give the Laplacean justification for this rule. They also derive the 18th-century rule for successive testimony.

Todhunter, in his famous history (1865 [78], pp. 71, 462), noted Prevost and Lhulier's memoir and used their report on Lambert's criticism of Bernoulli. Todhunter did not, however, quote Lambert's general rule or give any hint that Lambert had dealt with non-additive probabilities. For Todhunter, we may presume, non-additivity for probability made no sense at all.

That Bernoulli and Lambert contemplated non-additive probabilities will be obvious to anyone who carefully reads the passages presented in this essay. But after Prevost and Lhulier no one seems to have thought this non-additivity worth noting. In fact I know of no reference to it in the historical or mathematical literature from the time of Prevost and Lhulier until I pointed it out in 1972.[50]

7 Lessons for a Modern Theory

For 250 years our culture's conception of probability has been dominated by two ideas: the idea that probabilities concerning practical matters are obtained from frequencies, and the idea that probabilities are necessarily additive. These two ideas may well have been essential to the tremendous progress that probability has made during this period. But future progress may require that we lessen our dependence on them, and this, in turn, may require that we rediscover the alternatives provided by Bernoulli and Lambert.

The precise relation of probabilities to frequencies has been a matter of great scholarly debate during the past century. But the popular understanding of the matter has always been that the probability of a thing is to be found by observing the frequency with which similar things have been found to be

[49] These authors' last names appeared in this form in their papers on the probability of testimony, which appeared in the memoirs of the Berlin academy. The French spellings Prévost and L'Huilier are also sometimes seen.

[50] See Hacking 1975 [36], p. 144; Dempster 1974 [27], p. 58; and Shafer 1976 [72], p. 430.

true. It is also thought that we should take all our evidence into account. And, as I pointed out in Sect. 4.6 above, this produces a conundrum, for the more details of our evidence we take into account, the fewer similar cases are to be found. When we take all the evidence into account there are usually no similar cases at all. As John Venn (1888, p. 222) put it, what mortality table are we to apply to a consumptive Englishman living in Madeira?[51]

In order to escape this conundrum we must, I believe, return to Bernoulli and Lambert's conception of the combination of arguments. The detailed problems we face in practical life are always unique; they always present features and combinations of detail that are utterly new to us. But we can and do recognize familiar features in such problems, and we can and do refer these familiar features to our understanding and experience—we may even, occasionally, note the frequency with which some of these features have presaged various things in the past. And having made our judgments concerning these separate features of the evidence, we have no choice but to recombine them "at the epistemic level";[52] that is to say, we must weigh and combine the arguments they provide. Such a local approach to the assessment of evidence contrasts with the global approach most often favored by contemporary philosophers of probability, but it is not so foreign to common sense.

It should be mentioned that there has been at least one scholar of the past two centuries who attempted such a local assessment of evidence. The English mathematician George Boole, in his *Laws of Thought* (1854 [10]), maintained that probabilities always ultimately derive from frequencies. But he admitted that joint frequencies are sometimes unavailable, so that joint probabilities must sometimes be generated at the epistemic level; he prescribed that joint probabilities be obtained in these cases by treating the events *as if* they were independent. It is not clear to this author just how close Boole came to Lambert's rule of combination in his exploration of this idea.[53]

Whereas the relation of probabilities to frequencies has been subject to some debate during the past two centuries, the additivity of probability has not. Indeed, the study of epistemic probability has been pervaded by a nearly universal and almost unconscious acceptance of additivity. Most discussions of epistemic probability have been within the Bayesian theory, and even explicit attempts to generalize the Bayesian theory have tended to rely on its additive conception of probability. Even many of the proponents of "qualitative" probability, who would relinquish numerical probability altogether in favor of more general orderings on algebras of propositions, have insisted on the self-evidence of axioms whose appeal lies entirely in their mirroring of the rule of additivity.[54] Nonetheless, Bernoulli and Lambert's non-additive

[51] Cf. Dempster, 1968 [25], p. 34.

[52] See §11.3 of Shafer [70].

[53] Boole's treatment of probability baffled his contemporaries, and I do not pretend to understand him fully. See Theodore Hailperin's recent monograph (1976 [37]).

[54] See, e.g., Fine, 1973 [31], p. 17.

probabilities have gradually and awkwardly reemerged within the confines of the additive theory; they have re-emerged under the guise of "probability bounds" or "upper and lower probabilities."

It is easy to see how the mathematics that Lambert associated with non-additive probability can be translated into the language of "probability bounds." Instead of supposing, as Lambert in effect did, that a proposition might merit a probability p and its negation might merit a probability q, where $p+q<1$, one posits a lower bound p_* (which corresponds to p) and an upper bound p^*(which corresponds to $1-q$) for the proposition's unknown probability P. One supposes that this unknown probability P is additive, in the sense that the probability of the proposition's negation is taken to be the equally unknown number $1-P$. The possibility that $p+q<1$, which was the possibility of non-additivity for Lambert, becomes merely the possibility that $p_* < p^*$. Thus the mathematics of non-additivity is admitted but kept at arm's length from the concept of probability itself: one supposes there does in fact exist an additive probability P; it is just that our knowledge about P is limited to the knowledge that $p_* \leqq P \leqq p^*$.

The idea of probability bounds can be found in Boole's Laws of Thought. Boole sought, in general, to find probabilities from frequencies, either directly or indirectly. To find the probability of a proposition A, he would have us express A in terms of other propositions whose probabilities can be observed directly as frequencies. But if some of the frequencies remain unobserved, then the probability of A is determined only within limits. "Between these limits," he wrote (p. 268), "it is certain that the probability sought must lie independently of all new experience which does not absolutely contradict the past." Thus, though statements of probability may express an uncertainty of a "frequentist" character, bounds may express a second-order uncertainty as to what the frequencies are.

After a long period of relative dormancy, the idea of probability bounds reappeared in the work of Bernard O. Koopman (1940 [49]), who introduced the term "upper and lower numerical probabilities" to name numbers that emerged from his axiomatization of qualitative probability. In the early 1960's there were several discussions of upper and lower probabilities by statisticians and philosophers, including those by I.J. Good (1962 [34]), C.A.B. Smith (1961 [74], 1965 [75]), and Henry E. Kyburg (1961 [50]). And shortly after this work, A. P. Dempster (1966–1969 [20, 21, 22, 23, 24, 26]) used the language of upper and lower probabilities to formulate an original and remarkable new theory, a theory which includes the rule of combination repeatedly alluded to in this essay and a number of algorithms for statistical inference. More recently, upper and lower probabilities have been discussed by Beran (1970 [4]), Giles (1976 [33]), Huber (1973 [41]), and Suppes (1974 [77]).

This 20th century development of the idea of upper and lower probabilities has been predominantly epistemic, and because of this it has encountered a fundamental problem of interpretation: what is the meaning of the determined but not fully known probability P that is supposed to lie between the bounds p_* and p^*? If P can be interpreted as a frequency or as an aleatory probability, as in Boole's work, then we can make sense of the idea that P is unknown. But an unknown epistemic probability is a contradiction in terms—an unknown feature of our knowledge. Most of the recent writers on upper and lower probabilities more or less acknowledge the absence of a meaningful interpretation for the unknown additive epistemic probability P; they treat P as a metaphor and stress that one's knowledge is fully expressed by the pair (p_*, p^*). But they still struggle to place some significance on the additivity of this metaphor, and when they try to interpret the numbers (p_*, p^*) they reveal their puzzlement as to why one's knowledge should fall short of an additive probability. In Dempster's work, for example, the difference $p^* - p_*$ is called one's "confusion" about the proposition in question; this "confusion" is thought to reflect an uncertainty which somehow differs from the uncertainty reflected by additive probabilities.

I believe that Dempster's theory gains in clarity if one drops the language of upper and lower probability bounds in favor of the straightforward approach to non-additive epistemic probability that we find in Bernoulli and Lambert. And, as I have argued in *A Mathematical Theory of Evidence*, it also gains new mathematical power. By considering Dempster's p_* as a non-additive epistemic probability or degree of belief, one is able to understand the rule of combination as a rule for translating "weights of evidence" into degrees of belief. And this leads in turn to an understanding that additivity, rather than non-additivity, is anomalous for degrees of belief; additivity is a limiting case which is approached when one has extremely strong but discordant evidence.

The modern study of epistemic probability has a great deal to gain from a revival of the insights of Bernoulli and Lambert. From Bernoulli we can learn the importance of combining arguments and thereby rediscover a natural approach to the mathematical representation of probable reasoning. From Lambert we can learn how natural non-additivity is when probable reasoning amounts to deduction from probable premises. And the insights of both these scholars can help us correct the neglect of concepts of evidence that has led the 20th century theory of epistemic probability to its exaggerated emphasis on concepts of decision.

Acknowledgements

I wish to thank the teachers and colleagues who have encouraged the research for this paper. My interest in the history of probability was first aroused by two of my teachers, F. N. David and Ivo Schneider. I have benefited from conversation and correspondence with A. P. Dempster and Ian Hacking, and from comments by Russ Bradt, Ben Cobb, Paul Mostert, and Cora Lee Price. I have also been assisted by Stan Kelley, Mike Mahoney, Terry Shafer, and B. L. van der Waerden.

The translations from *Ars Conjectandi* were based on preliminary translations by Bing Sung and Thomas Drucker and were completed with advice from Cora Lee Price, who also assisted me with the other translations from Latin. The translations from *Neues Organon* were based on a preliminary translation by Jill Anderson and were completed with advice from Joe Van Zandt. I am, of course, responsible for any errors or other shortcomings in these translations.

This work was begun when I was supported by a U.S. National Science Foundation graduate fellowship, continued with support from U.S. National Science Foundation grant GP-43248, and completed with support from the General Research Fund of the University of Kansas, allocations # 3991-x038 and # 3315-x038.

References

[1] Anonymous (1699), A calculation of the credibility of human testimony. *Phil. Trans. Roy. Soc. London*, 21:359–365. The author was George Hooper; see §5.2 above.

[2] Antoine Arnauld and Pierre Nicole (1662), *la Logique ou l'art de penser.* Paris.

[3] Thomas Bayes (1764), An essay toward solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc, London* for 1763, 53:370–418. Reprinted in *Biometrika*, 45, 293–315 (1958), and in [63].

[4] Rudolph Beran (1970), Upper and lower risks and minimax procedures. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 1–16.

[5] Daniel Bernoulli (1778), Dijudicatio maxime probabilis plurium observationum discrepantium atque verisimillima inductio inde formanda. *Acta Acad. Petrop.* for 1777, pars prior:3–23. An English translation by G. D. Allen was published in *Biometrika*, 48:1–18 (1961), and reprinted in Pearson and Kendall [63].

[6] Jakob Bernoulli (1713), *Ars Conjectandi.* Basel. Reprinted in 1968 by Culture et Civilisation, Brussels, and in 1975 in [7], Volume 3. An English translation of Part IV, by Bing Sung, was issued as Research Report No. 2 of the Department of Statistics, Harvard University, Feb. 12, 1966. An English translation of the entire book, by Edith D. Sylla, was published by Johns Hopkins University Press in 2006. Sylla lists other translations, in various languages, on pp. 406–408.

[7] Jacob Bernoulli (1969–1999), *Die Werke von Jakob Bernoulli.* Basel. Five volumes have appeared: Volume 1 in 1969, Volume 2 in 1989, Volume 3 in 1975, Volume 4 in 1993, and Volume 5 in 1999. Volume 3, which is cited repeatedly in this article, was edited by B. L. van der Waerden.

[8] Nicolaus Bernoulli (1709), *Dissertationis de Usu Artis Conjectandi in Jure.* Basel. Excerpts were published on pp. 159–170 of Volume VI of *Acta Eruditorum, Supplementa*, 1711. The entire dissertation was reprinted in [7], Volume 3, pp. 287–326, with commentary by Karl Kohli on pp. 541–556.

[9] Kurt-Reinhard Biermann and Margot Faak (1957), G. W. Leibniz' "De incerti aestimatione". *Forschungen und Fortschritte*, 31:45–50.

[10] George Boole (1854), *An Investigation into the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities.* Macmillan, London. Reprinted by Dover, 1958.

[11] J. van Brakel (1976), Some remarks on the prehistory of the concept of statistical probability. *Archive for History of Exact Sciences*, 16:119–136.

[12] Edmund F. Byrne (1968), *Probability and opinion: A study in the medieval presuppositions of post-medieval theories of probability.* Martinus Nijhoff, The Hague.

[13] Marie Jean Condorcet (1785), *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Paris. Reprinted by Chelsea in 1973.

[14] Antoine-Augustin Cournot (1843), *Exposition de la théorie des chances et des probabilités.* Paris. Reprinted as Volume I of Cournot's *Oeuvres complètes* in 1984 by Vrin, Paris, with introduction, notes, and index by Bernard Bru.

[15] Louis Couturat (1901), *La Logique de Leibniz.* Paris. Reprinted by Georg Olms, Hildesheim, 1961.

[16] John Craig (1699), *Theologiae Christianae Principia Mathematica.* London. The part dealing with probability was reprinted by Mouton as *Craig's Rules of Historical Evidence*, Beiheft 4 of *History and Theory: Studies in the Philosophy of History*, 1964.

[17] F. N. David (1962), *Games, Gods and Gambling.* Griffin, London.

[18] Abraham De Moivre (1711), De mensura sortis, seu, de probabilitate eventum in ludis a casu fortuito pendentibus. *Phil. Trans. Roy. Soc. London*, 27: 213–264.

[19] Abraham De Moivre (1718), *The Doctrine of Chances, or a Method of Calculating the Probability of Events in Play.* London.

[20] A. P. Dempster (1966), New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics*, 37:355–374.

[21] A. P. Dempster (1967), Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339.

[22] A. P. Dempster (1967), Upper and lower probability inferences based on a sample from a finite univariate population. *Biometrika*, 54:515–528.

[23] A. P. Dempster (1968), Upper and lower probabilities generated by random closed interval. *Annals of Mathematical Statistics*, 39:957–966.

[24] A. P. Dempster (1968), A generalization of Bayesian inference (with discussion). *Journal of the Royal Statistical Society Series B*, 30:205–247.

[25] A. P. Dempster (1968), *Probability* (Chapter 2 of *The Theory of Statistical Inference: A Critical Analysis*). Department of Statistics, Harvard University, Research Report S-3, September 27.

[26] A. P. Dempster (1969), Upper and lower probability inferences for families of hypotheses with monotone density ratios. *Annals of Mathematical Statistics*, 40:953–969.

[27] A. P. Dempster (1974), Discussion of a paper by A. W. F. Edwards. In Blaesild, Barndorff-Nielsen, and Schou, editors, *Memoirs No. 1 of the Dept. of Theoretical Statistics*, pp. 57–58. Aarhus.

[28] Denis Diderot et al. (1765), *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers.* Paris.

[29] M. E. Eisenring (1942), *Johann Heinrich Lambert und die wissenschaftliche Philosophie der Gegenwart.* Zurich.

[30] Pierre de Fermat (1891–1922), *Oeuvres.* Paris. Edited by P. Tannery and C. Henry, 4 volumes.

[31] Terry Fine (1973), *Theories of Probability.* Academic Press.

[32] Bernard Fontenelle (1706), Eloge de Jacques Bernoulli. *Histoire de l'Académie royale des sciences* for 1705:148–149.

[33] Robin Giles (1976), A logic for subjective belief. In [38], Volume I, pp. 41–72.

[34] I.J. Good (1962), The measure of a non-measurable set. In Ernest Nagal, Patrick Suppes, and Alfred Tarski, editors, *Logic, Methodology and Philosophy of Science*, pp. 319–329. Stanford University Press, Stanford.

[35] Thomas Granger. *Divine Logike.* London, 1620.

[36] Ian Hacking (1975), *The Emergence of Probability.* Cambridge University Press, New York.

[37] Theodore Hailperin (1976), *Boole's Logic and Probability.* North-Holland, Amsterdam.

[38] William L. Harper and C. A. Hooker, editors (1976), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science.* Reidel, Dordrecht, 3 volumes.

[39] Julian Henny (1975), Niklaus und Johann Bernoullis Forschungen auf dem Gebiet der Wahrscheinlichkeitsrechnung in ihrem Briefwechsel mit Pierre Rémond de Montmort. In [7], Volume 3, pp. 457–507.

[40] Joseph E. Hofmann (1974), *Leibniz in Paris 1672–1676; His Growth to Mathematical Maturity.* Cambridge University Press.

[41] Peter Huber (1973), The use of Choquet capacities in statistics. *Bulletin of the International Statistical Institute*, 45, book 4:181–188.

[42] Christiaan Huygens (1657), De ratiociniis in aleae ludo. In F. Van Schooten, editor, *Exercitionum Mathematicorum.* Amsterdam. Incorporated into Part I of Bernoulli's *Ars Conjectandi.* The original Dutch version, published in 1660, is in Volume XIV of [43].

[43] Christiaan Huygens, (1888–1950), *Oeuvres complètes.* The Hague, 22 volumes.

[44] M. G. Kendall (1956), The beginnings of a probability calculus. *Biometrika*, 43:1–14, 1956. Reprinted in [63].

[45] Karl Kohli and B. L. van der Waerden (1975), Bewertung von Leibrenten. In [7], Volume 3, pp. 515–539.

[46] Karl Kohli (1975), Zur Publikationsgeschichte der Ars Conjectandi. In [7], Volume 3, pp. 391–401.

[47] Karl Kohli (1975), Spieldauer: Von Jakob Bernoullis Lösung der fünften Aufgabe von Huygens bis zu den Arbeiten von de Moivre. In [7], Volume 3, pp. 403–455.

[48] Karl Kohli (1975), Aus dem Briefwechsel zwischen Leibniz und Jakob Bernoulli. In [7], Volume 3, pp. 509–513.

[49] Bernard O. Koopman (1940), The bases of probability. *Bulletin of the American Mathematical Society*, 46:763–774.

[50] Henry E. Kyburg (1961), *Probability and the logic of rational belief.* Wesleyan University Press, Middletown.

[51] Johann Heinrich Lambert (1764), *Neues Organon, oder Gedanken uber die Erforschung und Bezeichnung des Wahren und dessen Unterscheidung Von Irrtum und Schein.* Leipzig. Reprinted in 1965 by Olms of Hildesheim as the first two volumes of Lambert's *Philosophische Schriften.*

[52] Pierre Simon Laplace (1774), Mémoire sur la probabilité des causes par les événements. *Mémoires de l'Académie royale des Sciences de Paris (Savants étrangers)*, 6:621–656. Reprinted in [54], Volume 8, pp. 27–65.

[53] Pierre Simon Laplace (1814), *Thèorie analytique des probabilités.* Paris, 2nd edition.

[54] Pierre Simon Laplace (1878–1912), *Oeuvres complètes.* Paris, 14 volumes.

[55] Gottfried Wilhelm Leibniz (1669), *De conditionibus.* Leipzig. Reprinted in [56], Volume IV, part III, p. 148ff.

[56] Gottfried Wilhelm Leibniz (1768), *Opera Omnia.* Geneva. Edited by L. Dutens, 6 volumes.

[57] Gottfried Wilhelm Leibniz (1849–1863), *Mathematische Schriften.* Halle. Edited by C. I. Gerhardt, 7 volumes.

[58] Geneviève Lewis (1952), *Lettres de Leibniz à Arnauld.* Paris.

[59] John Locke (1690), *An essay concerning human understanding.* London.

[60] John Locke (1936), *An early draft of Locke's essay.* Oxford. Edited by R. I. Aaron and J. Gibb.

[61] Pierre Rémond de Montmort (1708), *Essay d'analyse sur les jeux de hazard.* Paris.

[62] Oystein Ore (1953), *Cardano, the gambling scholar.* Princeton University Press. Includes Gould's translation of Cardano's *de Ludo Aleae.*

[63] E. S. Pearson and M. G. Kendall, editors (1970), *Studies in the History of Statistics and Probability.* Hafner, Darien.

[64] Siméon-Denis Poisson (1837), *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités.* Paris.

[65] Pierre Prevost and Simon Lhuilier (1800), Mémoire sur l'application du Calcul des probabilités à la valeur du témoignage. *Mémoires de l'Académie Royal des Sciences et Belles Lettres à Berlin* for 1797:120–152.

[66] Emile Ravier (1937), *Bibliographie des Oeuvres de Leibniz.* Paris.

[67] Alfred Renyi (1973), *Letters on probability.* Wayne State University Press.

[68] Joseph Saurin (1706), Eloge de Jacques Bernoulli. *Journal des Sçavans*, 34, Part I:126–139.

[69] Ivo Schneider (1968), Der Mathematiker Abraham de Moivre (1667–1754). *Archive for History of Exact Sciences*, 5:258–300.

[70] Glenn Shafer (1976), *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ.

[71] Glenn Shafer (1976), Review of Ian Hacking's *The Emergence of Probability. Journal of the American Statistical Association*, 71:519–521.

[72] Glenn Shafer (1976), A theory of statistical evidence. In [38], Volume II, pp. 365–436.

[73] O. B. Sheynin (1971), J. H. Lambert's work on probability. *Archive for History of Exact Sciences*, 7:244–256.

[74] C. A. B. Smith (1961), Consistency in statistical inference and decision (with discussion). *Journal of the Royal Statistical Society Series B*, 23:1–25.

[75] C. A. B. Smith (1965), Personal probability and statistical analysis (with discussion). *Journal of the Royal Statistical Society, Series A*, 128:469–499.

[76] Stephen M. Stigler (1975), Napoleonic statistics: The work of Laplace. *Biometrika*, 62:503–517.

[77] Patrick Suppes (1974), The measurement of belief. *Journal of Royal Statistical Society Series B*, 36:160–175.

[78] Isaac Todhunter (1865), *A History of the Mathematical Theory of Probability from the time of Pascal to that of Laplace.* Cambridge University Press. Reprinted by Chelsea 1949, 1965.

7

Allocations of Probability[1]

Glenn Shafer

Abstract. This paper studies *belief functions*, set functions which are normalized and monotone of order ∞. The concepts of *continuity* and *condensability* are defined for belief functions, and it is shown how to extend continuous or condensable belief functions from an algebra of subsets to the corresponding power set. The main tool used in this extension is the theorem that every belief function can be represented by an *allocation of probability*—i.e., by a $\cap$-homomorphism into a positive and completely additive probability algebra. This representation can be deduced either from an integral representation due to Choquet or from more elementary work by Revuz and Honeycutt.

Key words: Belief function, Allocation of probability, Capacity, Upper and lower probabilities, Condensability, Continuity

1 Belief Functions

In his pathbreaking "Theory of capacities," Gustave Choquet (1953) established the following definitions: a class $\mathcal{E}$ of subsets of a set Ω is a multiplicative subclass of $\mathcal{P}(\Omega)$ if $A \cap B$ is in $\mathcal{E}$ whenever A and B are in $\mathcal{E}$, an additive subclass of $\mathcal{P}(\Omega)$ if $A \cup B$ is in $\mathcal{E}$ whenever A and B are in $\mathcal{E}$. A real-valued function g on a multiplicative subclass $\mathcal{E}$ is *monotone of order n* if

$$g(A) \geqq \sum \left\{ (-1)^{|I|+1} g\left(\cap_{i \in I} A_i\right) | \varnothing \neq I \subset \{I, \cdots, n\} \right\}$$

for every collection $A, A_1, \cdots, A_n$ of elements of $\mathcal{E}$ such that $A \supset A_i$ for all i, *monotone of order* ∞ if it is monotone of order n for all $n \geqq 1$. A real-valued function g on an additive subclass $\mathcal{E}$ is *alternating of order n* if

$$g(A) \leqq \sum \left\{ (-1)^{|I|+1} g\left(\cup_{i \in I} A_i\right) | \varnothing \neq I \subset \{I, \cdots, n\} \right\}$$

[1] This research was partially supported by NSF Grant GP43248. *AMS* 1970 *subject classifications.* Primary 60A05; secondary 62A99.

for every collection $A, A_1, \cdots, A_n$ of elements of $\mathcal{E}$ such that $A \subset A_i$ for all i; *alternating of order* ∞ if it is alternating of order n for all $n \geqq 1$.

We call a function f on a multiplicative subclass $\mathcal{E}$ of $\mathcal{P}(\Omega)$ a *belief function* if $\varnothing$ and Ω are in $\mathcal{E}, f(\varnothing) = 0, f(\Omega) = 1$, and f is monotone of order ∞. The condition that f be monotone of order ∞ implies in particular that f is increasing; hence a belief function always takes values in the interval [0, 1]. The name "belief function" derives from the thought that these functions might be used to represent partial belief: if Ω is interpreted as a set of "possibilities" and A is a subset of Ω, then $f(A)$ might express one's degree of belief that the truth lies in A. In a recent monograph (1976a), I argue at length that belief functions are useful and appropriate for the representation of partial belief, and I study these functions in detail in the case where Ω is finite. This paper develops tools for extending that study to the case where Ω is infinite.

We call a function f^* on an additive subclass $\mathcal{E}^*$ of $\mathcal{P}(\Omega)$ an *upper probability function* if $\varnothing$ and Ω are in $\mathcal{E}, f^*(\varnothing) = 0, f^*(\Omega) = 1$, and f^* is alternating of order ∞. Notice that if f is a belief function on $\mathcal{E}$, then the function f^* defined on the additive subclass $\mathcal{E}^* = \{\overline{A} | A \in \mathcal{E}\}$ by $f^*(A) = 1 - f(\overline{A})$ is an upper probability function.

It will be shown in Sect. 5 below that a belief function f on a multiplicative subclass $\mathcal{E}$ of $\mathcal{P}(\Omega)$ can always be extended to a belief function on $\mathcal{P}(\Omega)$. In fact, it always has a *canonical extension* to $\mathcal{P}(\Omega)$: namely, the belief function $\bar{f}$ on $\mathcal{P}(\Omega)$ given by

$$\bar{f}(A) = \sup \Sigma \left\{ (-1)^{|I|+1} f(\cap_{i \in I} A_i) | \varnothing \neq I \subset \{1, \cdots, n\} \right\},$$

where the supremum is taken over all $n \geqq 1$ and all collections $A_1, \cdots, A_n$ of elements of $\mathcal{E}$ that are subsets of A. We call this extension canonical because it is minimal; i.e., $\bar{f} \leqq g$ for any other belief function g on $\mathcal{P}(\Omega)$ that extends f. (In fact, $\bar{f} \leqq g$ for any other belief function g on $\mathcal{P}(\Omega)$ such that $f \leqq g|\mathcal{E}$.) This can also be expressed by saying that $\bar{f}$'s upper probability function $(\bar{f})^*$ is the maximal extension of f^*; i.e., $(\bar{f})^* \geqq g$ for any other upper probability function g on $\mathcal{P}(\Omega)$ that extends f^*.

In this paper we consider two regularity conditions for a belief function over an infinite set Ω: *continuity* and *condensability.* We call a belief function f on $\mathcal{P}(\Omega)$ *continuous* if it satisfies

$$f(\cap_i A_i) = \lim_{i \to \infty} f(A_i) \tag{1}$$

for every decreasing sequence $A_1 \supset A_2 \supset \cdots$ of subsets of Ω, and we call a belief function on a proper multiplicative subclass of $\mathcal{P}(\Omega)$ *continuous* if it can be extended to a continuous belief function on $\mathcal{P}(\Omega)$. We call a belief function f on $\mathcal{P}(\Omega)$ *condensable* if

$$f(\cap \mathcal{Q}) = \inf_{A \in \mathcal{Q}} f(A) \tag{2}$$

for every downward net $\mathcal{Q}$ in $\mathcal{P}(\Omega)$, and we call a belief function on a proper multiplicative subclass of $\mathcal{P}(\Omega)$ *condensable* if it can be extended to a condensable belief function on $\mathcal{P}(\Omega)$. (A subset $\mathcal{Q}$ of $\mathcal{P}(\Omega)$ is called a *downward* *net* if for every pair A_1, A_2 of elements of $\mathcal{Q}$ there exists an element A of $\mathcal{Q}$ such that $A \subset A_1 \cap A_2$.)

Though condensability is a rather restrictive condition it is intimately related to the idea of "weights of evidence" (see Shafer (1976a)) and to Dempster's rule for combining belief functions (see Shafer (1978)), and hence it seems intuitively appropriate for belief functions that purport to represent empirical knowledge. The weaker condition of continuity seems appropriate in the case of partial beliefs arising from theoretical knowledge; it applies in particular to the partial beliefs arising from knowledge of chances or "objective probabilities."

The conditions of continuity and condensability can also be stated in terms of the upper probability function. A belief function f on $\mathcal{P}(\Omega)$ is continuous if

$$f^*(\cap_i A_i) = \lim_{i \to \infty} f^*(A_i)$$

for every increasing sequence $A_1 \subset A_2 \subset \cdots$ of subsets of Ω; it is condensable if

$$f^*(\cup\, \mathcal{Q}) = \sup_{A \in \mathcal{Q}} f^*(A)$$

for every upward net $\mathcal{Q} \subset \mathcal{P}(\Omega)$, or equivalently, if

$$f^*(A) = \sup\{f^*(B) | B \subset A; B \text{ is finite}\} \tag{3}$$

for all $A \subset \Omega$. This last expression shows how strong a condition condensability is; a condensable belief function on a power set is completely determined by its upper probabilities for finite subsets.

Suppose f is a belief function on an algebra $\mathcal{E}$ of subsets of Ω or, more generally, on a subset $\mathcal{E}$ of $\mathcal{P}(\Omega)$ that is both a multiplicative and an additive subclass. Then, as we see in Sect. 5 below, f is continuous if and only if it satisfies (1) for every decreasing sequence $A_1 \supset A_2 \supset \cdots$ of elements of $\mathcal{E}$ such that $\cap_i A_i$ is in $\mathcal{E}$. And f is condensable if and only if for every $A \in \mathcal{E}$ and every $\varepsilon > 0$ there exists a cofinite subset B of Ω such that $A \subset B$ and $f(C) - f(A) < \varepsilon$ for all $C \in \mathcal{E}$ such that $A \subset C \subset B$. These theorems are proven by showing how to extend a belief function satisfying one of these conditions to a continuous (or condensable) belief function on $\mathcal{P}(\Omega)$; the extensions exhibited are *canonical* in the sense that they award each subset of Ω the minimal degree of belief that is compelled by the adoption of f on $\mathcal{E}$ and by the hypothesis of continuity (or condensability).

The most important tool we use in our study of the extension of belief functions is the representation theorem presented in Sect. 3. This theorem is a direct consequence of an integral representation due to Choquet (1953), and it can also be deduced from more elementary work by Revuz (1955) and Honeycutt (1971). (These scholars' results are reviewed in Sect. 2.) The theorem says that every belief function can be represented by an *allocation of probability*: i.e., that for every belief function $f : \mathcal{E} \to [0, 1]$ there exists a complete Boolean algebra $\mathcal{M}$, a positive and completely additive measure μ on $\mathcal{M}$, and a mapping $\rho : \mathcal{E} \rightarrow \mathcal{M}$ that preserves finite meets and satisfies $f = \mu \circ \rho$. Notice the intuitive interpretation of this representation: the elements of $\mathcal{M}$ are portions of one's belief or "probability," and $\rho(A)$ is the portion of one's probability that is "allocated" or committed to A.

In addition to helping us extend belief functions, the representation of belief functions by allocations of probability also helps give intuitive content to the idea of condensability. It is also useful in the study of Dempster's rule of combination and in the study of particular belief functions that arise in connection with statistical inference.

2 $\cap$-homomorphisms

Suppose $\mathcal{E}$ is a multiplicative subclass of $\mathcal{P}(\Omega)$ containing both $\varnothing$ and Ω, and suppose $\mathcal{F}$ is a multiplicative subclass of $\mathcal{P}(\mathcal{X})$ containing both $\varnothing$ and $\mathcal{X}$. We call $r : \mathcal{E} \rightarrow \mathcal{F}$ a $\cap$-*homomorphism* if $r(\varnothing) = \varnothing, r(\Omega) = \mathcal{X}$, and $r(A \cap B) = r(A) \cap r(B)$ for all $A, B \in \mathcal{E}$. (Cf. Choquet (1953), p. 197.) It is easily seen that if f is a belief function and r is a $\cap$-homomorphism, then $f \circ r$ is also a belief function.

Since a finitely additive probability measure qualifies as a belief function, this implies in particular that $\mu \circ r$ is a belief function whenever $r : \mathcal{E} \rightarrow \mathcal{F}$ is a $\cap$-homomorphism, $\mathcal{F}$ is an algebra, and μ is a finitely additive probability measure on $\mathcal{F}$. Probability measures being abundant and $\cap$-homomorphisms being easy to construct, this fact enables us to construct an abundance of belief functions. In fact, all belief functions can be obtained in this way:

Theorem 1. *Suppose $\mathcal{E}$ is a multiplicative subclass of $\mathcal{P}(\Omega)$ and f is a belief function on $\mathcal{E}$. Then there exists a set $\mathcal{X}$, an algebra $\mathcal{F}$ of subsets of $\mathcal{X}$, a finitely additive probability measure μ on $\mathcal{F}$, and a $\cap$-homomorphism r: $\mathcal{E} \rightarrow \mathcal{F}$ such that $f = \mu \circ r$.*

This theorem is due to Choquet; it is a direct consequence of his integral representation theorem. It is also a direct consequence of a construction due to Revuz (1955) and Honeycutt (1971).

In its simplest version Choquet's integral representation theorem is merely a sharpening of the Krein-Milman theorem (see Choquet (1969), Vol. II, p. 117). It states that if $\mathcal{L}$ is a locally convex Hausdorff topological vector space, $\mathcal{U}$ is a compact convex subset of $\mathcal{L}$, and $f \in \mathcal{U}$, then there exists a Radon probability measure μ on $\mathcal{U}$ such that the support of $\mathcal{U}$ is contained in the closure $\mathcal{X}$ of the extreme points of $\mathcal{U}$ and f is the resultant of μ. (In other words, $\alpha(f) = \int_{\mathcal{X}} \alpha(g) d\mu(g)$ for every continuous linear function $\alpha : \mathcal{U} \rightarrow R$.) If we take $\mathcal{L}$ to be the vector space of all real-valued functions on $\mathcal{E}$, endowed with the topology of simple convergence, and let $\mathcal{U} \subset \mathcal{L}$ be the set of all belief functions on $\mathcal{E}$, then the set of extreme points of $\mathcal{U}$ consists of the *two-valued* belief functions—those that take only the values zero and one. (See Choquet (1953), pp. 260–261. Notice that the two-valued belief functions on $\mathcal{E}$ are in a one-to-one correspondence with the filters in $\mathcal{E}$; a filter $\mathcal{F} \subset \mathcal{E}$ corresponds to the belief function which assigns degree of belief one to all elements of $\mathcal{F}$ and degree of belief zero to all elements of $\mathcal{E} - \mathcal{F}$.) And this set is compact and hence equal to its closure $\mathcal{X}$. For each $A \in \mathcal{E}$, the mapping $\alpha_A : \mathcal{L} \to R : g \to g(A)$ is continuous and linear, and hence

$$\begin{aligned} f(A) = \alpha_A(f) &= \int_{\mathcal{X}} g(A)\ d\mu(g) \\ &= \mu\left(\{g \in \mathcal{X}\,|g(A) = 1|\}\right). \end{aligned}$$

That is to say, $f = \mu \circ r$, where r is the $\cap$-homomorphism given by $r(A) = \{g \in \mathcal{X} | g(A) = 1\}$.

In order to relate Theorem 1 to Revuz' construction, set $\mathcal{X} = \mathcal{P}(\mathcal{E}) - \varnothing$, define $r : \mathcal{E} \to \mathcal{P}(\mathcal{X})$ by $r(A) = \{B \in \mathcal{E} | \varnothing \neq B \subset A\}$, and let $\mathcal{F}$ be the algebra of subsets of $\mathcal{X}$ generated by the image $r(\mathcal{E})$. Revuz' work, as emended by Honeycutt, shows how to construct, for a given belief function f on $\mathcal{E}$, a unique finitely additive probability measure μ on $\mathcal{F}$ such that $f = \mu \circ r$.

The measure μ obtained in Choquet's proof is countably additive (in fact, it is a Radon measure), but the $\cap$-homomorphism r obtained in this proof need not preserve infinite intersections. In the Revuz-Honeycutt construction, on the other hand, the $\cap$-homomorphism r preserves arbitrary intersections (provided these intersections are in $\mathcal{E}$), but the measure μ need not be countably additive.

3 Allocations of Probability

As it turns out, it is both useful and intuitively appealing to replace the measure space $(\mathcal{X}, \mathcal{F}, \mu)$ of the preceding representation by a *probability* algebra: i.e., a complete Boolean algebra that has associated with it a positive and completely additive probability measure. In this section we show that every belief function can be represented by a $\cap$-homomorphism into a probability algebra. We call such $\cap$-homomorphisms *allocations of probability.*

Some notation and nomenclature: we denote a probability algebra $\mathcal{M}$'s zero by Λ, its unit by V. We use the symbols $\wedge, \vee$ and $\leqq$ to denote meet, join and majorization in $\mathcal{M}$, reserving the analogous symbols $\cap, \cup$ and $\subset$ for their set-theoretic roles. To say that the measure μ on $\mathcal{M}$ is positive is to say that $\mu(M) > 0$ for every nonzero element M of $\mathcal{M}$. To say that it is completely additive is to say that $\mu(\vee\mathcal{B}) = \Sigma_{M \in \mathcal{B}} \mu(M)$ whenever $\mathcal{B}$ is a collection of pairwise disjoint elements of $\mathcal{M}$. And when we say $\rho : \mathcal{E} \to \mathcal{M}$ is a $\cap$-homomorphism, we mean, of course, that $\rho(\varphi) = \Lambda, \rho(\Omega) = \mathrm{V}$, and $\rho(A \cap B) = \rho(A) \wedge \rho(B)$.

The condition that the measure μ on a probability algebra $\mathcal{M}$ be both positive and completely additive implies in particular that $\mathcal{M}$ must satisfy the *countable chain condition*: every collection of pairwise disjoint elements of $\mathcal{M}$ is countable. And using this statement one can further deduce that every subset $\mathcal{B}$ of $\mathcal{M}$ must have a countable subset $\mathcal{C}$ such that $\vee\mathcal{B} = \vee\mathcal{C}$, that $\mu(\vee\mathcal{B}) = \sup_{M\in\mathcal{B}}\mu(M)$ for every upward net $\mathcal{B}$ in $\mathcal{M}$, and that $\mu(\wedge\mathcal{B}) = \inf_{M\in\mathcal{B}}\mu(M)$ for every downward net $\mathcal{B}$ in $\mathcal{M}$. (See pp. 61–69 of Halmos (1963).)

Theorem 2. *Suppose f is a belief function on a multiplicative subclass $\mathcal{E}$. Then there exists an allocation of probability $\rho : \mathcal{E} \to \mathcal{M}$ such that $f = \mu \circ \rho$, where μ is the measure associated with the probability algebra $\mathcal{M}$.*

Proof. Recall that if $\mathcal{M}_0$ is a σ-algebra of subsets and μ_0 is a countably additive probability measure on $\mathcal{M}_0$, then a probability algebra can be constructed by taking the quotient of $\mathcal{M}_0$ by the σ-ideal $\mathcal{I}$ consisting of all sets in $\mathcal{M}_0$ of μ_0-measure zero; this quotient $\mathcal{M} = \mathcal{M}_0/\mathcal{I}$ is a complete Boolean algebra and the measure μ that μ_0 induces on $\mathcal{M}$ is positive and completely additive. The projection $\pi : \mathcal{M}_0 \to \mathcal{M}$ satisfies $\mu_0 = \mu \circ \pi$; and since it is a Boolean homomorphism, it is in particular a $\cap$-homomorphism. (For details, again see Halmos (1963).)

Since f is a belief function, Choquet's integral representation supplies us a σ-algebra $\mathcal{M}_0$, a countably additive probability measure μ_0 on $\mathcal{M}_0$, and a $\cap$-homomorphism $r : \mathcal{E} \to \mathcal{M}_0$ satisfying $f = \mu_0 \circ r$. Let $\mathcal{M}$ and π be defined as in the preceding paragraph, and set $\rho = \pi \circ r$. Then $f = \mu \circ \rho$, and ρ, being the composition of two $\cap$-homomorphisms, is itself a $\cap$-homomorphism and hence an allocation of probability.

(Notice that the appeal to Choquet's integral representation could be replaced by a more elementary approach based on Revuz' construction. That construction yields a $\cap$-homomorphism $r : \mathcal{E} \to \mathcal{M}_1$, where $\mathcal{M}_1$ is merely an algebra with a finitely additive probability measure μ_1. But the Stone representation theorem could be used to construct a σ-algebra $\mathcal{M}_0$, a countably additive measure μ_0, and a Boolean homomorphism $g : \mathcal{M}_1 \to \mathcal{M}_0$ such that $\mu_1 = \mu_0 \circ g$.)

The representation of a belief function f by an allocation ρ can be much more useful in theoretical discussions than the representation by a $\cap$-homomorphism into the algebra of a measure space, particularly if one is concerned with the conditions of continuity and condensability. For example:

Theorem 3. *Suppose $\rho : \mathcal{P}(\Omega) \to \mathcal{M}$ is an allocation for the belief function f. Then f is continuous if and only if*

$$\rho(\cap_i A_i) = \wedge_i \rho(A_i) \tag{4}$$

for every sequence $A_1, A_2, \cdots$ of subsets of Ω. And f is condensable if and only if

$$\rho(\cap\, \mathcal{Q}) = \wedge_{A\in\mathcal{Q}}\rho(A) \tag{5}$$

for every nonempty subset $\mathcal{Q}$ of $\mathcal{P}(\Omega)$.

The proof of this theorem is straightforward and directly yields a generalization to the case of an allocation ρ for a belief function on an arbitrary multiplicative subclass $\mathcal{E}$ of $\mathcal{P}(\Omega)$: in this case we may say that (1) holds for every decreasing sequence $A_1 \supset A_2 \supset \cdots$ of elements of $\mathcal{E}$ whose intersection is in $\mathcal{E}$ if and only if (4) holds for every sequence $A_1, A_2, \cdots$ of elements of $\mathcal{E}$ whose intersection is in $\mathcal{E}$; and that (2) holds for every downward net $\mathcal{Q} \subset \mathcal{E}$ whose intersection is in $\mathcal{E}$ if and only if (5) holds for every subset $\mathcal{Q}$ of $\mathcal{E}$ whose intersection is in $\mathcal{E}$.

The representation of a belief function by an allocation of probability ρ into a probability algebra $\mathcal{M}$ is intuitively meaningful because nonzero elements of $\mathcal{M}$ can be thought of as "probability masses" or "portions of belief," and $\rho(A)$ can be thought of as the (total) portion of belief one commits to A. The defining characteristics of an allocation of probability suit this interpretation; it seems reasonable to require that the measure of a portion of belief should always be positive, that the measures of disjoint portions should add, and that the portion committed to $A \cap B$ should include all of what is committed both to A and to B.

The notion of an allocation also lends itself to a geometric intuition. Suppose, for example, that ρ is an allocation from a power set $\mathcal{P}(\Omega)$ into a probability algebra $\mathcal{M}$. Then think of the probability represented by $\mathcal{M}$ as spread over the set Ω. But instead of distributing this probability in a fixed way, allow it a limited freedom of movement: require that a probability mass $M \in \mathcal{M}$ be constrained to remain inside a set $A \subset \Omega$ if and only if $M \leqq \rho(A)$. This makes geometric sense: if we write "M ct A" to indicate that M is constrained to A, then we find that M ct A and M ct B imply M ct $A \cap B$, that M ct A and N ct A imply $M \vee N$ ct A, etc.

Occasionally, it is convenient to shift our attention from an allocation $\rho : \mathcal{E} \to \mathcal{M}$ to the mapping $\zeta : \mathcal{E}^* \to \mathcal{M}$ defined by $\zeta(A) = \overline{\rho(\overline{A})}$. We call ζ an *allowment of probability* for $f = \mu \circ \rho$; it is dual to ρ in that it satisfies $f^* = \mu \circ \zeta$ and preserves joins rather than meets. Notice that in terms of the geometric intuition associated with an allocation, $\zeta(A) = \overline{\rho(\bar{A})}$ is the total probability mass that is not constrained to $\bar{A}$; i.e., the total probability mass that is *allowed* to move into A.

4 Condensability

The intuition associated with an allocation of probability on a power set $\mathcal{P}(\Omega)$ acquires its full force only when that allocation is condensable, for it is only in that case that a probability mass committed to each of a collection $\mathcal{B}$ of subsets of Ω is necessarily committed to the intersection $\cap\, \mathcal{B}$. Indeed, if f is a belief function on $\mathcal{P}(\Omega)$ with allocation $\rho : \mathcal{P}(\Omega) \to \mathcal{M}$ and allowment $\zeta : \mathcal{P}(\Omega) \to \mathcal{M}$, then the following conditions are all equivalent to the statement that f is condensable:

(1) $\rho(\cap\mathcal{B}) = \wedge_{\mathrm{B}\in\mathcal{B}}(B)$ for all $\mathcal{B} \subset \mathcal{P}(\Omega)$.

(2) If $\mathcal{B} \subset \mathcal{P}(\Omega), M \in \mathcal{M}$, and M ct B for each $B \in \mathcal{B}$, then M ct $\cap \mathcal{B}$.
(3) $\zeta(\cup\mathcal{B}) = \vee_{B\in\mathcal{B}}\zeta(B)$ for all $\mathcal{B} \subset \mathcal{P}(\Omega)$.
(4) If $\varnothing \neq A \subset \Omega$, then there exists a sequence $\omega_1, \omega_2, \cdots$ of elements of A and a countable disjoint partition $M_1, M_2, \cdots$ of $\zeta(A)$ such that $M_i \leqq \zeta(\{\omega_i\})$ for each i.
(5) There exists a mapping $\lambda : \mathcal{M} \to \mathcal{P}(\Omega)$ such that an element M of $\mathcal{M}$ and a subset A of Ω satisfy M ct A if and only if $\lambda(M) \subset A$.

Notice the geometric interpretation of (4) and (5). For each $M \in \mathcal{M}, \lambda(M)$ is the smallest subset of Ω to which all of M is constrained. And (4) demands sufficient freedom of movement for the probability mass $\zeta(A)$ to allow any diffusion, or “continuous” distribution, to be reversed: it must be possible for $\zeta(A)$ to “condense” into a countable number of discrete probability masses, each still located within A.

5 The Canonical Extension of Belief Functions

Given a belief function f on a multiplicative subclass $\mathcal{E}$ of $\mathcal{P}(\Omega)$, we define $\bar{f}$ on $\mathcal{P}(\Omega)$ by setting

$$\bar{f}(A) = \sup\left\{\Sigma(-1)^{|I|+1}f(\cap_{i\in I}A_i)|\varnothing \neq I \subset \{1,\cdots,n\}\right\}, \tag{6}$$

where the supremum is taken over all $n \geqq 1$ and all collections $A_1, A_2, \cdots A_n$ of elements of $\mathcal{E}$ that are subsets of A.

Notice that if $\mathcal{E}$ is an additive as well as a multiplicative subclass, then (6) reduces to

$$\bar{f}(A) = \sup\{f(B)|B \in \mathcal{E}; B \subset A\}. \tag{7}$$

In this case we define $\tilde{f}$ and $\hat{f}$ on $\mathcal{P}(\Omega)$ by

$$\tilde{f}(A) = \sup\{\lim_{i\to\infty} f(A_i)|A_1 \supset A_2 \supset \cdots \in \mathcal{E}; \cap\, A_i \subset A\} \tag{8}$$

and

$$\hat{f}(A) = \inf\left\{\bar{f}(B)|B \subset \Omega \text{ is cofinite; } A \subset B\right\}. \tag{9}$$

Theorem 4. *Suppose f is a belief function on a multiplicative subclass $\mathcal{E}$ of $\mathcal{P}(\Omega)$.*

(1) *$\bar{f}$ is a belief function, and $f = \bar{f}|\mathcal{E}$. Furthermore,*

$$\bar{f} = \inf\{g|g \text{ is a belief function on } \mathcal{P}(\Omega) \text{ and } g|\mathcal{E} = f\}.$$

(2) *Suppose $\mathcal{E}$ is an additive as well as a multiplicative subclass. Then f is continuous if and only if*

$$f(\cap_i A_i) = \lim_{i\to\infty} f(A_i) \tag{10}$$

for every decreasing sequence $A_1 \supset A_2 \supset \cdots$ *of elements of* $\mathcal{E}$ *such that* $\cap_i A_i \in \mathcal{E}$. *If* f *is continuous, then* $\tilde{f}$ *is a continuous belief function,* $f = \tilde{f}|\mathcal{E}$, *and*

$$\tilde{f} = \inf\{g|g \textit{ is a continuous belief function on } \mathcal{P}(\Omega) \textit{ and } g|\mathcal{E} = f\}.$$

If f *is continuous and* $\mathcal{E}$ *is closed under countable intersections, then* $\tilde{f} = \bar{f}$.

(3) *Suppose* $\mathcal{E}$ *is an additive as well as a multiplicative subclass. Then* f *is condensable if and only if for every* $A \in \mathcal{E}$ *and every* $\varepsilon > 0$ *there exists a cofinite subset* B *of* Ω *such that* $A \subset B$ *and* $f(C) - f(A) < \varepsilon$ *for all* $C \in \mathcal{E}$ *such that* $A \subset C \subset B$. *If* f *is condensable, then* $\hat{f}$ *is a condensable belief function,* $f = \hat{f}|\mathcal{E}$, *and*

$$\hat{f} = \inf\{g|g \textit{ is a condensable belief function on } \mathcal{P}(\Omega) \textit{ and } g|\mathcal{E} = f\}.$$

If f *is condensable and* $\mathcal{E}$ *is closed under arbitrary unions and intersections, then* $\hat{f} = \bar{f}$.

Proof. Let $\rho : \mathcal{E} \to \mathcal{M}$ be an allocation of probability for f, and let μ denote the measure on $\mathcal{M}$.

(1) Define $\bar{\rho} : \mathcal{P}(\Omega) \to \mathcal{M}$ by $\bar{\rho}(A) = \vee\{\rho(B)|B \in \mathcal{E}; B \subset A\}$. It is easily verified that $\bar{\rho}$ is an allocation and that $\bar{f} = \mu \circ \bar{\rho}$; hence $\bar{f}$ is a belief function. The other assertions in (1) are then obvious.

(2) It is clear that if f is continuous, then (10) holds. Suppose, on the other hand, that (10) holds.

For each $A \subset \Omega$, define $\mathcal{D}(A) \subset \mathcal{M}$ by

$$\mathcal{D}(A) = \{\wedge_{B \in \mathcal{B}} \rho(B) \,|\mathcal{B} \text{ is a countable subset of } \mathcal{E}; \cap \mathcal{B} \subset A\}.$$

Notice that $\mathcal{D}(A)$ is an upward net in $\mathcal{M}$. (If M_1 and M_2 are the elements of $\mathcal{D}(A)$ corresponding to subsets $\mathcal{B}_1$ and $\mathcal{B}_2$ of $\mathcal{E}$, then $\mathcal{B} \equiv \{B_1 \cup B_2 | B_1 \in \mathcal{B}_1; B_2 \in \mathcal{B}_2\}$ will also be countable subset of $\mathcal{E}$ with $\cap \mathcal{B} \subset A$, and the element of $\mathcal{D}(A)$ corresponding to $\mathcal{B}$ will majorize both M_1 and M_2.) Define $\tilde{\rho} : \mathcal{P}(\Omega) \to \mathcal{M}$ by $\tilde{\rho}(A) = \vee\mathcal{D}(A)$. We will show that $\tilde{\rho}$ is a continuous allocation, that $\tilde{f} = \mu \circ \tilde{\rho}$, and that $\tilde{\rho}|\mathcal{E} = \rho$; the assertions of (2) will then be obvious.

The relation $\tilde{\rho}|\mathcal{E} = \rho$ follows from the fact that

$$\rho(\cap\mathcal{B}) = \wedge_{B \in \mathcal{B}} \rho(B)$$

whenever $\mathcal{B} \in \mathcal{E}$ is countable and $\cap \mathcal{B} \in \mathcal{E}$. (See the comment following Theorem 3.2.) For in the case where $\mathcal{B} \subset \mathcal{E}$ and $\cap \mathcal{B} \subset A \in \mathcal{E}$, we therefore have

$$\wedge_{B \in \mathcal{B}} \rho(B) \leqq \wedge_{B \in \mathcal{B}} \rho(A \cup B) = \rho(\cap_{B \in \mathcal{B}}(A \cup B)) = \rho(A).$$

To verify that $\tilde{f} = \mu \circ \tilde{\rho}$, we must notice that for any sequence $A_1, A_2, \cdots$ in $\mathcal{E}$ there is a decreasing sequence $B_1, B_2, \cdots$, defined by

$$B_i = A_1 \cap \cdots \cap A_i,$$

which satisfies both $\cap_i B_i = \cap_i A_i$ and $\wedge_i \rho(B_i) = \wedge\ _i\rho(A_i)$. Hence

$$\mathcal{D}(A) = \{\wedge_i \rho\,(A_i)\,|A_1, A_2, \cdots \in \mathcal{E}; A_1 \supset A_2 \supset \cdots ; \cap_i A_i \subset A\}\,.$$

And since $\mathcal{D}(A)$ is an upward net, it follows that

$$\begin{aligned}\mu\,(\tilde{\rho}(A))) &= \mu(\vee\mathcal{D}(A))\\ &= \sup\nolimits_{M\in\mathcal{D}(A)} \mu(M)\\ &= \sup\{\mu\,(\wedge_i \rho\,(A_i))\,|A_1, A_2, \cdots \in \mathcal{E}; A_1 \supset A_2 \supset \cdots ; \cap_i A_i \subset A\}\\ &= \sup\{\lim\nolimits_{i\to\infty} f\,(A_i)|A_1, A_2, \cdots \in \mathcal{E}; A_1 \supset A_2 \supset \cdots ; \cap_i A_i \subset A\}\\ &= \tilde{f}(A).\end{aligned}$$

The fact that $\tilde{\rho}|\mathcal{E} = \rho$ means in particular that $\tilde{\rho}(\varnothing) = \Lambda$ and $\tilde{\rho}(\Omega) = \mathrm{V}$. So in order to show that $\tilde{\rho}$ is a continuous allocation, we need only show that it preserves countable meets; i.e., that

$$\tilde{\rho}\,(\cap_i A_i) = \wedge_i \tilde{\rho}(A_i),$$

or

$$\vee\mathcal{D}\,(\cap_i A_i) = \wedge \vee \mathcal{D}\,(A_i)$$

for any sequence $A_1, A_2, \cdots$ of subsets of Ω. To this end, we fix the sequence $A_1, A_2, \cdots$ and simplify our notation by setting $\mathcal{D} \equiv \mathcal{D}(\cap_i A_i), \mathcal{D}_i \equiv \mathcal{D}(A_i)$ and $M \equiv \wedge_i \vee \mathcal{D}_i$. Our task is then to show that $\vee\mathcal{D} = M$. And since $\mathcal{D} \subset \mathcal{D}_i$ for each i, the relation $\vee\mathcal{D} \leqq \wedge_i \vee \mathcal{D}_i = M$ is immediate, and it remains only to show that $\vee\mathcal{D} \geqq M$.

Since $\mathcal{D}_i$ is an upward net, it will include an element that arbitrarily nearly covers its meet $\vee\mathcal{D}_i$. In particular, if $\varepsilon > 0$ then we can choose $M_i \in \mathcal{D}_i$ such that

$$\mu\,(\vee\mathcal{D}_i - M_i) \leqq \frac{\varepsilon}{2i}.$$

(PROOF. By the countable chain condition, $\mathcal{D}_i$ has a countable subset $\mathcal{E}_i$ such that $\vee\mathcal{E}_i = \vee\mathcal{D}_i$. Since $\mathcal{D}_i$ is an upward net, $\mathcal{E}_i$ may be taken as an increasing sequence, and then the continuity of μ assures that an element sufficiently far along in this sequence will have measure within $\varepsilon/2i$ of the measure of $\vee\mathcal{D}_i$.) Since $M \leqq \vee\mathcal{D}_i$, we also have

$$\mu\,(M - M_i) \leqq \frac{\varepsilon}{2i}.$$

Fix $\varepsilon > 0$ and choose such an $M_i \in \mathcal{D}_i$ for each i. And let $\mathcal{B}_i$ be a countable subset of $\mathcal{E}$ such that $\cap\, \mathcal{B}_i \subset A_i$ and $M_i = \wedge_{B \in \mathcal{B}_i} \rho(B)$. Set $\mathcal{B}_\varepsilon = \cup_i \mathcal{B}_i$ and $M_\varepsilon = \wedge_i M_i$. Then $\cap\, \mathcal{B}_\varepsilon \subset \cap_i A_i$, and

$$M_\varepsilon = \wedge_i (\wedge_{B \in \mathcal{B}_i} \rho(B)) = \wedge_{B \in \mathcal{B}_\epsilon} \rho(B);$$

thus $M_\varepsilon \in \mathcal{D}$, so that $M_\varepsilon \leqq \vee \mathcal{D}$. Since

$$\mu\,(M - M_\varepsilon) = \mu\,(\vee_i (M - M_i)) \leqq \varepsilon$$

it follows that $\vee \mathcal{D}$ includes all but at most ε of M. And since ε is arbitrary, this yields the conclusion that $\vee \mathcal{D} \geqq M$.

(3) Suppose f is condensable. Then there exists a condensable belief function g on $\mathcal{P}(\Omega)$ such that $f = g|\mathcal{E}$. Since g is condensable,

$$g(A) = \inf \{g(B) | B \subset \Omega \text{ is cofinite;} \quad A \subset B\}. \tag{11}$$

(Cf. (3).) It follows that for all $A \in \mathcal{E}$ and all $\varepsilon > 0$, there exists a cofinite subset B of Ω such that $A \subset B$ and $f(C) - f(A) < \varepsilon$ for all $C \in \mathcal{E}$ such that $A \subset C \subset B$.

Suppose, on the other hand, that the condition of the preceding sentence is met. Then we define $\hat{\rho} : \mathcal{P}(\Omega) \to \mu$ by

$$\hat{\rho}(A) = \wedge \{\bar{\rho}(B) | B \subset \Omega \text{ in confinite}; A \subset B\}.$$

It is clear that $\hat{f} = \mu \circ \hat{\rho}$. We will show that $\hat{\rho}|\mathcal{E} = \rho$ and that $\hat{\rho}$ is a condensable allocation.

Suppose $A \in \mathcal{E}$. Clearly $\hat{\rho}(A) \geqq \rho(A)$. In order to show that $\hat{\rho}(A) = \rho(A)$, we fix $\varepsilon > 0$ and choose a cofinite subset B_ε of Ω such that $A \subset B_\varepsilon$ and $f(C) - f(A) < \varepsilon/2$ for all $C \in \mathcal{E}$ such that $A \subset C \subset B_\varepsilon$. Then

$$\begin{aligned}\rho(A) &= \wedge \{\vee \{\rho(C) | C \in \mathcal{E}; C \subset B\} | B \subset \Omega \text{ is confinite; } A \subset B\} \\ &= \wedge \{\vee \{\rho(C) | C \in \mathcal{E};\ A \subset C \subset B\} | B \subset \Omega \text{ is confinite; } A \subset B\} \\ &\leqq \vee \{\rho(C) | C \in \mathcal{E};\ A \subset C \subset B_\varepsilon\}.\end{aligned}$$

Denote this last element of $\mathcal{M}$ by M_ε. Since $\{\rho(C) | C \in \mathcal{E}; A \subset C \subset B_\varepsilon\}$ is an upward net, we may choose $C_\varepsilon \in \mathcal{E}$ such that $A \subset C \subset B_\varepsilon$ and $\mu(M_\varepsilon) - f(C_\varepsilon) = \mu(M_\varepsilon - \rho(C_\varepsilon)) < \varepsilon/2$. Since $M_\varepsilon \geqq \hat{\rho}(A) \geqq \rho(A)$, we have

$$\begin{aligned}\mu(\hat{\rho}(A)) - \rho(A)) &\leqq \mu(M_\varepsilon - \rho(A)) = |\mu(M_\varepsilon) - f(A)| \\ &= |\mu\,(M_\varepsilon) - f\,(C_\varepsilon) + f\,(C_\varepsilon) - f\,(A)| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

And since ε may be chosen arbitrarily small, this means $\mu(\hat{\rho}(A) - \rho(A)) = 0$, or $\hat{\rho}(A) = \rho(A)$. So $\hat{\rho}|\mathcal{E} = \rho$.

The fact that $\hat{\rho}|\mathcal{E} = \rho$ means in particular that $\hat{\rho}(\varnothing) = \Lambda$ and $\hat{\rho}(\Omega) = V$. So in order to show that $\hat{\rho}$ is a condensable allocation, we need only show that it preserves arbitrary meets. Fix a subset $\mathcal{B}$ of $\mathcal{E}$. A cofinite subset of Ω contains $\cap\, \mathcal{B}$ if and only if it contains some finite intersection of elements of $\mathcal{B}$, and it does this if and only if it itself is the intersection of a finite number of cofinite subsets of Ω, each of which contains some element of $\mathcal{B}$. Hence

$$\begin{aligned}\hat{\rho}(\cap\mathcal{B}) &= \wedge\,\{\bar{\rho}(C)|C \subset \Omega \text{ is cofinite}; \cap\, \mathcal{B} \subset C\}\\ &= \wedge\{\bar{\rho}(C_1 \cap \cdots \cap C_n)|n \geqq 1; C_1, \cdots, C_n \text{ are}\\ &\qquad \text{confinite subsets of } \Omega\text{, each containing some}\\ &\qquad \text{element of } \mathcal{B}\}\\ &= \wedge\{\bar{\rho}(C_1) \wedge \cdots \wedge \bar{\rho}(C_n)|n \geqq 1; C_1, \cdots, C_n\\ &\qquad \text{are confinite subsets of } \Omega\text{, each containing}\\ &\qquad \text{some element of } \mathcal{B}\}\\ &= \wedge\{\bar{\rho}(C)|C \subset \Omega \text{ is confinite; } C \text{ containing some}\\ &\qquad \text{element of } \mathcal{B}\}\\ &= \wedge_{B\in\mathcal{B}}\hat{\rho}(B).\end{aligned}$$

So $\hat{\rho}$ is a condensable allocation.

Suppose g is a condensable belief function on $\mathcal{P}(\Omega)$ and $g|\mathcal{E} = f$. Then $g \geqq \bar{f}$ by (1), and comparison of (9) and (11) shows that $g \geqq \hat{f}$.

Finally, suppose $\mathcal{E}$ is closed under arbitrary unions and intersections. Then a mapping $\theta : \mathcal{P}(\Omega) \to \mathcal{E}$ may be defined by $\theta(A) = \cup\{B|B \in \mathcal{E}, B \subset A\}$. This mapping satisfies $\bar{f} = f \circ \theta$ and preserves arbitrary intersections. So if $\mathcal{B}$ is a downward net in $\mathcal{P}(\Omega)$, then $\{\theta(B)|B \in \mathcal{B}\}$ is a downward net in $\mathcal{E}$. Using all these facts, together with the condensability of f, we obtain

$$\begin{aligned}\bar{f}(\cap\, \mathcal{B}) = f(\theta(\cap\, \mathcal{B})) &= f(\cap_{B\in\mathcal{B}}\theta(B))\\ &= \inf_{B\in\mathcal{B}} f(\theta(B)) = \inf_{B\in\mathcal{B}}\bar{f}(B)\end{aligned}$$

for any downward net $\mathcal{B}$ in $\mathcal{P}(\Omega)$. Thus $\bar{f}$ is condensable. It follows that $\bar{f} = \hat{f}$.

The belief function $\bar{f}$ assigns to each subset of Ω only the degree of belief that f forces it to assign, and it is therefore the belief function on $\mathcal{P}(\Omega)$ that we will adopt if our knowledge about Ω is limited to what f says about $\mathcal{E}$. (See Chap. 6 of Shafer (1976a) for further discussion.) Hence we may call $\bar{f}$ the *canonical extension* of f to $\mathcal{P}(\Omega)$.

Similarly, let us call a continuous belief function h on $\mathcal{P}(\Omega)$ the *canonical continuous extension* of f to $\mathcal{P}(\Omega)$ in the case where f is continuous and

$$h = \inf\,\{g|g \text{ is a continuous belief function on } \mathcal{P}(\Omega) \text{ and } g|\mathcal{E} = f\}\,.$$

And let us call a condensable belief function h on $\mathcal{P}(\Omega)$ the *canonical condensable extension* of f to $\mathcal{P}(\Omega)$ in the case where f is condensable and

$$h = \inf \{g | g \text{ is a condensable belief function on } \mathcal{P}(\Omega) \text{ and } g|\mathcal{E} = f\} .$$

Theorem 4 tells us that canonical continuous and condensable extensions always exist when $\mathcal{E}$ is an additive as well as a multiplicative subclass; it is an interesting open question whether they always exist when $\mathcal{E}$ is merely a multiplicative subclass.

The notion of canonical extension generalizes to the case of larger multiplicative subclasses that fall short of the whole power set; if $\mathcal{E}_1 \subset \mathcal{E}_2$ are both multiplicative subclasses of $\mathcal{P}(\Omega)$ and f is a belief function on $\mathcal{E}_1$, then it is evident that

$$\bar{f}|\mathcal{E}_2 = \inf \{g|g \text{ is a belief function on } \mathcal{E}_2 \text{ and } g|\mathcal{E}_1 = f\} ,$$

and hence we may call $\bar{f}|\mathcal{E}_2$ the canonical extension of f to $\mathcal{E}_2$.

Notice that this process of canonical extension is consistent: if $\mathcal{E}_2 \subset \mathcal{E}_3$, then the canonical extension to $\mathcal{E}_3$ of f is the canonical extension to $\mathcal{E}_3$ of the canonical extension to $\mathcal{E}_2$ of f. If $\mathcal{E}_1 \subset \mathcal{E}_2$ and a belief function f on $\mathcal{E}_2$ is the canonical extension to $\mathcal{E}_2$ of its restriction $f|\mathcal{E}_1$, we say that f is *discerned* by $\mathcal{E}_1$.

It should be pointed out that the "possibilities" in a set Ω can always be split into more fully described possibilities, so that $\mathcal{P}(\Omega)$ is rendered merely a complete subalgebra of a larger power set. (See Chap. 6 of Shafer (1976a).) Thus power sets must share with all complete algebras any special status they can claim as domains for belief functions. It is reassuring, therefore, that the canonical extension of a belief function f from a complete algebra coincides with the canonical continuous extension if f is continuous and with the canonical condensable extension if f is condensable.

As the reader may have noticed, the formula for $\tilde{f}$ in Theorem 4 gives the usual inner measure when applied to a continuous (i.e., countably additive) probability measure f on an algebra $\mathcal{E}$, and in particular gives the *unique* extension of f to a continuous probability measure on the σ-algebra $\tilde{\mathcal{E}}$ generated by $\mathcal{E}$. But the canonical continuous extension of a continuous belief function on an algebra $\mathcal{E}$ is not in general its only continuous extension, even to $\tilde{\mathcal{E}}$. To see that this is true, choose an algebra $\mathcal{E} \subset \mathcal{P}(\Omega)$ that contains no singletons, but such that $\tilde{\mathcal{E}}$ contains all the singletons in $\mathcal{P}(\Omega)$. (For example, set $\Omega = [0, 1)$ and let $\mathcal{E}$ consist of all finite unions of left-closed, right-open subintervals on Ω.) And let f be the *vacuous belief function* on $\mathcal{E}$; i.e., the belief function that assigns degree of belief zero to every proper subset of Ω in $\mathcal{E}$. Then the canonical continuous extension of $\mathcal{E}$ to $\tilde{\mathcal{E}}$ is simply the vacuous belief function on $\tilde{\mathcal{E}}$. But for every $\omega \in \Omega$, the two-valued belief function on $\tilde{\mathcal{E}}$ corresponding to the principal filter $(\Omega, \Omega - \{\omega\}) \subset \tilde{\mathcal{E}}$ is also a continuous extension of f.

The method of defining $\hat{f}$ will appear familiar to some readers; it is analogous to Choquet's method of extending a capacity. It does not appear, however, that (3) of Theorem 4 can be cast as a special case of Choquet's results on the extension of capacities. (See pp. 158–164 of Choquet (1969).)

If the multiplicative subclass $\mathcal{E}$ is not closed under countable intersections, then we can easily construct a continuous two-valued belief function f on $\mathcal{E}$ such that $\tilde{f} \neq \bar{f}$. We simply choose a sequence $A_1, A_2, \cdots$ in $\mathcal{E}$ such that $\cap_i A_i \notin \mathcal{E}$ and let f be the two-valued belief function corresponding to the principal filter $\{A \in \mathcal{E} | \cap_i A_i \subset A\}$, so that $\bar{f}(\cap_i A_i) = 0$ but $\tilde{f}(\cap_i A_i) = 1$. If $\mathcal{E}$ is not closed under arbitrary intersections, then one can similarly construct a condensable two-valued belief function f such that $\hat{f} \neq \bar{f}$.

Acknowledgment

I thank the many friends and colleagues who have encouraged the research for this paper. I am especially grateful to G. A. Hunt, Terry Shafer, and Hale Trotter.

References

[1] Choquet, Gustave (1953). Theory of capacities. *Ann. Inst. Fourier* (*Grenoble*) **5** 131–295.

[2] Choquet, Gustave (1969). *Lectures on Analysis*. Benjamin, New York.

[3] Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.* **38** 325–339.

[4] Dempster, A. P. (1968). A generalization of Bayesian inference. *J. Roy. Statist. Soc. Ser. B* **30** 205–247.

[5] Halmos, Paul R. (1963). *Lectures on Boolean Algebras.* Van Nostrand-Reinhold, London.

[6] Honeycutt, James E., Jr. (1971). On an abstract Stieltjes measure. *Ann. Inst. Fourier* (*Grenoble*) **21** 143–154.

[7] Revuz, Andre (1955). Fonctions croissantes et mesures sur les espaces topologiques ordonnés. *Ann. Inst. Fourier* (*Grenoble*) **6** 187–269.

[8] Shafer, Glenn (1976a). *A Mathematical Theory of Evidence.* Princeton Univ. Press.

[9] Shafer, Glenn (1976b). A theory of statistical evidence. In *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science* (W. L. Harper and C. A. Hooker, eds.) **2**, pp. 365–436.

[10] Shafer, Glenn (1978). Dempster's rule of combination. Unpublished manuscript.

[11] Sikorski, Roman (1969). *Boolean Algebras*, 3rd ed. Springer-Verlag, New York.

8

Computational Methods for A Mathematical Theory of Evidence *†

Jeffrey A. Barnett

Abstract. Many knowledge-based expert systems employ numerical schemes to represent evidence, rate competing hypotheses, and guide search through the domain's problem space. This paper has two objectives: first, to introduce one such scheme, developed by Arthur Dempster and Glen Shafer, to a wider audience; second, to present results that can reduce the computation-time complexity from exponential to linear, allowing this scheme to be implemented in many more systems. In order to enjoy this reduction, some assumptions about the structure of the type of evidence represented and combined must be made. The assumption made here is that each piece of the evidence either confirms or denies a single proposition rather than a disjunction. For any domain in which the assumption is justified, the savings are available.

1 Introduction

How should knowledge-based expert systems reason? Clearly, when domain-specific idiosyncratic knowledge is available, it should be formalized and used to guide the inference process. Problems occur either when the supply of easy-to-formalize knowledge is exhausted before our systems pass the "sufficiency" test or when the complexity of representing and applying the knowledge is beyond the state of our system building technology. Unfortunately, with the current state of expert-system technology, this is the normal, not the exceptional case.

At this point, a fallback position must be selected, and if our luck holds, the resulting system exhibits behavior interesting enough to qualify as a success.

* From *Proceedings of the Seventh International Conference on Artificial Intelligence*, 1981, 868–875.

† This research is supported by the Defense Advanced Research Projects Agency under Contract No. DAHC15 72 C 0308. Views and conclusions contained in this report are the author's and should not be interpreted as representing the official opinion or policy of DARPA, the US Government, or any person or agency connected with them.

Typically, a fallback position takes the form of a uniformity assumption allowing the utilization of a non-domain-specific reasoning mechanism: for example, the numerical evaluation procedures employed in mycin [17] and internist [14], the simplified statistical approach described in [10], and a multivalued logic in [18]. The hearsay-ii speech understanding system [13] provides another example of a numerical evaluation and control mechanism—however, it is highly domain-specific.

Section 2 describes another scheme of plausible inference, one that addresses both the problem of representing numerical weights of evidence and the problem of combining evidence. The scheme was developed by Arthur Dempster [3, 4, 5, 6, 7, 8, 9], then formulated by his student, Glen Shafer [15, 16], in a form that is more amenable to reasoning in finite discrete domains such as those encountered by knowledge-based systems. The theory reduces to standard Bayesian reasoning when our knowledge is accurate but is more flexible in representing and dealing with ignorance and uncertainty. Section 2 is a review and introduction. Other work in this area is described in [12].

Section 3 notes that direct translation of this theory into an implementation is not feasible because the time complexity is exponential. However, if the type of evidence gathered has a useful structure, then the time complexity issue disappears. Section 4 proposes a particular structure that yields linear time complexity. In this structure, the problem space is partitioned in several independent ways and the evidence is gathered within the partitions. The methodology also applies to any domain in which the individual experiments (separate components of the evidence) support either a single proposition or its negation.

Section 5 and 6 develop the necessary machinery to realize linear time computations. It is also shown that the results of experiments may vary over time, therefore the evidence need not be monotonic. Section 7 summarizes the results and notes directions for future work in this area.

2 The Dempster-shafer Theory

A theory of evidence and plausible reasoning is described in this section. It is a theory of evidence because it deals with weights of evidence and numerical degrees of support based upon evidence. Further, it contains a viewpoint on the representation of uncertainty and ignorance. It is also a theory of plausible reasoning because it focuses on the fundamental operation of plausible reasoning, namely the combination of evidence. The presentation and notation used here closely parallels that found in [16].

After the formal description of how the theory represents evidence is presented in Sect. 2.1, an intuitive interpretation is given in Sect. 2.2, then a comparison is made, in Sect. 2.3, to the standard Bayesian model and similarities and differences noted. The rule for combining evidence, Dempster's orthogonal sum, is introduced in Sect. 2.4 and compared to the Bayesians' method of conditioning in Sect. 2.5. Finally, Sect. 2.6 defines the simple and separable support functions. These functions are the theory's natural representation of actual evidence.

2.1 Formulation of the Representation of Evidence

Let Θ be a set of propositions about the exclusive and exhaustive possibilities in a domain. For example, if we are rolling a die, Θ contains the six propositions of the form 'the number showing is i' where $1 \leq i \leq 6$. Θ is called the *frame of discernment* and 2^{Θ} is the set of all subsets of Θ. Elements of 2^{Θ}, i.e., subsets of Θ, are the class of general propositions in the domain; for example, the proposition 'the number showing is even' corresponds to the set of the three elements of Θ that assert the die shows either a 2, 4, or 6.

The theory deals with refinings, coarsenings, and enlargements of frames as well as families of compatible frames. However, these topics are not pursued here—the interested reader should see [16] where they are developed.

A function Bel: $2^{\Theta} \to [0,1]$, is a *belief function* if it satisfies $\mathrm{Bel}(\emptyset) = 0$, and for any collection, $A_1, \ldots, A_n$, of subsets of Θ,

$$\mathrm{Bel}(A_1 \cup \cdots \cup A_n) \geq \sum_{\substack{I \subseteq \{1 \ldots n\} \\ I \neq \emptyset}} (-1)^{|I|+1} \mathrm{Bel}(\bigcap_{i \in I} A_i).$$

A belief function assigns to each subset of Θ a measure of our total belief in the proposition represented by the subset. The notation, $|I|$, is the cardinality of the set I.

A function m: $2^{\Theta} \to [0,1]$ is called a *basic probability assignment* if it satisfies $m(\phi) = 0$ and

$$\sum_{A \subseteq \Theta} m(A) = 1.$$

The quantity, $m(A)$, is called A's *basic probability number*; it represents our exact belief in the proposition represented by A. The relation between these concepts and probabilities are discussed in Sect. 2.3. If m is a basic probability assignment, then the function defined by

$$\mathrm{Bel}(A) = \sum_{B \subseteq A} m(B), \quad \text{for all } A \subseteq \Theta \tag{1}$$

is a belief function. Further, if Bel is a belief function, then the function defined by

$$m(A) = \sum_{B \subseteq A} (-1)^{|A-B|} \mathrm{Bel}(B) \tag{2}$$

is a basic probability assignment. If equations (1) and (2) are composed in either order, the result is the identity-transformation. Therefore, there corresponds to each belief function one and only one basic probability assignment. Conversely, there corresponds to each basic probability assignment one and only one belief function. Hence, a belief function and a basic probability assignment convey exactly the same information.

Other measures are useful in dealing with belief functions in this theory. A function $Q\colon 2^{\Theta} \to [0,1]$ is a *commonality function* if there is a basic probability assignment, m, such that

$$Q(A) = \sum_{A \subseteq B} m(B) \tag{3}$$

for all $A \subseteq \Theta$. Further, if Q is a commonality function, then the function defined by

$$\mathrm{Bel}(A) = \sum_{B \subseteq \neg A} (-1)^{|B|} Q(B)$$

is a belief function. From this belief function, the underlying basic probability assignment can be recovered using (2); if this is substituted into (3), the original Q results. Therefore, the sets of belief functions, basic probability assignments, and commonality functions are in one-to-one correspondence and each representation conveys the same information as any of the others.

Corresponding to each belief function are two other commonly used quantities that also carry the same information. Given a belief function Bel, the function $\mathrm{Dou}(A) = \mathrm{Bel}(\neg A)$, is called the *doubt function* and the function $P^{\star}(A) = 1 - \mathrm{Dou}(A) = 1 - \mathrm{Bel}(\neg A)$, is called the *upper probability function.*

For notational convenience, it is assumed that the functions Bel, m, Q, Dou, and $P^{\star}$ are each derived from one another. If one is subscripted, then all others with the same subscript are assumed to be derived from the same underlying information.

2.2 An Interpretation

It is useful to think of the basic probability number, $m(A)$, as the measure of a probability mass constrained to stay in A but otherwise free to move. This freedom is a way of imagining the noncommittal nature of our belief, i.e., it represents our ignorance because we can not further subdivide our belief and restrict the movement. Using this allusion, it is possible to give intuitive interpretations to the other measures appearing in the theory.

The quantity $\mathrm{Bel}(A) = \sum_{[B \subseteq A]} m(B)$ is the measure of the total probability mass constrained to stay somewhere in A. On the other hand, $Q(A) = \sum_{[A \subseteq B]} m(B)$ is the measure of the total probability mass that can move freely to any point in A. It is now possible to understand the connotation intended in calling m the measure of our exact belief and Bel the measure of our total belief. If $A \subseteq B \subseteq \Theta$, then this is equivalent to the logical statement that A implies B. Since $m(A)$ is part of the measure $\mathrm{Bel}(B)$, but not conversely, it follows that the total belief in B is the sum of the exact belief in all propositions that imply B plus the exact belief in B itself.

With this interpretation of Bel, it is easy to see that $\mathrm{Dou}(A) = \mathrm{Bel}(\neg A)$ is the measure of the probability mass constrained to stay out of A. Therefore, $P^\star(A) = 1 - \mathrm{Dou}(A)$ is the measure of the total probability mass that can move into A, though it is not necessary that it can all move to a single point, hence $P^\star(A) = \sum_{[A \cap B \neq \emptyset]} m(B)$ is immediate. It follows that $P^\star(A) \geq \mathrm{Bel}(A)$ because the total mass that can move into A is a superset of the mass constrained to stay in A.

2.3 Comparison with Bayesian Statistics

It is interesting to compare this and the Bayesian model. In the latter, a function $p\colon \Theta \to [0,1]$ is a *chance density function* if $\sum_{[a \in \Theta]} p(a) = 1$; and the function $\mathrm{Ch}\colon 2^\Theta \to [0,1]$ is a *chance function* if $\mathrm{Ch}(\emptyset) = 0$, $\mathrm{Ch}(\Theta) = 1$, and $\mathrm{Ch}(A \cup B) = \mathrm{Ch}(A) + \mathrm{Ch}(B)$ when $A \cap B = \emptyset$. Chance density functions and chance functions are in one-to-one correspondence and carry the same information. If Ch is a chance function, then $p(a) = \mathrm{Ch}(\{a\})$ is a chance density function; conversely, if p is a chance density function, then $\mathrm{Ch}(A) = \sum_{[a \in A]} p(a)$ is a chance function.

If p is a chance density function and we define $m(\{a\}) = p(a)$ for all $a \in \Theta$ and make $m(A) = 0$ elsewhere, then m is a basic probability assignment and $\mathrm{Bel}(A) = \mathrm{Ch}(A)$ for all $A \in 2^\Theta$. Therefore, the class of Bayesian belief functions is a subset of the class of belief functions. Basic probability assignments are a generalization of chance density functions while belief functions assume the role of generalized chance functions.

The crucial observation is that a Bayesian belief function ties all of its probability masses to *single points* in Θ, hence there is no freedom of motion. This follows immediately from the definition of a chance density function and its correspondence to a basic probability assignment. In this case, $P^\star = \mathrm{Bel}$ because, with no freedom of motion, the total probability mass that can move into a set is the mass constrained to stay there.

What this means in practical terms is that the user of a Bayesian belief function must somehow divide his belief among the singleton propositions. In some instances, this is easy. It we believe that a fair die shows an even number, then it seems natural to divide that belief evenly into three parts. If we don't know or don't believe the die is fair, then we are stuck.

In other words, there is trouble representing what we actually know without being forced to overcommit when we are ignorant. With the theory described here there is no problem—just let $m(\textsc{even})$ measure the belief and the knowledge that is available. This is not to say that one should not use Bayesian statistics. In fact, if one has the necessary information, I know of no other proposed methodology that works as well. Nor are there any serious philosophical arguments against the use of Bayesian statistics. However, when our knowledge is not complete, as is often the case, the theory of Dempster and Shafer is an alternative to be considered.

2.4 The Combination of Evidence

The previous sections describe belief functions, the technique for representing evidence. Here, the theory's method of combining evidence is introduced. Let m_1 and m_2 be basic probability assignments on the same frame, Θ, and define $m = m_1 \oplus m_2$, their *orthogonal sum*, to be $m(\emptyset) = 0$ and

$$\begin{aligned} m(A) &= K \sum_{X \cap Y = A} m_1(X) \cdot m_2(Y) \\ K^{-1} &= 1 - \sum_{X \cap Y = \emptyset} m_1(X) \cdot m_2(Y) \;=\; \sum_{X \cap Y \neq \emptyset} m_1(X) \cdot m_2(Y), \end{aligned}$$

when $A \neq \emptyset$. The function m is a basic probability assignment if $K^{-1} \neq 0$; if $K^{-1} = 0$, then $m_1 \oplus m_2$ does not exist and m_1 and m_2 are said to be *totally* or *flatly contradictory.* The quantity $\log K = \mathrm{Con}(\mathrm{Bel}_1, \mathrm{Bel}_2)$ is called the *weight of conflict* between Bel_1 and Bel_2. This formulation is called *Dempster's rule of combination.*

It is easy to show that if m_1, m_2, and m_3 are combinable, then $m_1 \oplus m_2 = m_2 \oplus m_1$ and $(m_1 \oplus m_2) \oplus m_3 = m_1 \oplus (m_2 \oplus m_3)$. If v is the basic probability assignment such that $v(\Theta) = 1$ and $v(A) = 0$ when $A \neq \Theta$, then v is called the *vacuous belief function* and is the representation of total ignorance. The function, v, is the identity element for $\oplus$, i.e., $v \oplus m_1 = m_1$.

Figure 1 is a graphical interpretation of Dempster's rule of combination. Assume $m_1(A), m_1(B) \neq 0$ and $m_2(X), m_2(Y), m_2(Z) \neq 0$ and that m_1 and m_2 are 0 elsewhere. Then $m_1(A) + m_1(B) = 1$ and $m_2(X) + m_2(Y) + m_2(Z) = 1$. Therefore, the square in the figure has unit area since each side has unit length. The shaded rectangle has area $m_1(B) \cdot m_2(Y)$ and belief proportional to this measure is committed to $B \cap Y$. Thus, the probability number $m(B \cap Y)$ is proportional to the sum of the areas of all such rectangles committed to $B \cap Y$. The constant of proportionality, K, normalizes the result to compensate for the measure of belief committed to $\emptyset$. Thus, $K^{-1} = 0$ if and only if the combined belief functions invest no belief in intersecting sets; this is what is meant when we say belief functions are totally contradictory.

Using the graphical interpretation, it is straightforward to write down the formula for the orthogonal sum of more than two belief functions. Let $m = m_1 \oplus \cdots \oplus m_n$, then $m(\emptyset) = 0$ and

$$\begin{aligned} m(A) &= K \sum_{\cap A_i = A} \prod_{1 \leq i \leq n} m_i(A_i) \qquad (4) \\ K^{-1} &= 1 - \sum_{\cap A_i = \emptyset} \prod_{1 \leq i \leq n} m_i(A_i) \;=\; \sum_{\cap A_i \neq \emptyset} \prod_{1 \leq i \leq n} m_i(A_i) \end{aligned}$$

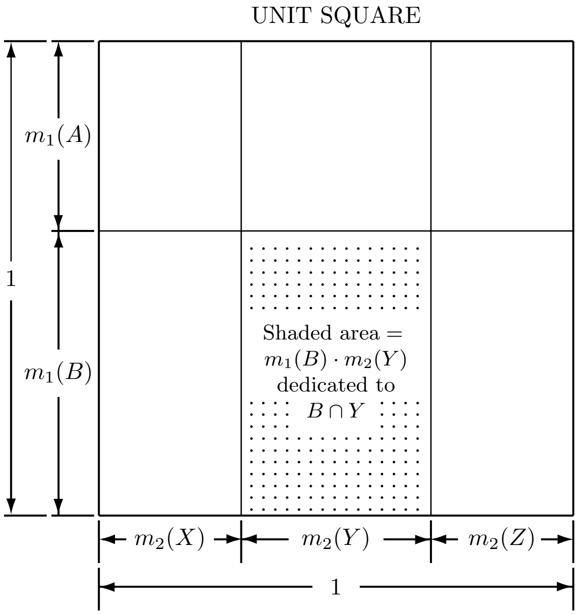


Fig. 1. Graphical representation of an orthogonal sum

when $A \neq \emptyset$. As above, the orthogonal sum is defined only if $K^{-1} \neq 0$ and the weight of conflict is $\log K$.

Since Bel, m, Q, Dou, and $P^\star$ are in one-to-one correspondence, the notation Bel = $\text{Bel}_1 \oplus \text{Bel}_2$, etc., is used in the obvious way. It is interesting to note that if $Q = Q_1 \oplus Q_2$, then $Q(A) = KQ_1(A)Q_2(A)$ for all $A \subseteq \Theta$ where $A \neq \emptyset$.

2.5 Comparison with Conditional Probabilities

In the Bayesian theory, the function $\text{Ch}(\cdot|B)$ is the *conditional chance* function, i.e., $\text{Ch}(A|B) = \text{Ch}(A \cap B)/\text{Ch}(B)$, is the chance that A is true given that B is true. $\text{Ch}(\cdot|B)$ is a chance function. A similar measure is available using Dempster's rule of combination.

Let $m_B(B) = 1$ and let m_B be 0 elsewhere. Then Bel_B, is a belief function that focuses all of our belief on B. Define $\text{Bel}(\cdot|B) = \text{Bel} \oplus \text{Bel}_B$. Then [16] shows that $P^\star(A|B) = P^\star(A \cap B)/P^\star(B)$; this has the same form as the Bayesians' rule of conditioning, but in general, $\text{Bel}(A|B) = (\text{Bel}(A \cup \neg B) - \text{Bel}(\neg B))/(1 - \text{Bel}(\neg B))$. On the other hand, if Bel is a Bayesian belief function, then $\text{Bel}(A|B) = \text{Bel}(A \cap B)/\text{Bel}(B)$.

Thus, Dempster's rule of combination mimics the Bayesians' rule of conditioning when applied to Bayesian belief functions. It should be noted, however, that the function Bel_B is *not* a Bayesian belief function unless $|B| = 1$.

2.6 Simple and Separable Support Functions

Certain kinds of belief functions are particularly well suited for the representation of actual evidence, among them are the classes of simple and separable support functions. If there exists an $F \subseteq \Theta$ such that $\mathrm{Bel}(A) = s \neq 0$ when $F \subseteq A$ and $A \neq \Theta$, $\mathrm{Bel}(\Theta) = 1$, and $\mathrm{Bel}(A) = 0$ when $F \not\subseteq A$, then Bel is a *simple support function*, F is called the *focus* of Bel, and s is called Bel's *degree of support.*

The vacuous belief function is a simple support function with focus Θ. If Bel is a simple support function with focus $F \neq \Theta$, then $m(F) = s$, $m(\Theta) = 1 - s$, and m is 0 elsewhere. Thus, a simple support function invests all of our committed belief on the disjunction represented by its focus, F, and all our uncommitted belief on Θ.

A separable support function is either a simple support function or the orthogonal sum of two or more simple support functions that can be combined. If it is assumed that simple support functions are used to represent the results of experiments, then the separable support functions are the possible results when the evidence from the several experiments is pooled together.

A particular case has occurred frequently. Let Bel_1 and Bel_2 be simple support functions with respective degrees of support s_1 and s_2, and the common focus, F. Let $\mathrm{Bel} = \mathrm{Bel}_1 \oplus \mathrm{Bel}_2$. Then $m(F) = 1 - (1 - s_1)(1 - s_2) = s_1 + s_2(1 - s_1) = s_2 + s_1(1 - s_2) = s_1 + s_2 - s_1 s_2$ and $m(\Theta) = (1 - s_1)(1 - s_2)$; m is 0 elsewhere.

The point of interest is that this formula appears as the rule of combination in mycin [17] and [11] as well as many other places. In fact, the earliest known development appears in the works of Jacob [2] circa 1713. For more than two and a half centuries, this formulation has had intuitive appeal to workers in a variety of fields trying to combine bodies of evidence pointing in the same direction. Why not use ordinary statistical methods? Because the simple support functions are not Bayesian belief functions unless $|F| = 1$.

We now turn to the problem of computational complexity.

3 The Computational Problem

Assume the result of an experiment—represented as the basic probability assignment, m—is available. Then, in general, the computation of $\mathrm{Bel}(A)$, $Q(A)$, $P^\star(A)$, or $\mathrm{Dou}(A)$ requires time exponential in $|\Theta|$. The reason[1] is the need to enumerate all subsets or supersets of A. Further, given any one of the functions, Bel, m, Q, $P^\star$, or Dou, computation of values of at least two of the others requires exponential time. If something is known about the structure of the belief function, then things may not be so bad. For example, with a simple support function, the computation time is no worse than $\mathrm{o}(|\Theta|)$.

[1] I have not proved this. However, if the formulae introduced in Sect. 2 are directly implemented, then the statement stands.

The complexity problem is exaggerated when belief functions are combined. Assume $\mathrm{Bel} = \mathrm{Bel}_1 \oplus \cdots \oplus \mathrm{Bel}_n$, and the Bel_i are represented by the basic probability assignments, m_i. Then in general, the computations of K, $\mathrm{Bel}(A)$, $m(A)$, $Q(A)$, $P^\star(A)$, and $\mathrm{Dou}(A)$ require exponential time. Once again, knowledge of the structure of the m_i may overcome the dilemma. For example, if a Bayesian belief function is combined with a simple support function, then the computation requires only linear time.

The next section describes a particularly useful structuring of the m_i. Following sections show that all the basic quantities of interest can be calculated in $\mathrm{o}(|\Theta|)$ time when this structure is used.

4 Structuring the Problem

Tonight you expect a special guest for dinner. You know it is important to play exactly the right music for her. How shall you choose from your large record and tape collection? It is impractical to go through all the albums one by one because time is short. First you try to remember what style she likes—was it jazz, classical, or pop? Recalling past conversations you find some evidence for and against each. Did she like vocals or was it instrumentals? Also, what are her preferences among strings, reeds, horns, and percussion instruments?

4.1 The Strategy

The problem solving strategy exemplified here is the well known technique of partitioning a large problem space in several independent ways, e.g., music style, vocalization, and instrumentation. Each partitioning is considered separately, then the evidence from each partitioning is combined to constrain the final decision. The strategy is powerful because each partitioning represents a smaller, more tractable problem.

There is a natural way to apply the plausible reasoning methodology introduced in Sect. 2 to the partitioning strategy. When this is done, an efficient computation is achieved. There are two computational components necessary to the strategy: the first collects and combines evidence within each partitioned space, while the second pools the evidence from among the several independent partitions.

In [16], the necessary theory for pooling evidence from the several partitions is developed using Dempster's rule of combination and the concept of refinings of compatible frames; in [1], computational methods are being developed for this activity. Below, a formulation for the representation of evidence within a single partitioning is described, then efficient methods are developed for combining this evidence.

4.2 Simple Evidence Functions

Let Θ be a partitioning comprised of n elements, i.e., $|\Theta| = n$; for example, if Θ is the set of possibilities that the dinner guest prefers jazz, classical, or pop music, then $n = 3$. Θ is a frame of discernment and, with no loss of generality, let $\Theta = \{i|1 \leq i \leq n\}$. For each $i \in \Theta$, there is a collection of basic probability assignments μ_{ij} that represents evidence in favor of the proposition i, and a collection, ν_{ij} that represents the evidence against i. The natural embodiment of this evidence is as simple support functions with the respective foci $\{i\}$ and $\neg\{i\}$.

Define $\mu_i(\{i\}) = 1 - \prod(1 - \mu_{ij}(\{i\}))$ and $\mu_i(\Theta) = 1 - \mu_i(\{i\})$. Then μ_i is a basic probability assignment and the orthogonal sum of the μ_{ij}. Thus, μ_i is the totality of the evidence in favor of i, and $f_i = \mu(\{i\})$ is the degree of support from this simple support function. Similarly, define $\nu_i(\neg\{i\}) = 1 - \prod(1 - \nu_{ij}(\neg\{i\}))$ and $\nu_i(\Theta) = 1 - \nu_i(\neg\{i\})$. Then $a_i = \nu_i(\neg\{i\})$ is the total weight of support against i. Note, $\neg\{i\} = \Theta - \{i\}$, i.e., set complementation is always relative to the fixed frame, Θ. Note also that j, in μ_{ij}, and ν_{ij}, runs through respectively the sets of experiments that confirm or deny the proposition i.

The combination of all the evidence directly for and against i is the separable support function, $e_i = \mu_i \oplus \nu_i$. The e_i formed in this manner are called the *simple evidence functions* and there are n of them, one for each $i \in \Theta$. The only basic probability numbers for e_i that are not identically zero are $p_i = e_i(\{i\}) = K_i \cdot f_i \cdot (1 - a_i)$, $c_i = e_i(\neg\{i\}) = K_i \cdot a_i \cdot (1 - f_i)$, and $r_i = e_i(\Theta) = K_i \cdot (1 - f_i) \cdot (1 - a_i)$, where $K_i = (1 - a_i f_i)^{-1}$. Thus, p_i is the measure of support pro i, c_i is the measure of support con i, and r_i is the measure of the residue, uncommitted belief given the body of evidence comprising μ_{ij} and ν_{ij}. Clearly, $p_i + c_i + r_i = 1$.

The goal of the rest of this paper is to find efficient methods to compute the quantities associated with the orthogonal sum of the n simple evidence functions. Though the simple evidence functions arise in a natural way when dealing with partitions, the results are not limited to this usage—whenever the evidence in our domain consists of simple support functions focused on singleton propositions and their negations, the methodology is applicable.

4.3 Some Simple Observations

In the development of computational methods below, several simple observations are used repeatedly and the quantity $d_i = 1 - p_i = c_i + r_i$ appears. The first thing to note is $K_i^{-1} = 0$ iff $a_i = f_i = 1$. Further, if $K^{-1} \neq 0$ and v is the vacuous belief function, then

$$
\begin{array}{ll}
p_i = 1 \text{ iff } f_i = 1 & c_i = 1 \text{ iff } a_i = 1 \\
p_i = 1 \Rightarrow c_i = r_i = 0 & c_i = 1 \Rightarrow p_i = r_i = 0 \\
f_i = 1 \text{ iff } \exists j\, \mu_{ij}(\{i\}) = 1 & a_1 = 1 \text{ iff } \exists j\, \nu_{ij}(\neg\{i\}) = 1
\end{array}
$$

$$\begin{array}{ll}
p_i = 0 \text{ iff } f_i = 0 \vee a_i = 1 & c_i = 0 \text{ iff } a_i = 0 \vee f_i = 1 \\
f_i = 0 \text{ iff } \forall j\, \mu_{ij} = v & a_i = 0 \text{ iff } \forall j\, \nu_{ij} = v \\
r_i = 1 \text{ iff } p_i = c_i = 0 & r_i = 0 \text{ iff } f_i = 1 \vee a_i = 1
\end{array}$$

5 Algorithms and Computations

The goal is to calculate quantities associated with $m = e_1 \oplus \cdots \oplus e_n$, where $n = |\Theta|$ and the e_i are the simple evidence functions defined in the previous section. All computations are achieved in $\mathrm{o}(n)$ time measured in arithmetic operations.

Figure 2 is a schematic of information flow in a mythical system. The μ_{ij} and ν_{ij} may be viewed as sensors, where a sensor is an instance of a knowledge source that transforms observations into internally represented evidence, i.e., belief functions. Each is initially v, the vacuous belief function. As time passes and events occur in the observed world, these sensors can update their state by increasing or decreasing their degree of support. The simple evidence function, e_i, recomputes its state, a_i and f_i, and changes the stored values of p_i, d_i, c_i, and r_i each time one of its sensors reports a change. From the definitions of μ_{ij}, ν_{ij}, and e_i it is evident that the effect of an update can be recorded in constant time. That is to say, the time is independent of both the ranges of j in μ_{ij} and ν_{ij} and of n.

A user asks questions about the current state of the evidence. One set of questions concerns the values of various measures associated with arbitrary

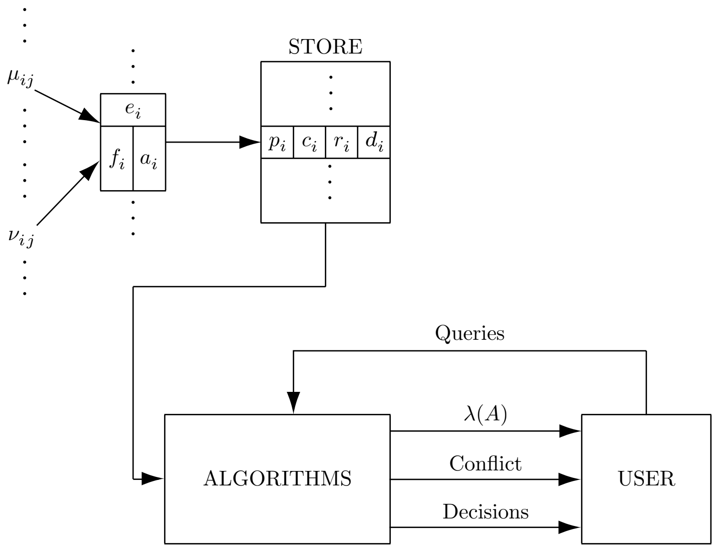


Fig. 2. Data flow model

$A \subseteq \Theta$. These questions take the form ‘what is the value of $\lambda(A)$?’, where λ is one of the functions Bel, m, Q, $P^\star$, or Dou. The other possible queries concern the general state of the inference process. Two examples are ‘what is the weight of conflict in the evidence?’ and ‘is there an A such that $m(A) = 1$; if so, what is A?’. The o(n) time computations described in this section and in Sect. 6 answer all these questions.

One more tiny detour is necessary before getting on with the business at hand: it is assumed that subsets of Θ are represented by a form with the computational nicety of bit-vectors as opposed to, say, unordered lists of elements. The computational aspects of this assumption are: (1) the set membership test takes constant time independent of n and the cardinality of the set; (2) the operators $\subseteq$, $\cap$, $\cup$, $=$, complementation with respect to Θ, null, and cardinality compute in o(n) time.

5.1 The Computation of K

From equation (4), $K^{-1} = \sum_{[\cap A_i \neq \emptyset]} \prod_{[1 \leq i \leq n]} e_i(A_i)$ and the weight of internal conflict among the e_i is $\log K$ by definition. Note that there may be conflict between the pairs of μ_i and ν_i that is not expressed because K is calculated from the point of view of the given e_i. Fortunately, the total weight of conflict is simply $\log[K \cdot \prod K_i]$; this quantity can be computed in o(n) time if K can be.

In order to calculate K, it is necessary to find the collections of A_i that satisfy $\cap A_i \neq \emptyset$ and $e_i(A_i) \neq 0$, i.e., those collections that contribute to the summation. If A_i is not $\{i\}$, $\neg\{i\}$, or Θ, then $e_i = 0$ identically from the definition of the simple evidence functions. Therefore, assume throughout that $A_i \in \{\{i\}\, \neg\{i\}\, \Theta\}$.

There are exactly two ways to select the A_i such that $\cap A_i \neq \emptyset$.

1. If $A_j = \{j\}$ for some j, and $A_i = \neg\{i\}$ or $A_i = \Theta$ for $i \neq j$, then $\cap A_i = \{j\} \neq \emptyset$. However, if two or more A_i are singletons, then the intersection is empty.
2. If none of the A_i are singletons, then the situation is as follows. Select any $S \subseteq \Theta$ and let $A_i = \Theta$ when $i \in S$ and $A_i = \neg\{i\}$ when $i \notin S$. Then $\cap A_i = S$. Therefore, when no A_i is a singleton, $\cap A_i \neq \emptyset$ unless $A_i = \neg\{i\}$ for all i.

Let J, K, L be predicates respectively asserting that exactly one A_i is a singleton, no A_i is a singleton, i.e., all $A_i \in \{\neg\{i\}\, \Theta\}$, and all $A_i = \neg\{i\}$. Then equation (4) can be written as

$$\begin{aligned} K^{-1} &= \sum_{\cap A_i \neq \emptyset} \prod_{1 \leq i \leq n} e_i(A_i) \\ &= \sum_J \prod_{1 \leq i \leq n} e_i(A_i) + \sum_K \prod_{1 \leq i \leq n} e_i(A_i) - \sum_L \prod_{1 \leq i \leq n} e_i(A_i). \end{aligned}$$

Now the transformation, below called transformation T,

$$\sum_{x_j \in S_j} \prod_{1 \le i \le n} f_i(x_i) = \prod_{1 \le i \le n} \sum_{x \in S_i} f_i(x) \tag{T}$$

can be applied to each of the three terms on the right; after some algebra, it follows that

$$K^{-1} = \sum_{1 \le q \le n} p_q \prod_{i \ne q} d_i + \prod_{1 \le i \le n} d_i - \prod_{1 \le i \le n} c_i, \tag{5}$$

where $p_i = e_i(\{i\})$, $c_i = e_i(\neg\{i\})$, and $d_i = e_i(\neg\{i\}) + e_i(\Theta)$ have been substituted. If $p_q = 1$ for some q, then $d_q = c_q = 0$ and $K^{-1} = \prod_{[i \ne q]} d_i$. On the other hand, if $p_i \ne 1$ for all i, then $d_i \ne 0$ for all i and equation (5) can be rewritten as

$$K^{-1} = \Big[\prod_{1 \le i \le n} d_i\Big]\Big[1 + \sum_{1 \le i \le n} p_i/d_i\Big] - \prod_{1 \le i \le n} c_i. \tag{6}$$

In either case, it is easy to see that the computation is achieved in o(n) time, as is the check for $p_i = 1$.

5.2 The Computation of m(A)

From equation (4), the basic probability numbers, $m(A)$ for the orthogonal sum of the simple evidence functions are

$$m(A) = K \sum_{\cap A_i = A} \prod_{1 \le i \le n} e_i(A_i),$$

for $A \ne \emptyset$ and by definition, $m(\emptyset) = 0$. Also, m can be expressed by

$$\begin{aligned} m(\emptyset) &= 0 \\ m(\{q\}) &= K\Big[p_q \prod_{i \ne q} d_i + r_q \prod_{i \ne q} c_i\Big] \\ M(A) &= K\Big[\prod_{i \in A} r_i\Big]\Big[\prod_{i \notin A} c_i\Big], \qquad \text{when } |A| \ge 2. \end{aligned} \tag{7}$$

It is easy to see that the calculation is achieved in o(n) time since $|A| + |\neg A| = n$.

Derivation of these formulae is straightforward. If $A = \cap A_i$, then $A \subseteq A_i$ for $1 \le i \le n$ and for all $j \notin A$, there is an A_i such that $j \notin A_i$. Consider the case in which A is a nonsingleton nonempty set; If $i \in A$, then $A_i = \Theta$—the only other possibilities are $\{i\}$ or $\neg\{i\}$, but neither contains A. If $i \notin A$, then both $A_i = \neg\{i\}$ and $A_i = \Theta$ are consistent with $A \subseteq A_i$. However, if $A_i = \Theta$ for some $i \notin A$, then $\cap A_i \supseteq A \cup \{i\} \ne A$. Therefore, the only choice is $A_i = \neg\{i\}$ when $i \notin A$ and $A_i = \Theta$ when $i \in A$. When it is noted that

$e_i(\Theta) = r_i$ and $e_i(\neg\{i\}) = c_i$ and, transformation T is applied, the formula for the nonsingleton case in equation (7) follows.

When $A = \{q\}$, there are two possibilities: $A_q = \Theta$ or $A_q = \{q\}$. If $A_q = \Theta$, then the previous argument for nonsingletons can be applied to justify the appearance of the term $r_q \prod_{[i \neq q]} c_i$. If $A_q = \{q\}$, then for each $i \neq q$ it is proper to select either $A_i = \Theta$ or $A_i = \neg\{i\}$ because, for both choices, $A \subseteq A_i$; actually, $\cap A_i = \{q\} = A$ because $A_q = A$. Using transformation T and noting that $e_q(\{q\}) = p_q$ and $d_i = c_i + r_i$ gives the term $p_q \prod_{[i \neq q]} d_i$ in the above and completes the derivation of equation (7).

5.3 The Computations of Bel(A), $P^\star$(A), and Dou(A)

Since $\mathrm{Dou}(A) = \mathrm{Bel}(\neg A)$ and $P^\star(A) = 1 - \mathrm{Dou}(A)$, the computation of $P^\star$ and Dou is $\mathrm{o}(n)$ if Bel can be computed in $\mathrm{o}(n)$ because complementation is an $\mathrm{o}(n)$ operation. Let Bel be the orthogonal sum of the n simple evidence functions. Then $\mathrm{Bel}(\emptyset) = 0$ by definition and for $A \neq \emptyset$,

$$\begin{aligned}\mathrm{Bel}(A) &= \sum_{B \subseteq A} m(B) = \sum_{\emptyset \neq B \subseteq A} K \sum_{\cap B_i = B} \prod_{1 \leq i \leq n} e_i(B_i) \\ &= K \sum_{\emptyset \neq \cap A_i \subseteq A} \prod_{1 \leq i \leq n} e_i(A_i).\end{aligned}$$

Bel is also expressed by

$$\mathrm{Bel}(A) = K\Big[\Big[\prod_{1 \leq i \leq n} d_i\Big]\Big[\sum_{i \in A} p_i/d_i\Big] + \Big[\prod_{i \notin A} c_i\Big]\Big[\prod_{i \in A} d_i\Big] - \prod_{1 \leq i \leq n} c_i\Big] \qquad (8)$$

when $d_i \neq 0$ for all i. If $d_q = 0$, then $p_q = 1$. Therefore, $m(\{q\}) = \mathrm{Bel}(\{q\}) = 1$. In all variations, $\mathrm{Bel}(A)$ can be calculated in $\mathrm{o}(n)$ time. Since the formula evaluates $\mathrm{Bel}(\emptyset)$ to 0, only the case of nonempty A needs to be argued.

The tactic is to find the collections of A_i satisfying $\emptyset \neq \cap A_i \subseteq A$ then apply transformation T. Recall that the only collections of A_i that satisfy $\emptyset \neq \cap A_i$ are those in which (1) exactly one A_i is a singleton or (2) no A_i is a singleton and at least one $A_i = \Theta$. To satisfy the current constraint, we must find the subcollections of these two that also satisfy $\cap A_i \subseteq A$.

If exactly one A_i is a singleton, say $A_q = \{q\}$, then $\cap A_i = \{q\}$. In order that $\cap A_i \subseteq A$ it is necessary and sufficient that $q \in A$. Thus, the contribution to $\mathrm{Bel}(A)$, when exactly one singleton A_i is permitted, is the sum of the contributions for all $i \in A$. A brief computation shows this to be $[\prod_{[1 \leq i \leq n]} d_i][\sum_{[i \in A]} p_i/d_i]$.

When no A_i is a singleton, it is clear that $A_i = \neg\{i\}$ for $i \notin A$; otherwise, $i \in A$ and $\cap A_i \not\subseteq A$. For $i \in A$, either $A_i = \neg\{i\}$ or $A_i = \Theta$ is permissible. The value of the contribution to Bel from this case is given by the term $[\prod_{[i \notin A]} c_i][\prod_{[i \in A]} d_i]$. Since at least one of the $A_i = \Theta$ is required, we must deduct for the case in which $A_i = \neg\{i\}$ for all i, and this explains the appearance of the term $-\prod_{[1 \leq i \leq n]} c_i$.

5.4 The Computation of $Q(A)$

The definition of the commonality function shows that $Q(\emptyset) = 1$ identically. For $A \neq \emptyset$

$$Q(A) = \sum_{A \subseteq B} m(B) = \sum_{A \subseteq B} K \sum_{\cap A_i = B} \prod_{1 \leq i \leq n} e_i(A_i) = K \sum_{A \subseteq \cap A_i} \prod_{1 \leq i \leq n} e_i(A_i).$$

Q can be expressed also by

$$\begin{aligned} Q(\emptyset) &= 1 \\ Q(\{q\}) &= K(p_q + r_q) \prod_{i \neq q} d_i \\ Q(A) &= K \Big[\prod_{i \in A} r_i\Big] \Big[\prod_{i \notin A} d_i\Big], \qquad \text{when } |A| \geq 2. \end{aligned}$$

In order that a collection, A_i, satisfy $A \subseteq \cap A_i$, it is necessary and sufficient that $A \subseteq A_i$ for all i. If $i \notin A$, then both $A_i = \neg\{i\}$ and $A_i = \Theta$ fill this requirement but $A_i = \{i\}$ fails. If $i \in A$, then clearly $A_i = \neg\{i\}$ fails and $A_i = \Theta$ works. Further, $A_i = \{i\}$ works iff $A = \{i\}$. It is now a simple matter to apply transformation T and generate the above result. It is evident that $Q(A)$ can be calculated in o(n) time.

6 Conflict and Decisiveness

In the previous section, a mythical system was introduced that gathered and pooled evidence from a collection of sensors. It was shown how queries such as 'what is the value of $\lambda(A)$?' could be answered efficiently, where A is an arbitrary subset of Θ and λ is one of Bel, m, Q, $P^\star$, or Dou. It is interesting to note that a sensor may change its value over time. The queries report values for the current state of the evidence. Thus, it is easy to imagine an implementation performing a monitoring task, for which better and more decisive data become available, as time passes, and decisions are reevaluated and updated on the bases of the most current evidence.

In this section, we examine more general queries about the combined evidence. These queries seek the subsets of Θ that optimize one of the measures. The sharpest question seeks the $A \subseteq \Theta$, if any, such that $m(A) = 1$. If such an A exists, it is said to be the *decision*. Vaguer notions of decision in terms of the other measures are examined too.

The first result is the necessary and sufficient conditions that the evidence be totally contradictory. Since the orthogonal sum of the evidence does not exist in this case, it is necessary to factor this out before the analysis of decisiveness can be realized. All queries discussed in this section can be answered in o(n) time.

6.1 Totally Contradictory Evidence

Assume there are two or more $p_i = 1$, say $p_a = p_b = 1$, where $a \neq b$. Then $d_j = c_j = r_j = 0$, for both $j = a$ and $j = b$. The formula for K is

$$K^{-1} = \sum_{1 \leq q \leq n} p_q \prod_{i \neq q} d_i + \prod_{1 \leq i \leq n} d_i - \prod_{1 \leq i \leq n} c_i,$$

and it is easy to see that $K^{-1} = 0$ under this assumption. Therefore, the evidence is in total conflict by definition.

Let $p_a = 1$ and $p_i \neq 1$ for $i \neq a$. Then $d_a = c_a = 0$, and $d_i \neq 0$ for $i \neq a$. Therefore. the above formula reduces to $K^{-1} = \prod_{[i \neq a]} d_i \neq 0$ and the evidence is not totally contradictory.

Now assume $p_i \neq 1$, hence $d_i \neq 0$, for all i. Can $K^{-1} = 0$? Since $d_i = c_i + r_i$, it follows that $\prod d_i - \prod c_i \geq 0$. If $K^{-1} = 0$, this difference must vanish. This can happen only if $r_i = 0$ for all i. Since $p_i \neq 0$, this entails $c_i = 1$ for all i. In this event the $p_i = 0$ and $K^{-1} = 0$.

Summary: The evidence is in total conflict iff either (1) there exists an $a \neq b$ such that both $p_a = p_b = 1$ or (2) $c_i = 1$ for all $i \in \Theta$.

6.2 Decisiveness in m

The evidence is *decisive* when $m(A) = 1$ for some $A \subseteq \Theta$ and A is called the *decision.* If the evidence is decisive and A is the decision, then $m(B) = 0$ when $B \neq A$ because the measure of m is 1. The evidence cannot be decisive if it is totally contradictory because the orthogonal sum does not exist, hence m is not defined. The determination of necessary and sufficient conditions that the evidence is decisive and the search for the decision is argued by cases.

If $p_q = 1$ for some $q \in \Theta$, then the evidence is totally contradictory if $p_i = 1$ for some $i \neq q$. Therefore, assume that $p_i \neq 1$ for $i \neq q$. From equation (7) it is easy to see $m(\{q\}) = K \prod_{[i \neq q]} d_i$ because $r_q = 0$. Further, it was shown directly above that $K^{-1} = \prod_{[i \neq q]} d_i$ under the same set of assumptions. Thus, $m(\{q\}) = 1$.

The other possibility is that $p_i \neq 1$, hence $d_i \neq 0$, for all $i \in \Theta$. Define $C = \{i | c_i = 1\}$, and note that if $|C| = n$, the evidence is totally contradictory. For $i \in C$, $p_i = r_i = 0$ and $d_i = 1$. If $|C| = n - 1$, then there is a w such that $\{w\} = \Theta - C$. Now $p_w \neq 1$ and $c_w \neq 1$ entails $r_w \neq 0$; therefore, from equation (7)

$$m(\{w\}) \ = \ K\Big[p_w \prod_{i \neq w} d_i + r_w \prod_{i \neq w} c_i\Big] \ = \ K[p_w + r_w] \ \neq \ 0.$$

If there is a decision in this case, it must be $\{w\}$. Direct substitution into equation (5) shows that, in this case, $K^{-1} = p_w + r_w$ and therefore, $m(\{w\}) = 1$.

Next, we consider the cases where $0 \leq |C| \leq n-2$ and therefore, $|\neg C| \geq 2$. Then, from equation (7)

$$m(\neg C) \;=\; K\Big[\prod_{i \notin C} r_i\Big]\Big[\prod_{i \in C} c_i\Big] \;=\; K\prod_{i \notin C} r_i \;\neq\; 0 \tag{9}$$

because $i \notin C$ iff $c_i \neq 1$ (and $p_i \neq 1$ for all $i \in \Theta$) has been assumed: hence, $r_i \neq 0$ for all $i \in \neg C$. Therefore, if the evidence is decisive, $m(\neg C) = 1$ is the only nonzero basic probability number. Can there be a $p_q \neq 0$? Obviously, $q \notin C$. The answer is no since $d_i \neq 0$, hence, $m(\{q\}) = K[p_q \prod_{[i \neq q]} d_i + r_q \prod_{[i \neq q]} c_i] \neq 0$, a contradiction. Thus, $p_i = 0$ for all $i \in \Theta$. From equation (5) it now follows that $K^{-1} = \prod_{[1 \leq i \leq n]} d_i - \prod_{[1 \leq i \leq n]} c_i$. Therefore, from (9), $\prod_{[i \notin C]} r_i = \prod_{[1 \leq i \leq n]} d_i - \prod_{[1 \leq i \leq n]} c_i$ if $m(\neg C) = 1$. Since $d_i = c_i = 1$ when $i \in C$, this can be rewritten as $\prod_{[i \notin C]} r_i = \prod_{[i \notin C]} d_i - \prod_{[i \notin C]} c_i$. But $d_i = c_i + r_i$. Therefore, this is possible exactly where $c_i = 0$ when $i \notin C$.

Summary: Assuming the evidence is not in total conflict, it is decisive iff either (1) exactly one $p_i = 1$; the decision is $\{i\}$. (2) There exists a w such that $c_w \neq 1$ and $c_i = 1$ when $i \neq w$; the decision is $\{w\}$. Or (3) there exists a $W \neq \emptyset$ such that $r_i = 1$ when $i \in W$ and $c_i = 1$ when $i \notin W$; the decision is W.

6.3 Decisiveness in Bel, $P^\star$, and Dou

If $\text{Bel}(A) = \text{Bel}(B) = 1$, then $\text{Bel}(A \cap B) = 1$ and it is always true that $\text{Bel}(\Theta) = 1$. The minimal A such that $\text{Bel}(A) = 1$ is called the *core* of Bel. If the evidence is decisive, i.e., $m(A) = 1$ for some $A \subseteq \Theta$, then clearly A is the core of Bel. Assume the evidence is not decisive, not totally contradictory, and $\text{Bel}(A) = 1$, then equations (8) and (6) can be smashed together and rearranged to show that

$$\sum_{q \notin A} p_q \prod_{i \neq q} d_i + \prod_{i \in A} d_i \Big[\prod_{i \notin A} d_i - \prod_{i \notin A} c_i\Big] = 0.$$

Since the evidence is not decisive, $d_i \neq 0$. Further, $d_i = c_i + r_i$ so that $r_i = 0$ when $i \notin A$; otherwise, the expression $\prod d_i - \prod c_i$ makes a nonzero contribution to the above. Similarly, $p_i = 0$ when $i \notin A$; hence $c_i = 1$ is necessary. Let $A = \{i | c_i \neq 1\}$, then substitution shows $\text{Bel}(A) = 1$ and A is clearly minimal.

Summary: The decision is the core when the evidence is decisive, otherwise $\{i | c_i \neq 1\}$ is the core.

$P^\star$ and Dou do not give us interesting concepts of decisiveness because $\text{Dou}(A) = \text{Bel}(\neg A) = 0$ would be the natural criterion. However this test is passed by any set in the complement of the core as well as others. Therefore, in general, no unique decision is found. A similar difficulty occurs in an attempt to form a concept of decisiveness in $P^\star$ because $P^\star(A) = 1 - \text{Dou}(A)$.

6.4 Decisiveness in Q

Since $Q(\emptyset) = 1$ and $Q(A) \leq Q(B)$ when $B \subseteq A$, it is reasonable to ask for the maximal N such that $Q(N) = 1$. This set, N, is called the *nucleus* of Bel. If

$m(A) = 1$, then the decision, A, is clearly the nucleus. If $i \in N$, then $i \in A$ for all $m(A) \neq 0$. Further, $Q(\{i\}) = 1$ iff i is an element of the nucleus.

Assume that the simple evidence functions are not totally contradictory and there is no decision. Then $d_i \neq 0$ and there is no w such that $c_i = 1$ whenever $i \neq w$. The necessary and sufficient conditions, then, that $Q(\{z\}) = 1$, and hence $z \in N$ are (1) $p_i = 0$ if $i \neq z$ and (2) $c_z = 0$. To wit,

$$\begin{aligned}
Q(\{z\}) &= 1 \\
K(p_z + r_z)\prod_{i \neq z} d_i &= 1 \\
(p_z + r_z)\prod_{i \neq z} d_i &= K^{-1} \\
(p_z + r_z)\prod_{i \neq z} d_i &= \sum_{1 \leq q \leq n} p_q \prod_{i \neq q} d_i + \prod_{1 \leq i \leq n} d_i - \prod_{1 \leq i \leq n} c_i \\
\sum_{q \neq z} p_q \prod_{i \neq q} d_i + (d_z - r_z)\prod_{i \neq z} d_i - \prod_{1 \leq i \leq n} c_i &= 0 \\
\sum_{q \neq z} p_q \prod_{i \neq q} d_i + c_z \prod_{i \neq z} d_i - \prod_{1 \leq i \leq n} c_i &= 0 \\
\sum_{q \neq z} p_q \prod_{i \neq q} d_i + c_z\Big(\prod_{i \neq z} d_i - \prod_{i \neq z} c_i\Big) &= 0
\end{aligned}$$

Since $d_i \neq 0$, it follows that $p_q = 0$ for $q \neq z$, else the first term makes a nonzero contribution. Since $d_i = c_i + r_i$, the quantity, $\prod d_i - \prod c_i$, can vanish only if $r_i = 0$ when $i \neq z$. However, this and $p_i \neq 1$ because there is no decision, entails $c_i = 1$ when $i \neq z$. Therefore, either $\{z\}$ is the decision or the evidence is contradictory. Thus, $c_z = 0$ so that the second term of the last equation vanishes. Since the steps above are reversible, these are sufficient conditions too.

Summary: If A is the decision, then A is the nucleus. If two or more $p_i \neq 0$, then the nucleus is $\emptyset$. If $p_z \neq 0$, $c_z = 0$, and $p_i = 0$ when $i \neq z$, then $\{z\}$ is the nucleus. If $p_i = 0$ for all i, then $\{i | c_i = 0\}$ is the nucleus. Clearly, this construction can be carried out in $\mathrm{o}(n)$ time.

6.5 Discussion

It has been noted that $p_i = 1$ or $c_i = 1$ if and only it there is a j such that respectively $\mu_{ij}(\{i\}) = 1$ or $\nu_{ij}(\neg\{i\}) = 1$, i.e., if and only if the result of some experiment is decisive within its scope. The above analyses show the effects occurring when $p_i = 1$ or $c_i = 1$; subsets of possibilities are irrevocably lost—most or all the nondecisive evidence is completely suppressed—or the evidence becomes totally contradictory.

Any implementation of this theory should keep careful tabs on those conditions leading to conflict and/or decisiveness. In fact, any decisive experiment (a degree of support of 1) should be viewed as based upon evidence so conclusive that no further information can change one's view. A value of 1 in this theory is indeed a strong statement.

7 Conclusion

Dempster and Shafer's theory of plausible inference provides a natural and powerful methodology for the representation and combination of evidence. I think it has a proper home in knowledge-based expert systems because of the need for a technique to represent weights of evidence and the need for a uniform method with which to reason. This theory provides both. Standard statistical methods do not perform as well in domains where prior probabilities of the necessary exactness are hard to come by, or where ignorance of the domain model itself is the case. One should not minimize these problems even with the proposed methodology. It is hoped that with the ability to directly express ignorance and uncertainty, the resulting model will not be so brittle.

However, more work needs to be done with this theory before it is on a solid foundation. Several problems remain as obvious topics for future research. Perhaps the most pressing is that no effective decision making procedure is available. The Bayesian approach masks the problem when priors are selected. Mechanical operations are employed from gathering evidence through the customary expected-value analysis. But our ignorance remains hidden in the priors.

The Dempster-Shafer theory goes about things differently—ignorance and uncertainty are directly represented in belief functions and remain through the combination process. When it is time to make a decision, should the estimate provided by Bel or the one provided by $P^\star$ be used? Perhaps something in between. But what? No one has a good answer to this question.

Thus, the difference between the theories is that the Bayesian approach suppresses ignorance up front while the other must deal with it after the evidence is in. This suggests one benefit of the Dempster-Shafer approach: surely, it must be right to let the evidence narrow down the possibilities, first, then apply some ad hoc method afterward.

Another problem, not peculiar to this theory, is the issue of independence. The mathematical model assumes that belief functions combined by Dempster's rule are based upon independent evidence, hence the name orthogonal sum. When this is not so, the method loses its feeling of inevitability. Also, the elements of the frame of discernment, Θ, are assumed to be exclusive propositions. However, this is not always an easy constraint to obey. For example, in the MYCIN application, it seems natural to make the frame the set of possible infections but the patient can have multiple infections. Enlarging the frame to handle all subsets of the set of infections increases the difficulty in obtaining rules and in their application; the cardinality of the frame grows from $|\Theta|$ to $2^{|\Theta|}$.

One more problem that deserves attention is computational efficiency. Above it is shown that, with a certain set of assumptions, it is possible to calculate efficiently. However, these assumptions are not valid in all or even most domains. A thorough investigation into more generous assumptions seems indicated so that more systems can employ a principled reasoning mechanism.

The computational theory as presented here has been implemented in SIMULA. Listings are available by writing directly to the author.

References

1. J. A. Barnett, Computational Methods for a Mathematical Theory of Evidence: Part II. Forthcoming.
2. Jacob, *Ars Conjectandi*, 1713.
3. A. P. Dempster, "On direct probabilities," *J. Roy. Statist. Soc. Ser.* B 25, 1963, 102–107.
4. A. P. Dempster, "New methods for reasoning toward posterior distributions based on sample data," *Ann. Math. Statis.* 37, 1967, 355–374.
5. A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Ann. Math. Statis.* 38, 1967, 325–339.
6. A. P. Dempster, "Upper and lower probability inferences based on a sample from a finite univariant population," *Biometrika* 54, 1967, 515–528.
7. A. P. Dempster, "Upper and lower probabilities generated by a random closed interval," *Ann. Math. Statis.* 39, (3), 1968, 957–966.
8. A. P. Dempster, "A generalization of Bayesian inference," *J. Roy. Statis. Soc. Ser. B* 30, 1968, 205–247.
9. A. P. Dempster, "Upper and lower probability inferences for families of hypotheses with monotone density ratios," *Ann. Math. Statis.* 40, 1969, 953–969.
10. R. O. Duda, P. E. Hart, and N. J. Nilsson, *Subjective Bayesian methods for rule-based inference systems.* Stanford Research Institute, Technical Report 124, January 1976.
11. L. Friedman, Extended plausible inference. These proceedings.
12. T. Garvey, J. Lowrance, M. Fischler, An inference technique for integrating knowledge from disparate sources. These proceedings.
13. F. Hayes-Roth and V. R. Lesser, "Focus of Attention in the Hearsay–II Speech-Understanding System," in *IJCAI77*, pp. 27–35, Cambridge, MA, 1977.
14. H. E. Pople. Jr.,"The formation of composite hypotheses in diagnostic problem solving: an exercise in synthetic reasoning," in *Proc. Fifth International Joint Conference on Artificial Intelligence*, pp. 1030–1037, Dept. of Computer Science, Carnegie-Mellon Univ., Pittsburgh, Pa., 1977.
15. G. Shafer, "A theory of statistical evidence," in W. L. Harper and C. A. Hooker (eds.), *Foundations and Philosophy of Statistical Theories in Physical Sciences*, Reidel, 1975.
16. G. Shafer, *A Mathematical Theory Of Evidence*, Princeton University Press, Princeton, New Jersey, 1976.
17. E. H. Shortliffe, *Artificial Intelligence Series*, Volume 2: *Computer-Based Medical Consultations: MYCIN*, American Elsevier, Inc., N. Y., chapter IV, 1976.
18. L. A. Zadeh, "Fuzzy sets," *Information and Control* 8, 1965, 338–353.

9

Constructive Probability*

Glenn Shafer

In a series of papers published in the 1960's, A. P. Dempster developed a generalization of the Bayesian theory of statistical inference. In *A Mathematical Theory of Evidence*, published in 1976, I advocated extending Dempster's work to a general theory of probability judgement. The central idea of this new general theory is that we might decompose our evidence into intuitively independent components, make probability judgements based on each component, and then extend, adapt, and combine these judgements using formal rules. In this way we might be able to construct numerical degrees of belief based on total evidence that is too complicated or confusing to deal with holistically. The systems of numerical degrees of belief that the theory helps us construct are called *belief functions*. Belief functions have a certain structure, but they are not, in general, additive like Bayesian probability distributions: a belief function Bel may assign a proposition A and its negation $\overline{A}$ degrees of belief $Bel(A)$ and $Bel(\overline{A})$ that add to less than one.

The theory of belief functions should be sharply distinguished from the ideas on "upper and lower probabilities" that have been developed by I. J. Good [11], C. A. B. Smith [28], and, more recently, Peter Williams [30, 31]. It is true that the theory's degrees of belief $Bel(A)$ have some properties in common with these authors' lower probabilities $P_*(A)$. And it is also true that Dempster, in his writing, used the vocabulary of upper and lower probabilities. But the conceptual structure of the theory of belief functions is quite different from the structure underlying Good, Smith, and Williams' work.

Since its publication, *A Mathematical Theory of Evidence* has been reviewed or discussed by several authors, including Persi Diaconis [4], Terry

* From *Synthese* 48 (1981) 1–60, © 1981 by D. Reidel Publishing Co., Dordrecht, Holland, and Boston, U.S.A. The author has corrected some typographical errors and updated some references. The original research was partially supported by grant MCS 78-01887 from the National Science Foundation. The basic ideas of Sect. 1 were formulated in discussions with Amos Tversky. The author also profited from discussions with Zeno Swijtink and correspondence with Peter Williams.

Fine [5], Isaac Levi [16], Dennis Lindley [17], Teddy Seidenfeld [20], and Peter Williams [32]. Most of these critics, being themselves dissatisfied with the Bayesian theory, have welcomed the new theory. But they have been troubled by the absence of a behavioral interpretation for the theory. The Bayesian theory can appeal to its "betting interpretation" to explain what its degrees of belief mean and to justify its rules for these degrees of belief. No such interpretation has been supplied for the theory of belief functions. So what do its degrees of belief mean? And why should we accept the theory's rules for these degrees of belief? Why, in particular, should we prefer these rules to the rules suggested by Good, Smith, and Williams?

In this paper, I argue that a constructive theory of probability judgment need not rely for its meaning and justification on any behavioral interpretation. My argument is based on an understanding of constructive probability judgment developed in recent unpublished work by Amos Tversky and myself. According to this understanding, numerical probability judgment amounts to comparing one's evidence to a scale of canonical examples, and a constructive theory of probability judgment must supply both the scale of canonical examples and methods of breaking the task of comparison down into simpler judgments. As I explain in Sect. 1 below, the Bayesian theory, the theory of belief functions, and a theory of lower probability functions can all be developed in this framework. All three of these constructive theories use the idea of chance in their scale of canonical examples. The theory of belief functions uses examples where the meaning of a message depends on chance, while the other two theories use examples where the truth is generated by chance.

In the course of the paper I give particular attention to Peter Williams' review of *A Mathematical Theory of Evidence*. Williams' writing is exceptionally lucid, and he is exceptionally explicit in relating his criticisms of the theory of belief functions to the betting interpretation of probability.

Williams treats both lower probabilities and Bayesian (i.e., additive) probabilities as betting rates. And he hints that his intuitions about lower probabilities are inherent in the very idea of betting. One of the purposes of this paper is to show that this is not so. The theory of belief functions is as consistent with the use of probability judgments as betting rates as the theory of lower probabilities Williams favors. It is especially important to recognize that one cannot choose between the different rules of conditioning used by belief functions and by Williams' theory (see Sect. 3 below) on the basis of the idea of betting alone.

1 The Meaning of Probability

Williams begins his review of *A Mathematical Theory of Evidence* with two questions: "(i) What is meant by 'degree of belief' and how might an individual determine his degrees of belief in a particular case? (ii) For what reasons are degrees of belief required to satisfy the conditions imposed?"

On a practical level, making a probability judgment means assessing the strength and significance of one's evidence by fitting it into a scale of canonical examples. And the probability judgment or "degree of belief" itself means that we have made the comparison—perhaps with the aid of some theory—and found our evidence to match a certain example on the scale best. Thus the meaning of a degree of belief depends on the scale we use and, more generally, the theory we use in arriving at it.

To make numerical probability judgments we need, of course, a numerical scale, and the obvious approach to constructing such a scale is to use examples involving chance. There is, however, more than one way of using the idea of chance to construct a scale of examples, and different ways correspond to different theories of probability judgment. It will be helpful, before going into Williams' questions more fully, to compare three such theories—the Bayesian theory, the theory of belief functions, and a theory of lower probabilities.

1.1 The Bayesian Theory

In the classical picture of chance, we imagine a game that can be played repeatedly and for which we know the chances. These chances are long-run frequencies, they can be thought of as propensities, and they also define fair betting rates—rates at which a bettor would break even in the long run. Since they are known and there is no other evidence, these chances give a measure of how much reason we have to believe that one or another of the game's outcomes will occur on a particular occasion. So we can call them numerical degrees of belief. If we imagine a number of different games, with different chances, then we have a scale of numerical degrees of belief.

The Bayesian theory uses this scale in a straightforward way. The Bayesian's task is to compare his problem to a scale of examples in which the truth is generated according to known chances and to decide which of these examples is most like his problem. And so when he makes the probability judgment $P(A) = p$, say, he is saying that his evidence provides support for A comparable to what would be provided by knowledge that the truth is generated by a chance setup that produces a result in A exactly p of the time. He is not saying that his evidence is just like such knowledge in all respects, nor that the truth is in fact a result of chance. But he is measuring the strength of his evidence by comparing it to a scale of chance setups.

How can the Bayesian accomplish his task? How can he make his scale of chances and the affinity of his evidence to this scale vivid enough to his imagination that he can meaningfully locate the evidence on the scale? This question does not, I believe, have a simple general answer. In any particular case the Bayesian must struggle to find ways of understanding his evidence that facilitate its comparison to the scale of chances. Perhaps he can understand his evidence in terms of a causal model and assess numerically the propensity of the model to produce various outcomes. Perhaps he can discern relevant frequencies in his evidence. And perhaps he can make enough well-founded judgments of these sorts to enable him to construct an overall probability distribution that seems well-founded to him. Or perhaps he cannot. There is nothing in the Bayesian theory that can guarantee its success.

The probability distributions of the Bayesian theory have, of course, exactly the same structure as chance distributions: a function P defined for all subsets of a finite set Θ (*the frame of discernment*) is a *Bayesian* (or *additive*) probability distribution if there exist non-negative numbers $p(\theta)$ for the elements θ of Θ such that

$$P(A) = \sum_{\theta \in A} p(\theta) \tag{1}$$

for all $A \subset \Theta$. (It is also required that $\sum_{\theta \in \Theta} p(\theta) = 1$.) In words: the degree of belief $P(A)$ that the truth lies in A is the sum over the elements θ of A of the degrees of belief $p(\theta)$ that the truth is θ.

1.2 The Theory of Belief Functions

A function *Bel* defined for all subsets of a frame Θ is called *a belief function* if it is of the form

$$Bel(A) = \sum_{B \subset A} m(B), \tag{2}$$

where $m(B)$ are non-negative numbers satisfying $m(\phi) = 0$ and $\sum_{B \subset \Theta} m(B) = 1$. Every Bayesian probability distribution is a belief function. (The m-values for a Bayesian probability distribution P are obtained by setting $m(\{\theta\}) = p(\theta)$ and $m(B) = 0$ for all B that contain more than one element.) But not every belief function is a Bayesian probability distribution.

The theory of belief functions is based on a way of comparing our evidence to the scale of chances that is quite different from that of the Bayesian theory. Instead of comparing our evidence to a scale of examples where the truth is generated according to known chances, we compare it to a scale of examples where the reliability and meaning of a message depends on known chances.

Here is a way to develop the scale of examples needed for belief functions. Suppose someone chooses a code at random from a list of codes, uses the chosen code to encode a message, and then sends us the result. We know the list of codes and the chance of each code being chosen—say the list is $c_1, \ldots, c_n$, and the chance of c_i being chosen is p_i. We decode the encoded message using each of the codes and find that this always produces a message of the form "the truth is in A" for some non-empty subset A of Θ. Let A_i denote the subset we get when we decode using c_i, and set

$$m(A) = \sum \{p_i \mid 1 \leq i \leq n; A_i = A\}$$

for each $A \subset \Theta$. Then $m(A)$ is, in a certain sense, the total chance that the true message was A.[1] And $Bel(A)$, given by (2), is the total chance that the

[1] This is not to say that we are dealing with a random mechanism that produces the message A with chance $m(A)$. It is just that $m(A)$ is the sum of the chances for those codes that decode our encoded message to A.

true message implies A. If the true message is infallible and the coded message is our only evidence, then we will want to call $Bel(A)$ our degree of belief that the truth lies in A.

We can tell this story with whatever values of the $m(A)$ we please, and so it provides us a canonical example corresponding to every possible belief function Bel. Of course we will seldom or never encounter in practice a situation in which our evidence really does consist of a coded message and all the assumptions of the canonical example are satisfied. But it is also rare that our evidence amounts to knowledge of a chance distribution according to which the truth has been or will be generated. In both cases the canonical examples are meant not as realistic examples but as standards for comparison.

Our task, when we assess evidence using belief functions, is to choose values of $m(A)$ that make the canonical "coded-message" example most like that evidence. But how do we do this? In complicated problems it is absurd, surely, to suppose that we can simply look at our evidence holistically and write down the best values for the $m(A)$. So we need a theory—a set of tools for constructing belief functions from simpler, more elementary judgments. *A Mathematical Theory of Evidence* suggests a number of such tools: assessment using simple support functions, assessment using consonance, discounting, minimal extension, and Dempster's rule of combination. All these tools are readily intelligible in terms of the canonical examples.

Dempster's rule of combination is the most important single tool of the theory. This rule tells us how to combine a belief function Bel_1 (with m–values $m_1(A)$, say) representing one body of evidence with a belief function Bel_2 (with m-values $m_2(A)$) representing an unrelated body of evidence so as to obtain a belief function Bel (with m-values $m(A)$) representing the pooled evidence. The idea underlying the rule is that the unrelatedness of the two bodies of evidence makes pooling them like combining two stochastically independent randomly coded messages. We should, that is to say, combine the canonical examples corresponding to the two bodies of evidence by supposing that the two random choices of codes are stochastically independent. It is easy to see how this leads to a rule for obtaining the $m(C)$ from the $m_1(A)$

Let us denote by C the set of codes that decode our encoded message to A. If we had not yet seen the encoded message, it would certainly be natural to adopt $m(A)$ as our degree of belief that the code used is in C. The suggestion here is that it is still natural to do so in the situation where we have seen the encoded message and thus know that the code used being in C is equivalent to A being the true message.

A similar tack is often taken by non-Bayesian statisticians when they make probability judgments based on probability sampling or on randomization. Here, as in those cases, one might refuse to adopt the suggested degrees of belief and adopt instead a parametric model. In this case the model would have the true message as its parameter and the encoded message as its observable given each value of the parameter. In the absence of other evidence about the true message, this model does not seem very useful. (Cf. Kempthorne, [15].)

and the $m_2(B)$. Denote by $c_1, \ldots, c_n$ and by $p_1, \ldots, p_n$ the codes and their chances in the case of the first message, and by $c'_1, \ldots, c'_m$ and $p'_1, \ldots, p'_m$ the codes and their chances in the case of the second. Then independence means that there is a chance $p_i p'_j$ that the pair $\left(c_i, c'_j\right)$ of codes will be chosen. But notice that decoding may now tell us something. If the message A_i we get by decoding the first message with c_i contradicts the message B_j we get by decoding the second message with c'_j (i.e., if $A_i \cap B_j = \phi$), then we know that $\left(c_i, c'_j\right)$ could not be the pair of codes actually used. So we must condition the chance distribution, eliminating such pairs and multiplying the chances for the others by K, where

$$\begin{aligned} K^{-1} &= \sum \{p_i p'_j \mid 1 \leq i \leq n; 1 \leq j \leq m; A_i \cap B_j \neq \phi\} \\ &= \sum \{m_1(A) m_2(B) \mid A \subset \Theta; B \subset \Theta; A \cap B \neq \phi\}. \end{aligned}$$

Notice also that if the first message is A and the second message is B, then the overall message is $A \cap B$. Thus the total chance of the overall message being C is

$$\begin{aligned} m\,(C) &= K \sum \{p_i p'_j \mid 1 \leq i \leq n; 1 \leq j \leq m; A_i \cap B_j = C\} \\ &= K \sum \{m_1\,(A)\, m_2\,(B) \mid A \subset \Theta; B \subset \Theta; A \cap B = C\}. \end{aligned} \tag{3}$$

Formula (3) is Dempster's rule.

The availability of Dempster's rule opens the possibility that we might construct a belief function based on complicated evidence by decomposing the evidence, breaking it down into small unrelated items whose message is relatively clear. The most convenient case, perhaps, is when each small item points clearly and unambiguously to a single subset of Θ. In this case the assessment of each item means the determination of a simple support function.

A *simple support function* focused on a subset A_0 of Θ and awarding it degree of support s is a belief function with m-values $m(A_0) = S$, $m(\Theta) = 1-s$ and $m(A) = 0$ for all other $A \subset \Theta$. This corresponds to a coded message which means A_0 with chance s and means Θ (i.e., means nothing at all) with chance $1 - s$. The values of the belief function are

$$Bel(A) = \begin{cases} 0 & \text{if } A_0 \not\subset A \\ s & \text{if } A_0 \subset A \neq \Theta \\ 1 & \text{if } A = \Theta. \end{cases}$$

In words: we have no positive beliefs beyond those implied by the degree of support s for A_0. Simple support functions are appropriate when the message of an argument or an item of evidence is clear and unambiguous, but its reliability must be assessed. The chance s corresponds, in such a case, to an assessment of that reliability. It is our assessment, so to speak, of the chance that the argument is sound.

The idea of the chance that an argument is sound (as opposed to the Bayesian idea of the chance that an assertion is true) is illustrated by the following example, which is essentially due to J. H. Lambert (see Shafer [22]) and which could be used to provide an alternative scale of canonical examples for simple support functions. Suppose we know all α's are β's, and we are told, by a randomizing device that tells the truth with chance s and lies with chance $1 - s$, that γ is an α. If the device told the truth (chance s), then we have a syllogism:

$$\begin{array}{c} \text{All } \alpha\text{'s are } \beta\text{'s.} \\ \underline{\quad \gamma \text{ is an } \alpha. \quad} \\ \gamma \text{ is a } \beta. \end{array}$$

If the device lied (chance $1 - s$), then we have nothing, for when the minor premise in the syllogism Barbara is negated, there is no conclusion:

$$\begin{array}{c} \text{All } \alpha\text{'s are } \beta\text{'s.} \\ \underline{\quad \gamma \text{ is not an } \alpha. \quad} \\ \text{Maybe } \gamma \text{ is a } \beta\text{; maybe not.} \end{array}$$

So the argument for the proposition "γ is a β" is sound with chance s and unsound with chance $1 - s$. As evidence, it amounts to the same thing as a message that asserts this proposition with chance s and says nothing with chance $1 - s$.

There is no guarantee that a satisfactory analysis of one's evidence will be achieved using belief functions, just as there is no guarantee of success with the Bayesian theory. I do believe, however, that the greater flexibility of belief functions will often be valuable. In many cases our deliberation needs to be directed towards the structure and reliability of the evidence rather than towards the nature of the process by which the truth is generated, and this means that a random model for the evidence may fit our needs better than a random model for the truth.

1.3 Lower Probabilities

Suppose we know a certain process is governed by chance, but instead of knowing precisely the chance law P governing it, we know only that P is in a class $\mathcal{P}$ of chance laws. Denote by Θ the set of possible outcomes for the process. Then we might set our degree of belief that the outcome of a given trial will be in a subset A of Θ equal to

$$P_*(A) = \inf \{P(A) | P \in \mathcal{P}\} . \tag{4}$$

This seems natural because we know the chance of A is at least $P_*(A)$. And so, in particular, we can expect to at least break even in the long run if we offer to bet (with others who have no more knowledge than we) on A at the odds $P_*(A) : 1 - P_*(A)$.

By varying the class $\mathcal{P}$ in this story we obtain a scale of examples. Perhaps we can construct a theory of probability judgment—a “theory of lower probabilities”—using this scale as the standard to which to compare our evidence. It will rarely if ever happen, of course, that our evidence really consists of knowledge that the truth is generated by chance and the chance law is in a class $\mathcal{P}$. But we have said the same thing about the canonical examples underlying the Bayesian theory and the theory of belief functions.

But what are the elements of this theory of lower probabilities? What tools do we have for locating our evidence on its scale of canonical examples? How, that is to say, do we break the task of constructing the class $\mathcal{P}$ down into simple judgments?

Here is an idea. Suppose we assess our evidence by making judgments of the form “our evidence is like knowing that the truth is generated by chance and that the chances have such-and-such a property.” Since there are many properties of chance distributions, this formulation permits a wide variety of judgments. We may say that our evidence is like knowing that the chance of A is greater than the chance of B, or like knowing that the conditional chance of A given C is greater than that of B given C, or like knowing that the mathematical expectation of some function of the truth is between certain bounds, etc. Our theory will ask us to make as many of these judgments as we think necessary to capture the message of the evidence, and $\mathcal{P}$ will consist of all the distributions that have all the properties we have specified.

Notice that this idea does not involve the decomposition of evidence. The task of constructing $\mathcal{P}$ is broken down into simple judgments by distinguishing different questions, not by distinguishing different items of evidence bearing on these questions. All the judgments are supposed to be based on the total evidence.

A class $\mathcal{P}$ of chance distributions determines, of course, more than the lower probabilities (4). It also determines *lower conditional probabilities*

$$P_* \left(A|B\right) = \inf \left\{P \left(A|B\right) |P \in \mathcal{P}; P \left(B\right) > 0\right\}, \tag{5}$$

which are defined whenever $P(B) > 0$ for some $P \in \mathcal{P}$,[2] and *lower expectations*

$$E_* \left(X\right) = \inf \left\{E_P \left(X\right) |P \in \mathcal{P}\right\},$$

which are defined (in the case where Θ is finite) for every real-valued function X on Θ. Since a lower unconditional probability is a special case of a lower conditional probability ($P_* \left(A\right) = P_* \left(A|\Theta\right)$) and a lower conditional probability can be determined from knowledge of lower expectations ($P_* \left(A|B\right) = p$ if

[2] De Finetti [8] assumes that $P(A|B)$ is defined for an additive probability distribution even if $P(B) = 0$, and Williams [30] accordingly supposes that $P_*(A|B)$ is always defined. But it is not necessary to explore these subtleties in the present discussion.

$E_*(X) = 0$, where $X(\theta) = 1 - p$ if $\theta \in A \cap B$, $-p$ if $\theta \in \overline{A} \cap B$, and 0 if $\theta \in \overline{B}$), we obtain more information about $\mathcal{P}$ as we pass from lower probabilities to lower conditional probabilities to lower expectations.

Example 1. Here are two classes $\mathcal{P}_1$ and $\mathcal{P}_2$ that have the same lower unconditional probabilities but can be distinguished by their lower conditional probabilities. Set $\Theta = \{a, b, c\}$, $\mathcal{P}_1 = \left\{P | P(\{a, b\}) \geq \frac{1}{2}\right\}$, and $\mathcal{P}_2 = \{P | P(\{b\} \mid \{b, c\}) \geq \frac{1}{2}\ \}$. Then $P_{*1}(A) = P_{*2}(A)$ for all $A \subset \Theta$. But $P_{*1}(\{b\} \mid \{b, c\}) = 0$, while $P_{*2}(\{b\} \mid \{b, c\}) = \frac{1}{2}$. (2) Here are two classes that have the same lower conditional probabilities but can be distinguished by other lower expectations. Set $\Theta = \{-2, -1, 1, 2\}$, set $\mathcal{P}_1 = \{P | E_P \geq 0\}$, where E_P denotes the mean of the distribution P, and set $\mathcal{P}_2 = \mathcal{P}_1 \cup \{P_2\}$, where P_2 is the distribution that puts mass $\frac{1}{2}$ on -2, $\frac{1}{3}$ on 1, and $\frac{1}{6}$ on 2. Then $P_{*1}(A|B) = P_{*2}(A|B)$ for all A and B, but the lower expectations of the identity function $X(\theta) = \theta$ are $E_{*1}(X) = 0$ and $E_{*2}(X) = -\frac{1}{3}$. (3) Here are two distinct classes that cannot be distinguished by their lower expectations. Set $\Theta = \{a, b\}$, $\mathcal{P}_1 = \{P | P(\{a\}) \geq .5\}$, and $\mathcal{P}_2 = \{P | .5 \leq P(\{a\}) \leq .6 \text{ or } P(\{a\}) \geq .9\}$.

Let us call a function P_*, defined for all $A \subset \Theta$, a *lower probability function* if it is given by (4) for some class $\mathcal{P}$. And let us call a function of two variables $P_*(A|B)$ a *lower conditional probability function* if it is given by (5) for some class $\mathcal{P}$; such a function is defined for $B = \Theta$ and for all other $B \subset \Theta$ such that $P_*(\overline{B}|\Theta) < 1$. In general, as we have seen, there are many classes that yield the same lower probability function or lower conditional probability function. But the largest class that yields a given lower probability function P_* is

$$\mathcal{P}(P_*) = \{P | P(A) \geq P_*(A) \text{ for all } A \subset \Theta\}, \tag{6}$$

and the largest class that yields a given lower conditional probability function $P_*(\cdot|\cdot)$ is

$$\mathcal{P}(P_*(\cdot|\cdot)) = \left\{P | \text{ if } P_*(\overline{B}|\Theta) < 1, \text{ then } P(B) > 0 \text{ and } P(A|B) > P_*(A|B)\right\}. \tag{7}$$

Lower probability functions have been characterized axiomatically by Williams [31], Huber [14], and Wolf [33]. I have not seen simple axioms for lower conditional probability functions, but see Williams [30].

Our "theory of lower probabilities," as I have described it so far, includes in its scale of canonical examples every possible class $\mathcal{P}$ of chance distributions over a frame Θ. For the theory allows us to specify an arbitrary property of a chance distribution and to say that our evidence is like knowing that the truth is generated according to chances having that property. Perhaps this is too rich a scale. In practice there will surely be a limit to the complexity and subtlety of properties that can sensibly be said to correspond to intuitive insights about our evidence. And it may be desirable, from a psychological point of view, for the theory to recognize this explicitly by specifying a somewhat sparser scale. It cannot help us in fitting our evidence to a scale of canonical examples to have that scale encumbered with confusing and superfluous possibilities.

Just what classes $\mathcal{P}$ should be included in the theory's scale? I see no definitive answer to this question, but it does seem that an adequate scale should include all $\mathcal{P}$ that can be defined by the sorts of constraints commonly placed on chance distributions—all that can be defined, say, by (1) bounds on chances, conditional chances, and expectations, (2) comparisons among chances and conditional chances, and (3) conditions of independence and conditional independence. This is a rich scale. It includes far more $\mathcal{P}$ than those of the form (6) or (7), and far more, even, that those that can be defined by bounds on expectations. (As we have already noted, bounds on chances and conditional chances can be reduced to bounds on expectations. Moreover, some comparisons can be reduced to bounds: the condition $P(A) > P(B)$, for example, is equivalent to $P\left(A \cap \overline{B} | A \triangle B\right) \geq \frac{1}{2}$, or simply to $P\left(A | A \cup B\right) \geq \frac{1}{2}$ if $A \cap B = \emptyset$. But conditions of independence and comparisons of the form $P(A|B) \geq P(A)$, say, go beyond bounds on expectations.)

Notice that if we were content with a scale consisting of $\mathcal{P}$ of the form (7), then the lower conditional probability function $P_*(\cdot|\cdot)$ would completely identify $\mathcal{P}$ and hence would be a complete report of our assessment of our evidence. If we agree, as I think we must, that a richer scale is necessary, then $P_*(\cdot|\cdot)$ cannot be regarded as a complete assessment. But it might be an adequate summary for some purposes.

1.4 The Literature on Lower Probabilities

The idea of constructing a class of distributions by comparing our evidence to knowledge that the truth is generated according to chances having certain properties is an adaptation of an idea developed by I. J. Good [11]. Good suggests that we pretend we have an additive probability distribution P in a black box. Initially we know nothing about P, except that it is defined for subsets of a frame Θ. But we make qualitative probability judgments about Θ, and we interpret these judgments as constraints on P. For example, we judge that A is more probable than B, and we interpret this as $P(A) > P(B)$. Or we judge that we would think A more probable than B if we knew C for certain, and we interpret this as $P(A|C) > P(B|C)$. If we manage to keep these constraints from conflicting, then they determine a non-empty set $\mathcal{P}$ of additive probability distributions.

Unfortunately, Good does not say that we are comparing our evidence with knowledge that the truth is generated by some chance law in $\mathcal{P}$. Instead he studiously avoids pinning down the nature of the unknown probability distribution P—he locates P in a "black box" precisely in order to avoid saying whether it is a chance law, a hidden subjective distribution, or something else. I believe this deliberate vagueness is untenable in a constructive theory. It leaves us uncertain about how to make the qualitative probability judgments and uneasy about whether we really want to interpret these judgments as constraints on P. We cannot make even qualitative probability judgments unless we have a definite language in which to work.

Most other recent literature on lower probabilities seems less relevant to our constructive view. Smith [28] and Williams [30, 31] study lower probabilities as betting rates, but as I argue in Sect. 2 below, it is difficult to relate talk about betting to constructive probability judgment. Huber's work on lower probabilities [13, 14] is mainly concerned with situations where the truth's being generated by chance is a serious hypothesis and not just a metaphor. For further references, see Shafer [22].

1.5 Belief Functions and Lower Probabilities

Mathematically, every belief function is a lower probability function. Every function of the form (2), that is to say, is also of the form (4). Here is one way to see this. Given a belief function Bel on a frame Θ, we can construct an additive probability distribution P such that $P(A) > Bel(A)$ for all $A \subset \Theta$ by choosing an element θ_B of every non-empty subset B of Θ and setting

$$p\left(\theta\right) = \sum \left\{ m\left(B\right) | \theta_B = \theta \right\}.$$

Let $\mathcal{P}$ denote the class of distributions obtained by varying the choice of the θ_B. Then $P(A)$ is smallest for those P in $\mathcal{P}$ that choose θ_B to be outside A whenever possible—i.e., whenever $B \not\subset A$. So

$$\inf \left\{ P\left(A\right) | P \in \mathcal{P} \right\} = \sum \left\{ m\left(B\right) | B \subset A \right\} = Bel\left(A\right).$$

Not every lower probability function, on the other hand, is a belief function; Williams exhibits an example of one that is not on page 380 of his review.

Does the fact that every belief function is a lower probability function mean that our theory of lower probabilities is more general than the theory of belief functions? Certainly not. For the theory of belief functions uses a belief function in a different way than our theory of lower probabilities would use it. The meaning is quite different in the two cases. One theory is comparing our evidence to knowledge provided by a randomly coded message; the other is comparing our evidence to knowledge about chances governing the truth. I will discuss some of the implications of this difference in meaning in Sects. 3 and 5 below.

Since it does retain the Bayesian idea that our evidence is like knowing that the truth is generated by chance, our theory of lower probabilities is much closer in spirit to the Bayesian theory than the theory of belief functions is. And, as we shall see in Sect. 3 below, it does not escape as thoroughly as one might think from the Bayesian emphasis on prior probabilities.

I will not surprise the reader when I say that I find belief functions more interesting and promising than lower probabilities. In many cases, I believe, our evidence is so unlike knowledge that the truth is generated by chance that it is misleading to liken a conviction that the evidence supports A better than B to knowledge that the chance of A is greater than the chance of B.

I hope, on the other hand, that the theory of lower probabilities I have sketched here is more than a straw man. It is quite possible that judgments of the kind the theory suggests will sometimes provide the most useful and insightful way to analyze one's evidence. And, as I shall try to show in this paper, the theory provides explicit motivation for assumptions that Good, Smith, and Williams have taken for granted in their writings on lower probabilities.

1.6 What is a Degree of Belief?

What is meant by "degree of belief," and how might an individual determine his degrees of belief in a particular case?

The meaning of an "epistemic probability" or "degree of belief" is very rich. It depends, I have argued, on the whole theory by which the probability judgment is made or, as we might put it, on the whole language in which it is expressed. A degree of belief of .3, say, means one thing in the Bayesian theory and something different in the theory of belief functions. It also depends on the canons of judgment that have been established in the particular field of inquiry. A historian's valuation of certain kinds of evidence may differ from a judge's.

There is room for ambiguity in the question about how an individual might "determine his degrees of belief." Some Bayesians give the impression of thinking that we have numerical probabilities for everything hidden in our psyche; they would interpret "determine" as a synonym for "elicit." Others take a more constructive view; for them probability judgment is a matter of assessing evidence and constructing reasonable numerical beliefs. As I have tried to make clear, I subscribe to the constructive view. Probability judgment is a matter of construction. We may come to the task with some vague beliefs, but these will not be numerically precise and will usually not even have any very definite structure. (It would be silly, for example, to argue about whether our unreflective beliefs have a structure more like belief functions or more like Bayesian probability distributions. There simply is not that much structure there.) And the process of construction should ideally be sufficiently fruitful in new insights and understanding as to render obsolete much of any rudimentary structure that might be in these initial vague beliefs.

1.7 Why Belief Functions?

For what reasons are degrees of belief required to satisfy the conditions imposed? Why, that is to say, should "belief functions" be required to be of the form (2) instead of, say, the more general form (4)?

As I see it, the theory of belief functions is a language in which one can construct and express probability judgments. Asking why the theory uses degrees of belief with a given structure is like asking why some aspect of a language's grammar is as it is. Explanations can be given, but they are inevitably internal explanations—explanations of how that aspect fits in with other aspects of the language. Challenged to explain why belief functions are required to be of the form (2), I might point out that only functions of this form can be combined by Dempster's rule. Or I might point out that functions of this form result when evidence is assessed using the scale of canonical examples involving randomly coded messages. But these are only internal explanations. They do not rule out the usefulness or even superiority of a different theory using a different and possibly more general structure for degrees of belief.

As I have tried to make clear, I do not deny the possibility of a theory superior to the theory of belief functions. I believe, though, that the superiority of one theory of probability judgment to another can be demonstrated only by a preponderance of examples where the best analysis using the one theory is more insightful than the best analysis using the other. As Amos Tversky puts it, the unit of comparison for theories of probability judgment is the individual analysis.

The individual analyses we compare should be complete analyses—analyses beginning with an intuitive account of one's actual evidence and building up formal judgments step by step. (Examples of such analyses using belief functions are given in Shafer [24] and Shafer and Breipohl [27].) It may be unfair to ask a theory to deal with a problem which has already been translated from actual experience into the language of another theory.

It would be unfair, for example, to argue that the very existence of a class $\mathcal{P}$ of chance distributions such that (4) is not a belief function is proof of the inadequacy of the theory of belief functions. For it is not the case that we can ever really know, in a concrete problem, that the truth is generated by chance in accordance with some distribution in a class $\mathcal{P}$. Rather, the determination of the class $\mathcal{P}$ must itself be regarded as the first step in one particular approach to constructing probability judgments. And so it proves nothing that the theory of belief functions may be unable to carry on from this first step. The important questions are: (1) Can a theory of lower probability functions show us how to carry out this first step insightfully? (2) In real examples where such a theory succeeds, can the theory of belief functions do as good or better using some other first step?

2 Betting

Since they use the picture of chance, our three constructive theories inevitably lead us to think about betting. But what exactly is the significance of betting for these theories?

Certainly we should not, in a constructive theory, interpret a probability judgment as an actual commitment to bet. Nor should we interpret it as a declaration that the person making the judgment has exactly the same attitude towards a bet in accordance with that judgment as he has towards a fair bet in a game of chance. Our relative equanimity about fair bets in games of chance is based on the assurance that the chances are objective facts and on the assurance that no possible opponent can gain an advantage over us through deeper understanding or knowledge of the game, and these elements are missing when we construct probability judgments on the basis of ordinary evidence. A probability judgment using the Bayesian theory, for example, is merely a judgment that our evidence is more similar in strength and significance to the evidence provided by knowledge of given chances than to the evidence provided by knowledge of different chances. We will not be happy unless we feel that the similarity is substantial and instructive and that our judgment is sound, but we will not pretend that the similarity is complete, nor that we are certain no one else could make a better judgment.

2.1 Long-Run Policies

So what are we saying about betting when we announce a probability judgment in one of our constructive theories? We are only saying, I think, that we judge our evidence to be similar to knowledge of a chance model where certain bets conform to a prudent long-run policy.

It is instructive to spell this out for each of our three theories.

- When we construct a Bayesian probability distribution P, we are judging our evidence to be like knowledge of a chance model where betting on A at the rate $P(A)$ conforms to a policy that breaks even in the long run. (If, for $i = 1, 2, \ldots$, a chance distribution P_i over Θ_i is used to generate an independent outcome $\theta_i \in \Theta_i$, and if on each occasion we choose a subset A_i of Θ and bet on it at rate $P_i(A_i)$, then we break even in the long run.)
- When we construct a belief function Bel, we are judging our evidence to be like knowledge of a chance model where betting on A at the rate $Bel(A)$ would conform to a policy that at least breaks even in the long run. (Consider a sequence of randomly and independently coded messages. Suppose the ith message bears on Θ_i. If we choose a subset A_i of each Θ_i, and if $Bel_i(A_i)$ turns out to be the total chance that the ith true message implies A_i, then we at least break even in the long run by betting on A_i at the rate $Bel_i(A_i)$.)
- When we construct a lower probability function P_*, we are judging our evidence to be like knowledge of a chance model where betting on A at the rate $P_*(A)$ would conform to a policy that at least breaks even in the long run. (If, for $i = 1, 2, \ldots$, a chance distribution P_i over Θ_i is used to generate an independent outcome $\theta_i \in \Theta_i$, and if on each occasion we choose a subset A_i of Θ_i and bet on it at a rate $P_{*i}(A_i) \leq P_i(A_i)$, then we at least break even in the long run.)

Notice that we can make statements for belief functions and lower probability functions that are identical on the surface. But in making these statements we have chance models and long-run policies in mind that are quite different in the two cases. A belief function and a lower probability function that are mathematically equivalent evoke the same bets in our actual problem, but they refer these bets to different chance models and embed them in different long-run policies.

Notice also that our statements about the long-run policies breaking even in the chance models are not quite theorems. They can be turned into theorems only by giving some mathematical form to the implicit assumption that our choice of the A_i is independent of the truth and of the random action of the model.

In formulating the statements about the models, I have been careful to embed each probability judgment in a sequence of judgments with different chance models and even different frames. For the chance model and the frame are constructed to represent the evidence in the problem at hand, and the next problem, and its evidence, will be different. If we were to allow ourselves to envision repeated trials using the same model (P, Θ), then we could make much stronger and more mathematically precise statements for the Bayesian and lower probability models. We could, for example, say the following:

- If a chance distribution P over Θ is used to generate a sequence $\theta_1, \theta_2, \ldots,$ of independent outcomes, and on each occasion we bet on $A \subset \Theta$ at the rate $P(A)$, then we will break even in the long run. In fact, we will break even even if we offer such bets for all $A \subset \Theta$ and let our opponents choose, on each occasion, which bets to accept.

But since P is a product of our particular problem, these strong statements are utterly irrelevant.

In the case of the chance model for belief functions there is no such temptation to talk about repetitions. For the belief function *Bel* is determined, in the model, by the random choice of a code and would vary even if the chance distribution for the code were kept fixed.

To summarize: Constructive probability judgments can be related to betting, but the relation is tenuous on two counts. It is tenuous because we are only comparing our evidence to a chance model. And it is tenuous because even in the model the bets can be justified only when embedded in a particular long-run policy involving other models.

2.2 The Dutch-Book Arguments

Williams must have a more intimate relation between probability and betting in mind when he writes about the "betting interpretation" of Bayesian degrees of belief and of lower probabilities and pleads for a similar "operational interpretation" for belief functions. But what more intimate relation can there be if we insist on a constructive understanding of probability judgment?

Williams' answer, apparently, is that our primary purpose in constructing probability judgments should be the setting of rates at which we will offer bets in accordance with some betting scheme.

There is, Williams reminds us, a betting scheme that seems to force a Bayesian structure on betting rates and another, looser one that seems to force the less restrictive structure of lower probability functions on them.

- Suppose we must choose, for each subset A of Θ, a betting rate $P(A)$ and then offer to take either side of a bet on A at odds $P(A) : 1 - P(A)$. Then an opponent can compile a book of bets from our offers that assures a net gain from us (a "Dutch book") if and only if the function P fails to be an additive probability distribution.
- Suppose we must choose a betting rate p_* for each A and then offer to bet on A at the odds $p_* : 1 - p_*$, but we are not required to offer to take the other side of the bet. Let $P_*(A)$ denote the greatest rate at which we have offered to bet on A—either explicitly or because such a bet can be compounded from our other offers. Then a Dutch book can be made against us if and only if P_* fails to be a lower probability function. (See Smith [28] or Williams [31]. Williams' proof of this result is especially elegant.)

But there does not seem to be a betting scheme in which the avoidance of Dutch book yields precisely the class of belief functions.

The Dutch-book arguments are interesting, but it is hard to accept the claim that the setting of betting rates in some particular betting scheme is the primary purpose of probability judgment.

It is often argued in this connection that every choice or action is like a bet and that probability judgments ultimately have no purpose other than to guide future choice and action.

But how well do human choices and actions fit the picture of a bet? How well, that is to say, do they fit the apparatus of "decision theory," where alternatives are weighed by the combination of probabilities and utilities? I believe that they do not fit very well. One way to understand why they do not fit is to recognize that utilities, like probabilities, do not simply exist. They are constructed. And in the case of utilities the construction is accomplished not so much by reflective thought as by our choices and actions themselves. It is only after a human being or a society of human beings has established a self-conception through crucial choices in a given domain that we can speak in any detail about his or its preferences in that domain. (For a review of some recent thinking about the inadequacy of decision theory, see March [18].)

Probability judgments should help guide our future choice and action, but it is also important to remember that the proximate purpose of probability judgment is always understanding. Human beings often seem to prize understanding for its own sake, and it is not easy to argue that this is always mere appearance. For it is only after we have gained understanding that we can formulate other goals.

Sometimes we are told that the Bayesian theory is a theory about the betting behavior of ideal rational agents, and that as such it is "normative"—it provides us with a definition of rationality that is so inherently attractive that we should try to conform to it, even if we cannot fully succeed. But surely this line of thought begs all the important questions. It is vacuous to call a mode of thinking or behavior an ideal unless it is appropriate to our needs and capabilities. And though the Bayesian theory is clearly a norm for behavior within a particular betting scheme, this does not make it a useful norm in ordinary thought and action.

I conclude that it is misleading to speak of a "betting interpretation" of probability. All three of our theories of probability judgment produce degrees of belief that can be used to set betting rates without fear of Dutch book. But this is only a minor aspect of their meaning.

2.3 Betting as a Tool in Probability Judgment

Another possible way of relating betting to probability might be to use introspection about betting as a tool in constructing probability judgments.

In the context of our three constructive theories, this would mean using such introspection to help us compare our evidence to canonical examples involving chance. We might try to locate the strength of our evidence on the scale of chances by asking ourselves at what odds our attitude towards a given bet would be comparable to our attitude towards a fair bet (Bayesian theory), or perhaps at what odds our attitude would be comparable to our attitude towards a bet we know to be at least fair (theory of lower probability functions). This might be more effective psychologically than trying to think about our evidence in terms of frequencies or propensities. The prospect of monetary loss or gain might concentrate our minds and thus permit a more honest and acute assessment of the strength of our evidence than we could obtain by thinking about it directly.

Here we have a reasonably sharp empirical question. Does it help people assess their evidence to think about betting? Or is it more helpful to think about frequencies or propensities? This question has not, perhaps, been investigated as directly as it might be. But the many empirical studies that have been made in this area do not seem to indicate that the betting metaphor is any more useful than the frequency metaphor, say, as a psychological aid in constructing degrees of belief.

I do not personally find that talk about betting concentrates my mind on my evidence; instead it tends to divert my mind to extraneous questions: my attitude towards the monetary and social consequences of winning or losing the bet, my assessment of the knowledge and astuteness of my opponent, etc. I find it inherently implausible, moreover, that I could better understand the strength of my evidence by asking myself about my willingness to bet. In a situation where I had somehow made a thorough and unimprovable but not fully conscious analysis of my evidence, it might be sensible for me to forget about the evidence and concentrate on my own hidden attitudes. But so far as I know, I do not make such unconscious analyses of evidence.

2.4 Lower Expectations

A function X which assigns a real number $X(\theta)$ to every $\theta \in \Theta$ can be thought of as a gamble: if $X(\theta) > 0$, then $X(\theta)$ is the amount we win; if $X(\theta) < 0$, then $-X(\theta)$ is the amount we lose. The idea of buying a gamble generalizes the idea of betting, for betting the amount p on A at the odds $p : 1-p$ means paying p to buy the gamble

$$X(\theta) = \begin{cases} 1 & \text{if } \theta \in A \\ 0 & \text{if } \theta \notin A. \end{cases}$$

Let us consider how each of our three theories would price a gamble.

- *Bayes.* If the truth is generated by chance in accordance with the chance distribution P, then the fair price for the gamble X is, of course, its expectation with respect to P, $E_P(X)$. Paying $E_P(X)$ for X is a policy that at least breaks even in the long run.
- *Belief Functions.* If we receive an infallible message that the truth is in $A \subset \Theta$, then we know the gamble $X(\theta)$ is worth at least $\inf\{X(\theta) | \theta \in A\}$ to us. So if we receive a randomly coded message and the chance of the message meaning A turns out to be $m(A)$ for each $A \subset \Theta$, then it is natural to price the gamble at the average value

$$\widehat{Bel}(X) = \sum_{A \subset \Theta} m(A) \left[\inf_{\theta \in A} X(\theta) \right]. \quad (8)$$

  Let us call $\widehat{Bel}(X)$ the *lower expectation* of X. It is a fair price to pay for X in the sense that we will at least break even if we pay such prices for gambles in a long run of independent randomly coded messages.
- *Lower probabilities.* Suppose we know the truth is generated by chance in accordance with some distribution in a class $\mathcal{P}$. Then we know the expectation of X is at least

$$E_*(X) = \inf_{P \in \mathcal{P}} E_P(X). \quad (9)$$

  And we will at least break even in the long run if we follow the policy of paying this price for X.

In Sect. 1 above I called (9) the lower expectation of X. Is it consistent to call both (8) and (9) by the same name? As it turns out, it is; if

$$\mathcal{P} = \{P | P(A) \geq Bel(A) \ \text{ for all } \ A \subset \Theta\},$$

then (8) and (9) will be equal. (See Huber [13] and Shafer [23].)

3 Conditioning

The idea of conditioning has its origin in the theory of chance.

Conditioning occurs most naturally, perhaps, in the case of a game of chance that unfolds step by step. When such a game has been only partly played out (when only the first die has been thrown, say), chance still has a role to play. And this role can be described by the conditional chance distribution. Suppose, indeed, that X denotes the set of complete outcomes for the game, and that the chance for each outcome x is denoted $p(x)$, so that the chance law P governing the game is given by

$$P(A) = \sum\{p(x)|x \in A\}$$

for all $A \subset X$. Say the partial playing out of the game determines only that the eventual outcome will be in the subset X_0 of X. Then the conditional chances $p'(x)$ governing the remainder of play are obtained by reducing the $p(x)$ for $x \notin X_0$ to zero and multiplying the $p(x)$ for $x \in X_0$ by the factor $P(X_0)^{-1}$. And the conditional chance distribution $P(\cdot|X_0)^{-1}$ is given by

$$P(A|X_0) = \sum\{p'(x)|x \in A\} = \frac{P\left(A \cap X_0\right)}{P\left(X_0\right)} \tag{10}$$

for all $A \subset X$. We can see that this is the right way to define the conditional chances by thinking about long-run frequencies: $P(A|X_0)$ is simply the proportion of the games that reduce to X_0 during the first stage of play that will go on to have their eventual outcome in A.

Conditioning can, of course, be applied in the case of any subset X_0 of X, even if X_0 does not correspond to a partial completion of the game. There are several ways of explaining what meaning conditioning might have in this more general case. One way is to turn our attention from the chances to the degrees of belief they justify. If we know the chance distribution P and have therefore adopted its values as our degrees of belief concerning how the game will turn out, then news that the outcome has fallen in X_0 will naturally lead us to revise our beliefs by (10). Of all the games in which this news is true, we will tell ourselves, $P(A|X_0)$ is the proportion in which the outcome is in A. And so adopting $P(A|X_0)$ as our new degree of belief seems reasonable, provided there is no trickery involved in our having received the news that the outcome is in X_0—provided, in other words, that our receipt of this news is not the result of some fiendish scheme to mislead us.

Now suppose we represent ordinary evidence by constructing degrees of belief over a frame Θ and then obtain new evidence whose direct effect on Θ is to establish with certainty that the truth is in a subset Θ_0. How should we change our degrees of belief to take this new evidence into account? Each of our constructive theories of probability has its own way of translating the rule of conditioning for chance distributions into an answer to this questions.

- *Bayes.* In the Bayesian case we have constructed an additive probability distribution P over Θ, with the understanding that our evidence is comparable to knowledge that the truth is generated by P. So we will simply adopt the conditional distribution $P(\cdot|\Theta_0)$ as our new additive probability distribution.
- *Belief functions.* In the case of belief functions, the chance distribution in our model is a distribution for the random choice of a code, and when we take the news that the truth is in Θ_0 into account, we have to condition this distribution on a subset of codes.

  Say we have represented our old evidence by a belief function Bel, corresponding to a randomly coded message with possible codes $c_1, \ldots, c_n$, where code c_i was used with chance p_i and decoding by code c_i produces the message $A_i \subset \Theta$. We can simply incorporate the news that the truth is in Θ_0 into the messages, thus changing A_i to $A_i \cap \Theta_0$. But we must also notice that the news may tell us something about which code was used: if $A_i \cap \Theta_0 = \emptyset$, then code c_i cannot be the code that was used. So in addition to changing A_i to $A_i \cap \Theta_0$ we must also condition the chance distribution for the codes on the subset $\{c_i | A_i \cap \Theta_0 \neq \emptyset\}$ of codes. This means we replace the p_i by p_i', where

$$p_i' = \begin{cases} 0 & \text{if } A_i \cap \Theta_0 = \emptyset \\ \frac{p_i}{\sum\{p_j | A_j \cap \Theta_0 \neq \emptyset\}} & \text{if } A_i \cap \Theta_0 \neq \emptyset. \end{cases}$$

  These two changes (replacing p_i with p_i' and A_i with $A_i \cap \Theta_0$) give us a new randomly coded message representing the total evidence. The belief function $Bel(\cdot|\Theta_0)$ corresponding to this randomly coded message has m-values

$$m(A|\Theta_0) = \sum \{p_i' | A_i \cap \Theta_0 = A\} = \frac{\sum\{p_i | A_i \cap \Theta_0 = A\}}{\sum\{p_i | A_i \cap \Theta_0 \neq \emptyset\}}$$

  for all $A \neq \emptyset$, and so

$$\begin{aligned} Bel(A|\Theta_0) &= \sum \{m(B|\Theta_0) | B \subset A\} \\ &= \frac{\sum\{p_i | A_i \cap \Theta_0 \subset A\} - \sum\{p_i | A_i \cap \Theta_0 = \emptyset\}}{1 - \sum\{p_i | A_i \cap \Theta_0 = \emptyset\}} \\ &= \frac{Bel\left(A \cup \overline{\Theta_0}\right) - Bel\left(\overline{\Theta_0}\right)}{1 - Bel\left(\overline{\Theta_0}\right)} \end{aligned}$$

  for all $A \subset \Theta$. This is the rule of conditioning for belief functions.
- *Lower probabilities.* Suppose we think the evidence bearing on a frame Θ is similar in strength to knowledge that the truth is generated by chance in accordance with some distribution in a class $\mathcal{P}$. Then we can take new evidence that the truth is in Θ_0 into account by saying that our total evidence is similar in strength to knowledge that the truth is generated

by chance in accordance with some distribution in the class $\mathcal{P}'$ obtained by conditioning on Θ_0 each element of $\mathcal{P}$ that can be so conditioned. In particular, we replace our lower probability P_* by P'_*, where

$$P'_*(A) = \inf \{P(A|\Theta_0)|P \in \mathcal{P}; P(\Theta_0) > 0\},$$

and we replace our lower conditional probability function $P_*(\cdot|\cdot)$ by

$$P'_*(A|B) = \inf \{P(A|B \cap \Theta_0)|P \in \mathcal{P}; P(B \cap \Theta_0) > 0\} = P_*(A|B \cap \Theta_0).$$

Notice that $P'_*(A|B)$ is undefined if $P_*(\overline{B \cap \Theta_0}) = 1$, in which case $P'_*(\overline{B}|\Theta_0) = 1$.

3.1 The Role of Conditioning

It should be emphasized that the decision to use the rule of conditioning in one of our constructive theories is itself a constructive judgment. We condition on B, as I have said, when the direct effect of new evidence on our frame Θ is to establish that the truth is in B. But whether this is the direct effect of the new evidence is a matter of judgment, not of fact. "The direct effect of the new evidence" is an idea that has reality only within our language of probability judgment. We learn the meaning of this idea by example, just as we learn the meaning of other elements of a language, and our application of the idea to particular evidence is, like other probability judgments, a comparison of that evidence with other examples.

The decision to condition is just one place where the idea of "the direct effect of given evidence" comes into play in the theory of belief functions. It also comes into play when we represent an item of evidence by a simple support function; in this case we must judge that the item's only direct effect on Θ is to support a given subset. And, as we shall see in Sect. 4 below, this is merely a special case of the judgment that the direct effect of given evidence on Θ is discerned by a given subalgebra.

The theory of belief functions is so concerned to identify the direct effect of given evidence because it often works with limited items of evidence. As I pointed out in Sect. 1 above, the fundamental strategy of the theory is to make judgments based on different items of evidence and then to combine these judgments. Conditioning is merely one example of such decomposition and recombination, and it is unusual only in that the message of one of the items of evidence is conclusive.

Theories which compare evidence to knowledge that the truth is generated by chance do not depend so extensively on the decomposition of evidence. Our theory of lower probabilities, for example, breaks the overall task of judgment down by distinguishing different questions, not by distinguishing different items of evidence bearing on those questions. We construct a lower probability function from many judgments of the form "our evidence is like knowing the chance of A to be greater than p," but it is "A" and "p" that vary from judgment to judgment, not the evidence; all the judgments are supposed to be based on the total evidence. In this theory, as in the Bayesian theory, it is only in the case of conditioning that we decompose our evidence, and so it is only in the case of conditioning that we are concerned with identifying the direct effect of a limited item of evidence.

How important is conditioning? Some Bayesians have given it a central role in their theory, perhaps because it is the only way their theory decomposes evidence and is hence the only way they can formally combine "new" evidence with old. (See, for example, de Finetti [8], p. 141.) But I am inclined to think of conditioning as a tool we will not use very often in a constructive theory. It will happen fairly often, no doubt, that we can formulate a frame and distinguish evidence whose direct effect is to establish that the truth is in a certain subset. But how often will this frame be the same as the one we have used or want to use in assessing the balance of our evidence? New evidence that we actually obtain after constructing numerical probability judgments over a frame Θ will seldom affect Θ so simply. And I also find it doubtful whether the assessment of a body of evidence already obtained will very often be best accomplished by singling out a part that establishes a subset B of a frame Θ, using the rest to construct degrees of belief over all of Θ, and then conditioning on B. It will usually, I think, be more sensible and efficient to treat knowledge of B as just another element of our background knowledge and to concentrate our probability judgments on matters that we really find uncertain. For a discussion of this point in the context of a detailed example, see Shafer [24].

One aspect of a decision to use conditioning in our constructive theories is the implicit judgment that the news that the truth is in B has not been selected from the many things we might be told just because it will interact with other evidence in such a way as to mislead us. This judgment can be translated into statements about the chance models used by the theories. In the Bayesian theory and the other theories that think of the truth as being generated by chance, the judgment comes down to saying that our new evidence is like learning the truth is in B by means of some mechanism that selects this message to send us without regard to the chances by which the truth was generated. In the theory of belief functions, the judgment comes down to saying that the selection of the message was without regard either to how the random coding of previous messages was set up or to how that random coding turned out. Notice that these statements assure, within the chance models, that betting in accord with the new degrees of belief remains a policy that at least breaks even in the long run.

3.2 A Comparison of Two Rules

The theory of belief functions and our theory of lower probabilities have very different rules of conditioning—rules that can give very different results even when applied to the same degrees of belief. We can gain insight into the difference between the two theories by studying a simple example of this divergence.

Let us first consider how the theory of belief functions conditions a simple support function. Suppose A_1 is a proper non-empty subset of Θ and we represent strong but inconclusive evidence that the truth is in A_1 by the simple support function

$$Bel\,(A) = \begin{cases} 0 & \text{if } A_1 \not\subset A \\ .95 & \text{if } A_1 \subset A \neq \Theta \\ 1 & \text{if } A = \Theta. \end{cases} \tag{11}$$

This belief function has m-values $m(A_1) = .95$, $m(\Theta) = .05$, and $m(A) = 0$ for all other A. In adopting it we are likening our evidence to a message that probably means A (chance .95) but might possibly (chance .05) mean nothing. Now suppose we obtain new evidence whose direct effect on Θ is to establish that the truth is in A_2, where A_2 is some other subset of Θ such that $A_1 \cap A_2 \neq \emptyset$. Then we condition Bel on A_2, obtaining

$$Bel\,(A|A_2) = \begin{cases} 0 & \text{if } A_1 \cap A_2 \not\subset A \\ .95 & \text{if } A_1 \cap A_2 \subset A \not\supset A_2 \\ 1 & \text{if } A_2 \subset A; \end{cases} \tag{12}$$

the news that the truth is in A_2 changes the message that it is probably in A_1 into the more specific message that it is probably in $A_1 \cap A_2$.

Let us make the story more concrete. Suppose a burglar is traced to a rooming house, in such a way as to make it highly probable that he is actually one of the roomers, though it is believed that he keeps his tools and loot elsewhere. A police detective searches the rooming house and interviews the five roomers, but on this first examination finds nothing that either exonerates or further incriminates any of them. At this point the detective might formulate a frame Θ which includes, for each roomer i, a subset B_i corresponding to the possibility that roomer i is the burglar. (See Fig. 1.) And he might adopt (11) as a representation of his evidence, where A_1 is the union of the B_i's.

Suppose now that roomers 4 and 5 produce airtight alibis, conclusively establishing that neither is the burglar. Such alibis, in order to be convincing,

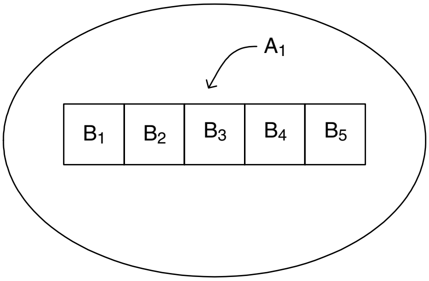


Fig. 1. Roomer i is the burglar

would have to involve great detail, and this detail would inevitably provide less conclusive evidence about other questions. But we may suppose that these other questions are not germane to the investigation and therefore need not be introduced into the frame Θ. Thus the detective may judge that the only direct effect of this new evidence on Θ is to eliminate B_4 and B_5 from consideration. In this case he will want to condition (11) on the set $A_2 = \overline{B_4 \cup B_5}$, which corresponds to the burglar being someone other than roomer 4 or roomer 5. The set $A_1 \cap A_2 = B_1 \cup B_2 \cup B_3$ corresponds to the burglar being one of the first three roomers. And according to the new belief function (12), the suspicion against the rooming house now points to these three.

Here is another way the story might go. Suppose the new evidence, instead of consisting of alibis, is evidence from the scene of the crime establishing that the burglar has blood type O. In this case the detective might introduce the question of the burglar's blood type into our frame Θ, so that there is a subset A_2 of Θ corresponding to its being type O. (This set A_2 is pictured in Fig. 2; since we do not yet know the roomers' blood types, A_2 intersects with each B_i.) And he will then condition (11) on A_2. The resulting belief function (12) awards degree of belief .95 to $A_1 \cap A_2$, which corresponds to the proposition that the burglar is one of the roomers and has blood type O. Under these circumstances the detective's next step will no doubt be to find out the blood type of each of the roomers and to condition (12) on this further information. I will refrain from illustrating this further conditioning graphically, because a very complicated picture arises when we introduce distinctions about each roomer's blood type into Θ. But the final result is obvious: if none of the roomers have type O blood then the suspicion against them is dispelled; otherwise it is focused on those that do.

One might challenge the adequacy of (11) and (12) as an analysis of this detective story on the grounds that there is probably other evidence that it does not take into account. Surely the detective acquired some hints and hunches in the course of interviewing the roomers. And might he not have some prior inclination to expect type O blood, given its high frequency in the

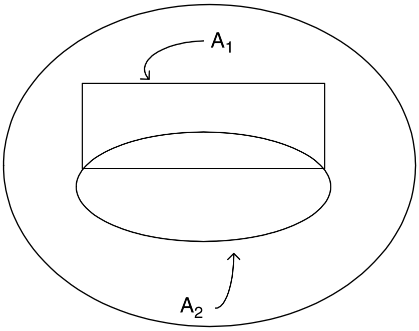


Fig. 2. The intersection of A_1 and A_2

population? The answer to this challenge is that the theory of belief functions can always accommodate further evidence, provided its relevance is identified and its value is assessed. The detective can decide he has further evidence worth introducing into the analysis, or he can decide he does not.

Let us now consider how to analyze the detective story using our theory of lower probabilities.

The most obvious approach is to liken the initial evidence in favor of A_1 to knowledge that the truth is generated by chance and that the chance of A_1 is at least .95. This means representing the evidence by the class $\mathcal{P} = \{P|P(A_1) \geq .95\}$ or by the lower probability function

$$P_*(A) = \begin{cases} 0 & \text{if } A_1 \not\subset A \\ .95 & \text{if } A_1 \subset A \neq \Theta \\ 1 & \text{if } A = \Theta, \end{cases} \tag{13}$$

which is mathematically identical to the belief function (11). But if we condition $\mathcal{P}$ on a subset A_2 that intersects both A_1 and $\overline{A_1}$, then we will obtain the new lower probability function

$$P'_*(A) = \begin{cases} 0 & \text{if } A_2 \not\subset A \\ 1 & \text{if } A_2 \subset A, \end{cases} \tag{14}$$

which indicates no particular support at all for $A_1 \cap A_2$. In fact, (14) seems to ignore the initial evidence. It is presumably the lower probability function we would adopt if we had only the new evidence establishing A_2.

It will be agreed, I think, that (14) is unsatisfactory. How is it to be avoided?

The natural move is to challenge the adequacy of the class $\mathcal{P} = \{P|P(A_1) \geq .95\}$ as a representation of our initial evidence. There is, one might argue, more to be said on the basis of the initial evidence than that the chance of A_1 is at least .95. In order to prepare for conditioning on the alibis of the two roomers for example, we might decide that the five roomers have equal chances of being the burglar, thus narrowing the class $\mathcal{P}$ down to the class

$$\mathcal{P}_1 = \{P|P(A_1) \geq .95; P(B_1) = P(B_2) = P(B_3) = P(B_4) = P(B_5)\}.$$

This already awards a lower probability of .57 to $B_1 \cup B_2 \cup B_3$. And when we condition $\mathcal{P}_1$ on $A_2 = \overline{B_4 \cup B_5}$, we obtain

$$\mathcal{P}'_1 = \left\{P|P(A_2) = 1, P(A_1 \cap A_2) \geq \frac{.57}{.62} \approx .92; P(B_1) = P(B_2) = P(B_3)\right\},$$

which awards a lower probability of .92 to $A_1 \cap A_2 = B_1 \cup B_2 \cup B_3$. This is nearly as great as the degree of belief .95 awarded by the belief function (12). Notice, though, that this analysis is sensitive to the number of roomers and the proportion with alibis in a way that the analysis using belief functions is not. If four out of the five roomers have alibis, then the final lower probability for the remaining one would be only $\frac{.19}{.24} \approx .79$; if there were 20 and 19 were similarly exonerated, then the final lower probability for the remaining one would be $\frac{.0475}{.0975} \approx .49$. And these figures could easily be altered if we claimed that our initial evidence justified unequal prior chances for the roomers.

The initial class $\mathcal{P}$ can also be adapted to give sensible results when conditioned on the burglar's blood type. In this case the natural move is to narrow $\mathcal{P}$ down to

$$\mathcal{P}_2 = \{P | P(A_1) \geq .95; P(A_1 \cap A_2) = P(A_1) P(A_2)\}.$$

We require, that is to say, that A_1 and A_2 be independent. This is reasonable; once we have decided to think of the truth as random, it is natural to think of the random determination of the burglar's blood type as stochastically independent of the random determination of whether he is one of the roomers. Conditioning $\mathcal{P}_2$ on A_2 yields

$$\mathcal{P}_2' = \{P | P(A_2) = 1; P(A_1 \cap A_2) \geq .95\},$$

which gives a lower probability function mathematically identical to the belief function (12).

This last analysis can be extended to an analysis incorporating further conditioning on the roomers' blood types that will continue to agree with the analysis using belief functions. Here is the set-up. Let T denote the burglar's blood type, let T_i denote the ith roomer's blood type, and set

$$X = \begin{cases} 0 & \text{if the burglar is not one of the roomers} \\ i & \text{if the burglar is the } i\text{th roomer.} \end{cases}$$

(Notice that $T = T_i$ when $X = i$. And "$X \neq 0$" is equivalent to A_1.) Replace the initial class $\mathcal{P}$ by the class $\mathcal{P}_3$ consisting of all P such that $P(A_1) \geq .95$, $(X, T_1, .., T_5)$ are jointly independent with respect to P, all the T_i have the same marginal distribution, and T has this same distribution conditional on $X = 0$. We may take the burglar's and the roomers' blood types into account by conditioning $\mathcal{P}_3$ on the values of T and the T_i, and if there is a subset of roomers whose blood type agrees with the burglar's they will inherit the full .95 suspicion against the rooming house.

To summarize: A basic idea of the theory of belief functions is the idea of evidence whose only direct effect on the frame Θ is to support a subset A_1, and an implicit aspect of this idea is that when this evidence is combined with further evidence whose only direct effect on Θ is to establish a compatible subset A_2, the support for A_1 is inherited by $A_1 \cap A_2$. The theory of lower probabilities does not have a fully equivalent idea. New evidence establishing A_2 *may* cause prior support for a subset A_1 to be inherited by $A_1 \cap A_2$ in the theory of lower probabilities, but whether this happens will depend, as in the Bayesian theory, on various "prior probabilities."

Indeed, the similarity between our theory of lower probabilities and the Bayesian theory in their dependence on prior probabilities is striking. Our theory of lower probabilities does not, apparently, always get us away from the Bayesian bemusement over how to assess prior probabilities when the evidence is weak. In the case of our five roomers there was a natural symmetry on which to pin "equal prior probabilities," but one could easily construct similar examples where there are no obvious symmetries or else competing ones, so that the prior probabilities needed in order to get sensible answers from conditioning seem much more arbitrary. This makes us wonder just how much is gained in the generalization from the Bayesian theory to the theory of lower probabilities.

However we answer this question, the drastically different results we get by conditioning (11) and (13) should bring home to us that a belief function can have quite a different meaning from a mathematically identical lower probability function. Saying our evidence is like a message that probably means A_1 but might mean nothing is quite different from saying it is like knowing that the truth is generated by chance and that the chance of A_1 is great. So we must decide when we make a probability judgment, just which formulation fits the significance of our evidence. We cannot simply make a vague judgment that the evidence supports A_1, express it numerically by (11), and then interpret (11) indifferently either as a belief function or as a lower probability function.

3.3 Conditional Bets

Consider again two proper subsets A_1 and A_2 of Θ such that $A_1 \neq A_2$ and $A_1 \cap A_2 \neq \emptyset$. Following de Finetti, let us call a gamble of the form

$$X(\theta) = \begin{cases} 1 - p & \text{if } \theta \in A_1 \cap A_2 \\ -p & \text{if } \theta \in \overline{A_1} \cap A_2 \\ 0 & \text{if } \theta \notin A_2, \end{cases} \tag{15}$$

where $0 < p \leq 1$, a "bet on A_1 conditional on A_2." The idea behind this name is that if we agree to this gamble (i.e., buy it for the price zero), then we will be betting on A_1 at odds $p : 1 - p$ and total stakes $p + (1 - p) = 1$, with the understanding that the bet will be called off if the truth turns out, when it is revealed, not to be in A_2.

In our constructive theories of probability judgment, our attitude towards a gamble depends, in the tenuous way discussed in Sect. 2 above, on the gamble's expectation or lower expectation. This is true in particular of a conditional bet. If the expectation or lower expectation of the conditional bet is nonnegative, then the bet conforms, in the chance model we have used to represent our evidence, to a policy that at least breaks even in the long run.

Our attitude towards any gamble will, in general, change as we acquire new evidence. And in the theory of belief functions, our attitude towards a conditional bet can change dramatically when we obtain new evidence establishing the condition of the bet. Suppose, for example, that we have represented our evidence about Θ by the belief function (11). Then our lower expectation for the conditional bet (15) is

$$\widehat{Bel}\,(X) = .95\left[\inf_{\theta\in A_1} X\,(\theta)\right] + .05\left[\inf_{\theta\in\Theta} X\,(\theta)\right] = -.05p.$$

Since this is negative, the theory gives no sanction to the bet. But if we obtain new evidence establishing A_2 and change our belief function to (12), then the lower expectation changes to

$$\widehat{Bel}\,(X|A_2) = .95\left[\inf_{\theta\in A_1\cap A_2} X\,(\theta)\right] + .05\left[\inf_{\theta\in A_2} X\,(\theta)\right] = .95\,(1-p) + .05\,(-p)\,.$$

If $p < .95$, then this will be positive and so the theory will sanction the bet as reasonable policy. It is easy to see intuitively why our attitude towards the bet changes in this way. The bet is essentially a bet on $A_1 \cap A_2$, and the original evidence, while supporting A_1, does not provide any particular support for $A_1 \cap A_2$ until it is conjoined with the evidence establishing A_2.

Neither the Bayesian theory nor the theory of lower probabilities, in contrast, ever changes its willingness to sanction a conditional bet because of new evidence whose direct effect is to establish the bet's condition. Indeed, when we condition a Bayesian probability distribution P on A_2, the expectation of (15) changes only from $E_P(X)$ to

$$E_P(X|A_2) = \frac{E_P(X)}{P(A_2)};$$

it cannot change in sign. And when we condition a class $\mathcal{P}$ of distribution on A_2, the lower expectation of (15) changes only from

$$E_*\,(X) = \inf\{E_P(X)|P \in \mathcal{P}\}$$

to

$$E_*\,(X|A_2) = \inf\left\{\frac{E_P(X)}{P(A_2)}|P \in \mathcal{P}; P(A_2) > 0\right\},$$

and while this may be a change from zero to a positive quantity it cannot be a change from a negative to a non-negative quantity or vice-versa.

This contrast can also be expressed in terms of maximum rates for conditional bets. *The maximum rate for betting on A_1 conditional on A_2* is defined as follows:

- In the case of a Bayesian probability distribution P such that $P(A_2) > 0$, it is

  $$\sup\{p|E_P\,(X) \geq 0\}\,,$$

  where X, which depends on p, is the conditional bet (15).

- In the case of a belief function Bel such that $Bel(\overline{A_2}) < 1$, it is

$$\sup \left\{ p | \widehat{Bel}\left(X\right) \geq 0 \right\}.$$

- In the case of a class $\mathcal{P}$ of distributions such that $p_*(\overline{A_2}) < 1$ (i.e., $P(A_2) > 0$ for some $P \in \mathcal{P}$), it is

$$\sup \left\{ p | E_* \left(X\right) \geq 0 \right\} = \sup \left\{ p | E_P \left(X\right) \geq 0 \text{ for all } P \in \mathcal{P} \right\}.$$

These definitions all say the same thing: the maximum rate is defined except when we are certain the truth is not in A_2 (in which case the conditional bet is of no interest), and it is defined to be the greatest value p for which the bet is sanctioned. In general, a bet on A_1 conditional on A_2 is sanctioned in one of the constructive theories only if the bet's value for p is less than or equal to this maximum rate. Thus the contrast between belief functions and the other two theories can be expressed by saying that the maximum rate for betting on A_1 conditional on A_2 may change when one conditions on A_2 in the theory of belief functions, but not in the other theories.

The picture becomes clearer, perhaps, then we notice that in the Bayesian theory the maximum rate for betting on A_1 conditional on A_2 happens to be equal to the conditional probability $P(A_1|A_2)$. This is because $E_P\left(X\right) \geq 0$ if and only if

$$P\left(A_1 \cap A_2\right)\left(1 - p\right) + P\left(\overline{A_1} \cap A_2\right)\left(-p\right) \geq 0$$

or

$$p \leq \frac{P\left(A_1 \cap A_2\right)}{P\left(A_2\right)} = P(A_1|A_2).$$

Bear in mind that though this maximum rate might be called a "conditional betting rate," it is the bet that is conditional; the rate itself is "unconditional" in the sense that it is our rate prior to obtaining new evidence and "conditioning" on A_2. But when we obtain this new evidence the conditional bet becomes, for practical purposes, unconditional—for we know its condition is satisfied. Thus our new maximum rate for the conditional bet will be the same as our new maximum rate for an unconditional bet on A_1—i.e., our new degree of belief in A_1. But this new degree of belief is $P(A_1|A_2)$. This is how it happens that our maximum rate for this particular conditional bet is unchanged.

The same thing happens in our theory of lower probabilities: the maximum rate for betting on A_1 conditional on A_2 happens to be equal to $P_*(A_1|A_2)$, and hence remains unchanged when we condition on A_2. But in the theory of belief functions this does not happen: our "prior" maximum rate for betting on A_1 conditional on A_2 is usually not equal to $Bel(A_1|A_2)$, our "posterior" maximum rate for betting on A_1.

3.4 The Dynamic Assumption of the Betting Theories

In this essay I have insisted on understanding both the Bayesian theory and the theory of lower probabilities as constructive theories. I have assumed that the degrees of belief given by both theories are the result of comparing one's evidence to knowledge about chances governing the truth. And I have used this assumption to derive the theories' methods for pricing gambles and their rules of conditioning.

In the literature that treats probability theory as a theory about the gambling behavior of "idealized rational agents," on the other hand, there is no possibility of appealing to chance models to derive rules of conditioning. And thus these rules for changing degrees of belief or betting rates become, to use Ian Hacking's eloquent phrase, *dynamic assumptions.*[3] And one faces the problem of making these assumptions plausible.

Here is how de Finetti tries to make the Bayesian rule of conditioning plausible. He begins by *defining* a Bayesian's "conditional probability of A given B," denoted $P(A|B)$, as his rate for betting "on A conditional on B"—his rate for betting, that is to say, on A with the understanding that the bet will be called off unless B is true. (See de Finetti [6], p. 109, [7], p. 82, [8], p. 135.) He then proceeds to interpret $P(A|B)$ as the probability of A conditional on B in the usual sense—i.e., as the Bayesian's degree of belief or betting rate for A after he has obtained new evidence establishing B. (See de Finetti [6], p. 119, [7], p. 210, [8], p. 141).

What are we to make of this procedure? It obviously takes for granted that *one's betting rate for a conditional bet should be unchanged when new evidence is obtained whose direct effect is to establish the truth of the bet's condition.* Let us call this *de Finetti's principle.* I have been unable to find a critical discussion of this principle in de Finetti's writing. He seems to consider the principle too self-evident to require such a discussion.

As one who finds the theory of belief functions, which does not obey de Finetti's principle, self-consistent and appealing; I find the idea that de Finetti's principle is self-evident baffling. I see the correctness of the principle when betting rates are based on knowledge of chances governing the truth. I am willing to accept the principle as part of a theory that compares our evidence to knowledge of chances. But I do not see that it is inherent to the idea of betting *per se.* It is clear enough that a bettor should change his betting rates when he learns that B is true, and that his new rate for an unconditional bet on A should be the same as his new rate for a bet on A conditional on B.

[3] See Hacking [12]. In this paper Hacking complains about the lack of any justification for the rule of conditioning in the Bayesian literature. The literature on lower probabilities is equally lacking. Since Hacking wrote, Teller [29] has given a Dutch-book argument for the Bayesian rule of conditioning, but this argument depends on the Bayesian rule of additivity and also on the assumption that we know before obtaining the new evidence that the subset established by it will be an element of a certain partition. See also Freedman and Purves [10].

Moreover, these new rates should be the same as the new rate for a bet on A conditional on any B' such that $B \subset B' \subset \Theta$. All these bets are equivalent for someone who knows that the truth is in B. But why should the new rates for all these bets be the same as the old rate for the bet conditional on B? Why should this particular rate remain unchanged while the others change?

De Finetti's principle can similarly serve as the dynamic assumption of a betting theory of lower probabilities. Smith [28] seems to use it in this way, for he gives the name "lower conditional probability" to a bettor's maximum rate for a bet on A which is to be called off unless B is true (p. 6) and then takes it for granted that this should become his betting rate for an unconditional bet on A when he obtains new evidence establishing B. Williams [30] similarly identifies lower conditional probabilities as betting rates for conditional bets but does not discuss changes in betting rates resulting from new evidence.

3.5 Williams' Argument on Conditional Bets

On p. 381 of his review, Williams discusses the pricing of conditional bets in the theory of belief functions. He casts his argument in terms of a numerical example, but we can easily recast it in general terms. It begins, essentially, with the following fact: *offers to bet on A_2 at rate p and on A_1 conditional on A_2 at rate q entail an offer to bet on $A_1 \cap A_2$ at rate pq.* (Proof: If the bet on A_2 has total stakes q, then it is the gamble

$$X_1\left(\theta\right) = \begin{cases} \left(1-p\right)q & \text{if } \theta \in A_2 \\ \left(-p\right)q & \text{if } \theta \notin A_2. \end{cases}$$

If the conditional bet has unit stakes, then it is the gamble

$$X_2\left(\theta\right) = \begin{cases} 1-q & \text{if } \theta \in A_1 \cap A_2 \\ -q & \text{if } \theta \in \overline{A_1} \cap A_2 \\ 0 & \text{if } \theta \notin A_2. \end{cases} \qquad (16)$$

Taking both these gambles means taking the gamble

$$X_1\left(\theta\right) + X_2\left(\theta\right) = \begin{cases} 1-pq & \text{if } \theta \in A_1 \cap A_2 \\ -pq & \text{if } \theta \in A_1 \cap A_2, \end{cases}$$

which is merely a bet on $A_1 \cap A_2$ at the rate pq.)

Suppose we price gambles using a belief function Bel, so that $Bel(A_2)$ and $Bel(A_1 \cap A_2)$ are the greatest rates at which we will bet on A_2 and $A_1 \cap A_2$, respectively. If q is a rate at which we bet on A_1 conditional on A_2, then our willingness to bet on A_2 at the rate $Bel(A_2)$ implies, by the italicized sentence, a willingness to bet on $A_1 \cap A_2$ at the rate $Bel(A_2)q$. So the assertion that $Bel(A_1 \cap A_2)$ is the greatest rate at which we will bet on $A_1 \cap A_2$ will be valid only if

$$Bel(A_1 \cap A_2) \geq Bel(A_2)q. \qquad (17)$$

Williams asks, in effect, whether the pricing of conditional gambles in the theory of belief functions guarantees that (17) will be true.

In fact, the theory of belief function does guarantee (17). For it sanctions the conditional bet (16) only if (16) has a non-negative lower expectation—i.e., only if

$$(1-q)\sum\{m(A)\,|A\subset A_1\cap A_2\}\geq q\sum\left\{m(A)\,|A\cap\overline{A_1}\cap A_2\neq\emptyset\right\},$$

which implies

$$(1-q)\sum\{m(A)\,|A\subset A_1\cap A_2\}\geq q\sum\{m(A)\,|A\subset A_2;A\not\subset A_1\cap A_2\},$$

or

$$(1-q)\,Bel(A_1\cap A_2)\geq q\,(Bel\,(A_2)-Bel\,(A_1\cap A_2))\,,$$

which is equivalent to (17).

There is, of course, a more general issue here. The question is whether interpreting $Bel(A)$, for each $A\subset\Theta$, as the greatest rate at which a bet on A is sanctioned is consistent with sanctioning every gamble with non-negative lower expectation. We easily see that a bet on A at rate p has non-negative lower expectation if and only if $p\leq Bel(A)$. But perhaps it is possible, in some cases, to build up a bet on A at a rate higher than $Bel(A)$ by compounding other sanctioned gambles. In fact, it is not possible. One way to verify this is to check directly that the lower expectation $\widehat{Bel}$ obeys $\widehat{Bel}(X_1+X_2)\geq\widehat{Bel}(X_1)+\widehat{Bel}(X_2)$ and $\widehat{Bel}(aX)=a\widehat{Bel}(X)$ for $a\geq 0$. Another way is to apply the general theory developed by Smith and Williams.

The relation (17) *would* be a problem for belief functions if we interpreted the conditional degree of belief $Bel(A_1|A_2)$ as a sanctioned rate for a bet on A_1 conditional on A_2. For then (17) would imply

$$Bel\,(A_1\cap A_2)\geq Bel\,(A_2)\,Bel(A_1|A_2), \tag{18}$$

and, as Williams shows using a numerical example, this relation can easily be violated by belief functions.

Unfortunately, Williams finds the identification of conditional degrees of belief with betting rates for conditional bets so compelling that he takes the failure of (18) to be a shortcoming of the theory of belief functions. He concedes (p. 381) that one might say that "$Bel(A|B)$ as defined by Shafer should be interpreted as the largest rate at which the subject would be prepared to bet on A if B were discovered to be true (whatever this means), whereas the interpretation given is in terms of the subject's prior readiness to accept conditional bets." But he evidently finds this too bizarre to take seriously, for he concludes (p. 387) that the theory's rule of conditioning "excludes the possibility of interpreting degrees of belief in terms of acceptable betting rates."

I have, I hope, adequately explained why the theory of belief functions does not identify conditional degrees of belief with betting rates for conditional bets. And I think we may conclude from the example provided by the theory of belief functions that such an identification is not inherent in the idea of betting itself. So if we apply to Williams' ideas on lower probabilities the same standards of justification that he has applied to the rules for belief functions, we must ask him to justify this identification. Perhaps the best justification is the one I have developed in this essay: the identification holds if our model for evidence is partial knowledge of chances governing the truth.

4 Minimal Extension

A lower probability function defined only on a restricted class of subsets of a frame Θ can always be extended in a minimal way to a lower probability function defined on all subsets of Θ. Belief functions can be extended in a similar way provided that the restricted class is closed under intersections but not, in general, otherwise. And this, Williams argues, makes it "difficult, in certain cases, to find a belief function which might adequately express a subject's opinions."

Here, as elsewhere in his review, it is not clear whether Williams is taking a constructive point of view. His talk about "expressing a subject's opinion" could be construed to mean that we are concerned not so much with constructive probability judgment as with the task of eliciting opinions already determined. I shall, however, respond to Williams' criticism within the constructive framework of this essay.

4.1 Minimal Extension for Belief Functions

Consider a detective who is trying to find out who stabbed a man to death. Many questions will engage his interest: the circumstances of the killing, the circumstances of the victim, etc. But few of his sources of evidence will bear directly on more than a few of these questions. A medical specialist might, for example, give evidence that bears directly only on the time of death and the nature of the struggle. Evidence that bears on the time of death may, of course, ultimately point to the killer, but only indirectly, through its interaction with other evidence.

It may be the case, as I suggested in Sect. 3 above, that the idea of "direct effect of evidence" cannot be reduced to simpler ideas and so must be learned by example. Be this as it may, it is a clear and commonplace idea, and one that is fundamental in the theory of belief functions. The use of the idea is quite simple. When we judge that given evidence bears directly only on certain questions, we formulate a frame that deals only with these questions and then construct a belief function *Bel* over this frame to represent the evidence. We then think of this frame as a coarsening of a finer frame Θ that takes into account the other questions with which we are concerned. (See Chap. 6 of *A Mathematical Theory of Evidence.*) Or, to use a more familiar vocabulary, we think of the subsets of the first frame as forming a subalgebra $\mathcal{B}$ of the algebra of all subsets of the finer frame $\varTheta$. And we adopt the belief function $\overline{Bel}$ over $\varTheta$, where

$$\overline{Bel}\,(A) = \sup\left\{Bel\,(B)\,|B \in \mathcal{B}, B \subset A\right\} \tag{19}$$

for each $A \subset \varTheta$. The belief function $\overline{Bel}$ is called the *minimal* (or *vacuous*, or *canonical*) *extension* of Bel; it gives each element of $\mathcal{B}$ the same degree of belief as Bel does, and it gives the other subsets the smallest degrees of belief consistent with these. (See Sect. 7.3 of *A Mathematical Theory of Evidence.*)

The subalgebra $\mathcal{B}$ may be more or less detailed. The detective and medical specialist, for example, may judge that the direct significance of certain medical evidence is exhausted by saying that it is highly probable that death took place between 5 and 10 hours ago. Or they may think this evidence also provides some support for a more exact time of death. Or they may think it provides both this and also some indication of the nature of the struggle. In the first case they might set $\mathcal{B} = \{\emptyset, B_0, \overline{B_0}, \varTheta\}$, where B_0 corresponds to the death taking place between 5 and 10 hours ago, set $Bel(\emptyset) = Bel(\overline{B_0}) = 0$, $Bel(B_0) = .95$, and $Bel(\varTheta) = 1$, and thus obtain for $\overline{Bel}$ a simple support function focused on B_0. But in the other cases $\mathcal{B}$ will be more detailed and $\overline{Bel}$ will be more complicated.

The idea of minimal extension can be generalized to the case where the initial belief function Bel is defined not on a subalgebra but merely on a collection $\mathcal{E}$ of subsets of $\varTheta$ that is closed under intersections. (A function on such a collection is called a belief function if there is at least one way to extend it to a belief function over $\varTheta$.) As it turns out, there always exists in this general case a belief function $\overline{Bel}$ over $\varTheta$ that extends such a belief function Bel (i.e., agrees with it on $\mathcal{E}$) and gives all subsets of $\varTheta$ the smallest degrees of belief given to them by any belief function that extends Bel. To put it another way, the function $\overline{Bel}$ defined by

$$Bel(A) = \inf\{Bel'(A)|Bel' \text{ is an extension of } Bel\} \tag{20}$$

for all $A \subset \varTheta$ is a belief function. If $\mathcal{E}$ is not an algebra, then the formula (19) for $\overline{Bel}$ may not be valid, but a more complicated formula can be given. (See Shafer [26].)

The notion of minimal extension breaks down for belief functions if the collection $\mathcal{E}$ on which Bel is initially defined is not even closed under intersections. For in this case there may not be a single extension of Bel which assigns smallest degree of belief to all subsets of $\varTheta$. To put it another way, the function $\overline{Bel}$ given by (20) may fail to be a belief function. The practical implication of this is that probability judgments based on a single item of evidence should include direct judgments about $A \cap B$ whenever they include direct judgments about A and about B. If, for example, our medical specialist judges given evidence to indicate both that the death occurred within the last ten hours and that the victim resisted, then his numerical judgments should include not only judgments about the support for each of these propositions but also a judgment about the support for their conjunction. If the specialist judges that the support for the two propositions comes from intuitively independent items of evidence or aspects of the evidence, then he can use Dempster's rule to determine the degree of support for the conjunction, but otherwise he must make a direct judgment.

In practice, the theory of belief functions applies minimal extension mainly to the case where initial judgments determine a belief function on a subalgebra. For the intuitive judgment that given evidence bears directly only on certain questions seems to translate naturally into the idea that it bears directly only on a subalgebra. And most of the theory's relevant tools (assessment relative to a single dichotomy, consonant assessment, discounting of frequencies) are readily understood as tools for constructing belief functions on subalgebras. The generalization to the case of a collection of subsets closed only under intersection seems to be of interest only as a technical tool in a theoretical context. (See Shafer [23].)

4.2 Minimal Extension for Lower Probabilities

As Williams points out, minimal extension can be applied to lower probabilities defined on an arbitrary collection $\mathcal{E}$. Suppose, indeed, that we make direct judgments that give us lower probabilities $P_*(A)$ for A in such a collection $\mathcal{E}$ and then make the judgment that those lower probabilities exhaust the impact of the evidence. If we have arranged the judgments $P_*(A)$ for $A \subset \mathcal{E}$ so that there is at least one extension to a lower probability function over Θ (i.e., so that there is at least one lower probability function P'_* defined for all subsets of Θ such that $P'_*(A) = P_*(A)$ for all $A \subset \mathcal{E}$; this may be a difficult condition to check), then there exists a *minimal extension*—a lower probability function $\overline{P'_*}$ defined for all $A \subset \mathcal{E}$ and awarding all subsets the least values awarded by any P'_* that extends P_*. In other words,

$$\overline{P'_*}(A) = \inf \{P'_*(A) \,|P'_* \text{ is an extension of } P_*\}$$

defines a lower probability function. This is obviously the same concept of minimal extension as the one used by the theory of belief functions. The only difference is that it works for all $\mathcal{E}$, not just for $\mathcal{E}$ that are closed under intersections.

The matter can be put most concisely by saying that there always exists a minimum element in the class of those lower probability functions assigning given values to given subsets. Notice, however, that there are many other properties such that there does not exist a minimum element in the class of lower probability functions having the property. If, for example, $\Theta = \{-1, 0, 1\}$, then there is no minimum element in the class of lower probability functions having lower expectation zero. Thus even lower probability functions are limited in this respect. One cannot specify arbitrary properties for a lower probability function, decide that these specifications exhaust the impact of the evidence, and then adopt the minimum lower probability function having the properties.

Williams' notion of minimal extension finds a place in the general constructive theory of lower probabilities that I developed in Sect. 1 above, but only as a rather special case. For in that theory we make judgments that impose a rather wide variety of constraints on a supposed chance distribution P before judging that we have exhausted the impact of the evidence and proceeding to derive a lower probability function P_* from the class $\mathcal{P}$ of distributions satisfying the constraints. And only if the constraints are all of the particular form "$P(A) > c$" can we think of each judgment as establishing a particular value $P_*(A)$.

4.3 Williams' Example

The tool of minimal extension is more widely available for lower probabilities than for belief functions. But what significance does this have? It seems to me that it has little immediate significance, and that its ultimate significance can only emerge from comparing the two theories as a whole in the context of actual examples. Discussing the question in isolation is rather like comparing two tool boxes on the basis of the weight of their hammers without regard for the different roles the two hammers play.

Williams does give an example to support his belief that minimal extension for arbitrary $\mathcal{E}$ is needed. He writes as follows:

> ... suppose there is evidence relating to the unknown outcomes of two tosses of a coin giving rise, for each toss, to a belief function
>
> $$Bel(\{H\}) = \frac{1}{2}, \qquad\qquad Bel(\{T\}) = 0.$$
>
> The upper and lower probabilities of heads, on either toss are therefore $\frac{1}{2}$ and 1, respectively. Now consider which belief function might be chosen to express the impact of the evidence on the set of possible joint outcomes $\{HH, HT, TH, TT\}$. We must have
>
> $$Bel(\{HH, HT\}) = Bel(\{HH, TH\}) = \frac{1}{2}, \tag{5}$$
>
> $$Bel(\{TH, TT\}) = Bel(\{HT, TT\}) = 0 \tag{6}$$
>
> since the arguments in (5) are respectively the events 'heads on the first toss' and 'heads on the second toss', whilst the arguments in (6) refer correspondingly to tails. Furthermore, one can imagine situations in which it would seem reasonable to say that no more support accrues to the remaining sets of possibilities than is required by (5) and (6). That is to say, we should look for a minimum element in the set of belief functions satisfying these conditions. ...

But, as Mr. Williams points out, there is no minimum in the class of belief functions over the frame $\Theta = \{HH, HT, TH, TT\}$ satisfying (5) and (6). (Here we have, in effect, $\mathcal{E} = \{\emptyset, \{HH, HT\}, \{HH, TH\}, \{TH, TT\}, \{HT, TT\}, \emptyset\}$, and this is not closed under intersection. We have made judgments about the degree of support for $\{HH, HT\}$ and about the degree of support for $\{HH, TH\}$, but not about the degree of support for $\{HH\} = \{HH, HT\} \cap \{HH, TH\}$.)

What are we to make of this example? Does it demonstrate that the wider availability of minimal extension can enable a theory of lower probabilities to do better than the theory of belief functions? No. The deficiency of the example in this respect is its abstract starting point. To compare theories fairly we need to compare complete analyses—analyses beginning with a full intuitive account of one's evidence and then building up the formal judgments step by step. Williams begins with the assumption that his evidence is best represented by the judgments (5) and (6) and the further judgment that $\mathcal{E}$ exhausts the impact of the evidence, and this assumption begs the real questions. If we do begin with an intuitive account of the evidence, then it may emerge that these judgments provide one sensible analysis, but it is unlikely that they will provide the only one. It is quite possible that there will be sensible analyses using belief functions that take quite different tacks. We might even choose to make a direct judgment about $\{HH\}$.

The only gesture Williams makes towards giving an intuitive basis to his example is the following:

> ...Suppose the evidence to consist of the outcome of a single toss of the coin. It is hard to see how this could provide evidence for or against any particular correlation...

And this, to my mind, says nothing about the real evidence. It seems to indicate that we have dreamed up a statistical model as one approach to analyzing the evidence. Apparently we are regarding two possible events (here called coin tosses) as repeatable experiments, with some joint chance distribution governing the pair of outcomes (X_1, X_2), say. And apparently our statistical model consists of those chance laws with identical marginals for X_1 and X_2. We are to observe another toss independent of (X_1, X_2) but governed by the same marginal and to infer what we can about the joint distribution and hence about how (X_1, X_2) will turn out. This is a parametric statistical problem. But where does it come from? What is the evidence for the model? A sensible analysis using belief functions would require answers to these questions.

5 The Independence of Evidence

Both the Bayesian theory and the theory of belief functions have a concept of independence for evidence. Both recognize different items of evidence as intuitively independent and model this intuitive independence in terms of stochastic independence. But since the two theories use the picture of chance in different ways, their concepts of independence are different.

In the theory of belief functions we liken evidence to a message whose meaning is random, or to a randomly valid argument—one whose validity depends on chance. We call different items of evidence intuitively independent when they can be likened to stochastically independent randomly valid arguments.

In the Bayesian theory, on the other hand, we liken our evidence to knowledge that the truth is generated by certain chances. Thus we do not, in general, think of the evidence itself as random. If, however, we single out a few items of our evidence, imagine that we have not yet obtained these items of evidence, and include the question of whether we will obtain them among the questions about which we are making probability judgments, then whether or not these items will occur becomes part of the truth which we are modeling as random, and so it becomes possible to think of these items of evidence as stochastically independent.

The two theories' concepts of independence have much in common. In many cases, the two theories can agree on calling certain items of evidence independent. And in both theories independence is relative to a given frame of discernment. In the theory of belief functions, this is expressed by saying that different arguments should be treated as independent only relative to a frame that discerns the interactions of their conclusions, while in the Bayesian theory it is expressed by saying that different items of evidence may be independent only conditionally given certain hypotheses.

We should not be misled, however, into thinking that the two concepts of independence are practically identical—that the two theories will always agree on whether given items of evidence are independent.[4] The fact is that they will often disagree. As we shall see in this section, the theory of belief functions may allow us to discern independent items of evidence in situations where the Bayesian theory suggests dependent items of evidence or even suggests that we need not distinguish separate items of evidence at all.

Confusion between the two theories' concepts of independence can be held responsible for the suggestion, made by Williams in his review, that the theory of belief functions cannot do as well as the Bayesian theory in taking dependencies in evidence into account. One goal of this section is to understand the thinking behind this claim and to explain why it is wrong.

5.1 Independence in the Theory of Belief Functions

The concurrence of many independent arguments can justify a high degree of belief. And it is natural to account for this by reasoning about chances. There may be a substantial chance, we tell ourselves, for any single one of the arguments to be invalid, but there is a much smaller chance that they should

[4] In Shafer [24] I suggested, wrongly, that there was such a practical identity.

all be invalid. If p_i is the chance that the ith argument is invalid, and the arguments are independent, then the chance that they are all invalid is the product of the p_i.[5]

This is a sensible account, but it must be rightly understood. When we say that the chance of an argument's validity is p_i we do not mean that the argument is literally a repeatable experiment, sometimes valid, sometimes not, and that we know the chance p_i in the way we might know the chance of heads when tossing a well-studied coin. We mean rather that we judge the force of the argument to be comparable to the force of such a randomly valid argument. And when we say that the arguments are independent, we do not mean that their validities are literally stochastically independent random events. We mean rather that we judge the arguments to be independent in an intuitive sense that is well-represented by stochastic independence[6]—i.e., that we judge the uncertainties in the arguments to be sufficiently unrelated that the combination of the arguments should have the force of the concurrence of two stochastically independent randomly valid arguments.

Dempster's rule of combination is merely an extension of this simple idea of combining the force of independent arguments by multiplication. As I explained in Sect. 1 above, the rule pools two bodies of evidence by treating the two randomly coded messages representing them as stochastically independent. When one uses the rule, one is making a judgment that the two bodies of evidence are sufficiently unrelated that pooling them is like pooling stochastically independent randomly coded messages.

Consider a simple example from *A Mathematical Theory of Evidence.* A detective investigating a burglary turns up one argument indicating that the burglar was lefthanded and another argument indicating that the burglary was an inside job. Suppose these two arguments are intuitively independent, in the sense that they involve different uncertainties and that the evaluation of each depends on a different small world of experience. Say the argument for the burglar being left-handed is based on smudges on the door of the safe, and thus depends for its evaluation on the detective's experience and insight into the question of how safes are forced open, whereas the argument for the burglary being an inside job is based on the detective's understanding of the possibilities

[5] This rule was discussed by James Bernoulli in his *Ars Conjectandi*, published posthumously in 1713. Bernoulli also gave several other rules for combining probabilities based on independent arguments. Since most of these rules are special cases of Dempster's rule of combination, Bernoulli can be regarded as the founder of the theory of belief functions. Though Bernoulli's account of the combination of arguments was popular during the 17th century, it was eventually displaced by the Bayesian account developed by Condorcet and Laplace. See pp. 345–349 of Shafer [22] and pp. 465–469 of Pearson [19].

[6] We should bear in mind that chance is never an objective fact but is always an abstract picture that we impose on nature to aid our understanding. Stochastic independence, in particular, is an abstract concept that we use to model situations where we have first perceived a causal or intuitive independence.

for entering the building. It might, in such a case, be quite reasonable for the detective to treat the two arguments as if they were stochastically independent randomly coded messages. It is not that his train of thought in forming each argument is an independent chance process and that he knows the chance that each process has to produce a valid result; it is just that he can evaluate his confidence in each argument by comparing it with the scale of randomly coded messages and he can judge that there is no important common element in the uncertainties in the two arguments.

We might, of course, challenge the detective's judgment. We might discover a soft spot which is common to both arguments and which the detective failed to notice—perhaps he is too readily ruling out some hypothesis that could explain both the smudge on the safe door and an unnoticed entry into the building. But the possibility of challenge is not peculiar to judgments of independence. Every probability judgment is open to challenge.

One point that emerges from this example is that the idea of independence applies not to isolated facts or propositions but to whole small worlds of experience and human interaction with experience. When we explain what arguments we are combining, it is natural to identify each by a proposition: Argument 1 = "there were smudges on the door of the safe;" Argument 2 = "the building was being watched." But these propositions are only tags. We are really combining whole "bodies of evidence"—whole bodies of concrete experience and interactive human evaluation of that experience.

It is inherent in the idea of analyzing our evidence into independent arguments that the force of each argument is evaluated in abstraction from the other arguments. Each argument is evaluated, that is to say, in abstraction from the other evidence bearing on its conclusions. But when we combine arguments we must take the interaction of the conclusions into account—we must take into account whether the arguments concur, what they support when they are combined, and whether they conflict, either in pairs or in more complicated interactions. Since conflict modifies our evaluation of the weight of the arguments (through the renormalizing constant K in (3)) even when the conflict is not on a point of substantive interest to us, we must take all conflict in conclusions into account. So we should apply Dempster's rule to belief functions representing different arguments only if the frame Θ over which these belief functions are defined is fine enough to take all conflict and other relevant interaction into account.

So we have two requirements for the use of Dempster's rule of combination: (i) The bodies of evidence must be entirely distinct. The uncertainties in the arguments being combined, that is to say, must be independent when the arguments are viewed abstractly—i.e., before the interactions of their conclusions are taken into account. (ii) The frame Θ must be fine enough to discern all relevant interaction of the conclusions.

5.2 Is There an Objective Criterion for Independence?

Peter Williams is not satisfied with the preceding explanation of the conditions for the legitimate use of Dempster's rule of combination. It is not clear, he tells us,

> that this formulation is sufficient to distinguish unambiguously between permissible and impermissible applications of the rule. To begin with, the identity criteria for bodies of evidence are unclear if these cannot be expressed as propositions. Indeed, even if they can be, do two propositions which are not logically equivalent, but are nonetheless equivalent by virtue of natural laws, express 'entirely distinct bodies of evidence'? Or again, suppose that two bodies of evidence are distinct, taken as wholes, but nonetheless partly overlap. ... [H]ow is one to extract the common part, given that bodies of evidence are not necessarily expressible as propositions?

In this passage Williams seems to be demanding some objective criterion for deciding when two bodies of evidence are independent and, more generally, some mechanical way of analyzing evidence into distinct or independent items. Do these demands make sense?

It seems to me that the idea of an objective criterion for the independence of evidence—the idea of a criterion exterior to the judgment—is a chimera. The judgment that two bodies of evidence are independent is a probability judgment, and the appropriateness of probability judgments can never be justified on the basis of criteria that do not themselves demand the application of judgment.

The analysis of evidence into distinct and independent arguments is, moreover, always a constructive act of judgment. Williams is quite right to suggest that there is no unambiguous formula telling us how to do it. It is usually the most creative and the most difficult part of our effort to understand a problem.

There is, in short, no royal road. The analysis of evidence is difficult, and foolish mistakes are always possible. As James Bernoulli put it, "many things will happen which can cause one to err frequently and shamefully unless one proceeds cautiously in discerning arguments. For sometimes arguments can seem distinct which are in fact one and the same argument. Or, vice versa, those which are distinct can seem identical..." (See p. 337 of Shafer [22]).

As Williams' comments indicate, one concomitant of the desire for a mechanical approach to the analysis of evidence is a desire to express evidence as sentences or as propositions. If we could translate all our evidence into statements of fact, then we could, it would seem, give rules for mechanically analyzing this evidence using symbolic logic together with background knowledge encoded as prior probabilities. But we cannot usually translate our evidence into statements of fact.

We can always describe our evidence, the reader may protest. This is true. But the description will usually have to include not only statements of fact but also statements of probability judgment. How might the detective describe the evidence that convinces him that a person cannot enter the building without being seen by the watchman? The evidence consists, in a very real sense, of mental experiments that the detective carried out on the scene. He tried everything he could think of, and nothing seemed plausible. Perhaps he can describe some of this mental experimentation—at least if you allow him to draw pictures. But how can he reduce his conviction that a certain trick will not work to statements of fact? How can he formulate statements of fact to express his degree of conviction that he has tried everything? In the end he will simply have to supplement his statements of fact with probability judgments.[7]

5.3 Independence in the Bayesian Theory

The Bayesian theory can combine intuitively independent items of evidence, but it does not do so, as the theory of belief functions does, by regarding each as an independent argument. Instead it asks us to think of the occurrence of each item of evidence as a random event and to assess the probabilities of these events under various hypotheses. And it asks us to model the intuitive independence of the different items of evidence by stochastic independence, conditional on the various hypotheses, of the events that these items of evidence will occur.

The idea is that we should single out certain items of evidence and then imagine ourselves assessing, before these items of evidence occur, both the probabilities of the hypotheses on which we want to bring these items of evidence to bear and also the probability that these items of evidence will occur, given each of the hypotheses. Suppose, for example, that we are considering an exhaustive list of mutually exclusive hypotheses $H_1, \ldots, H_k$ and we single out two items of evidence E_1 and E_2. Then our task is to use "old evidence" (evidence other than the occurrence of E_1 and E_2) to construct Bayesian probabilities $P(H_i)$ and $P\left(E_1 \text{ and } E_2 | H_i\right)$. And if we judge E_1 and E_2 to be like independent random events given H_i—if, that is to say, our old evidence together with knowledge of H_i can be compared to knowledge that E_1 and E_2 are stochastically independent—then we can construct $P\left(E_1 \text{ and } E_2 | H_i\right)$ by making separate probability judgments $P\left(E_1 | H_i\right)$ and $P\left(E_2 | H_i\right)$ and then setting

$$P\left(E_1 \text{ and } E_2 | H_i\right) = P\left(E_1 | H_i\right) P\left(E_2 | H_i\right). \tag{21}$$

[7] In another passage, Williams coments on my insistence on the "hazy and non-propositional nature of evidence." While standing by the claim that evidence cannot usually be reduced to statements of fact, I would like to withdraw any suggestion (see, for example, p. 120 of *A Mathematical Theory of Evidence*) that evidence is "vague" or "hazy." These epithets are themselves vague, and no useful idea is conveyed when they are applied to evidence. (Cf. Austin [1], pp. 125–127.)

Notice that making all these probability judgments amounts to constructing a Bayesian probability distribution P over a certain frame of discernment Θ. We can suppose, indeed that the H_i and $\overline{E_i}$ are subsets of this frame, and that the $4k$ subsets $H_i \cap E_1 \cap E_2$, $H_i \cap E_1 \cap \overline{E_2}$, $H_i \cap \overline{E_1} \cap E_2$ and $H_i \cap \overline{E_1} \cap \overline{E_2}$ are disjoint and each contain exactly one element.

The point of constructing this probability distribution P is that we may then take the "new evidence"

$$E_1 \text{ and } E_2 = E_1 \cap E_2$$

into account by conditioning. We can calculate, in particular, the probability

$$P\left(H_i|E_1 \cap E_2\right) = \frac{P\left(H_i\right) P\left(E_1|H_i\right) P\left(E_2|H_i\right)}{\sum\limits_{j=1}^{k} P\left(H_j\right) P\left(E_1|H_j\right) P\left(E_2|H_j\right)}, \tag{22}$$

our probability for H_i based on the total evidence. Formula (22) is known as *Bayes' Theorem.*

Consider, for example, the detective who has evidence that the burglar was lefthanded and evidence that the burglary was an inside job. Give names to these two items of evidence—say E_1 and E_2. The propositions of substantive interest are

$$I = \text{ an insider was involved in the burglary,}$$

and

$$L = \text{ the safe was opened by a left-hander,}$$

and so the hypotheses are $H_1 = I \cap L$, $H_2 = I \cap \overline{L}$, $H_3 = \overline{I} \cap L$, and $H_4 = \overline{I} \cap \overline{L}$. And formula (22) provides a way of constructing probability judgments concerning the H_i using the total evidence.

We must always ask, of course, whether the independent judgment (21) is reasonable. Is it reasonable to think of the evidence E_1 involving the smudge on the safe and the evidence E_2 involving access to the building as random events that are stochastically independent given the H_i?

A more fundamental question is whether it is reasonable or helpful to think of E_1 and E_2 as random events at all. In our belief-function analysis we regarded E_1 and E_2 as arguments involving independent uncertainties. Here the perspective is different. Here we think of E_1 and E_2 not as arguments but as facts. And we transfer all the uncertainties to the hypothetical question of whether these facts would have occurred, given each of the hypotheses. But does this make sense? Can we, for example, intelligibly translate the question of how strongly E_2, the detective's study of access to the building supports I into the question of how likely his study would have been to turn out as it did, given that I is true and given that it is false?

In my opinion, we often cannot intelligibly translate our understanding of the significance of given evidence into answers to the question of how likely that evidence would be to occur. And this, I believe, is the fundamental objection to the version of the Bayesian theory that would have us assess all new evidence using Bayes' theorem. For a detailed discussion, see Shafer [24].

It should be noted, in any case, that the Bayesian theory, like the theory of belief functions, has no objective criterion for independence. In both theories the judgment that two items of evidence should be treated as independent is itself a probability judgment.[8]

5.4 Dependent Evidence?

Bayesian assessment of two items of new evidence does not necessarily require a judgment that the items are conditionally independent. Even if E_1 and E_2 are judged dependent, we can still construct the probability judgment $P\left(E_1 \cap E_2 | H_i\right)$ through the formula

$$P\left(E_1 \cap E_2 | H_i\right) = P\left(E_1 | H_i\right) P\left(E_2 | E_1 \cup H_i\right),$$

where $P\left(E_2 | E_1 \cap H_i\right)$ is a judgment as to how likely E_2 would be to occur based on the old evidence together with knowledge that E_1 has occurred and that H_i is true. And thus we can still use Bayes' theorem, in the form

$$P\left(H_i | E_1 \cap E_2\right) = \frac{P\left(H_i\right) P\left(E_1 | H_i\right) P\left(E_2 | E_1 \cap H_i\right)}{\sum_{j=1}^{k} P\left(H_j\right) P\left(E_1 | H_j\right) P\left(E_2 | E_1 \cap H_j\right)}.$$

So if we do use the Bayesian idea of assessing new evidence in terms of its likelihood to occur, it is not very important whether two items of evidence are independent or not.

The independence of different items of evidence is much more important in the theory of belief functions. For Dempster's rule of combination can be used to combine arguments only if those arguments are judged independent.

There seems to be a paradox here. The Bayesian theory can be understood as a special case of the theory of belief functions, and then Bayesian conditioning is seen as a special case of Dempster's rule of combination. (See p. 20 of *A Mathematical Theory of Evidence.*) But how can Bayesian conditioning be a special case of Dempster's rule if it can be used with dependent evidence and Dempster's rule cannot be?

The paradox is quickly resolved when we remind ourselves that "independence" does not have the same meaning in the two theories. The fact is that two items of evidence that are taken into account by conditioning are necessarily independent in the sense of the theory of belief functions, even though they may be either independent or dependent in the sense of the Bayesian theory.

[8] Seidenfeld [20] seems to think otherwise. The Bayesian theory, he writes, "provides the machinery for deciding whether the data are mutually independent." What machinery?

Let us recall the relation between conditioning and Dempster's rule. We explained conditioning in Sect. 3 above by saying that we condition a belief function Bel on a subset E_1 of its frame Θ in order to take into account new evidence whose direct effect on the frame Θ is to establish for certain that the truth is in E_1. But we can also treat such new evidence as an argument for E_1 whose validity is certain and represent it by a belief function Bel_1 with m-values $m_1(E_1) = 1$ and $m_1(A) = 0$ for all other $A \subset \Theta$. And it is because combining Bel with Bel_1 by Dempster's rule gives the same result as conditioning Bel on E_1 that we say that conditioning is a special case of Dempster's rule.

Now consider a second item of new evidence whose direct effect on Θ is to establish for certain that the truth is in $E_2 \subset \Theta$. This evidence can be represented by a belief function Bel_2 with m-values $m_2(E_2) = 1$ and $m_2(A) = 0$ for all other $A \subset \Theta$. Are the uncertainties in the two new items of evidence independent? Yes, for there are no uncertainties; we are modeling each item of evidence as a randomly valid argument in which the chance of validity is one, and so stochastic independence is automatic and it is legitimate to combine Bel_1 and Bel_2 by Dempster's rule. When we do combine Bel_1 and Bel_2, we obtain a belief function $Bel_1 \oplus Bel_2$ that gives $E_1 \cap E_2$ the m-value one, and combining Bel with $Bel_1 \oplus Bel_2$ by Dempster's rule amounts to conditioning Bel on $E_1 \cap E_2$.

One way of putting the matter is to say that the only decompositions of evidence recognized by the Bayesian theory are decompositions into items of evidence that are, from the point of view of the theory of belief functions, independent. The Bayesian theory permits the combination of evidence only through conditioning, and this means that only one of the bodies of evidence being combined, the "old evidence," can involve uncertainties. The other items of evidence must amount to certainties relative to the frame Θ and hence will be trivially independent of each other and of the old evidence.

When we assign names ("E_1" and "E") to new items of evidence and incorporate them into our frame of discernment, we are, in effect, reducing them from uncertain arguments to facts. We are stripping them of their uncertainties and putting all these uncertainties into what we call the "old evidence," the evidence on which the probability distribution P over the frame Θ must be based.

From the point of view of the theory of belief functions, the concentration of all our uncertainties in the "old evidence" does not, of course, solve the problem of probability judgment. Nor does it necessarily exhaust our interest in the combination of evidence. For we face a new problem of assessment of evidence, the problem of constructing a Bayesian probability distribution P (or, more generally, a belief function Bel) over the frame Θ based on this old evidence. And one way of doing this may be to decompose the old evidence into independent items that can be recombined by Dempster's rule.

It may deepen our understanding of the differences between the Bayesian and belief-function concepts of independence to recognize that Bayesian dependence of E_1 and E_2 may be compatible with belief-function independence not only of the items of evidence provided by the occurrence of E_1 and E_2 but also of the components of the old evidence that bear on E_1 and E_2. It is possible, that is to say, for the combination of belief functions representing intuitively independent components of the old evidence to produce a belief function over Θ which happens to be Bayesian and in terms of which E_1 and E_2 are dependent in the Bayesian sense. In fact, any Bayesian probability distribution P over Θ can, in theory, be produced by such a combination of belief functions.

5.5 Sorting out the Uncertainties

The preceding comments should not be construed as a denial of the practical problems that dependent arguments cause in the theory of belief functions. In many problems it will be easy to analyze the evidence into dependent arguments and more difficult to analyze it into independent arguments.

How do we go about analyzing our evidence into independent arguments? How, to put it another way, do we sort our evidence into arguments that involve distinct uncertainties? Perhaps there is no general answer to this question. But we can gain some insight by thinking about examples.

Suppose we are charged with deciding whether an aerial sprayer has allowed insecticide to drift onto the property of a neighboring landowner. Two arguments are presented by the prosecution: (1) The homeowner testifies that spray billowed across the road from the field being sprayed and settled onto her house and that this drift was significant enough to cause her and her family to suffer from headaches and burning eyes and lips. (2) A government bee inspector testifies that he found dead honey bees lying around the homeowner's beehive, that in his judgment they were killed by insecticide, and that the availability of flowering plants indicates that the bees must have been on the homeowner's property rather than on the field being sprayed when they were exposed.

Both items of evidence seem to directly support the charge of negligence. But one can argue that they involve overlapping uncertainties. The main uncertainties are distinct. The main uncertainty in the first item of evidence is how precise and reliable the homeowner is—how well she remembers and how much she exaggerates. The main uncertainty in the second item of evidence is the reliability of the bee inspector's judgment. But suppose the homeowner, out of pure malice, made up the story about drift and poisoned the bees herself. This possibility constitutes, it would seem, an uncertainty common to both items of evidence. And so if we take the possibility seriously we must count the two items as dependent.

There is, however, an obvious way of getting this common uncertainty out of the two items of evidence: incorporate it into the frame of discernment. We might, for example, consider a frame of discernment Θ consisting of three possibilities:

θ_1: The sprayer was not negligent; the homeowner was inaccurate, and the bee inspector was mistaken.

θ_2: The sprayer was not negligent; the homeowner is lying, and she poisoned the bees herself.

θ_3: The sprayer was negligent.

Relative to this frame of discernment we might describe our two items of evidence a little differently. The first item is our evidence for the reliability and probity of the homeowner (we have listened to her testify, etc.), and it supports θ_3 to some extent, and $\{\theta_1, \theta_3\}$ to a stronger extent. The second item is our evidence from the bee inspector, and it supports $\{\theta_2, \theta_3\}$. Notice that though the two items no longer both directly support negligence (θ_3), they still interact to support it. And they can now be regarded as independent arguments.

This example illustrates a reasonably general idea: often two arguments which seem dependent because of common uncertainties can be understood as independent once the common uncertainties are incorporated into the frame of discernment as explicit possibilities. This idea is the basis for saying that Dempster's rule should be used only when the frame "discerns the relevant interaction" of the different arguments.

The task of sorting our uncertainties into distinct arguments is not always so easy, of course. But I would argue that a theory that directs us to this task is grappling with the real problems in the assessment of evidence.

References

[1] J. L. Austin, 1962. *Sense and Sensibilia.*Oxford.

[2] A. P. Dempster, 1966. New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics*, 37:355–374.

[3] A. P. Dempster, 1968. A generalization of Bayesian inference (with discussion). *Journal of the Royal Statistical Society Series B*, 30:205–247.

[4] Persi Diaconis, 1978. Review of *A Mathematical Theory of Evidence. Journal of the American Statistical Association*, 73:677–678.

[5] Terrence L. Fine, 1977. Review of *A Mathematical Theory of Evidence. Bulletin of the American Mathematical Society*, 83:667–672.

[6] Bruno de Finetti, 1964. Foresight: Its logical laws, its subjective sources. In Kyburg and Smokler, editors, *Studies in Subjective Probability*, pp. 93–158. Wiley.

[7] Bruno de Finetti, 1972. *Probability, Induction, and Statistics.* Wiley.

[8] Bruno de Finetti, 1974. *Theory of Probability.* Vol. 1, Wiley.

[9] Bruno de Finetti, 1975. *Theory of Probability.* Vol. 2, Wiley.

[10] D. A. Freedman and R. A. Purves, 1969. Bayes' methods for bookies. *Annals of Mathematical Statistics*, 40:1177–1186.

[11] I. J. Good, 1962. The measure of a non-measurable set. In E. Nagel, P. Suppes, and A. Tarski, editors, *Logic, Methodology and Philosophy of Science*, pp. 319–329. Stanford University Press, Stanford.

[12] Ian Hacking, 1967. Slightly more realistic personal probability. *Philosophy of Science*, 34:311–325.

[13] Peter Huber, 1973. The use of Choquet capacities in statistics. *Bulletin of the International Statistical Institute*, 45, Book 4:181–188.

[14] Peter Huber, 1976. Kapazitäten statt Wahrscheinlichkeiten? Gedanken zur Grundlegung der Statistik. *Jber. Deutsch. Math. Verein*, 78(2):81–92.

[15] Oscar Kempthorne, 1975. Inference from experiments and randomization. In J. N. Srivastava, editor, *A Survey of Statistical Design and Linear Models*. North-Holland.

[16] Isaac Levi, 1981. Dissonance and consistency according to Shackle and Shafer. In *Proceedings of the Biennial Meeting of Philosophy of Science Association (PSA-78)*, Vol. 2. Philosophy of Science Association, East Lansing, Michigan. Asquith and Hacking, editors.

[17] Dennis V. Lindley, 1977. Review of *A Mathematical Theory of Evidence*. *Bulletin of the London Mathematical Society*, 9:237–238.

[18] James G. March. Bounded rationality, ambiguity, and the engineering of choice. *The Bell Journal of Economics*, 9:586–608, 1978.

[19] Karl Pearson, 1978. *The History of Statistics in the 17th and 18th Centuries*. Griffin.

[20] Teddy Seidenfeld, 1981. Statistical evidence and belief functions. In *Proceedings of the Biennial Meeting of Philosophy of Science Association (PSA-78)*, Vol. 2. Philosophy of Science Association, East Lansing, Michigan. Asquith and Hacking, editors.

[21] Glenn Shafer, 1976. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ.

[22] Glenn Shafer, 1978. Non-additive probabilities in the work of Bernoulli and Lambert. *Archive for History of Exact Sciences*, 19:309–370.

[23] Glenn Shafer, 1978. Dempster's rule of combination. Unpublished paper.

[24] Glenn Shafer, 1981. Two theories of probability. In *Proceedings of the Biennial Meeting of Philosophy of Science Association (PSA-78)*, Vol. 2. Philosophy of Science Association, East Lansing, Michigan. Asquith and Hacking, editors.

[25] Glenn Shafer, 1979. Lindley's paradox. Technical Report No. 125, Department of Statistics, Stanford University. (Later appeared in *Journal of the American Statistical Association*, 77:325–351, 1982.)

[26] Glenn Shafer, 1979. Allocations of probability. *Annals of Probability*, 7: 827–839.

[27] Glenn Shafer and A. M. Breipohl, 1979. Reliability described by belief functions. In *Proceedings of the 1979 Reliability and Maintainability Symposium*, pp. 23–27.

[28] C. A. B. Smith, 1961. Consistency in statistical inference and decision (with discussion). *Journal of the Royal Statistical Society Series B*, 23:1–25.

[29] Paul Teller, 1973. Conditionalization and observation. *Synthese*, 26:218–258.

[30] Peter M. Williams, 1975. Coherence, strict coherence and zero probabilities. *Contributed Papers*, Fifth International Congress of Logic, Methodology and Philosophy of Science, VI, 29,30.

[31] Peter M. Williams, 1976. Indeterminate probabilities. Pp. 229–246 of M. Przelecki, K. Szaniawski, and R. Wójciki, editors, *Formal Methods in the Methodology of Empirical Sciences*. Ossolineum and D. Reidel.

[32] Peter M. Williams, 1978. On a new theory of epistemic probability (Review of *A Mathematical Theory of Evidence*). *The British Journal for the Philosophy of Science*, 29:375–387.

[33] Guus Wolf, 1977. *Obere und Untere Wahrscheinlichkeiten*. PhD thesis, Eidgenössische Technische Hochschule, Zürich. (Diss. ETH 5884).

10

Belief Functions and Parametric Models

Glenn Shafer

Abstract. The theory of belief functions assesses evidence by fitting it to a scale of canonical examples in which the meaning of a message depends on chance. In order to analyse parametric statistical problems within the framework of this theory, we must specify the evidence on which the parametric model is based. This article gives several examples to show how the nature of this evidence affects the analysis. These examples also illustrate how the theory of belief functions can deal with problems where the evidence is too weak to support a parametric model.

Key words: Bayesian statistics, Belief functions, Constructive probability, Dempster's rule, Exchangeability, Fiducial inference, Lower probabilities, Likelihood, Parametric models

1 Constructive Probability

In *A Mathematical Theory of Evidence* (1976), I discussed the possibility that the mathematical structure for upper and lower probabilities that Dempster developed in his attempt to deal with parametric models might be used more widely as a structure for probability judgements. I suggested that we call set functions that have the structure of Dempster's lower probabilities *belief functions*, and I developed the implications of Dempster's rule for combining belief functions based on different bodies of evidence.

The central role of Dempster's rule of combination in the theory of belief functions is merely one aspect of the theory's emphasis on the decomposition and description of evidence. In general, the theory allows probability judgements to depend not only on the overall strength of the evidence on which they are based but also on the structure of that evidence.

In this paper I turn this general emphasis on evidence back onto the problem of parametric models. I argue that belief-function analyses of these models should depend not just on the models themselves but also on the nature of the evidence for them. I give several examples of this dependence.

Before taking up the problem of parametric models, I briefly review the theory of belief functions and its relation to other constructive theories of probability judgement.

The exposition that follows is based on the idea, first developed in work with Amos Tversky and subsequently published as Shafer and Tversky (1985) (see Chap. 13), that all theories of probability judgement, including both the theory of belief functions and the Bayesian theory, should be thought of in terms of canonical examples to which the theories compare evidence. For a further development of this theme, see Shafer (1981a, b).

1.1 Three Constructive Theories

Probability judgement, like all judgement, involves comparison. In order to judge whether given evidence makes something practically certain, very probable, fairly probable, or not at all probable, say, we must compare this evidence to examples where it is agreed that these adjectives fit. We must, in other words, fit our evidence to a scale of canonical examples. Numerical probability judgement similarly involves fitting our evidence to a scale of canonical examples. Different choices of this scale produce different constructive theories of probability.

Here are three such theories.

The Bayesian theory. Suppose our scale consists exclusively of examples where the truth is generated according to known chances. Then when we make a probability judgement $P(A) = p$ we are saying that our evidence provides support for A comparable to what would be provided by knowledge that the truth is generated by a chance set-up that produces a result in A exactly p of the time. And these probability judgements will obey the usual Bayesian rules.

If we are working with a set of possibilities Ω, then our scale of canonical examples will include, for each chance distribution P over Ω, an example where the truth is generated according to the chances given by P. Usually we will not, of course, be able to fit our evidence to this scale by means of a single holistic judgement. Instead we will break the overall comparison down into many simpler comparisons and then construct P from these simpler judgements.

Lower probabilities. Suppose we know that a certain process is governed by chance, but instead of knowing precisely the chance distribution P governing it, we know only that P is in a class $\mathcal{P}$ of chance distributions. Denote by Ω the set of possible outcomes for the process. Then we might set our probability or degree of belief that the outcome of a particular trial will be in a subset A of Ω equal to

$$P_*(A) = \inf\{P(A)|P \in \mathcal{P}\}.$$

This seems natural because we know the chance of A is at least $P_*(A)$. Notice that the probabilities or degrees of belief obtained in this way will, in general, be non-additive: $P_*(A)$ and $P_*(\bar{A})$ may add to less than one.

By varying the class $\mathcal{P}$ in this story we obtain a scale of canonical examples. Let us call the constructive theory that uses this scale the theory of lower probabilities.

Belief functions. Suppose someone chooses a code at random from a list of codes, uses the code to encode a message, and then sends us the result. We know the list of codes and the chance of each code being chosen—say the list is $c_1, \ldots, c_n$, and the chance of c_i being chosen is p_i. We decode the encoded message using each of the codes and find that this always produces a message of the form "the truth is in A" for some non-empty subset A of the set of possibilities Ω. Let A_i denote the subset we get when we decode using c_i, and set

$$m(A) = \sum \{p_i | 1 \leq i \leq n; A_i = A\} \tag{1}$$

for each $A \subset \Omega$. The number $m(A)$ is the sum of the chances for those codes that indicate A was the true message; it is, in a sense, the total chance that the true message was A. Notice that $m(\phi) = 0$ and that the $m(A)$ sum to one. The quantity

$$\mathrm{Bel}(A) = \sum_{B \subset A} m(B) \tag{2}$$

is, in a sense, the total chance that the true message implies A. If the true message is infallible and the coded message is our only evidence, then it is natural to call Bel (A) our probability or degree of belief that the truth lies in A.

A function Bel is called a *belief function* if it is of the form (2), with the $m(A)$ non-negative and summing to one and with $m(\phi) = 0$. The subsets A of Ω for which $m(A) > 0$ are called the *focal elements* of the belief function.

It is easily seen from (2) that Bel $(A)+\mathrm{Bel}(\bar{A}) \leq 1$, or Bel $(A) \leq 1-\mathrm{Bel}(\bar{A})$. The quantity $1-\mathrm{Bel}(\bar{A})$ is called the *plausibility* of A and is denoted by $Pl(A)$; it can be large even if the evidence for A is slight, provided that the evidence against A is also slight.

The equation $\mathrm{Bel}(A)+\mathrm{Bel}(\bar{A}) = 1$, which is equivalent to $\mathrm{Bel}(A) = Pl(A)$, holds for all subsets A of Ω if and only if Bel's focal elements are all singletons. In this case, Bel is an additive probability distribution.

We can tell the story of the coded message with any values for the $m(A)$ we please. So this story provides a canonical example corresponding to each possible belief function. It is sometimes helpful to vary the story slightly; what is essential is that some chance experiment with outcomes $c_1, \ldots, c_n$ has been carried out, that we know these outcomes had chances $p_1, \ldots, p_n$, and that we receive a message that means A_i if c_i was the outcome.

1.2 Elements of the Theory of Belief Functions

Belief functions, we have suggested, are obtained by fitting evidence to a certain scale of canonical examples. In order to turn this idea into a practical tool, we need rules for breaking the fitting task down into simpler judgements, and techniques for making these simpler judgements feasible. Here we will review some of these rules and techniques for the case where Ω is finite. For an introduction to the case where Ω is infinite, see Shafer (1979).

The vacuous belief function. Consider the belief function Bel obtained by setting $m(\Omega) = 1$ and $m(A) = 0$ for every proper subset A of Ω. We see by (2) that Bel also satisfies Bel $(A) = 0$ for every proper subset A; Bel indicates no positive beliefs at all as to where in Ω the truth lies. This belief function is appropriate when the evidence being considered does not, by itself, tell us anything about which element of Ω is the truth.

Simple support functions. Consider the following variation on the story of the randomly coded message. A certain mechanism that produces messages has two modes of operation: reliable and unreliable. It is in its reliable mode with chance p_1, and then it produces only true messages. It is in its unreliable mode with chance $p_2 = 1 - p_1$, and then it is completely unpredictable; we have no idea whether or how often the messages it produces will be true or false. Suppose this mechanism produces the message that the truth is in the subset of E of Ω. Then we will say that the message has a chance p_1 of meaning E and a chance p_2 of meaning nothing—i.e. meaning only that the truth is in Ω. And so we will adopt a belief function with focal elements E and Ω, with $m(E) = p_1$ and $m(\Omega) = p_2$. This belief function, given by

$$\mathrm{Bel}(A) = \begin{cases} 0 & \text{if } E \not\subset A, \\ p_1 & \text{if } E \subset A \neq \Omega, \\ 1 & \text{if } A = \Omega, \end{cases}$$

is called a *simple support function.*

It is often natural to compare evidence to a mechanism that is only sometimes reliable and thus to represent it by a simple support function. The reliability of a witness can obviously be taken into account in this way. The strength of an argument can often be assessed in the same way; this means we compare the argument to one that has a definite and known chance of being reliable.

Dempster's rule of combination. One of the basic strategies of the theory is to decompose our evidence into two or more unrelated bodies of bodies of evidence, make probability judgements separately on the basis of each of these bodies of evidence, and then combine these judgements by Dempster's rule. This rule tells us how to combine a belief function Bel_1 representing one body of evidence with a belief function Bel_2 representing an unrelated body of evidence so as to obtain a belief function $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ representing the pooled evidence. The rule is most easily stated in terms of m-values: If the m-values for Bel_1 and Bel_2 are denoted by $m_1(A)$ and $m_2(B)$, respectively, then $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ is the belief function with m-values $m(C)$, where $m(\phi) = 0$ and

$$m(C) = \frac{\sum\{m_1(A)m_2(B) \mid A \subset \Omega; B \subset \Omega; A \cap B = C\}}{\sum\{m_1(A)m_2(B) \mid A \subset \Omega; B \subset \Omega; A \cap B \neq \phi\}} \tag{3}$$

for all non-empty subsets C of Ω. (Notice that the focal elements of $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ consist of all the non-empty intersections of focal elements of Bel_1 with focal elements of Bel_2.)

The idea underlying Dempster's rule is that the unrelatedness of two bodies of evidence makes pooling them like combining two stochastically independent randomly coded messages. Suppose Bel_1 and Bel_2 do correspond to two such messages. Denote by $c_1, \ldots, c_n$ and $p_1, \ldots, p_n$ the codes and their chances in the case of the first message, and by $c'_1, \ldots, c'_m$ and $p'_1, \ldots, p'_m$ the codes and their chances in the case of the second. Independence means that there is a chance $p_i p'_j$ that the pair (c_i, c'_j) of codes will be chosen. But decoding may tell us something. If the message A_i we get by decoding the first message with c_i contradicts the message B_j we get by decoding the second message with c'_j (i.e. if $A_i \cap B_j = \phi$), then we know that (c_i, c'_j) cannot be the pair of codes actually used. So we condition the chance distribution, eliminating such pairs and multiplying the chances for the others by K, where

$$\begin{aligned} K^{-1} &= \sum \left\{ p_i p'_j \mid 1 \le i \le n; 1 \le j \le m; A_i \cap B_j \ne \phi \right\} \\ &= \sum \left\{ m_1(A) m_2(B) \mid A \subset \Omega; B \subset \Omega; A \cap B \ne \phi \right\}. \end{aligned}$$

If the first message is A and the second message is B, then the overall message is $A \cap B$. So the total chance of the overall message being C is

$$\begin{aligned} m(C) &= K \sum \left\{ p_i p'_j | 1 \le i \le n; 1 \le j \le m; A_i \cap B_j = C \right\} \\ &= K \sum \left\{ m_1(A) m_2(B) | A \subset \Omega; B \subset \Omega; A \cap B = C \right\}, \end{aligned}$$

which is indeed equal to (3).

We may call $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ the "orthogonal sum" of Bel_1 and Bel_2. Here are some elementary properties of the operation $\oplus$: (i) $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ exists unless there is a subset A of Ω such that $\mathrm{Bel}_1(A) = 1$ and $\mathrm{Bel}_2(\bar{A}) = 1$. (ii) Commutativity: $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2 = \mathrm{Bel}_2 \oplus \mathrm{Bel}_1$, (iii) Associativity: $(\mathrm{Bel}_1 \oplus \mathrm{Bel}_2) \oplus \mathrm{Bel}_3 = \mathrm{Bel}_1 \oplus (\mathrm{Bel}_2 \oplus \mathrm{Bel}_3)$. (iv) In general: $\mathrm{Bel} \oplus \mathrm{Bel} \ne \mathrm{Bel}$; $\mathrm{Bel} \oplus \mathrm{Bel}$ will favour the same subsets as Bel but with, as it were, twice the weight of evidence. (v) If Bel_1 is Bayesian, then so is $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$. (vi) If Bel_1 is vacuous, then $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2 = \mathrm{Bel}_2$.

Dempster's rule can be seen as a generalization of rules formulated in the eighteenth century by James Bernoulli and Johann Heinrich Lambert. (See Shafer, 1978.)

Conditioning. Consider evidence which establishes conclusively that the truth is in a subset E of Ω but which does not tell us anything more specific. Such evidence can be compared to a randomly coded message which has chance one of meaning E, and so we can represent it by a belief function Bel_E whose m-value for E is one. The values of Bel_E are

$$\mathrm{Bel}_E(A) = \begin{cases} 0 & \text{if } A \not\supset E, \\ 1 & \text{if } A \supset E. \end{cases}$$

An important property of Bel_E is its idempotence: $\mathrm{Bel}_E \oplus \mathrm{Bel}_E = \mathrm{Bel}_E$.

If Bel is a belief function satisfying $\mathrm{Bel}(E) < 1$, then $\mathrm{Bel} \oplus \mathrm{Bel}_E$ exists. It is natural to call $\mathrm{Bel} \oplus \mathrm{Bel}_E$ the result of conditioning Bel on E and to denote $(\mathrm{Bel} \oplus \mathrm{Bel}_E)(A)$ by $\mathrm{Bel}(A|E)$. Notice that conditioning an orthogonal sum is equivalent to conditioning each term in the sum before combining: since Bel_E is idempotent,

$$(\mathrm{Bel}_1 \oplus \mathrm{Bel}_2) \oplus \mathrm{Bel}_E = (\mathrm{Bel}_1 \oplus \mathrm{Bel}_E) \oplus (\mathrm{Bel}_2 \oplus \mathrm{Bel}_E).$$

The process of conditioning can be described directly in terms of focal elements: to condition Bel on E, reduce the focal elements of Bel to their intersections with E and then renormalize the m-values to take into account the elimination of those focal elements that have been reduced to ϕ. If Bel is an additive probability distribution, then this reduces to the usual Bayesian rule of conditioning.

Minimal extension. Suppose the set of possibilities Ω has n elements: $\Omega = \{\omega_1, \ldots, \omega_n\}$. And suppose Λ is a finer set of possibilities. This means that the elements $\omega_1, \ldots, \omega_n$ of Ω correspond to a partition $E_1, \ldots, E_n$ of Λ: "ω_i is the truth" means the same as "the truth is in E_i", and, more generally, a subset $\{\omega_{i_1}, \ldots, \omega_{i_k}\}$ of Ω has the same meaning as the subset $E_{i_1} \cup \ldots \cup E_{i_k}$ of Λ.

Given a belief function Bel over Λ we can speak of its *marginal* over Ω: the belief function $\mathrm{Bel}|\Omega$ given by

$$(\mathrm{Bel}|\Omega)(\{\omega_{i_1}, \cdots, \omega_{i_k}\}) = \mathrm{Bel}(E_{i_1} \cup \ldots \cup E_{i_k}).$$

Marginalization can be described in terms of focal elements by saying that a focal element A of Bel is reduced to the subset $\{\omega_i | E_i \cap A \neq \phi\}$ of Ω. In general, there will be many belief functions over Λ having a given marginal over Ω. Or, to put the matter another way, a belief function over Ω will extend in many ways to a belief function over Λ.

Suppose we use a given body of evidence to construct a belief function Bel_0 over Ω. And suppose we judge that this evidence bears on the questions discerned by Λ only insofar as it bears on those already discerned by Ω. In terms of the randomly coded message to which we are comparing our evidence, this says that if $\{\omega_{i_1}, \ldots, \omega_{i_k}\}$ is the meaning of the message relative to Ω, then $E_{i_1} \cup \ldots \cup E_{i_k}$ is its meaning relative to Λ. This suggests that Bel_0 should be extended to the belief function Bel over Λ whose m-values are given by $m(E_{i_1} \cup \ldots \cup E_{i_k}) = m_0(\{\omega_{i_1}, \ldots, \omega_{i_k}\})$ and $m(A) = 0$ for all $A \subset \Lambda$ which are not unions of elements of the partition $E_1, \ldots, E_n$. This belief function Bel does have Bel_0 as its marginal. And for each $A \subset \Lambda$, $\mathrm{Bel}(A)$ is less than or equal to the degree of belief given to A by any other extension of Bel_0 to Λ. So we call Bel the *minimal extension* of Bel_0.

Conditional embedding. Sometimes we rule out some of the possibilities in a set of possibilities Λ, thus reducing it to a smaller set of possibilities $\Omega \subset \Lambda$. If we have constructed a belief function Bel over Λ and we then reduce Λ to Ω

because of new evidence that establishes that the truth is in Ω without saying anything more specific, then we will, of course, replace Bel by its *conditional given* Ω—i.e. by the belief function over Ω that assigns to each $A \subset \Omega$ the degree of belief Bel $(A|\Omega)$. In general, there will be many belief functions over Λ having a given conditional given Ω.

Suppose we begin by taking it for granted that the truth is in Ω and construct a belief function Bel_0 over Ω, but we later decide that all the elements of Λ must be admitted as possibilities. And suppose we judge that the evidence on which Bel_0 is based does not impugn any of the possibilities in $\Lambda - \Omega$. In terms of the randomly coded message to which we are comparing the evidence, this means that if $A \subset \Omega$ is the meaning of the message relative to Ω, then $A \cup (\Lambda - \Omega)$ is its meaning relative to Λ. This suggests that Bel_0 should be replaced by the belief function Bel over Λ whose m-values are given by $m(A \cup (\Lambda - \Omega)) = m_0(A)$ for all $A \subset \Omega$ and $m(A) = 0$ for all subsets A of Λ that do not contain $\Lambda - \Omega$. This belief function Bel has Bel_0 as its conditional given Ω. And for each $A \subset \Lambda$, $\text{Bel}(A)$ is less than or equal to the degree of belief given to A by any other belief function over Ω that has Bel_0 as its conditional given Ω. We call Bel the conditional embedding of Bel_0 in Λ.

The idea of conditioning embedding was first developed by Smets (1978).

Discounting. Suppose that after observing a randomly coded message and calculating the belief function Bel by (1) and (2) we discover that our understanding of the process producing the message is not fully reliable; say there is a chance $1-\alpha$ that our understanding is correct, so that the message is indeed the result of choosing among the codes $c_1, \ldots, c_n$ with chances $p_1, \ldots, p_n$, but a chance α that the message was produced in some other way about which we know nothing and must therefore be counted as meaning nothing. Then we must change the chance associated with the code c_i from p_i to $(1-\alpha)p_i$, and we must, in effect, introduce a new "code" that is used with chance α and which decodes any message to the non-informative statement that the truth is in Ω. This means reducing each m-value $m(A)$ to $(1-\alpha)m(A)$ and then increasing the m-value for Ω by α. The result is a belief function Bel^α related to Bel by $\text{Bel}^\alpha(A) = (1-\alpha)\text{Bel}(A)$ for all proper subsets A of Ω. ($\text{Bel}^\alpha(\Omega) = \text{Bel}(\Omega) = 1$, of course.) We say that Bel^α is the result of *discounting* Bel. Discounting is the natural way to take account of doubts or second thoughts about belief functions constructed by ourselves or others.

1.3 The Constructive View of Probability

By saying that probability judgements are made by fitting given evidence to a scale of canonical examples, we are able to bring together two ideas that have sometimes been set up in opposition to one another: the idea that probabilities are subjective judgements, and the idea that probabilities can be based on a limited body of evidence.

The idea that probability judgements can be based on limited evidence is essential, of course, to a proper understanding of the theory of belief functions.

Ultimately, we are always interested in judgements based on our total evidence. But the motivation for using Dempster's rule of combination is the idea that we might gain in clarity of thought by weighing different items of evidence separately before thinking about how they reinforce or contradict each other.

I do not wish to suggest that the idea of basing subjective probability judgements on limited evidence is utterly new. But consider the typology of views on the interpretation of probability that Savage presents in *The Foundation of Statistics* (1954, p. 3). Savage distinguishes three main classes of views: objectivistic, personalistic and necessary. Objectivistic views hold that probability is an objective property of certain repetitive events; personalistic views hold that probability measures the confidence that a particular individual has in the truth of a particular proposition; necessary views hold that probability measures the extent to which one set of propositions, out of logical necessity and apart from human opinion, confirms the truth of another. This typology obviously does not accommodate the idea of probability judgement based on limited evidence. Personalistic views focus on the attitudes a person actually has towards a proposition, and these attitudes are presumably based on his total evidence. Necessary views allow us to delimit the evidence, but they insist that this evidence be cast in the form of propositions, and they exclude any role for judgement in assessing it.

I would like to suggest that our *constructive* view of probability—the view that probability judgement amounts to fitting given evidence to a scale of canonical examples—should be recognized as a fourth view of probability, distinct from and on a par with the objectivistic, personalistic and necessary views.

2 Generalizations of Bayesian Parametric Inference

Let us adopt the now standard general notation for parametric statistical models: Θ denotes the set of possible values for the parameter, $\theta, \mathcal{X}$ denotes the set of possibilities for the data x and $\{P_\theta : \theta \in \Theta\}$ denotes the model. How do we make probability judgements about θ after observing x?

The Bayesian answers this question by representing prior evidence about θ by an additive probability distribution P_0 over Θ and by using this distribution, together with the P_θ, to construct a distribution, say P, over $\Theta \times \mathcal{X}; P$ is the unique probability distribution over $\Theta \times \mathcal{X}$ that has P_0 as its marginal for θ and the P_θ as its conditionals given θ. Once the Bayesian has observed x, he will condition P on x to obtain posterior probabilities for θ.

How should our constructive generalizations of the Bayesian theory generalize this Bayesian treatment of parametric statistical inference?

Lower probabilities. The natural lower-probability generalization is to replace the prior distribution P_0 by a class $\mathcal{P}_0$ of additive probability distributions. This leads in turn to a class $\mathcal{P}$ of additive probability distributions over $\Theta \times \mathcal{X}$, and conditioning this class on the observed data x gives posterior lower probabilities for θ. The weakness of this approach is that if $\mathcal{P}_0$ is a reasonably broad class, then the posterior lower probabilities are not very informative. If, for example, we judge that we have no cogent prior evidence about θ and so allow $\mathcal{P}_0$ to be the class of all additive probability distributions on Θ, then our posterior degrees of belief will not indicate any evidence for any proper subset of the set of P_θ which are possible in light of the observed data. For a review of the literature on this lower-probability approach to parametric inference, see DeRobertis (1978).

Belief functions. Suppose we represent our prior evidence about θ by a belief function Bel_0 over Θ. Then it seems natural to generalize the Bayesian approach by asking for a belief function over $\Theta \times \mathcal{X}$ that has Bel_0 as its marginal for θ and P_θ as its conditional given Θ. Such a belief function could then be conditioned on the observed data x to yield a posterior belief function over Θ.

This line of thought brings us immediately to a fundamental difference between additive probability measures and more general belief functions: a belief function is not, in general, uniquely determined by its marginal for a given partition and its conditionals given elements of that partition. There may be many belief functions over $\Theta \times \mathcal{X}$ having a given marginal Bel_0 and given conditionals P_θ. And there may be no reason to prefer one to the others. In his original work on "generalized Bayesian inference", Dempster (1968) proposed a particular method of constructing a belief function with a given marginal Bel_0 and given conditionals P_θ, but both he and his critics were uncomfortable with the seemingly arbitrary character of the method. (There are general principles from which Dempster's method can be derived (see Shafer, 1976b) but I now believe the method is appropriate only in the case where the evidence about a random experiment is limited to evidence for its randomness; see Sect. 4 below.)

But it is no embarrassment to the general theory of belief functions that a belief function is not fully determined by a given marginal and corresponding conditionals. Belief functions are not meant, in general, to be constructed from such elements. They are meant to be constructed from analyses of evidence. And so long as we are working within the theory of belief functions we expect to represent individual items of evidence by belief functions, not by objects like conditional belief functions or parametric models.

So the general spirit of the theory of belief functions leads us to look beyond the parametric model $\{P_\theta : \theta \in \Theta\}$ to the evidence on which the model is based. Our goal should be to represent this evidence directly by a belief function over $\Theta \times \mathcal{X}$, and it will be this belief function, say Bel, that we should regard as a full account of the effect of this evidence on $\Theta \times \mathcal{X}$. The model $\{P_\theta : \theta \in \Theta\}$ will be only a partial account: P_θ will be Bel's conditional given θ.

Once we have constructed such a belief function, Bel, we can take the prior evidence about θ into account by combining Bel with Bel_0's minimal extension to $\Theta \times \mathcal{X}$, which we may denote by $\overline{\text{Bel}_0}$. If the evidence for the parametric model does not by itself give any indication as to the value of θ (so that Bel's marginal for θ is vacuous), resulting belief function Bel $\oplus$ Bel_0 Bel_0 will satisfy the conditions formulated above: Bel_0 will be its marginal for θ, and P_θ will be its conditional given θ.

3 Some Examples of Evidence for Parametric Models

Here we shall consider three possible ways a parametric model $\{P_\theta : \theta \in \Theta\}$ might arise:

(1) Perhaps the values of the parameter θ have a substantive significance, and our knowledge of each P_θ derives from actual observations, the observations affording our knowledge of one P being distinct and independent of those affording our knowledge of another. In a problem of medical diagnosis, for example, each θ might correspond to the hypothesis that the patient has a particular disease, with P_θ giving the frequencies with which that disease has been observed to give rise to various symptoms.
(2) Perhaps the model arises from a single empirical frequency distribution—an "error distribution". This possibility is often mentioned in textbooks.
(3) Perhaps we are convinced that a phenomenon is random without having any evidence as to the frequency distribution of its outcomes, so that the model includes all additive probability distributions on $\mathcal{X}$.

These three ways suggest, as we shall see, quite different belief functions on $\Theta \times \mathcal{X}$, though in each case the belief function has the P_θ as its conditionals and has a vacuous marginal for θ. For another example of the use of belief functions in statistical problems see Shafer (1982).

3.1 Models Composed of Independent Frequency Distributions

Suppose our model consists of finitely many P_θ and each is based on independent empirical data—i.e. each P_θ is an empirical frequency distribution which we would be willing to translate into degrees of belief about x if we knew θ to be true, and the P_θ for different θ are based on independent observations. Then how should we combine them to obtain a belief function Bel on $\Theta \times \mathcal{X}$?

Smets (1978, pp. 145–190) has pointed out that the method of conditional embedding can be used to answer this question. We represent each P_θ by its conditional embedding, say Bel_θ, in $\Theta \times \mathcal{X}$, and then we set Bel equal to the orthogonal sum of all the Bel_θ.

Let us show that Bel is vacuous for θ and has P_θ for its conditional given θ. We begin with the fact that Bel_θ's focal elements are in one-to-one correspondence with the elements of $\mathcal{X}$; corresponding to $x \in \mathcal{X}$ is the focal element

$$\{(\theta, x)\} \cup ((\Theta - \{\theta\}) \times \mathcal{X}), \tag{4}$$

with m-value equal to $P_\theta(x)$. (i) A focal element for Bel is obtained by intersecting focal elements from the different Bel_θ's; in other words, it is of the form

$$\bigcap_{\theta\epsilon\Theta}[\{(\theta, x_\theta)\} \cup ((\Theta - \{\theta\}) \times \mathcal{X})] = \bigcup_{\theta\epsilon\Theta}\{(\theta, x_\theta)\} \tag{5}$$

for some choice of x_θ's. But any subset of $\Theta \times \mathcal{X}$ of the form (5) has a nonempty intersection with every cylinder set $\{\theta\} \times \mathcal{X}$. So Bel has a vacuous marginal for θ. (ii) Intersecting the focal element (4) with $\{\theta\} \times \mathcal{X}$ yields $\{(\theta, x)\}$, while intersecting it with $\{\theta'\} \times \mathcal{X}$, where $\theta' \neq \theta$, yields $\{\theta'\} \times \mathcal{X}$. So Bel_θ yields P_θ when conditioned on θ and yields the vacuous belief function on $\mathcal{X}$ when conditioned on $\theta' \neq \theta$. Since the conditioning of an orthogonal sum can be achieved by conditioning each component before combining, it follows that Bel yields P_θ when conditioned on θ.

It should be stressed that Smet's method depends on the assumption that Θ is finite. Moreover, it gives sensible results only when the number of elements in Θ is fairly small, for enlarging Θ has the effect of weakening the posterior degrees of belief. It is only when Θ is small, of course, that we could hope to satisfy the assumption that each P_θ be based on independent empirical data.

Example 1. Consider, for simplicity, the case where $\mathcal{X}$ and Θ have only two elements; say $\mathcal{X} = \{0, 1\}$, $\Theta = \{\theta_1, \theta_2\}$, $P_{\theta_1}(1) = p_1$ and $P_{\theta_2}(1) = p_2$. Then the belief function Bel on $\Theta \times \mathcal{X}$ has the m-values given in Table 1. Conditioning Bel on the observation $x = 1$ yields the degrees of belief

$$\begin{aligned} \mathrm{Bel}\,(\theta_1|x = 1) &= \frac{p_1(1 - p_2)}{1 - (1 - p_1)(1 - p_2)} \text{ and} \\ \mathrm{Bel}(\theta_2|x = 1) &= \frac{(1 - p_1)p_2}{1 - (1 - p_1)(1 - p_2)} \end{aligned} \tag{6}$$

Some insight into these formulae may be gained by fixing p_2 at some value equal neither to 0 nor to 1 and considering extreme values of p_1. If $p_1 = 0$, then the observation $x = 1$ tells us that $\theta = \theta_2$; we have $\mathrm{Bel}(\theta_1|x = 1) = 0$ and Bel $(\theta_2|x = 1) = 1$. If $p_1 = 1$, then the observation $x = 1$ is evidence in favour of $\theta = \theta_1$; we have Bel $(\theta_1|x = 1) = 1 - p_2$ and Bel $(\theta_2|x = 1) = 0$.

The combination of observations. Smet's method can be applied, of course, to the case of multiple observations. If we expect to make n independent

Table 1.

Focal element	*m-value*
$\{(\theta_1,\ 1),\ (\theta_2,\ 1)\}$	$p_1\ p_2$
$\{(\theta_1,\ 1),\ (\theta_2,\ 0)\}$	$p_1\ (1 - p_2)$
$\{(\theta_1,\ 0),\ (\theta_2,\ 1)\}$	$(1 - p_1)\ p_2$
$\{(\theta_1,\ 0),\ (\theta_2,\ 0)\}$	$(1 - p_1)\ (1 - p_2)$

observations from P_θ, then we simply construct the product distributions P_θ^n on $\mathcal{X}^n$, conditionally embed these to obtain belief functions Bel_θ^n on $\Theta \times \mathcal{X}^n$, and then combine by Dempster's rule to obtain a belief function Bel^n on $\Theta \times \mathcal{X}^n$ that can be conditioned on the observations $x_1, \ldots, x_n$ to yield a posterior belief function on Θ.

An alternative approach to assessing independent observations $x_1, \ldots, x_n$ is to use each x_i to construct a posterior belief function Bel $(\cdot|x_i)$ on Θ and then to combine these posterior belief functions by Dempster's rule. This, it turns out, gives the same result (Smets, private communication).

Proof. For each $(x_1, \ldots, x_n) \in \mathcal{X}^n$, Bel_θ^n assigns the m-value $P_\theta(x_1) \ldots P_\theta(x_n)$ to the focal element

$$\{(\theta, x_1, \cdots, x_n)\} \cup ((\Theta - \{\theta\}) \times \mathcal{X}^n) . \tag{7}$$

Let Bel_θ denote, as before, the conditional embedding of P_θ in $\Theta \times \mathcal{X}$. And let $\mathrm{Bel}_{i\theta}$ denote the result of conditionally embedding Bel_θ in $\Theta \times \mathcal{X}^n$, with the $\mathcal{X}$ in $\Theta \times \mathcal{X}$ identified with the ith copy of $\mathcal{X}$ in $\Theta \times \mathcal{X}^n$. Then $\mathrm{Bel}_{i\theta}$ assigns, for each $x_i \in \mathcal{X}$, the m-value $P_\theta(x_i)$ to the focal element

$$(\{\theta\} \times \mathcal{X}_1 \times \ldots \times \mathcal{X}_{i-1} \times \{x_i\} \times \mathcal{X}_{i+1} \times \ldots \times \mathcal{X}_n) \cup ((\Theta - \{\theta\}) \times \mathcal{X}^n) . \tag{8}$$

We see, by comparing (7) and (8), that $\mathrm{Bel}_\theta^n = \mathrm{Bel}_{1\theta} \oplus \ldots \oplus \mathrm{Bel}_{n\theta}$. So

$$\mathrm{Bel}_n = \oplus_\theta \mathrm{Bel}_\theta^n = \oplus_\theta(\mathrm{Bel}_{1\theta} \oplus \ldots \oplus \mathrm{Bel}_{n\theta}) = (\oplus_\theta \mathrm{Bel}_{1\theta}) \oplus \ldots \oplus (\oplus_\theta \mathrm{Bel}_{n\theta}) .$$

But $\oplus_\theta \mathrm{Bel}_{i\theta}$ is the conditional embedding in $\Theta \times \mathcal{X}^n$ of $\mathrm{Bel} = \oplus_\theta \mathrm{Bel}_\theta$. So conditioning $\oplus_\theta \mathrm{Bel}_{i\theta}$ on $(x_1, \ldots, x_n)$ yields the same belief function on Θ as conditioning Bel on x_i. So

$$\mathrm{Bel}_n (\cdot|x_1, \ldots, x_n) = \mathrm{Bel}\,(\cdot|x_1) \oplus \ldots \oplus \mathrm{Bel}\,(\cdot|x_n)$$

for all $(x_1, \ldots, x_n) \in \mathcal{X}^n$.

Example 1 continued. Suppose k of our observations $x_1, \ldots, x_n$ are equal to 1 and $n-k$ are equal to 0. Then $\mathrm{Bel}_n(\cdot|x_1, \ldots, x_n)$ is obtained by using Dempster's rule to combine k copies of $\mathrm{Bel}(\cdot|x = 1)$ and $n-k$ copies of $\mathrm{Bel}(\cdot|x = 0)$. The result is

$$\mathrm{Bel}_n\,(\theta_1|x_1, \ldots, x_n) = \frac{p_1^k(1-p_1)^{n-k} - (p_1p_2)^k\,((1-p_1)\,(1-p_2))^{n-k}}{p_1^k\,(1-p_1)^{n-k} + p_2^k(1-p_2)^{n-k} - (p_1p_2)^k\,((1-p_1)\,(1-p_2))^{n-k}}$$

and

$$\mathrm{Bel}_n\,(\theta_2|x_1, \ldots, x_n) = \frac{p_2^k(1-p_2)^{n-k} - (p_1p_2)^k\,((1-p_1)\,(1-p_2))^{n-k}}{p_1^k\,(1-p_1)^{n-k} + p_2^k(1-p_2)^{n-k} - (p_1p_2)^k\,((1-p_1)\,(1-p_2))^{n-k}} .$$

Notice that for large values of n and $n-k$,

$$\mathrm{Bel}_n\,(\theta_1|x_1, \ldots, x_n) + \mathrm{Bel}_n\,(\theta_2|x_1, \ldots, x_n) \approx 1,$$

and

$$\frac{\mathrm{Bel}_n\left(\theta_1|x_1,\ldots,x_n\right)}{\mathrm{Bel}_n\left(\theta_2|x_1,\ldots,x_n\right)} \approx \left(\frac{p_1}{p_2}\right)^k \left(\frac{1-p_1}{1-p_2}\right)^{n-k} . \tag{9}$$

This agrees with the posterior Bayesian odds that would result from equal prior probabilities for θ_1 and θ_2.

Medical diagnosis. Smets' work was inspired by the problem of medical diagnosis. Here Θ is a list of possible diseases from which a patient might be suffering, $\mathcal{X}$ is a list of symptoms he might exhibit, and we assume that study of each disease θ has resulted in a distribution P_θ that gives the frequency with which that disease produces the various symptoms. Conditional embedding seems reasonable because P_θ bears on the set of possibilities $\Theta \times \mathcal{X}$ regarding our patient only conditionally on his having disease θ, and the use of Dempster's rule seems reasonable because the different frequency distributions can be regarded as independent items of evidence.

The assumption that one's evidence in a problem of medical diagnosis consists of complete and clearly relevant frequency distributions of symptoms is, of course, very unrealistic. But, as Smets points out (p. 160), the method of conditional embedding can still be used when the evidence about each disease justifies only a relatively weak belief function instead of a full frequency distribution. The following example illustrates some of the possibilities.

Example 2. Imagine a disorder called "ploxoma", which comprises two distinct "diseases": $\theta_1 =$ "virulent ploxoma", which is invariably fatal, and $\theta_2 =$ "ordinary ploxoma", which varies in severity and can be treated. Virulent ploxoma can be identified unequivocally at the time of a victim's death, but the only way to distinguish between the two diseases in their early stages seems to be a blood test with three possible outcomes, labelled x_1, x_2 and x_3. The following evidence is available: (i) Blood tests of a large number of patients dying of virulent ploxoma showed the outcomes x_1, x_2 and x_3 occurring 20, 20 and 60 per cent of the time, respectively. (ii) A study of patients whose ploxoma had continued so long as to be almost certainly ordinary ploxoma showed outcome x_1 to occur 85 per cent of the time and outcomes x_2 and x_3 to occur 15 per cent of the time. (The study was made before methods for distinguishing between x_2 and x_3 were perfected.) There is some question whether the patients in the study represent a fair sample of the population of ordinary ploxoma victims, but experts feel fairly confident (say 75 per cent) that the criteria by which patients were selected for the study should not affect the distribution of test outcomes. (iii) It seems that most people who seek medical help for ploxoma are suffering from ordinary ploxoma. There have been no careful statistical studies, but physicians are convinced that only 5–15 per cent of ploxoma patients suffer from virulent ploxoma.

We can represent each of these three items of evidence by a belief function on $\Theta \times \mathcal{X} = \{\theta_1, \theta_2\} \times \{x_1, x_2, x_3\}$. (i) The first item of evidence can be represented by the conditional embedding in $\Theta \times \mathcal{X}$ of the frequency distribution

Table 2.

Focal element	m-value	Focal element	m-value
$\{(\theta_2,\ x_1)\}$	0.541875	$\{(\theta_1, x_1)\}$	0.01
$\{(\theta_2, x_1), (\theta_2, x_2), (\theta_2, x_3)\}$	0.2125	$\{(\theta_1, x_2)\}$	0.01
$\{(\theta_2, x_2), (\theta_2, x_3)\}$	0.095625	$\{(\theta_1, x_3), (\theta_2, x_2), (\theta_2, x_3)\}$	0.00675
$\{(\theta_1, x_3), (\theta_2, x_1)\}$	0.03825	$\{(\theta_1, x_1), (\theta_2, x_1), (\theta_2, x_2), (\theta_2, x_3)\}$	0.005
$\{(\theta_1, x_3)\}$	0.03	$\{(\theta_1, x_2), (\theta_2, x_1), (\theta_2, x_2), (\theta_2, x_3)\}$	0.005
$\{(\theta_1, x_3), (\theta_2, x_1), (\theta_2, x_2), (\theta_2, x_3)\}$	0.015	$\{(\theta_1, x_1), (\theta_2, x_2), (\theta_2, x_3)\}$	0.00225
$\{(\theta_1, x_1), (\theta_2, x_1)\}$	0.01275	$\{(\theta_1, x_2), (\theta_2, x_2), (\theta_2, x_3)\}$	0.00225
$\{(\theta_1, x_2), (\theta_2, x_1)\}$	0.01275		

P_{θ_1}, where $P_{\theta_1}(x_1) = 0.2$, $P_{\theta_1}(x_2) = 0.2$ and $P_{\theta_1}(x_3) = 0 \cdot 6$. (ii) For the second item of evidence, we begin with a belief function Bel_{θ_2} on $\mathcal{X}$ that has focal elements $\{x_1\}$ and $\{x_2,\ x_3\}$ with m-values 0.85 and 0.15, respectively. We discount this belief function at rate $\alpha = 0.25$, and then conditionally embed it in $\Theta \times \mathcal{X}$. (iii) For the third item of evidence we begin with a belief function Bel_0 on Θ that has m-values $m_0(\{\theta_1\}) = 0.05$, $m_0(\{\theta_2\}) = 0.85$ and $m_0(\Theta) = 0.10$, and we minimally extend Bel_0 to $\Theta \times \mathcal{X}$

Combining these three belief functions by Dempster's rule results in the belief function on $\Theta \times \mathcal{X}$ with the m-values given in Table 2. Table 3 shows the posterior degrees of belief that result when this belief function is conditioned on the result of the patient's blood test. As these numbers indicate, the blood test is not as informative as one might hope. The physician's initial 85 per cent degree of belief that a given ploxoma is ordinary is raised only to 96·5 per cent by a test that comes out x_1 and lowered only to 78.2 per cent by a test that comes out x_3.

3.2 Models derived from a Single Frequency Distribution

Let us turn from the case where there is a different frequency distribution underlying each P_θ to an opposite extreme: the case where all the P_θ are derived from a single frequency distribution. And let us think about the tritest example: the parametric model generated by an error distribution.

Table 3.

	$\mathrm{Bel}(\theta_1 \mid x)$	$\mathrm{Bel}(\theta_2 \mid x)$
x_1	0.014	0.965
x_2	0.062	0.918
x_3	0.165	0.782

Consider a measuring instrument whose propensities to err are thoroughly known to us; we have used it to measure many known quantities and recorded its errors in these cases so as to obtain a frequency distribution $P(e)$ which we are willing to translate into degrees of belief about what our error $e = x - \theta$ will be when we shortly use the instrument to obtain a measurement x of an unknown quantity θ. Consider $\Theta \times \mathcal{X}$, where Θ is the set of possible values of θ and $\mathcal{X}$ is the set of possible values of x; we assume that $\Theta = \mathcal{X}$. Each possible error e will correspond to a subset $\Theta \times \mathcal{X}$; namely, $\{(\theta, x)|x - \theta = e\}$. So we can accomplish the translation of the error distribution $P(e)$ into degrees of belief about $x - \theta$ by minimally extending P to $\Theta \times \mathcal{X}$. This means adopting the belief function Bel on $\Theta \times \mathcal{X}$ that assigns the m-value $P(e)$ to the focal element $\{(\theta, x)|x - \theta = e\}$. It is evident that Bel is vacuous for θ. And its conditional on $\mathcal{X}$ given θ is given by $\mathrm{Bel}(x|\theta) = P(x - \theta)$. The belief function $\mathrm{Bel}(\cdot|\theta)$ is an additive probability distribution, and so it may denote it by P_θ, thus obtaining a parametric model $\{P_\theta : \theta \in \Theta\}$ on $\mathcal{X}$.

The preceding paragraph merely translates into the language of belief functions a traditional account of how a parametric model arises from an error distribution. Moreover, the result of conditioning the belief function Bel on the actual measurement x is the additive probability distribution Bel $(\cdot|x)$ on Θ given by $\mathrm{Bel}(\theta|x) = P(\theta - x)$, and this is the familiar fiducial solution to the problem of inference for this model. Notice, however, that the belief-function argument depends on the model having really arisen from the error distribution; the argument gives no sanction to fiducial methods in cases where one begins with an abstract model $\{P_\theta : \theta \in \Theta\}$ and then notices a pivotal quantity $x - \theta$. (This belief-function treatment of the fiducial method was given by Dempster (1966). The only novelty in the present exposition is my insistence that the criterion for the method's validity should be sought in the origin of the parametric model.) This lack of sanction for the use of arbitrary pivotal quantities appears to rule out marginalization paradoxes of the type discussed by Dawid *et al.* (1973).

The belief function Bel on $\Theta \times \mathcal{X}$ is non-additive, even though its conditionals $\mathrm{Bel}(\cdot|\theta)$ and $\mathrm{Bel}(\cdot|x)$ are all additive. Notice also that Bel can, in some circumstances, lead to posterior probabilities for θ that are non-additive. If instead of observing the measurement x we observe only that x is in some subset A of $\mathcal{X}$, then we will condition Bel on $\Theta \times A$, and the resulting conditional belief function will have a non-additive marginal for θ.

The combination of observations. Here, as in the case of Smets' method, there are two approaches to combining independent observations. We can construct the product distribution P^n, conditionally embed it in $\Theta \times \mathcal{X}^n$, and then condition on the observations $(x_1, \ldots, x_n)$. Or we can construct a posterior belief function $\mathrm{Bel}(\cdot|x_i)$ for each observation and then combine these by Dempster's rule. It can be shown, here as in the case of Smets' method, that both approaches give the same final belief function $\mathrm{Bel}_n(\cdot|x_1, \ldots, x_n)$ on Θ. In this case, $\mathrm{Bel}_n(\cdot|x_1, \ldots, x_n)$ is an additive probability distribution.

Table 4.

e	$P(e)$	e	$P(e)$
−6	0.00009	1	0.21321
−5	0.00101	2	0.10916
−4	0.00750	3	0.03577
−3	0.03577	4	0.00750
−2	0.10916	5	0.00101
−1	0.21321	6	0.00009
0	0.26651		

Example 3. Suppose $\mathcal{X}$ and Θ are both equal to the set of all integers, and P is given by $P(e) = c0.8^{e^2}$, where $c \approx 0.26651$. Table 4 gives the values of $P(e)$ that exceed 10^{-5}. If we observe $(x_1, \ldots, x_n)$,then $\mathrm{Bel}(\theta|x_i) = c0{\cdot}8^{(\theta-x_i)^2}$, and $\mathrm{Bel}_n(\cdot|x_1, \ldots, x_n) = \mathrm{Bel}(\cdot|x_i) \oplus \ldots \oplus \mathrm{Bel}(\cdot|x_n)$ is the additive probability distribution specified by

$$\mathrm{Bel}_n(\theta|x_1, \ldots, x_n) \propto \prod_{i=1}^{n} \mathrm{Bel}(\theta|x_i) \propto (0.8)^{\mathrm{n}(\theta-\bar{x})^2}$$

If, for example, $n = 4$ and $(x_1, \ldots, x_4) = (-2, 1, 0, 9)$, then we obtain

$$\mathrm{Bel}_4\,(\theta| - 2, 1, 0, 9) \propto 0.8^{4(\theta-2)^2}.$$

Table 5 gives the values of $\mathrm{Bel}_4(\theta| - 2, 1, 0, 9)$ that exceed 10^{-5}.

Example 4. Let us suppose, in order to construct an example that is comparable to Example 1 above, that $\Theta = \{0, 1\}$, that 0 and 1 are also the possible errors, with frequencies $P(0) = p$ and $P(1) = 1 - p$, and that the addition to obtain $x = \theta + e$ is modulo 2. This means that $\mathcal{X} = \{0, 1\}$, and that P_θ assigns 0 and 1 the frequencies p and $1 - p$, respectively, when $\theta = 0$ and the frequencies $1 - p$ and p, respectively, when $\theta = 1$.

Table 5.

θ	$\mathrm{Bel}_4(\theta| - 2, 1, 0, 9)$
−1	0·00017
0	0·01500
1	0·21832
2	0·53300
3	0·21832
4	0·01500
5	0·00017

The belief function Bel on $\Theta \times \mathcal{X}$ has focal elements {(0,0),(1,1)} and {(0,1),(1,0)} with m-values p and $1 - p$, respectively. So conditioning on $x = 1$ yields $\mathrm{Bel}(\theta = 0|x = 1) = 1 - p$ and $\mathrm{Bel}(\theta = 1|x = 1) = p$. Notice that these posterior degrees of belief do not agree with the posterior degrees of belief that we obtained using Smets' method in Example 1. In order to make the comparison, we set $\theta_1 = 1$, $\theta_2 = 0$, $p_1 = p$ and $p_2 = 1 - p$ in (6), thus obtaining

$$\mathrm{Bel}(\theta = 1|x = 1) = p^2 \left\{1 - p\,(1 - p)\right\}^{-1} \quad \text{and}$$
$$\mathrm{Bel}(\theta = 0|x = 1) = (1 - p)^2 \left\{1 - p\,(1 - p)\right\}^{-1}.$$

There is asymptotic agreement, however. If we have measurements $x_1, \ldots,$ x_n, k of which equal 1 and $n - k$ of which equal 0, then we obtain

$$\mathrm{Bel}_n\left(\theta = 1|x_1 \ldots, x_n\right) = p^k \left(1 - p\right)^{n-k} \left(p^k \left(1 - p\right)^{n-k} + \left(1 - p\right)^k p^{n-k}\right)^{-1},$$
$$\mathrm{Bel}_n\left(\theta = 0|x_1 \ldots, x_n\right) = \left(1 - p\right)^k p^{n-k} \left(p^k \left(1 - p\right)^{n-k} + \left(1 - p\right)^k p^{n-k}\right)^{-1},$$

and

$$\frac{\mathrm{Bel}_n\left(\theta = 1|x_1, \ldots, x_n\right)}{\mathrm{Bel}_n\left(\theta = 0|x_1, \ldots, x_n\right)} = \frac{p^k (1 - p)^{n-k}}{\left(1 - p\right)^k p^{n-k}},$$

which agrees with (9).

Practical complications. The premises for our justification of the fiducial method through belief functions will rarely be fully satisfied. Usually our experience with a measuring instrument will be inadequate for us to credit fully a frequency distribution and the possibility of systematic errors will always limit the extent to which we are willing to treat successive errors as independent. However, these complications, though they do push us away from the fiducial method, need not push us away from the use of belief functions.

Example 3 *continued.* Suppose we take seriously the possibility of outliers and therefore discount the frequency distribution $P(e)$, using the discount rate $\alpha = 0.01$. This results in the $\mathrm{Bel}(\cdot|x_i)$ also being discounted at this rate. When we combine these four discounted belief functions by Dempster's rule, we obtain a belief function $\mathrm{Bel}_4^{0.01}(\theta|-2, 1, 0, 9)$ that is very nearly an additive probability distribution; the whole set Θ is a focal element, but its m-value is only $0 \cdot 00005$, and all the other focal elements are singletons. Values of $\mathrm{Bel}_4^{0.01}(\theta| - 2, 1, 0, 9)$ that exceed 10^{-5} are shown in Table 6. Notice the sharp disagreement with the values of $\mathrm{Bel}_4(\theta| - 2, 1, 0, 9)$ given in Table 5. When we do not discount, we obtain a probability of 0.53300 for $\theta = 2$, but when we do discount, we obtain a probability of only 0.02517 for $\theta = 2$ and a probability of 0.87302 for $-1 \leq \theta \leq 1$. This disagreement can be explained by saying that discounting leads us to treat the measurement $x_4 = 9$ as a probable outlier. (See pp. 251–255 of Shafer, 1976.)

Now suppose we admit the possibility that there may be a systematic error f affecting all our measurements. And suppose we make the following probability judgements about f, based on our knowledge of the measuring

Table 6.

	$\mathrm{Bel}_4^{0.01}(\theta \mid -2, 1, 0, 9)$		$\mathrm{Bel}_4^{0.01}(\theta \mid -2, 1, 0, 9)$
−7	0.00001	4	0.00042
−6	0.00004	5	0.00013
−5	0.00022	6	0.00022
−4	0.00119	7	0.00059
−3	0.00961	8	0.00116
−2	0.08154	9	0.00144
−1	0.32160	10	0.00116
0	0.39843	11	0.00059
1	0.15299	12	0.00019
2	0.02517	13	0.00004
3	0.00320	14	0.00001

instrument and process: we consider it certain that $|f| \leq 2$, and we feel there is a chance $0\cdot 8$ that $|f| \leq 1$ and a chance 0.6 that $f = 0$. In other words, we adopt a belief function Bel_f that has focal elements $\{0\}$, $\{-1, 0, 1\}$ and $\{-2, -1, 0, 1, 2\}$, with m-values 0.6, 0.2 and 0.2, respectively.

We are now assuming that $x_i = \theta + f + e_i$, or $\theta + f = x_i - e_i$. So the belief function $\mathrm{Bel}_4^{0\cdot 01}(\cdot \mid -2, 1, 0, 9)$ must now be interpreted as giving degrees of belief about $\theta + f$ rather than about θ. When we combine these degrees of belief about $0 + f$ with the degrees of belief about f given by Bel_f, we obtain a belief function Bel^* with the m-values given (to the nearest 0.00001) in Table 7. A few values of Bel^* are given in Table 8.

3.3 Pure Randomness

Suppose we know an unknown quantity X must take one of a finite set, say $\mathcal{X} = \{1, \ldots, k\}$, of possible values, and we feel it does so randomly. We can express this by saying that X is governed by some frequency distribution. But there are only so many frequency distributions on $\mathcal{X}$—so many as there are vectors $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ of non-negative numbers that add to one. Setting Θ equal to the set of all these vectors and letting P_θ denote the frequency distribution corresponding to $\boldsymbol{\theta}$ (i.e. $P_\theta(x) = \theta_x$ for all $x \in \mathcal{X}$), we obtain a parametric model $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. This model, it seems fair to say, arises solely from the idea that X is random.

As a result of work by de Finetti (1964), Hewitt and Savage (1955) and others, many Bayesians subscribe to a purely subjective interpretation of the idea that X is random and is governed by one of the frequency distributions P. This interpretation involves thinking of X as one of a sequence $X = (X_1, X_2, \ldots)$ of unkown quantities, each of which takes values in $\mathcal{X}$, and considering a countably additive[1] probability distribution P that represents a Bayesian's

[1] De Finetti prefers the weaker condition of finite additivity. But we can neglect this subtlety in the present brief exposition.

Table 7.

Focal element	m-value	Focal element	m-value
$\{-7\}$	0.00000	$\{3,4,5\}$	0.00008
$\{-6\}$	0.00003	$\{4,5,6\}$	0.00003
$\{-5\}$	0.00013	$\{5,6,7\}$	0.00004
$\{-4\}$	0.00071	$\{6,7,8\}$	0.00012
$\{-3\}$	0.00577	$\{7,8,9\}$	0.00023
$\{-2\}$	0.04892	$\{8,9,10\}$	0.00029
$\{-1\}$	0.19297	$\{9,10,11\}$	0.00023
$\{0\}$	0.23906	$\{10,11,12\}$	0.00012
$\{1\}$	0.09179	$\{11,12,13\}$	0.00004
$\{2\}$	0.01510	$\{12,13,14\}$	0.00001
$\{3\}$	0.00192	$\{13,14,15\}$	0.00000
$\{4\}$	0.00025	$\{-9,-8,-7,-6,-5\}$	0.00000
$\{5\}$	0.00008	$\{-8,-7,-6,-5,-4\}$	0.00001
$\{6\}$	0.00013	$\{-7,-6,-5,-4,-3\}$	0.00004
$\{7\}$	0.00036	$\{-6,-5,-4,-3,-2\}$	0.00024
$\{8\}$	0.00069	$\{-5,-4,-3,-2,-1\}$	0.00192
$\{9\}$	0.00087	$\{-4,-3,-2,-1,0\}$	0.01631
$\{10\}$	0.00069	$\{-3,-2,-1,0,1\}$	0.06432
$\{11\}$	0.00035	$\{-2,-1,0,1,2\}$	0.07969
$\{12\}$	0.00012	$\{-1,0,1,2,3\}$	0.03060
$\{13\}$	0.00002	$\{0,1,2,3,4\}$	0.00503
$\{14\}$	0.00000	$\{1,2,3,4,5\}$	0.00064
Θ	0.00005	$\{2,3,4,5,6\}$	0.00008
$\{-8,-7,-6\}$	0.00000	$\{3,4,5,6,7\}$	0.00003
$\{-7,-6,-5\}$	0.00001	$\{4,5,6,7,8\}$	0.00004
$\{-6,-5,-4\}$	0.00004	$\{5,6,7,8,9\}$	0.00012
$\{-5,-4,-3\}$	0.00024	$\{6,7,8,9,10\}$	0.00023
$\{-4,-3,-2\}$	0.00192	$\{7,8,9,10,11\}$	0.00029
$\{-3,-2,-1\}$	0.01631	$\{8,9,10,11,12\}$	0.00023
$\{-2,-1,0\}$	0.06432	$\{9,10,11,12,13\}$	0.00012
$\{-1,0,1\}$	0.07969	$\{10,11,12,13,14\}$	0.00004
$\{0,1,2\}$	0.03060	$\{11,12,13,14,15\}$	0.00001
$\{1,2,3\}$	0.00503	$\{12,13,14,15,16\}$	0.00000
$\{2,3,4\}$	0.00064		

beliefs about $\mathbf{X}$ and that is symmetric—i.e. invariant under permutations of finitely many of the X_i's. As it turns out, the countable additivity and symmetry of P imply that for each $x \in \mathcal{X}$, $P(\lim_{n\to\infty} f(x,n) \text{ exists}) = 1$, where $f(x,n)$ is the proportion of the quantities $X_1, \ldots, X_n$ that equal x. The vector $\lim_{n\to\infty}(f(1,n), \ldots, f(k,n))$ can be identified, of course, with the unknown parameter $\boldsymbol{\theta}$; conditioning P on this vector being equal to $\boldsymbol{\theta}$ reduces P to the product distribution $P_{\boldsymbol{\theta}}^{\infty}$. The Bayesian's prior distribution for $\boldsymbol{\theta}$ is implicitly contained in P; it is P's marginal for the vector $\lim_{n\to\infty}(f(1,n), \ldots, f(k,n))$.

Table 8.

A	$\mathrm{Bel}^*(A)$
$\{0\}$	0.23906
$\{-1\}$	0.19297
$\{1\}$	0.09179
$\{-2\}$	0.04892
$\{2\}$	0.01510
$\{-1,0,1\}$	0.60351
$\{-2,-1,0\}$	0.54527
$\{0,1,2\}$	0.37655
$\{-2,-1,0,1,2\}$	0.84214
$\{-3,-2,-1,0,1,2,3\}$	0.96609

The distribution P is fully determined, moreover, by this prior distribution; there is only one symmetric and countably additive distribution for X having a given marginal for $\lim_{n\to\infty}(f(1,n),\ldots,f(k,n))$.

How might we give a treatment of randomness via belief functions which is analogous to this Bayesian treatment? The obvious goal is to capture the aspects of our idea of randomness (belief in the existence of limiting frequencies and recovery of $\{P_\theta : \theta \in \Theta\}$ by conditioning on the limiting frequencies) captured by the Bayesian treatment while avoiding opinions about the value of the limiting frequency. This means we should try to construct a symmetric belief function Bel for $X = (X_1, X_2, \ldots)$ that satisfies

$$\mathrm{Bel}\left(\lim_{n\to\infty} f(x,n) \text{ exists}\right) = 1 \tag{10}$$

for all $x \in \mathcal{X}$,

$$\mathbf{Bel}(\boldsymbol{X_1 = x_1, \ldots, X_n = x_n} | \lim_{n\to\infty}(\boldsymbol{f(1,n), \ldots, f(k,n)}) = \theta) = \boldsymbol{P_\theta(x_1), \ldots, P_\theta(x_n)} \tag{11}$$

for all $x_1, \ldots, x_n \in \mathcal{X}$, and

$$\mathrm{Bel}(\lim_{n\to\infty}(f(1,n),\ldots,f(k,n)) \in A) = 0 \tag{12}$$

for every proper subset A of Θ. As it turns out, this goal can be achieved; there are belief functions satisfying these conditions.

The dichotomous case. The construction of a belief function Bel satisfying (10), (11) and (12) is most easily carried out in the case where $\mathcal{X}$ has only two elements. In this case it is convenient to use {0,1} rather than {1,2} to label the elements of $\mathcal{X}$ and to use [0,1] as the parameter space Θ, with $P_\theta(1) = \theta$ and $P_\theta(0) = 1 - \theta$. Let us also write $S_n = \Sigma_{i=1}^n X_i$. Then (10), (11) and (12) become

$$\mathrm{Bel}(\lim_{n\to\infty}(S_n/n) \text{ exists}) = 1, \tag{13}$$

$$\mathrm{Bel}(X_1 = x_1, \ldots, X_n = x_n | \lim_{n\to\infty}(S_n/n) = \theta) = P_\theta(x_1)\ldots P_\theta(x_n) \tag{14}$$

and

$$\mathrm{Bel}(\lim_{n\to\infty}(S_n/n) \in A) = 0 \tag{15}$$

for all $A \subset [0,1]$.

The construction of a belief function Bel satisfying (13), (14) and (15) begins with the construction of a belief function Bel_n for the finite sequence $(X_1, X_2, \ldots, X_n)$. We construct Bel_n, which is a belief function over $\{0,1\}^n$, by assigning m-values $1/n!$ to each of the $n!$ subsets of $\{0,1\}^n$ of the form

$$A_\sigma = \left\{(x_1, \ldots, x_n) \in \{0,1\}^n \,|\, x_{\sigma(1)} \geq x_{\sigma(2)} \geq \ldots \geq x_{\sigma(n)}\right\},$$

where σ is a permutation of $\{1, \ldots, n\}$. (Here is an example of a set A_σ. If $n = 3$ and $(\sigma(1), \sigma(2), \sigma(3)) = (1,3,2)$, then

$$A_\sigma = \{(0,0,0), (1,0,0), (1,0,1), (1,1,1)\}.)$$

A permutation of $(X_1, \ldots, X_n)$ merely permutes the A_σ. So Bel_n is symmetric —i.e. it satisfies

$$\mathrm{Bel}_n\left((X_1, \ldots, X_n) \in A\right) = \mathrm{Bel}_n\left(\left(X_{\sigma(1)}, \ldots, X_{\sigma(n)}\right) \in A\right) \tag{16}$$

for all permutations σ. It is also easy to see that each A_σ has exactly one representative for each possible frequency of ones—i.e. for each $k, 0 \leq k \leq n$, there is exactly one element $(x_1, \ldots, x_n) \in A_\sigma$ such that $\Sigma_{i=1}^n x_i = k$. This means that Bel_n's marginal for S_n is vacuous—i.e.

$$\mathrm{Bel}_n\left(S_n \in A\right) = 0 \tag{17}$$

for every proper subset A of $\{0, 1, \ldots, n\}$. It also means that conditioning on $S_n = k$ reduces the A_σ to singletons and hence reduces Bel_n to an additive (i.e. Bayesian) belief function. Thus, by the symmetry of Bel_n,

$$\mathrm{Bel}_n\left(X_1 = x_1, \ldots, X_n = x_n | S_n = k\right) = 1/\binom{n}{k}, \tag{18}$$

provided that $\Sigma_{i=1}^n x_i = k$.

The belief functions Bel_n "cohere", in the sense that if $m < n$ then Bel_m is Bel_n's marginal for marginal for $X_1, \ldots, X_m$. And it is fairly easy to show that this set of coherent belief functions is the only one satisfying (16), (17) and (18). All the Bel_n together can be regarded as a belief function for the infinite sequence $(X_1, X_2, \ldots)$. More precisely, they can be regarded as defining a belief function on the algebra of subsets of $\{0,1\}^\infty$ consisting of all "finite cylinder sets". This belief function can then be minimally extended to a belief function on the algebra of all subsets of $\{0,1\}^\infty$. It turns out that if we use a form of minimal extension that preserves "sequential continuity" (a condition equivalent to countable additivity in the presence of finite additivity), then the resulting belief function Bel on $\{0,1\}^\infty$ does indeed satisfy (13). Since

Bel_n is Bel's marginal, (16) says that Bel is symmetric. And, as it turns out, (17) implies (15) and (18) implies (14). (The proofs of the assertions in this paragraph have not been published. But the concepts of continuity and minimal continuous extension are discussed in Shafer, 1979.)

Since the belief function Bel, like the Bayesian's additive probability distribution P, gives degree of belief one to the existence of the limit $\theta = \lim_{n\to\infty}(S_n/n)$, we can examine Bel's marginal for (θ, X_1), which is a belief function on $\Theta \times \mathcal{X}$. By (15), this belief function has a vacuous marginal for θ. And by (14), its conditional given θ is P_θ. Thus the construction of Bel yields a solution to our general problem of constructing a belief function on $\Theta \times \mathcal{X}$—a solution which seems appropriate when the specification is based purely on the idea of randomness. As it turns out, this solution is Dempster's original "generalized Bayesian" method. (See Dempster, 1968, or Shafer, 1976b).

Instead of considering the marginal just for (θ, X_1), we could also consider the marginal for $(\theta, X_1, \ldots, X_n)$, thus obtaining a belief function on $\Theta \times \mathcal{X}^n$ which is vacuous for θ and has P_θ^n as its conditional given θ. It is also true, here as in the case of Smets' method and the fiducial $\mathrm{Bel}_{\mathrm{x}}$ on method, that the belief function Θ obtained by conditioning on a vector $x = (x_1, \ldots, x_n)$ of actual observations is the same as the belief function $\mathrm{Bel}_{x_1} \oplus \ldots \oplus \mathrm{Bel}_{x_n}$, where Bel_{x_i} is the belief function on obtained by conditioning on a single observation x_i. See Sect. 4 of Dempster (1966) for some calculations of values of $\mathrm{Bel}_{\mathrm{x}}$.

The general case. The results in the dichotomous case generalize to the case where $\mathcal{X} = \{1, \ldots, k\}$ in that there does exist a symmetric belief function on $\mathcal{X}^\infty$ that satisfies (10), (11) and (12) and has a marginal for $\mathcal{X} \times \Theta$ corresponding to Dempster's generalized Bayesian method. It appears, however, that when $k > 2$ there are other symmetric belief functions on $\mathcal{X}^\infty$ that satisfy (10), (11) and (12) but have different marginals for $\mathcal{X} \times \Theta$. It would be interesting to obtain an understanding of these belief functions.

It should be noted, in any case, that the justification for Dempster's generalized Bayesian method offered here depends on the idea of pure randomness and hence only applies when the parametric model consists of all the distributions on $\mathcal{X}$. This rules out Aitchison's counter-example to the method. (See Aitchison, 1968, or Lindley, 1972, p. 9.)

4 Parametric Models not Based on Evidence

In Chap. 11 of *A Mathematical Theory of Evidence* I suggested a general belief-function treatment of statistical evidence which, in contrast to the methods just discussed, does not depend on the nature of the evidence establishing the parametric model and does not condition on the observations. This method simply translates each observation x into the consonant belief functions on Θ given by

$$\mathrm{Bel}_x(A) = \sup\{s | f_x(\theta) \geq 1 - s \text{ implies } \theta \in A\}, \tag{19}$$

where $f_x(\theta)$ is the normalized likelihood function:

$$f_x(\theta) = P_\theta(x) / \sup_{\theta' \in \Theta} P_{\theta'}(x).$$

(Bel_{x_i} is determined by the conditions that it be consonant and that it award degree of belief s to each "likelihood interval" $\{\theta | f_x(\theta) \geq 1 - s\}$.)

Many statisticians have discussed the idea of determining degrees of belief by (19). (See, for example, Hudson, 1971, and Edwards, 1972.) But the usefulness of the idea seems to be limited, for one can construct examples where the likelihood function cannot be normalized, or where the normalized likelihood function seems to be misleading. (See Lindley, 1972, pp. 12–13.) I emphasized likelihood intervals in *A Mathematical Theory of Evidence* because of their simple relation to the idea of weights of evidence. But I now think (19) should be rejected as a general method of statistical inference because it does not take into account the origin of the model.

If we do use (19), then how should we combine physically independent observations $x_1, \ldots, x_n$? For each of the three methods we considered above (Smets' method, the fiducial method and the model of pure randomness) there are two different ways of combining observations: (1) A belief function can be constructed on $\Theta \times \mathcal{X}^n$ that has P_θ^n as its marginal given θ, and this belief function can be conditioned on $x = (x_1, \ldots, x_n)$ to yield a belief function $\mathrm{Bel}_{\mathrm{x}}$ on Θ. (2) A belief function can be constructed on $\Theta \times \mathcal{X}$ that has P_θ as its marginal given θ, for each x_i this belief function can be conditioned on x_i to yield a belief function Bel_{x_i} on Θ, and Dempster's rule can be used to obtain the orthogonal sum $\mathrm{Bel}_{x_1} \oplus \ldots \oplus \mathrm{Bel}_{x_n}$. These two ways of combining $x_1, \ldots, x_n$ give the same final result for all three methods: we always find that $\mathrm{Bel}_{\mathrm{x}} = \mathrm{Bel}_{x_1} \oplus \ldots \oplus \mathrm{Bel}_{x_n}$. In the case of (19) we are not conditioning belief functions constructed on $\Theta \times \mathcal{X}$ or $\Theta \times \mathcal{X}^n$, but we can still distinguish two ways of combining observations: (1) We can represent the physical independence of $x_1, \ldots, x_n$ by constructing the product model $\{P_\theta^n : \theta \in \Theta\}$ and apply (19) directly to this model to obtain a belief function $\mathrm{Bel}_{\mathrm{x}}$. (2) We can apply (19) for each x_i and then combine the resulting belief functions, obtaining the orthogonal sum $\mathrm{Bel}_{x_1} \oplus \ldots \oplus \mathrm{Bel}_{x_n}$. And in this case $\mathrm{Bel}_{\mathrm{x}}$ and $\mathrm{Bel}_{x_1} \oplus \ldots \oplus \mathrm{Bel}_{x_n}$ will, in general, be different.

Several reviewers of *A Mathematical Theory of Evidence* (see Diaconis, 1977, p. 678; Fine, 1978, p. 671; and Williams, 1978, pp. 384–385) have found the divergence between $\mathrm{Bel}_{\mathrm{x}}$ and $\mathrm{Bel}_{x_1} \oplus \ldots \oplus \mathrm{Bel}_{x_n}$ in the case of (19) unacceptable. I am now inclined to agree with them. The choices that a theory of evidence asks us to make ought always to be judgements based on our evidence—i.e. choices for which we can look to our evidence for guidance. And it is not clear how we can use our evidence to choose between $\mathrm{Bel}_{\mathrm{x}}$ and $\mathrm{Bel}_{x_1} \oplus \ldots \oplus \mathrm{Bel}_{x_n}$.

The use of likelihood intervals, though unacceptable where the evidence for a parametric model can be spelled out, may still be of interest in cases where there is no evidence for the model—in cases, that is to say, where one is merely trying out the model to see how it fits and what it suggests about θ. Here the arbitrariness of the choice between $\mathrm{Bel}_{\mathrm{x}}$ and $\mathrm{Bel}_{x_1} \oplus \ldots \oplus \mathrm{Bel}_{x_n}$ can be be seen as a consequence of the arbitrariness of the model itself.

5 Beyond the Parametric Model

In the preceding pages we have seen several examples where evidence conventionally used to justify parametric models can further be used to justify belief-function analyses of those models. The purpose of presenting these examples was to illustrate how the choice of a belief-function analysis depends on the nature of the evidence for the model, not just on the model itself. But a second lesson also emerged from our discussion—the lesson that the evidence for a parametric model often does not justify the model very well and that a belief-function analysis that makes weaker claims on behalf of the evidence may often be appropriate.

It is here, I believe, that the theory of belief functions has the most to offer. There is no great need for new methods of statistical inference for traditional problems where we have well-supported parametric models involving few parameters. But there is a need for new methods for problems where such models are not available. Some Bayesians have sought to address this need by constructing models that have so many parameters that they could not possibly fail to fit the data and then pretending to have prior beliefs about these parameters. The theory of belief function offers an approach that better respects the realities and limitations of our knowledge and evidence.

Acknowledgement

Research supported by National Science Foundation Grant No. MCS 8002213.

References

Aitchison, J. (1968). In discussion of “A generalization of Bayesian inference”, by A. P. Dempster. *J. R. Statist. Soc.* B, **30**, 234–237.

Dawid, A. P., Stone, M. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *J. R. Statist. Soc.* B, **35**, 189–233.

Dempster, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Ann. Math. Statist.*, **37**, 355–374.

—— (1967a). Upper and lower probabilities induced by a multivariate mapping. *Ann. Math. Statist.*, **38**, 325–339.

—— (1967b). Upper and lower probability inferences based on a sample from a finite univariate population. *Biometrika*, **54**, 515–528.

—— (1968a). Upper and lower probabilities generated by a random closed interval. *Ann. Math. Statist.*, **39**, 957–966.

—— (1968b). A generalization of Bayesian inference (with discussion). *J. R. Statist. Soc.* B, **30**, 205–247.

—— (1969). Upper and lower probability inferences for families of hypotheses with monotone density ratios. *Ann. Math. Statist.*, **40**, 953–969.

DeRobertis, L. (1978). The use of partial prior knowledge in Bayesian inference. Ph.D. Dissertation, Yale University.

Diaconis, P. (1978). Review of *A Mathematical Theory of Evidence. J. Amer. Statist. Ass.*, **73**, 677–678.

Edwards, A. W. F. (1972). *Likelihood.* Cambridge: University Press.

Fine, T. L. (1977). Review of *A Mathematical Theory of Evidence. Bull. Amer. Math. Soc.*, **83**, 667–672.

de Finetti, B. (1964). Foresight: its logical laws, its subjective sources. In *Studies in Subjective Probability* (H. E. Kyburg and H. E. Smokler, eds). New York: Wiley.

Hewitt, E. and Savage, L. J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.*, **80**, 470–501.

Hudson, D. J. (1971). Interval estimation from the likelihood function. *J. Roy. Statist. Soc.* B, 256–262.

Lindley, D. V. (1972). *Bayesian Statistics, A Review.* Philadelphia: SIAM.

Savage, L. J. (1954). *The Foundations of Statistics.* New York: Wiley.

Shafer, G. (1976a). *A Mathematical Theory of Evidence*, Princeton: University Press.

—— (1976b). A theory of statistical evidence. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* (W. Harper and C. A. Hooker, eds), Vol. II, 365–436.

—— (1978). Non-additive probabilities in the work of Bernoulli and Lambert. *Archive for History of Exact Sciences*, **19**, 309–370.

—— (1979). Allocations of probability. *Ann. Prob.*, **7**, 827–839.

—— (1981a). Constructive probability. *Synthese*, **48**, 1–60.

—— (1981b). Jeffrey's rule of conditioning. *Philos. Sci.*, **48**, 337–362.

—— (1982). Lindley's paradox. *J. Amer. Statist. Assoc.* **77**, 325–351.

Shafer, Glenn and Amos, Tversky, (1985) "Languages and Designs for Probability Judgment," Cognitive Science Society, **9**, 309–339.

Smets, P. (1978). Un modèle mathématico-statistique simulant le processus du diagnostic médical. Doctoral Dissertation at the Free University of Brussels, Presses Universitaires de Bruxelles.

Williams, P. M. (1978). On a new theory of epistemic probability. (Review of *A Mathematical Theory of Evidence.*) *The British Journal for the Philosophy of Science*, **29**, 375–387.

11

Entropy and Specificity in a Mathematical Theory of Evidence

Ronald R. Yager

Abstract. We review Shafer's theory of evidence. We then introduce the concepts of entropy and specificity in the framework of Shafer's theory. These become complementary aspects in the indication of the quality of evidence.

Key words: Entropy, Fuzzy sets, Specificity, Belief, Plausibility.

1 Introduction

In [1] Shafer presents a comprehensive theory of evidence. The problem of concern to Shafer is the location of some special element in a set X, called the frame of discernment or base set. In Shafer's framework he is provided with evidence as to the identity of this special element in terms of a mapping from the power set of X (set of all subsets of X) into the unit interval. This mapping which Shafer calls the basic assignment, associates with each subset A of X, the degree of belief that the special element is located in the set A with the understanding that he can't make any more precise statement with regards to the location of the element.

A significant aspect of Shafer's structure is the ability to represent in this common framework various different types of uncertainty, i.e. probabilistic uncertainty and possibilistic uncertainty. Our purpose here is to take some concepts developed in these individual frameworks and generalize them to the comprehensive framework of Shafer. In particular we shall generalize the idea of entropy from the probabilistic framework and specificity from the possibilistic framework. We shall find that these two measures of uncertainty provided complementary measures of the quality of a piece of evidence.

2 Shafer's Theory of Evidence

In Ref. 1 Shafer presents a comprehensive theory of evidence based on the concept of belief. The theory begins with the idea of using a number between zero and one to indicate the degree of support a body of evidence provides for a proposition. The fundamental concept in Shafer's theory is the basic assignment.[1]

Definition 1. *Assume m is a set mapping from subsets of the finite set X into the unit interval*

$$m : 2^X \to [0,1]$$

such that

1) $m(\varnothing) = 0$
2) $\sum_{A \subset X} m(A) = 1$

m is then called a basic assignment.

The interpretation of m consistent with Shafer's theory is that there exists in the base set X some special unknown element u and $m(A)$ is the degree of belief that this element lies in the set A and nothing smaller than A. In order to help in the understanding of this concept I quote several attempts at clarification from Shafer [1].

"$m(A)$ is the belief that the smallest set that the outcome is in is A."

"$m(A)$ measures the total portion of belief that is confined to A yet none of which is confined to any proper subset of A."

"m(A) measures the belief mass that is confined to A but can move to every point of A."[2]

Note — The formulation of m leads us to the following observations:

1) $m(X)$ is *not* necessarily one.
2) $A \subset B$ does *not* necessarily imply $m(A) \leqq m(B)$.
3) It allows that belief not be committed to either A or not A.

Having introduced the idea of the basic assignment Shafer next introduces the concept of a belief function.

Definition 2. *Given a basic assignment m we can define a belief function*

$$\text{Bel} : 2^X \to [0,1]$$

such that for any $A \subset X$

$$\text{Bel}(A) = \sum_{B \subseteq A} m(B).$$

[1] I have chosen to use the term basic assignment where Shafer uses the term basic probability assignment. I feel that the use of the word probability conjures up certain preconceived notions in the reader which I want to avoid.

[2] If the special element u is the age of some person, then $m(A)$ may measure the degree to which we believe that u is contained in the set *young*, where $A =$ young is defined as a subset of X.

Bel(A) *measures the belief that the special element is a member of A. Whereas $m(A)$ measures the amount of belief that one commits exactly to A alone, Bel(A) measures the total belief that the special element is in A.*

A subset A of X is called a focal element of a belief function Bel if $m(A) > 0$.

Shafer shows that $\mathrm{Bel}(\varnothing) = 0$, $\mathrm{Bel}(X) = 1$ and that for every collection $A_1, A_2, \ldots, A_n$ of subsets of X

$$\begin{aligned}&\mathrm{Bel}(A_1 \cup A_2 \ldots \cup A_n)\\ &\geqq \sum_{\substack{I \subset \{1,2,\ldots,n\}\\ I \neq \varnothing}} (-1)^{|I|+1}\mathrm{Bel}\left(\bigcap_{i\in I} A_i\right),\end{aligned}$$

where $|I|$ denotes the cardinality of the set I.

Shafer also shows that a belief function uniquely determines an underlying basic assignment,

$$m(A) = \sum_{B\subset A} (-1)^{|A-B|}\mathrm{Bel}(B),$$

$|A - B|$ indicates the cardinality of the elements in A not in B.

Shafer next defines the plausibility associated with A.

Definition 3. *Given a belief function* $\mathrm{Bel} : 2^X \rightarrow [0, 1]$ *we define a plausibility function Pl as,*

$$\mathrm{Pl} : 2^X \rightarrow [0, 1]$$

such that for any $A \subset X$

$$\mathrm{Pl}(A) = 1 - \mathrm{Bel}(\bar{A}).$$

Note — The following observations can be made with respect to P1:

1) Pl(A) measures the degree to which one fails to doubt A, where dou $(A) = \mathrm{Bel}(\bar{A})$
2) Pl(A) measures the total belief mass that can move into A, whereas Bel(A) measures the total belief mass that is constrained to A.
3) $\mathrm{Pl}(A) = \sum_{B\cap A\neq\varnothing} m(B)$
4) $\mathrm{Bel}(A) \leqq \mathrm{Pl}(A)$

An important aspect of Shafer's theory involves the combination of belief functions to form a resulting belief function, that is, the combining of various sources of evidence. Shafer accomplishes this by use of Dempster's Rule of Combination. Zadeh [2] has raised some questions as to the appropriateness of this rule. Prade [3] has shown the relationship between Dempster's rule and the intersection of fuzzy sets. Smets [4] has used Shafer belief functions in medical diagnosis. Nguyen [5] has discussed the relationship between belief functions and random sets.

While we shall not in this paper be concerned with the question of the combination of evidence, we shall use a concept developed by Shafer in his approach to combining evidence.

Definition 4. *Assume* Bel_1 *and* Bel_2 *are two belief functions over* 2^X *with their associated basic assignments* m_1 *and* m_2*. The weight of conflict between* Bel_1 *and* Bel_2*, denoted* $\mathrm{Con}(\mathrm{Bel}_1, \mathrm{Bel}_2)$*, is defined as*

$$\mathrm{Con}\,(\mathrm{Bel}_1, \mathrm{Bel}_2) = -\ln(1-k)$$

where

$$k = \sum_{\substack{i,j \\ A_i \cap B_j = \varnothing}} m_1(A_1) \cdot m_2(B_j).$$

The situation of no conflict occurs when $k = 0$ and hence $\mathrm{Con}(\mathrm{Bel}_1, \mathrm{Bel}_2) = 0$. If Bel_1 and Bel_2 are flatly contradictory $k = 1$ and $\mathrm{Con}(\mathrm{Bel}_1, \mathrm{Bel}_2) = \infty$. Thus $\mathrm{con}(\mathrm{Bel}_1, \mathrm{Bel}_2) \geqq 0$ and increases with increasing conflict.

3 Types of Belief Functions

Shafer introduces various classes of belief functions. We shall discuss some of these in the following.

Definition 5. *A belief function over* 2^X *is called a vacuous belief function if*

$$\mathrm{Bel}(X) = 1 \text{ and } \mathrm{Bel}(A) = 0 \text{ for } A \neq X.$$

Note

1) If Bel is a vacuous belief function, then $m(X) = 1$ and $m(A) = 0$ for $A \neq X$.
2) Vacuous belief functions are used in situations where there is no evidence.

Definition 6. *A belief function is called a simple support function focused at* A *if*

$$\mathrm{Bel}(B) = \begin{cases} 0 & \text{if} \quad A \not\subset B \\ 1 & \text{if} \quad B = X \\ s & \text{if} \quad A \subset B, B \neq X. \end{cases} \qquad \text{for} \quad 0 < s < 1$$

Note If Bel is a simple support function focussed at A, then its basic assignment function m is:

$$m(A) = \mathrm{Bel}(A) = s$$
$$m(X) = 1 - \mathrm{Bel}(A) = 1 - s$$
$$m(B) = 0 \text{ for all others.}$$

The simple support function focused at A is used to indicate the situation that we think the special outcome is in A with belief s.

We shall call the simple support function focused at A with $m(A) = 1$ the *certain support function focused* at A.

Definition 7. *A belief function on 2^X is said to be a Bayesian belief function if*

$$\text{Pl}(A) = \text{Bel}(A) \text{ for all } A \subset X.$$

Note The following are two equivalent formulations of a Bayesian belief function.

I) $\text{Bel}(\varnothing) = 0$
$\text{Bel}(X) = 1$
$\text{Bel}(A \cup B) = \text{Bel}(A) + \text{Bel}(B)$, whenever $A \cap B = \varnothing$

II) $\text{Bel}(A) + \text{Bel}(\bar{A}) = 1$

Theorem 1. *If Bel is a Bayesian belief function, then the basic assignment m is such that m takes non-zero values for only subsets of X that are singletons. Hence*

$$\sum_{x \in X} m(\{x\}) = 1$$

The Bayesian structure implies that none of the evidence mass has freedom of movement.

The Bayesian structure forms the prototype in Shafer's theory for probabilistic uncertainty in which the basic assignment function m plays the role of the probability distribution function p. That is, every probability distribution $p : 2^X \to [0, 1]$ can be associated with a Bayesian belief function in which $p(x) = m(\{x\})$.

We note that a Bayesian structure is fully defined by a point function of X equal to $m(\{x\})$.

Since

$$\text{Bel}(A) = \sum_{B \subseteq A} m(B) = \sum_{x \in A} m\left(\{x\}\right),$$

and since $\text{Pl}(A) = \text{Bel}(A)$ for Bayesian belief structure,

$$\text{Pl}(A) = \sum_{x \in A} m\left(\{x\}\right).$$

Furthermore,

$$\text{Bel}\left(\{x\}\right) = \text{Pl}\left(\{x\}\right) = m\left(\{x\}\right).$$

Hence

$$\text{Bel}(A) = \sum_{x \in A} \text{Bel}\left(\{x\}\right) = \text{Pl}(A) = \sum_{x \in A} \text{Pl}\left(\{x\}\right).$$

Definition 8. *A belief function Bel: $2^X \to [0, 1]$ is said to be consonant if*

1) $\text{Bel}(\varnothing) = 0$
2) $\text{Bel}(X) = 1$
3) $\text{Bel}(A \cap B) = \text{Min}(\text{Bel}(A),\ \text{Bel}(B))$ for all $A,\ B \subset C$

Note — The following are two equivalent formulations of a consonant belief function:

1) $\text{Pl}(A \cup B) = \text{Max}(\text{Pl}(A),\ \text{Pl}(B))$
2) $\text{Pl}(A) = \text{Max}_{x \in A}[\text{Pl}(\{x\})]$ for all $A \neq \varnothing$

Note — Every simple support function is consonant.

Note — If Bel is a consonant belief function, then for all $A \subset X$ either $\text{Bel}(A) = 0$ or $\text{Bel}(\bar{A}) = 0$.

The characterization of a consonant belief function is expressed by the following theorem (Shafer).

Theorem 2. *A belief function is consonant if the focal elements of its basic assignment function m are nested. That is, if there exists a family of subsets of X, A_i, $i = 1, 2, \ldots, n$, such that $A_i \subset A_j$ for $i < j$ and $\Sigma_i m(A_i) = 1$.*

Note — A consonant belief structure is completely determined by a point function

$$f : X \to [0, 1]$$

such that $f(x) = \text{Pl}(\{x\})$. At least one element $x \in X$, has $f(x) = 1$. This follows since for any $A \subset X$, $\text{Pl}(A) = \text{Max}_{x \in A}[\text{Pl}(x)]$. Hence Pl is completely determined by Pl defined over the point set X. Since $\text{Bel}(A) = 1 - \text{Pl}(\bar{A})$, $\text{Bel}(A)$ is also uniquely determined. Since $\text{Bel}(A)$ uniquely determines m we have completely defined the structure from this mapping.

This relationship can be made even clearer with the following construction suggested by Prade [2].

Assume we have a consonant belief structure. We can always build a nested sequence of sets

$$\{x_1\} \subset \{x_1, x_2\} \subset \{x_1, x_2, x_3\} \subset \ldots \subset X,$$

indicating these sets as $A_1 \subset A_2 \subset A_3 \ldots \subset A_n = X$ such that $\Sigma_{i=1}^{n} m(A_i) = 1$. Hence all the belief mass lies in this nested sequence. (Some of the elements in the sequence may have zero basic assignment but any subset not in the sequence definitely has zero basic assignment.)

Since

$$\text{Pl}(B) = \sum_{B \cap A \neq \varnothing} m(A),$$

$$\text{Pl}(\{x\}) = \sum_{\{x\} \cap A \neq \varnothing} m(A) = \sum_{\substack{i \\ \{x\} \cap A_{i \neq \varnothing}}} m(A_i) = \sum_{\substack{i \\ \text{such that} \\ x \in A_i}} m(A_l).$$

Therefore

$$\text{Pl}(\{x_1\}) = m(A_1) + m(A_2) + \ldots m(A_{n-1}) + m(X)$$

$$\begin{array}{lccc}
\mathrm{Pl}\left(\{x_2\}\right) = & m\left(A_2\right) + & m\left(A_{n-1}\right) + m(X) & \\
\mathrm{Pl}\left(\{x_n\}\right) = m\left(X\right) & \vdots & \vdots & \vdots
\end{array}$$

Conversely

$$\begin{gathered}
m(A_i) = \mathrm{Pl}\left(\{x\}\right) - \mathrm{Pl}\left(\{x_{i+1}\}\right) \\
m(X) = \mathrm{Pl}\left(\{x_n\}\right) \\
m(A) = 0 \text{ for all else.}
\end{gathered}$$

The consonant belief structure forms the prototype for the possibilistic type of uncertainty introduced by Zadeh [6] in which the plausibility measure in Shafer's theory plays the role of the possibility measure π in Zadeh's theory. Furthermore, since $\mathrm{Bel}(A) = 1 - \mathrm{Pl}(\bar{A})$, the belief function is analagous to Zadeh's measure of certainty [6].

The representations of both these common types of uncertainty in a similar format allows for a comparison of the two types of uncertainty. We see that in a certain respect possibilistic and probabilistic (consonant and Bayesian) uncertainty are opposite extremes. Whereas possibilistic uncertainty assigns its beliefs mass m to a nested sequence of sets, probabilistic uncertainty assigns its belief mass to a collection of disjoint sets. There exists only one type of belief structure which satisfies both structures.

Theorem 3. *The certain support function focused at $\{x\}$, i.e., such that $m(\{x\}) = 1$ for some $x \in X$ is the only belief function that is both a Bayesian and a consonant belief function.*

Note — This structure is a certainty structure in that we know that the special element is x.

4 Entropy Like Measure

An important concept in the theory of probability is Shannon's measure of entropy for a probability distribution. This is a measure of the discordance associated with a probability distribution. We shall introduce here a measure of entropy associated with a basic assignment function m.

Definition 9. *Assume that m is a basic assignment over 2^X with associated belief function Bel.*

We define the entropy of m as

$$Em = \sum_{A \subset X} m(A) \cdot \mathrm{Con}\left(\mathrm{Bel},\ \mathrm{Bel}_A\right)$$

where Bel_A is the certain support function focused at A. The next theorem justifies our use of the term entropy.

Theorem 4. *Assume that m is a Bayesian structure. Then*

$$Em = -\sum_{x\in X} m(x)\cdot \ln m(x).$$

Proof.

$$Em = \sum_{A\subset X} m(A)\cdot \mathrm{con}\left(\mathrm{Bel},\ \mathrm{Bel}_A\right).$$

Since for a Bayesian structure $m(A) = 0$ for all non-singletons,

$$Em = \sum_{x\in X} m(\{x\})\cdot \mathrm{con}\left(\mathrm{Bel},\ \mathrm{Bel}_A\right).$$

We shall denote the basic assignment function associated with the certain support function at $\{x\}$, by g_x. Then

$$g_x\left(\{x\}\right) = 1$$

$g_x(B) = 0$ for all other $B \subset X$, and Con $(\mathrm{Bel},\ \mathrm{Bel}_{\{x\}}) = -\ln(1-k)$, where

$$k = \sum_{\substack{i,j \\ \text{for } A_i\cap B_j=\varnothing}} m(A_i)\cdot g_x\left(B_j\right).$$

Since $g_x(B) = 0$ for $B \neq \{x\}$ and elsewhere equals 1,

$$k = \sum_{\substack{i \\ \text{for } A_i\cap \{x\}=\varnothing}} m(A_i)$$

Since m is Bayesian,

$$k = \sum_{\substack{i \\ \{x_i\}\cap\{x\}=\varnothing}} m\left(\{x_i\}\right) = \sum_{\substack{i \\ \text{for } x_i\neq x}} m\left(\{x_i\}\right) = 1 - m\left(\{x\}\right).$$

Thus

$$\begin{aligned}\mathrm{Con}(\mathrm{Bel},\mathrm{Bel}_{\{x\}}) &= -\ln\left(1-\left(1-m\left(\{x\}\right)\right)\right)\\ &= -\ln\left(m\left(\{x\}\right)\right),\end{aligned}$$

hence

$$Em = -\sum_{x\in X} m\left(\{x\}\right).\ln m\left(\{x\}\right).$$

Thus this definition reduces to the Shannon entropy when the belief structure is Bayesian.

As a simplification for our further work we note that

$$\mathrm{Con}(\mathrm{Bel},\ \mathrm{Bel}_A) = -\ln(1-k)$$
$$\text{and } k = \sum_{\substack{i,j \\ \text{for } A_i \cap B_j = \varnothing}} m(A_i)\cdot m_A(B_j).$$

But since m_A is such that $m_A(A) = 1$ and elsewhere it is zero,

$$k = \sum_{\substack{i \\ A_i \cap A = \varnothing}} m(A_i).$$

However, since

$$1 = \sum_{A_i \subset A} m(A_i) = \sum_{A_i \cap A = \varnothing} m(A_i) + \sum_{A_i \cap A \neq \varnothing} m(A_i)$$

and since

$$\sum_{A_i \cap A \neq \varnothing} m(A_i) = \mathrm{Pl}(A)$$

it follows that

$$1 - k = \mathrm{Pl}(A),$$

where Pl(A) is the plausibility function associated with A under m. Thus

$$\mathrm{Con}(\mathrm{Bel},\ \mathrm{Bel}_A) = -\ln(\mathrm{Pl}(A)).$$

Hence

$$Em = -\sum_{A \subset X} m(A)\cdot\ln(\mathrm{Pl}(A)) - \sum_{A \subset X} \ln(\mathrm{Pl}(A))^{m(A)}$$

Thus we have proved the following.

Theorem 5. *For a belief structure with basic assignment m and plausibility Pl the entropy of this structure is*

$$Em = -\sum_{A \subset X} \ln\left(\mathrm{Pl}(A)^{m(A)}\right) = -\sum_{A \subset X} m(A)\cdot\ln \mathrm{Pl}(A) :$$

Corollary 1.

$$e^{E_m} = \prod_{A \subset X}\left(\mathrm{Pl}(A)^{-m(A)}\right).$$

Proof.

$$e^{E_m} = e^{-\left(\Sigma \mathrm{Pl}(A)^{m(A)}\right)} = \prod_{A \subset X} e^{-\ln\left(\mathrm{Pl}(A)^{m(A)}\right)}$$
$$= \prod_{A \subset X}\left(\mathrm{Pl}(A)^{-m(A)}\right)$$

Since $\mathrm{Pl}(A) \in [0,1]$ for all $A \subset X$ then $\ln \mathrm{Pl}(A) \leqq 0$ and since $m(A) \in [0,1]$ then

$$Em = -\sum_{A \subset X} m(A) \cdot \ln\left(\mathrm{Pl}\left(A\right)\right) \geqq 0.$$

Thus Em assumes as its minimal value the value zero.

Let us look at the belief structures which take this minimal value for Em.

Theorem 6. *For any simple support belief structure $Em = 0$.*

Proof. Assume our simple support structure is focused at B, with $m(B) = b$. Then since

$$Em = -\sum_{A \subset X} m(A) \cdot \ln \mathrm{Pl}\left(A\right),$$

and since for this type of belief function $m(B) = b,\ m(X) = 1 - b$ and for all sets A not equal to B or $X,\ m(A) = 0$, it then follows that

$$Em = -\left(b \cdot \ln\ \mathrm{Pl}\left(B\right)\right) + \left((1-b) \cdot \ln\ \mathrm{Pl}\left(X\right)\right).$$

Since

$$\begin{aligned} \mathrm{Pl}(A) &= 1 - \mathrm{Bel}(\bar{A}) \text{ we have} \\ \mathrm{Pl}(X) &= 1 - \mathrm{Bel}(\varnothing) = 1 - 0 = 1 \\ \mathrm{Pl}(B) &= 1 - \mathrm{Bel}(\bar{B}) = 1 - 0 = 1 \end{aligned}$$

from which we get $E_m = -(b \ln 1 + (1-b) \ln 1) = 0$.

A more general classification of belief structures with zero entropy can be obtained.

Lemma 1. *Any belief structure for which the plausibility is one at all focal elements has $Em = 0$.*

Proof. This follows directly from

$$Em = -\sum_{A \subset X} m(A) \cdot \ln\ \mathrm{Pl}\left(A\right)$$

and the fact that $\ln 1 = 0$.

Lemma 2. *In a consonant belief structure the plausibility function is one at focal elements.*

Proof. Because of the nested nature of the focal elements of this structure there exists at least one $x \in X$ contained in all the focal elements, denote this x^*.

From the definition of plausibility it follows that

$$\mathrm{Pl}\left(\{x^{*}\}\right) = \sum_{A\cap\{x^{*}\}\neq\varnothing} m\left(A\right)$$

Since x^{*} is contained in all focal elements then $\mathrm{Pl}\{x^{*}\} = \sum_{i} m(A_i) = 1$, where Ai are all the focal elements.

We note that for any $A \subset X$

$$\mathrm{Pl}\left(A\right) = \underset{x\in A}{\mathrm{Max}}\left[\mathrm{Pl}\left\{x\right\}\right].$$

Hence if A_i is a focal element of m, then $x^{*} \in A_i$ and hence $\mathrm{Pl}(A_i) = 1$. Thus we have shown the following theorem.

Theorem 7. *For every consonant believe structure $Em = 0$.*

Since consonant belief structures are isomorphic to possibility distributions and normalized fuzzy subsets, the concept of Shannon like entropy proves to be a meaningless or empty concept in a theory dealing with only normal fuzzy sets.

While it would be nice if only consonant belief structures had zero entropy this is not the case as seen from the following example [10].

Example 1. $X = \{x_1, x_2, x_3\}$

Let

$$A = \{x_1, x_2\} \quad B = \{x_2, x_3\}$$

Assume

$$m(A) = 1/2 \quad m\left(B\right) = 1/2$$

Since neither $A \subset B$ nor $B \subset A$, this is not a consonant belief structure. Our definition for entropy implies for this situation

$$Em = -\left[m\left(A\right)\cdot\ln\ \mathrm{Pl}\left(A\right) + m\left(B\right)\cdot\ln\ \mathrm{Pl}\left(B\right)\right].$$

But

$$\mathrm{Pl}\left(A\right) = \sum_{\substack{D\\ D\cap A\neq\varnothing}} m\left(D\right) = m\left(A\right) + m\left(B\right) = 1$$

and

$$\mathrm{Pl}\left(B\right) = \sum_{\substack{D\\ \cap B\neq\varnothing}} m\left(A\right) = m\left(A\right) + m\left(B\right) = 1$$

Hence $Em = 0$.

Actually the class of zero entropic belief structures can be classified as follows.

From our definition of Em, in order that $Em = 0$, any A where $m(A) \neq 0$ requires that $\ln \mathrm{Pl}(A) = 0$, which requires $\mathrm{Pl}(A) = 1$. Since

$$\mathrm{Pl}\,(A) = \sum_{\substack{B \\ B \cap A \neq \varnothing}} m\,(B)$$

this means that every pair of focal elements must have at least one element in common. Thus we have proved the following.

Theorem 8. *A belief structure has zero entropy if $A_i \cap A_j \neq \varnothing$ for each pair of focal elements.*

Thus we can see that this measure of entropy is related in some way to the disjointedness of the sets containing the evidence mass. We note that disjointedness in the focal elements is related to the discordance in the evidence.

We further note that Bayesian structures, while not the only ones, are prototypical examples of disjoint belief structures.

We now turn to belief structures which produce maximal type values for the entropy.

Theorem 9. *Em is finite.*

Proof. From our definition of Em and the fact that for non focal elements $m(A) = 0$, we get

$$Em = -\sum_{A_i} m\,(A_i) \cdot \ln\ \mathrm{Pl}\,(A_i)\,,$$

where A_i are the focal elements.

Since there are at most a finite number of focal elements, $Em = \infty$ iff $\ln \mathrm{Pl}(A_i) = -\infty$, for some i, hence $\mathrm{Pl}(A_i) = 0$ for some i. However, since $A_i \cap A_i \neq \varnothing \cdot m(A_i) > 0$ implies that $\mathrm{Pl}(A_i) > 0$.

Theorem 10. *Assume we have k focal elements with the values $m(A_i) = a_i$. Then Em is maximal if the focal sets A_i are disjoint, i.e., if $A_i \cap A_j = \varnothing$ for all $i \neq j$.*

Proof.

$$Em = -\sum_{i=1}^{K} m\,(A_i) \cdot \mathrm{Pl}\,(A_i)$$

$$\mathrm{Pl}\,(A_i) = \sum_{\substack{j \\ \text{for } A_i \cap A_j \neq \varnothing}} m\,(A_j) = m\,(A_i)$$

$$+ \sum_{\substack{A_j \\ \text{for } A_i \cap A_j \neq \varnothing \\ i = j}} m\left(A_j\right) = a_i + d_i$$

$$Em = -\sum_{i=1}^{K} a_i \ln\left(a_i + d_i\right).$$

As d_i increases $\ln(a_i + d_i)$ increases and $-\sum_{i=1}^{K} a_i \ln(a_i + d_i)$ decreases hence *Em* is maximal when $d_i = 0$ for all i. This occurs when all the A_j are disjoint.

Theorem 11. *Assume we have k disjoint focal elements. Then Em is maximal if* $m(A_i) = 1/K$ *for all elements and in this case*

$$E_m = -\sum_{i=1}^{K} \frac{1}{K} \ln \frac{1}{K} = \ln\ K$$

Proof.

$$Em = -\sum_{i=1}^{K} a_i \ln a_i,$$

where

$$\sum_{i=1}^{n} a_i = 1, a_1 \geqq 0$$

A proof that this well known situation produces a maximal *Em* when $a_i = 1/k$ can be found in Ref. 7.

Assume that we have a belief structure defined over the set X with cardinality N. The maximal number of disjoint subsets of X consist of the N disjoint sets of singletons and this has a value of ln N when the belief mass is equally divided. It appears that this situation induces the largest entropy for a situation where the cardinality of X is N. We say "it appears" since to be certain that this is so, we must prove that there is no non-disjoint collection of *RN* focal elements which have more entropy than the best situation with N disjoint focal elements. We are not ready at this time to prove this theorem.

5 Specificity Like Measure

Yager[8, 9] has introduced a measure of specificity associated with a possibility distribution.

If $\Pi : X \rightarrow [0, 1]$ is a possibility distribution over the finite set X, then Yager[8, 9] has defined the measure of specificity associated with Π as

$$S\left(\Pi\right) = \int_{0}^{\alpha_{\max}} \frac{1}{\text{card } \Pi_\alpha} d\alpha.$$

$\Pi_\alpha = \{x|\Pi(x) \geqq \alpha, x \in X\}$ is a crisp set called the α level set of Π, card Π_α is the number of elements in Π_α and $\alpha_{\max} = \mathrm{Max}_{x\in X}\Pi(x)$.

Yager[8, 9] has shown $S(\Pi)$ to have the following properties:

1) $0 \leqq S(\Pi) \leqq 1$.
2) $S(\Pi) = 1$ iff there exists one and only one $x \in X$ such that $\Pi(x) = 1$ and $\Pi(y) = 0$ for all $y \neq x$.
3) if Π and Π^* are such that $\mathrm{Max}_{x\in X}\Pi(X) = 1$ and $\Pi(x) \leqq \Pi^*(x)$ for all $x \in X$, when
$$S\left(\Pi\right) \geqq S\left(\Pi^*\right).$$

This measure is an indication of the specificity of a possibility distribution in the sense that it indicates the degree to which Π points to one and only one element as its manifestation.

Example 2. Let $X = \{a, b, c, d\}$ and let

$$\Pi(a) = 1$$
$$\Pi(b) = 0.7$$
$$\Pi(c) = 0.5$$
$$\Pi(d) = 0.2$$

$$\begin{array}{lll}
0 \leqq \alpha \leqq 0.2 & \Pi_\alpha = \{a, b, c, d\} & \text{card } \pi_\alpha = 4 \\
0.2 < \alpha \leqq 0.5 & \Pi_\alpha = \{a, b, c\} & \text{card } \pi_\alpha = 3 \\
0.5 < \alpha \leqq 0.7 & \Pi_\alpha = \{a, b\} & \text{card } \pi_\alpha = 2 \\
0.7 < \alpha \leqq 1 & \Pi_\alpha = \{a\} & \text{card } \pi_\alpha = 1
\end{array}$$

$$S(\Pi) = \int_0^1 \frac{1}{\text{card } \Pi_\alpha} d\alpha$$

$$S(\Pi) = \int_0^{0.2} \frac{1}{4} d\alpha + \int_{0.2}^{0.5} \frac{1}{3} d\alpha + \int_{0.5}^{0.7} \frac{1}{2} d\alpha + \int_{0.7}^{1} d\alpha$$

$$S(\Pi) = (0.2)\frac{1}{4} + (0.3)\frac{1}{3} + (0.2)\,\frac{1}{2} + 0.3\,(1) = 0.55.$$

We now generalize this measure from possibilistic belief structures to any belief structure.

Definition 10. *Assume m is a belief structure defined over the set X the generalized specificity measure, denoted Sm, is defined as*

$$S_m = \sum_{\substack{A \subset X \\ A \neq \varnothing}} \frac{m(A)}{n_A}.$$

n_A is the number of elements in the set A, i.e., $n_A = \text{Card } A = |A|$.

First we show that this generalized measure reduces to the particular measure suggested by Yager for possibility distributions, i.e., for consonant belief structures.

Assume that X has n elements with membership grades

$$a_n \leqq a_{n-1} \leqq a_{n-2} \ldots \leqq a_1 = 1$$

Then

$$\begin{aligned} S(\Pi) = &\int_0^{a_n} \frac{1}{n} d\alpha + \int_{a_n}^{a_{n-1}} \frac{1}{n-1} d\alpha \\ &+ \int_{a_{n-1}}^{a_{n-2}} \frac{1}{n-2} d\alpha + \ldots \int_{a_2}^{a_1=1} \Bigg| \, 1 d\alpha \end{aligned}$$

hence

$$\begin{aligned} S(\Pi) = &\frac{1}{n} a_n + \frac{1}{n-1}(a_{n-1} - a_n) + \frac{1}{n-2}(a_n - a_{n-1}) \\ &+ \ldots (a_1 - a_2) \end{aligned}$$

More generally

$$S(\Pi) = \sum_{i=1}^{n} \frac{1}{i}(a_i - a_{i+1}),$$

with $a_{n+1} = 0$ by definition.

Now assume that m is a consonant belief structure.

As Prade [3] has shown, if m is consonant, then there exists a nested family of subsets $A_i \subset X$ such that card $A_i = i$ and $\sum_{i=1}^{n} m(A_i) = 1$, where n is the cardinality of X.

Thus

$$S_m = \sum_{\substack{A \subset X \\ A \neq \varnothing}} \frac{m(A)}{n_A} = \sum_{i=1}^{n} \frac{m(A_i)}{i}$$

Furthermore, it was shown by Prade [3] that if $a_n \leqq a_{n-1} \leqq \cdots \leqq a_1$ are the plausibilities of the singletons, the possibilities of the individual elements, then $m(A_i) = a_i - a_{i+1}$. Thus

$$S_m = \sum_{i=1}^{n} \frac{a_i - a_{i+1}}{i}$$

in the consonant case. This shows that our generalized definition captures the original case.

Theorem 12. *Assume that m is a belief structure over X, where the cardinality of X is n. Then*

$$\frac{1}{n} \leqq S_m \leqq 1.$$

Proof. (1) For any A, $n_A \leqq n$, hence

$$S_m \geqq \frac{1}{n} \sum_A m(A)$$

and since $\sum m(A) = 1$, then $S_m \geqq 1/n$.

2) For any $A \neq \varnothing$, $n_A \geqq 1$, hence

$$S_m \leqq \sum_{A \subset X} m\,(A) \leqq 1$$

Let us look at the situations which attain these extremal values for S_m

Theorem 13. *S_m assumes its minimal value for a given X iff m is a vacuous belief structure. This minimal value is $1/n$ where n is the cardinality of X.*

Proof. (1) If m is vacuous $m(X) = 1$ hence $S_m = 1/n$

2) If m is not vacuous then there exists some A, such that $m(A) > 0$ and $n_A < n$ hence

$$S_m \geqq \frac{1}{n}.$$

Theorem 14. *S_m assumes its maximal value of 1 iff m is a Bayesian belief structure.*

Proof. (1) Assume that m is Bayesian. Then the sets having $m(A) > 0$ are only the singletons. Thus

$$S_m = \sum_{i=1}^{n} m[\{x_i\}] = 1$$

2) Assume that m is not Bayesian. Then there exists some A such that $m(A) > 0$ and $n_A > 1$ hence $S_m < 1$.

Thus whereas the *entropy* measure is *minimized* for *consonant* belief structures the *specificity* is *maximized* for *Bayesian* belief structures.

To get further insight into this measure we consider its evaluation on simple support structures.

Theorem 15. *Assume that m is a simple support structure focused at B, with $m(B) = b$. Then*

$$S_m = \frac{b}{n_B} + \frac{1-b}{n}.$$

Proof. For a simple support structure

$$m(B) = b$$
$$m(X) = 1 - b$$
$$S_m = \frac{b}{n_B} + \frac{1-b}{n}.$$

If b increases S_m increases. Furthermore as n_B decreases, without becoming vacuous, S_m increases.

Let us now examine the workings of this measure on consonant belief structures.

Theorem 16. *Assume that m_1 and m_2 are consonant belief structures generating plausibility measures Pl_1 and Pl_2 such that, for each $x \in X$,*

$$\mathrm{Pl}_1(x) \leqq \mathrm{Pl}_2(x)$$

Then.

$$S_{m_1} \geqq S_{m_2}.$$

Proof. For consonant belief structures

$$S_m = \int\limits_0^1 \frac{1}{\text{Card } \Pi_\alpha} d\alpha$$

Since $\mathrm{Pl}_2(x) > \mathrm{Pl}_1(x)$, card $\prod_{2_\alpha} \geqq$ card $\prod_{1_\alpha}$.

As a special case of this situation consider two consonant belief structures m_1 and m_2 defined over the same nested sets $A_1 \subseteq A_2 \subset \cdots \subset A_n$ where

$$\mathrm{Pl}_1(x) \leqq \mathrm{Pl}_2(x)$$
$$m_2(A_n) = \mathrm{Pl}_2(x_n) > \mathrm{Pl}_1(x_n) = m_1(A_n)$$

so $m_2(A_n) \geqq m_1(A_n)$.

In the same manner for all $j > 1$,

$$m_2(A_n) + m_2(A_{n-1}) + \dots m_2(A_j) > m_1(A_j) + \dots m_1(A_n).$$

But

$$\sum_{i=1}^n m_2(A_i) = \sum_{i=1}^n m_1(A_i),$$

hence $m_1(A_1) \geqq m_2(A_1)$ and $m_2(A_n) \geqq m_1(A_n)$.

Thus the higher the specificity the more of the evidence mass lies in the one element set and the less in the set X.

The meaning of the measure S_m appears to relate to the degree to which the evidence is pointing to a one element realization. When one considers that the total amount of plausibility assigned to the elements in X is

$$\sum_{x_i \in X} \mathrm{Pl}\left(X_i\right) = \sum_{A \subset X} n_A \cdot m\left(A\right)$$

it appears that S_m is a measure of the reduction of excess plausibility. We can also see that as the total plausibility value, which is always greater than the belief, gets closer to the belief value than S_m increases. Hence S_m appears inversely related to excess of plausibility over belief. In bringing the plausibility in a structure closer to the belief ascertained in the structure we are getting more specific in our allocation of evidence. This interpretation is reinforced by the fact that for Bayesian structures in which the plausibility always equals the belief, the value of S_m is maximum.

Since obtaining evidence involves a process of reducing possibilities, specificity thus seems to be measuring the effect of the evidence in that direction.

6 Using Both Measures

We feel that the two measures developed herein provide a complementary approach to measuring the certainty with which a belief structure is pointing to a unique outcome.

As noted, the entropy measure provides a measure of the dissonance of the evidence. This is illustrated by the fact that consonant belief structures have lowest entropic measures, while the highly dissonant type of Bayesian structures have high entropic measures.

The specificity measure provides an indication of the dispersion of the belief. We note that in this situation the Bayesian structure gets the highest grades, while the vacuous case gets the lowest.

As we noted earlier the only structure that is Bayesian, specific and consonant is the structure which $m(x) = 1$ for some $x \in X$. However this structure corresponds to the certain situation where the evidence points precisely to x as the special element.

Thus we see the following: the lower the E_m, the more consistent the evidence; and the higher S_m, the less diverse. Ideally we want low E_m and high S_m for certainty. Thus by using a combination of the two measures we feel that we can have a good indication of the quality of a belief structure with respect to suggesting one element as the outcome.

In particular the measure E_m indicates the success of the structure in reducing plausibilities, which is a desired quality in a belief structure up to a point. This point will be that where the reduction is so great that everything appears not possible, which implies an inconsistency in the evidence. The entropy measure thus indicates the success of the belief structure in being consistent. On the other hand, consistency is also desirable up to a point, this being where we leave everything as possible in order to obtain this consistency. The success with which we are able to satisfy both these criteria therefore provides a good procedure for judging the quality of evidence.

We here suggest as a measure of quality of a belief structure the two tuple (S_m, E_m). As we have noted, the ideal situation, certain knowledge, occurs only when $(S_m, E_m) = (1, 0)$. The closer a belief structure is to this point, the better quality of evidence it is supplying.

7 Conclusion

We have extended Shafer's theory of evidence to include a measure of entropy and specificity to be associated with a belief structure. These measures taken together provide an indication of the quality of the evidence supplied by a belief structure.

8 Acknowledgement

I would like to acknowledge my debt to Didier Dubois whose personal tutorial opened my eyes to Shafer's theory.

References

1. G. Shafer, *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, 1976.
2. L. A. Zadeh, "On the validity of Dempster's rule of combination of evidence." Memo No. ERL M79/24, U. of California, Berkeley, 1979.
3. H. Prade, "On the link between Dempster's rule of combination of evidence and fuzzy set intersection." *Busefal*, **8**, 1981, pp. 60–64.
4. P. Smets, "Medical diagnosis: fuzzy sets and degrees of belief." *Fuzzy Sets and Systems*, **5**, 1981, pp. 259–266.
5. H. T. Nguyen, "On random sets and belief structures." *J. Math. Anal. and Appl.*, **65**, 1978, pp. 531–542.
6. L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility." *Fuzzy Sets and Systems*, **1**, 1978, pp. 3–28.
7. A. I. Khinchin, *Mathematical Foundations of Information Theory.* Dover Publications, New York, 1957.
8. R. R. Yager, "Measuring tranquility and anxiety in decision making: An application of fuzzy sets." *Int. J. of General Systems*, **8**, 1982, pp. 139–146.
9. R. R. Yager, "Measurement of properties of fuzzy sets and possibility distributions." *Proc. Third International Seminar on Fuzzy Sets*, Linz, 1981, pp. 211–222.
10. M. Higashi and G. J. Klir, "Measures of uncertainty and information based on possibility distributions." *Int. J. of General Systems*, **9**, No. 1, 1982, pp. 43–58.

12

A Method for Managing Evidential Reasoning in a Hierarchical Hypothesis Space*

Jean Gordon and Edward H. Shortliffe

Abstract. Although informal models of evidential reasoning have been successfully applied in automated reasoning systems, it is generally difficult to define the range of their applicability. In addition, they have not provided a basis for consistent management of evidence bearing on hypotheses that are related hierarchically. The Dempster–Shafer (D-S) theory of evidence is appealing because it does suggest a coherent approach for dealing with such relationships. However, the theory's complexity and potential for computational inefficiency have tended to discourage its use in reasoning systems. In this paper we describe the central elements of the D-S theory, basing our exposition on simple examples drawn from the field of medicine. We then demonstrate the relevance of the D-S theory to a familiar expert-system domain, namely the bacterial-organism identification problem that lies at the heart of the MYCIN system. Finally, we present a new adaptation of the D-S approach that achieves computational efficiency while permitting the management of evidential reasoning within an abstraction hierarchy.

1 Introduction

The representation and manipulation of incomplete and imperfect knowledge are issues central to the design of reasoning systems. Drawbacks in traditional probabilistic approaches to the management of such uncertainty led us to develop the *certainty factor* (CF) model of inexact reasoning [15]. The initial CF model was implemented in the medical advice program known as MYCIN and subsequently adapted for use in similar (EMYCIN) systems [3]. However, despite the model's good performance in many task domains, its restrictive assumptions [1] and its inability to deal consistently with hierarchical relationships among values of parameters have left us dissatisfied with

* This research was supported by the National Library of Medicine under grants LM-03395 and LM-00048. Dr. Gordon was supported by the Stanford Medical Alumni Scholars Program. Dr. Shortliffe was a Henry J. Kaiser Family Foundation Faculty Scholar in General Internal Medicine. Computing resources were provided by the SUMEX-AIM facility under NIH grant RR-00785.

the generality of the approach. We have accordingly been attracted to the mathematical theory of evidence developed by Arthur Dempster. Although it also makes assumptions that do not hold in all problem-solving domains, its coherent approach to the management of uncertainty among hierarchically related hypotheses merits careful study and interpretation in the context of automated reasoning systems.

This theory was first set forth by Dempster in the 1960s and subsequently extended by Glenn Shafer when he published *A Mathematical Theory of Evidence* [14]. The theory's relevance to the issues addressed in the CF model was not immediately recognized [19], but recently researchers have begun to investigate applications of the theory to artificial intelligence systems [2, 6, 7, 10, 11, 16].

An advantage of the Dempster–Shafer (D-S) theory over previous approaches is its ability to model the narrowing of the hypothesis set with accumulation of evidence, a process which characterizes diagnostic reasoning in medicine and expert reasoning in general. An expert uses evidence which may apply not only to single hypotheses but also to sets of hypotheses that together comprise a concept of interest. The functions and combining rule of the D-S theory are well suited to represent this type of evidence and its aggregation.

We believe there are several reasons why the D-S theory is not yet well appreciated by the artificial intelligence research community. One problem has been the mathematical notation used in most of the books and papers that discuss it. In addition, the discussions generally lack simple examples that could add clarity to the theory's underlying notions. Finally, the D-S theory is widely assumed to be impractical for computer-based implementation due to an evidence-combination scheme that assures computational complexity with exponential-time requirements. Although we could not totally avoid mathematical notation in this paper, we do address all three of the issues cited here, paying particular attention to methods for applying the theory in ways that are computationally tractable.

In 1981, Barnett showed that apparent exponential-time requirements of the D-S model could be reduced to simple polynomial time if the theory were applied to single hypotheses, and to their negations, and if evidence were combined in an orderly fashion [2]. However, Barnett's proposal did not solve the larger problem of how to allow evidential reasoning about *sets of hypotheses* in a way that is computationally tractable for complex domains.

In this paper we propose a technique that permits adapting the D-S theory so that hierarchical relationships among hypotheses are handled in a consistent manner. The method builds on Barnett's approach, augmenting it to provide the additional features in a computationally efficient manner. We shall show that the technique requires an assumption (that the hypothesis space can be reduced to a strict hierarchy) and an approximation (it assigns disconfirmatory evidence only to hypotheses with 'meaning' in the domain), but it does manage to capture the major strengths of the D-S theory while achieving a computationally tractable execution time and, hence, a practical method for its implementation.

We accordingly have three goals in this paper. First, in Sect. 2 we wish to describe for an AI audience the central elements of the D-S theory, avoiding excessive mathematical notation and basing our exposition on simple examples drawn from the field of medicine. In Sect. 3 we demonstrate the relevance of the D-S theory to a familiar expert-system domain, namely the bacterial-organism identification problem that lies at the heart of MYCIN [3]. Since MYCIN's identification rules deal with single hypotheses and ignore hierarchical relationships, the Barnett technique is directly relevant to the program's task. In Sect. 4 we present an adaptation of the D-S approach that allows computationally efficient reasoning within abstraction hierarchies.

The importance of hierarchical relationships among hypotheses can best be appreciated in the setting of a simple example. Consider MYCIN's task of bacterial-organism identification. Here the hypothesis set is a group of over 100 organisms known to the program. By focusing on single organisms (hypotheses), MYCIN's rules and CF model are unable to deal with *groups of organisms* as hypotheses that have explicit relationships to the single bacteria about which knowledge is available. Such relationships, if they exist, must be specified in MYCIN using additional rules; they are not reflected automatically in the structure of the hypothesis space for the domain. When searching for the identity of an infecting organism, however, microscopic examination of a smear showing gram-negative (pink-staining) organisms narrows the hypothesis set of the 100 or so possible organisms to a proper subset. This subset can also be thought of as a new hypothesis: *the organism is one of the gram-negative organisms*. However, this piece of evidence gives no information concerning the relative likelihoods of the individual organisms in the subset. Bayesians[1] might assume equal prior probabilities and distribute the weight of this evidence equally among the gram-negative organisms but, as Shafer points out, they would thus fail to distinguish between uncertainty, or lack of knowledge, and equal certainty. Because the D-S approach allows one to attribute belief to subsets, as well as to individual elements of the hypothesis set, we believe that it is similar to the evidence-gathering process observed when human beings reason at varying levels of abstraction.

A second piece of evidence, such as the morphology (shape) of the organism, narrows the original hypothesis set (the 100 or so bacterial organisms) to a different subset. How does the D-S theory pool this new piece of evidence with the first? Each is represented by a belief function, and the two belief functions thus must be merged using a combination rule to yield a new function. Belief functions assign numerical measures of belief to hypotheses based on observed evidence. In a rule-based expert system, for example, each inferential rule would have its own belief function associated with it, a function

[1] A thorough discussion of Bayesian theory and its application to medical diagnostic reasoning may be found in [17].

that assigns belief to the consequent based on the evidence in the premise. The combination rule proposed by Dempster, like the Bayesian and CF combining functions, is independent of the order in which evidence is gathered. In fact, the D-S combination rule includes the Bayesian and CF functions as special cases.

Another consequence of the generality of the D-S belief functions is avoidance of the Bayesian restriction that commitment of belief to a hypothesis implies commitment of the remaining belief to its negation, i.e., the assumption that belief in H is equivalent to $P(H)$ so that the resulting belief in NOT-H is $1 - P(H)$. The concept that, in many situations, evidence partially in favor of a hypothesis should *not* be construed as evidence partially against the same hypothesis (i.e., in favor of its negation) was one of the desiderata in the development of the CF model [15]. As in that model, the D-S measures of belief assigned to each hypothesis in the original set need not sum to 1 but may sum to a number less than 1; some of the remaining belief can be allotted to sets of hypotheses that comprise higher-level concepts of interest.

Although the D-S theory includes many of the features of the CF model, its derivation is based on set-theoretic notions which allow explicit and consistent handling of subset and superset relationships in a hierarchy of hypotheses. As we shall show, this feature provides a conceptual clarity that is lacking in the CF model. In the next sections, we motivate the exposition of the theory with a medical example and then discuss the relevance of the theory to systems that reason in hierarchically organized hypothesis spaces.

2 Basics of the Dempster-Shafer Theory

2.1 A Simple Example of Medical Reasoning

Suppose a physician is considering a case of cholestatic jaundice, i.e., the development of a yellow hue to a patient's skin (jaundice) due to elevated blood levels of bilirubin (a pigment produced by the liver). This problem is caused by an inability of the liver to excrete bile normally, often due to a disease within the liver itself (intrahepatic cholestasis) or blockage of the bile ducts outside the liver (extrahepatic cholestasis). In a typical case of this type, the diagnostic hypothesis set might well include two types of intrahepatic cholestasis, hepatitis (Hep) and cirrhosis (Cirr), and two types of extrahepatic cholestasis, gallstones (Gall) and pancreatic cancer (Pan). There are actually more than four causes of jaundice, but we have simplified the example here for illustrative purposes. In the D-S theory, this set of four disorders is called a frame of discernment, denoted Θ or {Hep, Cirr, Gall, Pan}. As noted earlier, the hypotheses in Θ are assumed mutually exclusive and exhaustive.

One piece of evidence considered by the physician might lend support to the diagnosis of intrahepatic cholestasis rather than to a single disease, i.e., it might support the two-element subset of Θ, {Hep, Cirr}. Note that this subset corresponds to the hypothesis which is the disjunction of its elements, viz. the hypothesis HEP-OR-CIRR. Similarly, the hypothesis extrahepatic cholestasis = {Gall, Pan} = GALL-OR-PAN. Evidence confirming intrahepatic cholestasis to some degree will cause the physician to allot belief to the subset {Gall, Pan}.

Subsequently a new piece of evidence might help the physician exclude hepatitis to some degree. Evidence disconfirming HEP (i.e., disconfirming the set {Hep}) is equivalent to evidence confirming the hypothesis NOT-HEP, which corresponds to the hypothesis CIRR-OR-GALL-OR-PAN or the subset {Cirr, Gall, Pan}. Thus, evidence disconfirming {Hep} to some degree will cause the physician to allot belief to this three-element subset. Note, however, that although evidence disconfirming the set {Hep} may be seen as confirming the set {Cirr, Gall, Pan}, it says nothing about how the belief in the three-element subset should be allocated among the singleton hypotheses {Cirr}, {Gall}, and {Pan}.

As illustrated above, any subset of hypotheses in Θ gives rise to a new hypothesis, which is equivalent to the disjunction of the hypotheses in the subset. Each element in Θ corresponds to a one-element subset (called a singleton). By considering all possible subsets of Θ, denoted 2^{Θ}, the set of hypotheses to which belief can be allotted is enlarged. Henceforth, we use the term 'hypothesis' in this enlarged sense to denote any subset of the original hypotheses in Θ. We shall also hereafter use set notation to refer to the corresponding hypothesis, e.g., {Cirr, Hep} refers to the hypothesis HEP-OR-CIRR, {Pan} refers to the hypothesis PAN, etc.

A diagrammatic representation of 2^{Θ} for the cholestasis example is given in Fig. 1. Note that a set of size n has 2^n subsets. (The empty set, $\emptyset$, is one of these subsets, but is not shown in Fig. 1; it corresponds to a hypothesis known to be false since the hypotheses in Θ are exhaustive.)

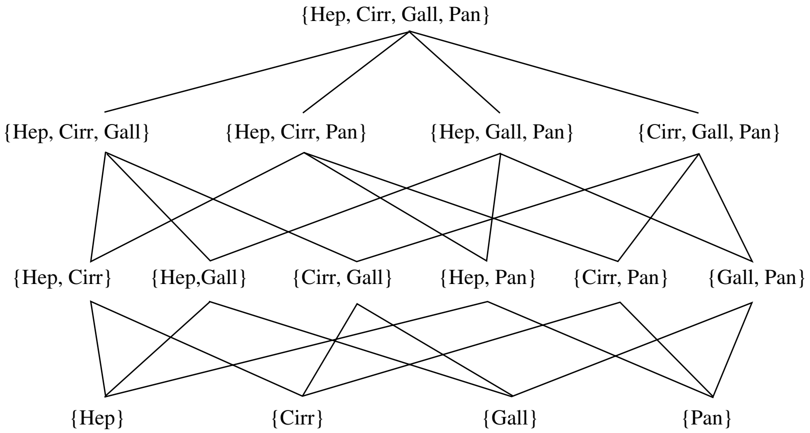


Fig. 1. The subsets of the set of causes of cholestasis

2.2 Basic Probability Assignments

The D-S theory uses a number in the range [0, 1] inclusive to indicate belief in a hypothesis given a piece of evidence. This number is the degree to which the evidence supports the hypothesis.[2] Recall that evidence against a hypothesis is regarded as evidence for the negation of the hypothesis, i.e., for the complement in the set-theoretic interpretation of hypotheses introduced in the previous section. Thus, unlike the CF model, the D-S model avoids the use of negative numbers to represent disconfirming evidence.

The impact of each distinct piece of evidence on the subsets of Θ is represented by a function called a basic probability assignment (*bpa*). A *bpa* is a generalization of a probability mass distribution; the latter assigns a number in the range [0, 1] to every *singleton* of Θ such that the numbers sum to 1. Using 2^{Θ}, the enlarged domain of all subsets of Θ, a *bpa*, denoted m, assigns a number in [0, 1] to every *subset* of Θ such that the numbers sum to 1. (By definition, the number 0 must be assigned to the empty set, since this set corresponds to a false hypothesis.) Thus, m_i allows assignment of a portion of the total belief of 1, based on a given piece of evidence i, to every element in the hierarchy of Fig. 1, not just to those elements on the bottom row as is the case for a probability density function.

The quantity $m(A)$ is a measure of that portion of the total belief committed exactly to A, where A is an element of 2^{Θ}. This portion of belief cannot be further subdivided among the subsets of A and does not include portions of belief committed to subsets of A. Since belief in A certainly entails belief in all subsets of Θ containing A (i.e., nodes 'higher' up in the network of Fig. 1), it would be useful to define a function which computes a *total* amount of belief for each subset in Θ. This function applied to a subset in 2^{Θ}, A, would include not only belief committed exactly to A but to all subsets of A. Such a function, called a belief function in the D-S model, is defined in the next section.

The quantity, $m(\Theta)$, is a measure of that portion of the total belief which is committed to Θ, i.e., which remains unassigned after commitment of belief to various proper subsets of Θ. For example, evidence favoring a single subset A need not say anything about belief in the other subsets. If $m(A) = s$ and m assigns no belief to other subsets of Θ, then $m(\Theta) = 1 - s$. Thus, the remaining belief is assigned to Θ and *not* to the negation of the hypothesis (equivalent to A^{c}, the set-theoretic complement of A), as would be assumed in the Bayesian model.

Example 1. Suppose there is no evidence concerning the specific diagnosis in a patient with known cholestatic jaundice, i.e., a patient for whom $\Theta = \{\text{Cirr, Hep, Gall, Pan}\}$. The *bpa* representing ignorance, called the vacuous *bpa*, assigns 1 to $\Theta = \{\text{Hep, Cirr, Gall, Pan}\}$ and 0 to every other subset

[2] Note that this definition corresponds to the notion of a *measure of belief* (MB) in the CF model.

of Θ. Bayesians might attempt to represent ignorance by a function assigning 0.25 to each singleton hypothesis ({Hep}, {Cirr}, {Gall}, and {Pan}), or by a function apportioning the total belief in accordance with information regarding prevalence of the four disorders in the population. As remarked before, however, such functions would imply more information given by the evidence than is truly the case.

Example 2. Suppose that the evidence supports, or confirms, the diagnosis of intrahepatic cholestasis = {Hep, Cirr} to the degree 0.6, but does not support a choice between cirrhosis and hepatitis. The remaining belief, $1-0.6 = 0.4$, is assigned to Θ. The hypothesis corresponding to Θ is known to be true under the assumption of exhaustiveness. Thus, $m(\{\text{Hep, Cirr}\}) = 0.6$, $m(\Theta) = m(\{\text{Hep, Cirr, Gall, Pan}\}) = 0.4$ and the value of m for every other subset of Θ is 0. Bayesians might have assigned the remaining belief to extrahepatic cholestasis= {Gall, Pan}, the negation (complement) of intrahepatic cholestasis, rather than to Θ.

Example 3. Suppose that the evidence disconfirms the diagnosis of {Hep} to the degree 0.7. This is equivalent to confirming that of {Cirr, Gall, Pan} to the degree 0.7. Thus, $m(\{\text{Cirr, Gall, Pan}\}) = 0.7$, $m(\Theta) = 0.3$ and the value of m for every other subset of Θ is 0. Note that the notion of disconfirmation does not have a clear correlate in classical probability theory; the CF theory, for example, was developed largely in an effort to address the need to define relationships between confirmation and disconfirmation.

Example 4. Suppose that the evidence confirms the diagnosis of {Hep} to the degree 0.8. Then, $m(\{\text{Hep}\}) = 0.8$, $m(\Theta) = 0.2$, and m is 0 elsewhere.

2.3 Belief Functions

A belief function, denoted *Bel*, corresponding to a specific *bpa, m*, assigns to every subset A of Θ the sum of the beliefs committed exactly to every subset of A by m. For example:

$$\begin{aligned} Bel\left(\{\text{Hep, Cirr,Pan}\}\right) = & m\left(\{\text{Hep, Circ, Pan}\}\right) + m\left(\{\text{Hep, Cirr}\}\right) \\ & + m\left(\{\text{Hep, Pan}\}\right) + m\left(\{\text{Cir, Pan}\}\right) \\ & + m\left(\{\text{Hep}\}\right) + m\left(\{\text{Cirr}\}\right) + m\left(\{\text{Pan}\}\right). \end{aligned}$$

Thus, $Bel(A)$ is a measure of the *total* amount of belief in A and not the amount committed precisely to A by the evidence corresponding to the *bpa m*.

This relationship may be clarified by referring to Fig. 1. Note that the following observations follow from the definition given:

(1) Bel and m are equal for singletons. For example, $Bel(\{\text{Hep}\}) = m(\{\text{Hep}\})$.
(2) $Bel(A)$, where A is any other subset of Θ, is the sum of the values of m for every subset in the subhierarchy formed by using A as root. For example,

$$\begin{aligned} Bel\,(\text{intrahepatic cholestasis}) &= Bel\,(\{\text{Hep, Cirr}\}) \\ &= m\,(\{\text{Hep, Cirr}\}) + m\,(\{\text{Hep}\}) \\ &\quad + m\,(\{\text{Cirr}\})\,. \end{aligned}$$

(3) $Bel(\Theta)$ is always equal to 1 since $Bel(\Theta)$ is the sum of the values of m for every subset of Θ. This sum must be 1 by definition of a *bpa*. Clearly, the total amount of belief in Θ should be equal to the total amount of belief, 1, since the singletons are exhaustive. In Fig. 1, this means that Bel(cholestatic jaundice)= $Bel(\Theta) = 1$.

To further illustrate, the belief function corresponding to the *bpa* of Example 2 above is given by $Bel(\Theta) = 1$, $Bel(A) = 0.6$, where A is any proper subset of Θ containing {Hep, Cirr}, and the value of Bel is 0 for every other subset of Θ.

2.4 Combination of Belief Functions

The evidence-gathering process for diagnosis requires a method for combining the support for a hypothesis, or for its negation, based upon multiple, accumulated observations [15]. The D-S model also recognizes this requirement and provides a formal proposal for its management. Given two *bpa*'s, each with the same frame of discernment Θ but based on two different observations (e.g., two different inferential rules lending positive or negative support to the same or competing hypotheses in an expert system), Dempster's combination rule shown below computes a new *bpa* which represents the impact of the combined evidence.

Concerning the validity of this rule, Shafer writes that although he can provide "no conclusive *a priori* argument,. . .it does seem to reflect the pooling of evidence." In the special case of a frame of discernment containing two elements, Dempster's rule can be found in Johann Heinrich Lambert's book, *Neues Organon*, published in 1764. In another special case where the two *bpa*'s assign evidential support to exactly one and the same hypothesis, the rule reduces to that found in the MYCIN CF model and in *Ars Conjectandi*, the work of the mathematician Jean Bernoulli in 1713. It is based on intuition of how evidence should combine, however, and not on any formal underlying theory.

The Dempster combination rule differs from the CF combining function in the pooling of evidence supporting mutually exclusive hypotheses. For example, evidence supporting {Hep} reduces belief in each of the singleton hypotheses—{Cirr}, {Gall}, {Pan}—and in any disjunction (subset of Θ) not containing {Hep}, e.g., {Cirr, Gall, Pan}, {Cirr, Pan}, etc. As we discuss later, if the D-S model were adapted for use in an EMYCIN system, each new piece of evidence would have an indirect impact on competing hypotheses, a feature not provided by the CF model. The Dempster combination rule also differs from the CF model in its approach to the assignment of belief in a hypothesis when confirming and disconfirming evidence is pooled.

Let Bel_1, Bel_2 and m_1, m_2 denote two belief functions and their corresponding *bpa*'s, respectively. The D-S combination rule defines a new *bpa*, denoted $m_1 \oplus m_2$, which represents the combined effect of m_1 and m_2. The corresponding belief function, denoted $Bel_1 \oplus Bel_2$, may then be computed from $m_1 \oplus m_2$ by definition of a belief function.

The Dempster combining function, also known as Dempster's rule, suggests that $m_1 \oplus m_2$ may be calculated from m_1 and m_2 by considering all products of the form $m_1(X)m_2(Y)$ where X and Y are individually varied over all subsets of Θ. It can be shown that the resulting function is itself a *bpa* since the result of summing all such products is 1 by elementary algebra and the definition of a *bpa*:

$$\sum m_1(X)\, m_2(Y) = \sum m_1(X) \sum m_2(Y) = 1 \times 1 = 1.$$

Dempster's rule states that the *bpa* representing the combination of m_1 and m_2 apportions the total amount of belief among the subsets of Θ by assigning $m_1(X)m_2(Y)$ to the set intersection of X and Y. Note that there are typically several different subsets of Θ whose intersection yields the same subset of Θ. In the cholestatic jaundice example of Fig. 1, for example, the set {Hep, Cirr} will be obtained by intersecting {Hep, Cirr} with any superset of {Hep, Cirr}, by intersecting {Hep, Cirr, Pan} with {Hep, Cirr, Gall}, etc. Thus, for every subset A of Θ, Dempster's rule defines $m_1 \oplus m_2(A)$ to be the sum of all products of the form $m_1(X)m_2(Y)$ where X and Y are selected from the subsets of Θ in all possible ways such that their intersection is A. The commutativity of multiplication ensures that the rule yields the same value regardless of the order in which the functions are combined. This is an important property since evidence aggregation should be independent of the order of its gathering. The following two examples illustrate the combination rule.

Example 5. As in Examples 2 and 3, suppose that for a given patient, one observation supports intrahepatic cholestatic = {Hep, Cirr} to degree 0.6 (m_1) whereas another disconfirms hepatitis (i.e., confirms {Cirr, Gall, Pan}) to degree 0.7 (m_2). Then our net belief based on both observations is given by $m_1 \oplus m_2$. For illustrative purposes, an 'intersection tableau' with values assigned by m_1 and m_2 along the rows and columns, respectively, is a helpful device. Only nonzero values assigned by m_1 and m_2 need be considered since if $m_1(X)$ and/or $m_2(Y)$ is 0, then the product $m_1(X)m_2(Y)$ contributes 0 to $m_1 \oplus m_2(A)$, where A is the intersection of X and Y. Entry i, j in the tableau is the intersection of the subsets in row i and column j. Clearly, a given subset of Θ may occur in more than one location of the tableau. The product of the *bpa* values is shown below in parentheses next to the subset. The value of

$m_1 \oplus m_2(A)$ is computed by summing the products in the tableau that are noted in parentheses adjacent to each occurrence of A.

m_2 / m_1	$\{\text{Cirr, Gall, Pan}\}\,(0.7)$	$\Theta(0.3)$
$\{\text{Hep, Cirr}\}\,(0.6)$	$\{\text{Cirr}\}\,(0.42)$	$\{\text{Hep, Cirr}\}\,(0.18)$
$\Theta\,(0.4)$	$\{\text{Cirr, Gall, Pan}\}\,(0.28)$	$\Theta\,(0.12)$

In this example, each subset appears only once in the tableau and $m_1 \oplus m_2$ is easily computed:

$$\begin{aligned}
&m_1 \oplus m_2\,(\{\text{Cirr}\}) = 0.42,\\
&m_1 \oplus m_2\,(\{\text{Hep, Cirr}\}) = 0.18,\\
&m_1 \oplus m_2\,(\{\text{Cirr, Gall, Pan}\}) = 0.28,\\
&m_1 \oplus m_2\,(\Theta) = 0.12,\\
&m_1 \oplus m_2 \text{ is 0 for all other subsets of } \Theta.
\end{aligned}$$

Since $Bel_1 \oplus Bel_2$ is fairly complex, we give only a few sample values:

$$\begin{aligned}
Bel_1 \oplus Bel_2\,(\{\text{Hep, Cirr}\}) &= m_1 \oplus m_2\,(\{\text{Hep, Cirr}\}) + m_1 \oplus m_2\,(\{\text{Hep}\})\\
&\quad + m_1 \oplus m_2\,(\{\text{Cirr}\})\\
&= 0.18 + 0 + 0.42\\
&= 0.60,\\
Bel_1 \oplus Bel_2\,(\{\text{Cirr, Gall, Pan}\}) &= m_1 \oplus m_2\,(\{\text{Cirr, Gall, Pan}\})\\
&\quad + m_1 \oplus m_2\,(\{\text{Cirr, Gall}\})\\
&\quad + m_1 \oplus m_2\,(\{\text{Cirr, Pan}\})\\
&\quad + m_1 \oplus m_2\,(\{\text{Gall, Pan}\})\\
&\quad + m_1 \oplus m_2\,(\{\text{Cirr}\})\\
&\quad + m_1 \oplus m_2\,(\{\text{Gall}\}) + m_1 \oplus m_2\,(\{\text{Pan}\})\\
&= 0.28 + 0 + 0 + 0 + 0.42 + 0 + 0\\
&= 0.70,\\
Bel_1 \oplus Bel_2\,(\{\text{Hep, Cirr, Pan}\}) &= Bel_1 \oplus Bel_2\,(\{\text{Hep, Cirr}\}) = 0.60,
\end{aligned}$$

since

$$\begin{aligned}
m_1 \oplus m_2\,(\{\text{Hep, Cirr, Pan}\}) &= m_1 \oplus m_2\,(\{\text{Hep, Pan}\})\\
&= m_1 \oplus m_2\,(\{\text{Cirr, Pan}\}) = 0.
\end{aligned}$$

In this example, the reader should note that $m_1 \oplus m_2$ satisfies the definition of a *bpa*: $\Sigma m_1 \oplus m_2(X) = 1$ where X varies over all subsets of Θ, and $m_1 \oplus m_2(\emptyset) = 0$. We have already shown that the first condition in the definition of a *bpa* is always fulfilled, i.e., the sum of the beliefs assigned to all subsets in Θ by the Dempster rule will always sum to 1. However, the second condition (viz. that a *bpa* assign 0 to the empty set) is problematic in cases where the 'intersection tableau' contains $\emptyset$. This situation did not occur in Example 5 because every two sets with nonzero *bpa* values always had at least one element in common. In general, nonzero products of the form $m_1(X)m_2(Y)$ will be assigned to $\emptyset$ whenever X and Y have nonzero *bpa* values but their intersection is the empty set.

The D-S model deals with this problem by setting $m_1 \oplus m_2(\emptyset)$ equal to 0 and normalizing the remaining *bpa* assignments so that they continue to sum to 1.[3] This behavior is achieved by defining κ as the sum of all nonzero values assigned to $\emptyset$ in a given case ($\kappa = 0$ in Example 5). Dempster then divides all other values of $m_1 \oplus m_2$ by $1 - \kappa$. The revised values still sum to 1 and hence satisfy that condition in the definition of a *bpa*. This approach is illustrated by the following example.

Example 6. Suppose now that, for the same patient as in Example 5, a third belief function (m_3) corresponds to a new observation which confirms the diagnosis of hepatitis to the degree 0.8 (i.e., suppose we have a combination of Examples 4 and 5). We now need to compute $m_3 \oplus m_4$, where $m_4 = m_1 \oplus m_2$ of Example 5.

m_4 m_3	{Cirr} (0.42)	{Hep, Cirr} (0.18)	{Cirr, Gall, Pan} (0.28)	Θ (0.12)
{Hep} (0.8)	$\emptyset$(0.336)	{Hep} (0.144)	$\emptyset$ (0.224)	{Hep} (0.096)
Θ (0.2)	{Cirr} (0.084)	{Hep, Cirr} (0.036)	{Cirr, Gall, Pan} (0.056)	Θ (0.024)

In this example, there are two null entries in the tableau, one assigned the value 0.336 and the other 0.224. Thus:

$$\begin{aligned}
&\kappa = 0.336 + 0.224 = 0.56 \quad \text{and} \quad 1 - \kappa = 0.44,\\
&m_3 \oplus m_4\,(\{\text{Hep}\}) = (0.144 + 0.096)\,/0.44 = 0.545,\\
&m_3 \oplus m_4\,(\{\text{Cirr}\}) = 0.084/0.44 = 0.191.\\
&m_3 \oplus m_4\,(\{\text{Hep, Cirr}\}) = 0.036/0.44 = 0.082,\\
&m_3 \oplus m_4\,(\{\text{Cirr, Gall, Pan}\}) = 0.056/0.44 = 0.127,\\
&m_3 \oplus m_4\,(\Theta) = 0.024/0.44 = 0.055,\\
&m_3 \oplus m_4 \text{ is 0 for all other subsets of } \Theta.
\end{aligned}$$

[3] This convention is intuitive in that it maintains the relative beliefs among the rest of the hypotheses in 2^{Θ}. It should be noted, however, that the normalization convention is not supported in any theoretic sense and can lead to paradoxical behavior of the model in certain settings [19]. Some have argued that it would be just as rational to move the belief originally assigned to $\emptyset$ to Θ.

Note that $\sum m_3 \oplus m_4(X) = 1$, as is required by the definition of a *bpa.*

2.5 Belief Intervals

After combining all *bpa*'s with the same frame of discernment and then computing the belief function *Bel* defined by this new *bpa*, how should the information given by *Bel* be used? *Bel(A)* gives the total amount of belief committed to the subset A after all evidence bearing on A has been pooled. However, the function *Bel* contains additional information about A, namely $Bel(A^c)$, the extent to which the evidence supports the negation of A. The quantity $1 - Bel(A^c)$ expresses the plausibility of A, i.e., the maximum extent to which the current evidence could allow one to believe A (note that this is *not* the same as *Bel(A)*, the extent to which the current evidence specifically supports A).

The information contained in *Bel* concerning a given subset A may be conveniently expressed by the interval:

$$[Bel(A), 1 - Bel\,(A^c)].$$

It is not difficult to see that the left endpoint is always less than or equal to the right: $Bel(A) \leq 1 - Bel(A^c)$, or equivalently, $Bel(A) + Bel(A^c) \leq 1$. Since $Bel(A)$ and $Bel(A^c)$ are the sum of all values of m for subsets of A and A^c, respectively, and since A and A^c have no subsets in common, $Bel(A) + Bel(A^c) \leq \Sigma m(X) = 1$ where X varies over all subsets of Θ.

In the Bayesian situation, in which $Bel(A){+}Bel(A^c){=}1$, the two endpoints of the belief interval are equal and the width of the interval, $1 - Bel(A^c) - Bel(A)$, is 0. In the D-S model, however, the width is usually not 0 and is a measure of the belief which, although not committed to A, is also not committed to A^c. It may be seen that the width is the sum of belief committed exactly to subsets of Θ which intersect A but which are not subsets of A. If A is a singleton, all such subsets are supersets of A, but this is not true for a nonsingleton A. To illustrate, let $A = \{\text{Hep}\}$ and refer to Fig. 1:

$$\begin{aligned}1 - Bel\,(A^c) - Bel\,(A) &= 1 - Bel\,(\{\text{Cirr, Gall, Pan}\}) - Bel\,(\{\text{Hep}\})\\ &= 1 - [m\,(\{\text{Cirr, Gall, Pan}\}) + m\,(\{\text{Cirr, Gall}\})\\ &\quad + m\,(\{\text{Cirr, Pan}\}) + m\,(\{\text{Gall, Pan}\}) + m\,(\{\text{Cirr}\})\\ &\quad + m\,(\{\text{Gall}\}) + m\,(\{Pan\})] - m\,(\{\text{Hep}\})\\ &= m\,(\{\text{Hep, Cirr}\}) + m\,(\{\text{Hep, Gall}\})\\ &\quad + m\,(\{\text{Hep, Pan}\}) + m\,(\{\text{Hep, Cirr, Gall}\})\\ &\quad + m\,(\{\text{Hep, Cirr, Pan}\}) + m\,(\{\text{Hep, Gall, Pan}\})\\ &\quad + m\,(\Theta)\,.\end{aligned}$$

Belief committed to a superset of {Hep} might, upon further refinement of evidence, result in belief committed to {Hep}. Thus, the width of the belief interval is a measure of that portion of the total belief, 1, which could be added to that committed to {Hep} by a physician willing to ignore all but the disconfirming effects of the evidence.

The width of a belief interval can also be regarded as the amount of uncertainty with respect to a hypothesis given the evidence. It is belief which is committed to neither the hypothesis nor the negation of the hypothesis by the evidence. The vacuous belief function results in width 1 for all belief intervals and Bayesian functions result in width 0. Most evidence leads to belief functions with intervals of varying widths where the widths are numbers between 0 and 1.

3 The Dempster-Shafer Theory Applied to Singleton Hypotheses

Despite the intuitive appeal of many aspects of the D-S theory outlined above, the enumeration of all subsets of Θ in the application of the Dempster combining rule becomes computationally intractable when there are a large number of elements in Θ (as is true for many real-world problems in which the evidence-gathering scheme could otherwise be employed). If we restrict the hypotheses of interest in 2^{Θ} to the mutually exclusive singletons and their negations, however, Barnett has shown that a linear-time algorithm will permit rigorous application of the Dempster rule [2]. In this section we show that one expert system, MYCIN, can be viewed as a reasoning program in which the principal hypotheses are restricted to singletons. MYCIN will therefore be discussed to illustrate the applicability of the D-S theory in general and the relevance of the Barnett formulation in particular.

MYCIN's representation may be simply recast in terms of the D-S theory we have outlined. A frame of discernment in MYCIN, for example, is a clinical parameter (attribute) which may take on a range of values. The possible values are mutually exclusive and may therefore be seen as the competing hypotheses that make up the elements in Θ.[4] This condition may be a stumbling block to the model's implementation in systems where mutual exclusivity does not generally hold.

The belief functions which represent evidence in MYCIN correspond to the individual rules in the system's knowledge base. These are of a particularly simple form (the CF in a rule corresponds to the value assigned by a *bpa* to the hypothesis in the rule's conclusion based on the evidence in its premise). These features will now be discussed and illustrated with examples.

[4] Some parameters in MYCIN can take on multiple values, e.g., the patient's drug allergies [3], but we will be focussing here on the central inferences in the system, such as an organism's identity, which satisfy the mutual exclusivity requirement.

3.1 Frames of Discernment

How should the frames of discernment for a reasoning system be chosen? Shafer points out [14] that:

> It should not be thought that the possibilities that comprise Θ will be determined and meaningful independently of our knowledge. Quite to the contrary: Θ will acquire its meaning from what we know or think we know; the distinctions that it embodies will be embedded within the matrix of our language and its associated conceptual structures and will depend on those structures for whatever accuracy and meaningfulness they possess.

The 'conceptual structures' in MYCIN, for example, are the associative triples found in the conclusions of the rules [3]. These have the form (object, attribute, value), i.e., each triple corresponds to a singleton hypothesis of the form 'the attribute of object is value'. As mentioned previously, a frame of discernment would then consist of all triples with the same object and attribute.

For example, one frame of discernment is generated by the set of all triples of the form (Organism-1, Identity, X), where X ranges over all possible identities of organisms known to MYCIN—Klebsiella, E.coli, Pseudomonas, etc. Another frame is generated by replacing 'Organism-1' with 'Organism-2'. A third frame is the set of all triples of the form (Organism-1, Morphology, X), where X ranges over all known morphologies—coccus, rod, pleomorph, etc.

Although it is true that a patient may be infected by more than one organism, these organisms are represented as separate objects in MYCIN (not as separate values of the same parameter for a single object). Thus MYCIN'S representation scheme for the parameter that corresponds to its major classification task (i.e., the identity of an organism) complies with the mutual-exclusivity demand for frames of discernment in the D-S theory. Many other expert systems meet this demand less easily. Consider, for example, how the theory might be applicable in a system which gathers and pools evidence concerning a patient's diagnosis. Then there is often the problem of multiple, coexistent diseases, i.e., the hypotheses in the frame of discernment may not be mutually exclusive. One way to overcome this difficulty is to choose Θ to be the set of all subsets of all possible diseases. The computational implications of this choice are harrowing since if there are 600 possible diseases (the approximate scope of the INTERNIST-1 knowledge base [12]), then $|\Theta| = 2^{600}$ and $|2^{\Theta}| = 2^{2^{600}}$! However, since the evidence may actually focus on a small subset of 2^{Θ}, the computations need not be intractable because the D-S theory need not depend on explicit enumeration of all subsets of 2^{Θ} when many have a belief value of zero. An alternative would be to apply the D-S theory after partitioning the set of diseases into groups of mutually exclusive diseases and considering each group as a separate frame of discernment. The latter approach would be similar to that used in INTERNIST-1 [12], where scoring and comparison of hypotheses is undertaken only after a partitioning algorithm has separated evoked hypotheses into subsets of mutually exclusive diagnoses.

3.2 Rules as Basic Probability Assignments

In the most general situation, a given piece of evidence supports many of the subsets of Θ, each to varying degrees. However, the simplest situation is that in which the evidence supports or disconfirms only one singleton subset to a certain degree and the remaining belief is assigned to Θ. Because of the modular way in which knowledge is captured and encoded in MYCIN, this latter situation applies in the case of its rules.

If the premises confirm the conclusion of a rule with degree s, then the rule's effect on belief in the subsets of Θ can be represented by *bpa*. This *bpa* would assign s to the singleton corresponding to the hypothesis in the conclusion of the rule, call it A, and $1 - s$ to Θ. In the language of MYCIN, the CF associated with this conclusion is s. Since there is no concept equivalent to Θ in MYCIN, however, the remaining belief, $1 - s$, is left unassigned. If the premise of a rule disconfirms the conclusion with degree s, then the corresponding *bpa* would assign s to the subset corresponding to the negation of the conclusion, A^{c}, and $1 - s$ to Θ. The CF associated with his conclusion is $-s$. Thus, we are suggesting that the CF's associated with rules in MYCIN, and other EMYCIN systems, can be viewed as *bpa*'s in the D-S sense. Note, however, that MYCIN's rules do not permit inferences regarding nonsingleton hypotheses in 2^{Θ}, e.g., the conclusion that an organism is either an E.coli or a Klebsiella, which corresponds to the two-element subset {E.coli, Klebsiella}. Our suggested solution to this problem is outlined in Sect. 4.

3.3 Dempster's Rule Applied to Singleton Hypotheses

If we continue the analogy between CF's in MYCIN's rules and *bpa*'s in the D-S theory, we can consider the use of Dempster's rule for combining belief when two or more rules succeed and assign belief to the same or competing singleton hypotheses. To illustrate, we consider a frame of discernment Θ consisting of all associative triples of the form (Organism-1, Identity, X) where X ranges over all possible identities of organisms known to MYCIN. The triggering of two rules that affect belief in such triples can be categorized in one of three ways:

(1) they may both confirm or both disconfirm the same hypothesis;
(2) one may confirm and the other may disconfirm the same hypothesis;
(3) each may bring evidence to bear on different competing hypotheses.

We describe the approach to each of these possibilities below.

Theorem 1. *Two rules are both confirming or both disconfirming of the same triple, or conclusion. For example, both rules confirm Pseudomonas (Pseu),* *one to degree 0.4 and the other to degree 0.7. The effect of triggering the rules is represented by bpa's,* m_1 *and* m_2*, where* $m_1((\text{Pseu}\}) = 0.4,\ m_1(\Theta) = 0.6$*, and* $m_2(\{\text{Pseu}\}) = 0.7,\ m_2(\Theta) = 0.3$*. The combined effect on belief is given by* $m_1 \oplus m_2$*, computed using the tableau below:*

m_2 m_1	$\{\text{Pseu}\}\ (0.7)$	$\Theta(0.3)$
$\{\text{Pseu}\}\ (0.4)$	$\{\text{Pseu}\}\ (0.28)$	$\{\text{Pseu}\}\ (0.12)$
$\Theta\ (0.6)$	$\{\text{Pseu}\}\ (0.42)$	$\Theta\ (0.18)$

Note that $k = 0$ in this example, so normalization is not required (i.e., $1-k = 1$).

$$m_1 \oplus m_2\ (\{\text{Pseu}\}) = 0.28 + 0.12 + 0.42 = 0.82,$$
$$m_1 \oplus m_2\ (\Theta) = 0.18.$$

Note that $m_1 \oplus m_2$ is a *bpa* which, like m_1 and m_2, assigns some belief to a certain subset of Θ, $\{\text{Pseu}\}$, and the remaining belief to Θ. For two confirming rules, the subset is a singleton; for disconfirming rules, the subset is a set of size $n-1$, where n is the size of Θ.[5]

Theorem 2. *One rule is confirming and the other disconfirming of the same singleton hypothesis. For example, one rule confirms {Pseu} to degree 0.4 and the other disconfirms {Pseu} to degree 0.8. The effect of triggering these two rules is represented by bpa's* m_1, m_3 *where* m_1 *is defined in the previous example and* $m_3(\{\text{Pseu}\}^c) = 0.8,\ m_3(\Theta) = 0.2$*. The combined effect on belief is given by* $m_1 \oplus m_3$*.*

m_2 m_1	$\{\text{Pseu}\}^c\ (0.8)$	$\Theta(0.2)$
$\{\text{Pseu}\}\ (0.4)$	$\emptyset\ (0.32)$	$\{\text{Pseu}\}\ (0.08)$
$\Theta\ (0.6)$	$\{\text{Pseu}\}^c\ (0.48)$	$\Theta\ (0.12)$

This time the tableau does contain the empty set as an entry; therefore $k = 0.32$ and $1-k = 0.68$.

$$m_1 \oplus m_3\ (\{\text{Pseu}\}) = 0.08/0.68 = 0.118,$$
$$m_1 \oplus m_3\ (\{\text{Pseu}\}^c) = 0.48/0.68 = 0.706,$$
$$m_1 \oplus m_3\ (\Theta) = 0.12/0.68 = 0.176,$$
$$m_1 \oplus m_3 \text{ is 0 for all other subsets of } \Theta.$$

[5] Note that in this case Dempster's rule has provided the same result as would the original CF combining function (MYCIN would also combine 0.4 and 0.7 to get 0.82; see [15]).

Given m_1 above, the belief interval of {Pseu} is initially $[Bel_1(\{\text{Pseu}\}), 1 - Bel_1(\{\text{Pseu}\}^c)] = [0.4,\ 1]$. After combination with m_3, it becomes [0.118, 0.294]. Similarly, given m_3 alone, the belief interval of {Pseu} is [0, 0.2]. After combination with m_1, it becomes [0.118, 0.294].

As is illustrated in this example, an essential aspect of Dempster's rule is the effect of evidence that supports a hypothesis in 2^{Θ} in reducing belief in other hypotheses in 2^{Θ} that are disjoint from the supported hypothesis. Thus, evidence confirming $\{\text{Pseu}\}^c$ will reduce the effect of evidence confirming {Pseu}; in this case the degree of support for {Pseu}, 0.4, is reduced to 0.118. Conversely, evidence confirming {Pseu} will reduce the effect of evidence confirming $\{\text{Pseu}\}^c$; 0.8 is reduced to 0.706. These two effects are reflected in the modification of the belief interval of {Pseu} from [0.4, 1] to [0.118, 0.294], where $0.294 = 1 - Bel(\{\text{Pseu}\}^c) = 1 - 0.706$.

Consider the application of the CF combining function ($\text{CF}_{\text{combine}}$) to this same situation.[6] If CF_p is the positive (confirming) CF for {Pseu}, and CF_n is the negative (disconfirming) CF:

$$\begin{aligned}\text{CF}_{\text{COMBINE}}(\text{CF}_\text{p}, \text{CF}_\text{n}) &= (\text{CF}_\text{p} + \text{CF}_\text{n}) \,/\, (1 - \min\{|\text{CF}_\text{p}|\,, |\text{CF}_\text{n}|\}) \\ &= (s_1 - s_3) \,/\, (1 - \min\{s_1, s_3\}) \\ &= (0.4 - 0.8) \,/\, (1 - 0.4) \\ &= -0.667.\end{aligned}$$

Adapting this certainty factor to the language of the D-S theory, the result of the CF combining function is belief in {Pseu} and $\{\text{Pseu}\}^c$ to the degree 0 and 0.667, respectively. The larger disconfirming evidence of 0.8 completely negates the smaller confirming evidence of 0.4. The confirming evidence reduces the effect of the disconfirming from 0.8 to 0.667.

If one examines $\text{CF}_{\text{combine}}$ applied to combinations of confirming and disconfirming evidence as shown here, it is clear that it results in a CF whose sign is that of the CF with the greater magnitude. Thus, support for A and A^c is combined into reduced support for one or the other. In contrast, the D-S function results in reduced support for both A and A^c, a behavior that may more realistically reflect the competing effects of conflicting pieces of evidence.

The difference in the two approaches is most evident in the case of aggregation of two pieces of evidence, one confirming A to degree s and the other disconfirming A to the same degree. The CF function yields CF = 0 whereas

[6] The CF combining function shown here has been used in MYCIN systems for several years but is slightly different from the formula described in the original CF model [15]. The revised empirically derived function prevents single pieces of positive or negative evidence from overwhelming the effect of several pieces of evidence in the opposite direction. The combining function remains unchanged from its original form, however, when applied to two pieces of evidence that are either both confirming or both disconfirming. See [3, Chap. 10] for a more detailed discussion of these points.

the D-S rule yields reduced but nonzero belief in each of A and A^c. We believe that the D-S rule's behavior in this case is preferable on the grounds that the notion of applying confirming and disconfirming evidence of the same weight should be different from that of having no evidence at all.

We now examine the effect on belief of combination of two pieces of evidence supporting mutually exclusive singleton hypotheses. The CF combining function results in no interaction between the beliefs in the two hypotheses and differs most significantly from the D-S rule in this case.

Theorem 3. *The rules involve different hypotheses in the same frame of discernment. For example, one rule confirms {Pseu} to degree 0.4 (see m_1 in the examples from Categories 1 and 2) and the other disconfirms {Strep} to degree 0.7. The application of the second rule corresponds to m_4, defined by* $m_4(\{\text{Strep}\}^c) = 0.7$, $m_4(\Theta) = 0.3$. *The combined effect on belief is given by* $m_1 \oplus m_4$.

m_1 \ m_4	$\{\text{Strep}\}^c\ (0.7)$	$\Theta(0.3)$
{Pseu} (0.4)	{Pseu} (0.28)	{Pseu} (0.12)
Θ (0.6)	$\{\text{Strep}\}^c$ (0.42)	Θ (0.18)

In this case $\kappa = 0$ since the empty set does not occur in the tableau.

$$m_1 \oplus m_4\,(\{\text{Pseu}\}) = 0.28 + 0.12 = 0.40,$$
$$m_1 \oplus m_4\,(\{\text{Strep}\}^c) = 0.42,$$
$$m_1 \oplus m_4\,(\Theta) = 0.18,$$
$$m_1 \oplus m_4 \text{ is 0 for all other subsets of } \Theta.$$
$$Bel_1 \oplus Bel_4\,(\{\text{Pseu}\}) = 0.40,$$
$$\begin{aligned} Bel_1 \oplus Bel_4\,(\{\text{Strep}\}^c) &= m_1 \oplus m_4\,(\{\text{Strep}\}^c) + m_1 \oplus m_4\,(\{\text{Pseu}\}) \\ &= 0.42 + 0.40 \\ &= 0.82, \end{aligned}$$
$$Bel_1 \oplus Bel_4\,(\{\text{Pseu}\}^c) = Bel_1 \oplus Bel_4\,(\{\text{Strep}\}) = 0.$$

Before combination, the belief intervals for {Pseu} and $\{\text{Strep}\}^c$ are [0.4, 1] and [0.7, 1], respectively. After combination, they are [0.4, 1] and [0.82, 1], respectively. Note that evidence confirming {Pseu} has also confirmed $\{\text{Strep}\}^c$, a superset of {Pseu}, but that evidence confirming $\{\text{Strep}\}^c$ has had no effect on belief in {Pseu}, a subset of $\{\text{Strep}\}^c$. This kind of interaction among competing hypotheses is ignored by the CF model.

3.4 Evidence-combination Scheme

Although the calculations in Categories 1–3 in the previous section were straightforward, their simplicity is misleading. As the number of elements in Θ increases, Barnett [2] has shown that direct application of the D-S theory, without attention to the order in which the *bpa*'s representing rules are combined, results in exponential increases in the time for computations. This is due to the need to enumerate all subsets or supersets of a given set. For settings in which it is possible to restrict the hypotheses of interest to singletons and their negations, Barnett has proposed a scheme for reducing the D-S computations to polynomial time by combining the functions in an order that simplifies the calculations. We outline this scheme as it could be adapted to reasoning system (such as MYCIN) in which evidence bears on mutually exclusive singleton hypotheses.

Step 1. For each triple (i.e., singleton hypothesis), combine all *bpa*'s representing rules confirming that value of the parameter. If s_1, $s_2, \ldots,$ s_k represent different degrees of support derived from the triggering of k rules confirming a given singleton, then the combined support is $1-(1-s_1)(1-s_2)\cdots(1-s_k)$. (Refer to the example in Theorem 1 above for an illustration of this kind of combination. The formula shown here may be easily derived and is identical to the combining function used in the original CF model). Similarly, for each singleton, combine all *bpa*'s representing rules disconfirming that singleton. The same combining function is used for this calculation, and the numerical beliefs can simply be associated with the negation of the singleton hypotheses; it is not necessary to enumerate explicitly the elements in the set of size $n - 1$ (where n is the size of Θ) that corresponds to the complement of the singleton hypothesis in question. Thus, all evidence confirming a singleton is pooled and represented by a *bpa* and all evidence disconfirming the singleton (confirming the hypothesis corresponding to the set complement of the singleton) is pooled and represented by another *bpa*. We thus have $2n$ *bpa*'s, half of which assign belief to a singleton hypothesis and Θ (and which assign zero to all other hypotheses), the other half of which assign belief to the negation of a singleton hypothesis and Θ. Except for the notion of Θ, this step is identical to the original CF model's approach for gathering positive and negative evidence into the total confirming and disconfirming evidence respectively (MB and MD; see [15]).

Step 2. For each triple (singleton hypothesis), combine the two *bpa*'s computed in Step 1. Such a computation is a Theorem-2 combination and has been illustrated. Formulae that permit this calculation without the enumeration of any but the singleton subsets in 2^{Θ} are derived in [2] and described with examples in [9]. This step results in the definition of n *bpa*'s, one for each of the n singleton hypotheses. Each *bpa* that results assigns belief to a singleton hypothesis, its complement, and Θ while assigning zero to all other hypotheses.

Step 3. The final task is to blend all n *bpa*'s from Step 2 into a single belief function. This can be accomplished by combining the *bpa*'s derived in Step 2 in one computation, using formulae developed by Barnett to obtain the final belief function *Bel* [2]. Since these formulae allow computation of both the net belief in a singleton A and in its negation A^c, the belief interval $[Bel(A),\ 1 - Bel(A^c)]$ for each singleton hypothesis can then be computed.

The details of Barnett's approach are described in [2]. In another publication, we have also provided the form of the required computation and have shown an example based on a small MYCIN rule set [9]. Since the new method proposed in the next section borrows only on Step 1 of the Barnett approach, we will not show the details of Steps 2 and 3 here.

4 The Dempster–Shafer Theory Applied to a Hierarchical Hypothesis Space

In a system in which all evidence either confirms or disconfirms singleton hypotheses, the combination of evidence via the D-S scheme with Barnett's formulae can be computationally simple as outlined in the previous section. As we have shown, a program such as MYCIN could be easily recast to use the D-S approach rather than the CF model.[7]

What attracted us to the D-S theory, however, and left us dissatisfied with the approach to singleton hypotheses proposed by Barnett, is the theory's potential for handling evidence bearing on categories of diseases as well as on specific disease entities. We are unaware of another model that suggests how evidence concerning hierarchically-related hypotheses might be combined coherently and consistently to allow inexact reasoning at whatever level of abstraction is appropriate for the evidence that has been gathered. The pure D-S model provides such a method for handling the aggregation of evidence gathered at varying levels of detail or specificity. Much of our frustration with the original MYCIN representation scheme and the CF model resulted from their inability to handle such hierarchical relationships cleanly. In recent years, a recurring theme in AI has been the explicit representation of hierarchic relationships among hypotheses (e.g., [8, 13]). Thus the D-S scheme might be especially suitable for handling uncertainty in such hierarchically organized networks. The problem, as we have emphasized, is the theory's computational complexity due to the potential need to enumerate all subsets in 2^{Θ}. Thus we have sought a technique that allows the model's use in a hierarchical hypothesis space while avoiding the exponential-time requirements that the theory

[7] Additional conventions similar to those adopted in the CF model would be needed before the D-S approach could be used, however. For example, it would be necessary to adopt some mechanism for propagation of uncertainty in a rule-chaining environment. Barnett's suggestion [2] that MYCIN is ill-suited to such as implementation (due to its failure to satisfy the mutual exclusivity requirement) reflects a misunderstanding of the program's representation and control mechanisms. Multiple diseases are handled by instantiating each as a separate context (object); within a given context, the requirements of single-valued parameters (attributes assumed to take on precisely one value) maintain mutual exclusivity [3].

otherwise would entail. Since Barnett's approach is applicable only when the space is limited to singleton hypotheses and their negations, it will not serve our purposes.

To illustrate the need for such a capability, consider the way in which hierarchic relationships in the MYCIN domain were handled in that program. An example would be evidence suggesting that an organism was one of the Enterobacteriaceae (a family of gram-negative rods). The triple (hypothesis) for this conclusion was handled as (Organism Class Enterobacteriaceae), i.e., the frame of discernment (the Class parameter) was different from that normally used for concluding the identity of an organism (the Ident parameter). There was no way for the system to reach conclusions about both singleton hypotheses (e.g., Ident = E.coli) and supersets (e.g., Ident = Enterobacteriaceae) within the single Ident frame of discernment. Thus the Class parameter was introduced to handle the latter case. The relationship between Class Enterobacteriaceae and the individual organisms that make up that class was handled using rules in which evidence for Enterobacteriaceae was effectively transferred to Ident. This was accomplished by assigning as the values of the Ident parameter each of the bacteria on the list of gram negative organisms in that Class. The CF's assigned to the individual organism identities in this way were based more on guesswork than on solid data. The evidence really supported the higher-level concept, Enterobacteriaceae, and further breakdown may have been unrealistic. In actual practice, decisions about treatment are often made on the basis of high level categories rather than specific organism identities (e.g., "I'm quite sure that this is one of the enterics (i.e., the Enterobacteriaceae), and would therefore treat with an aminoglycoside and a cephalosporin (i.e., two types of antibiotic), but I have no idea which of the enteric organisms is causing the disease.").

Problems such as this would be better handled if experts could specify rules which refer to semantic concepts at whatever level in the domain hierarchy is most natural and appropriate. They should ideally not be limited to the most specific level—the singleton hypotheses in the frame of discernment—but should be free to use more unifying concepts. Because of the complexity in the D-S theory's approach to handling evidence, then, the challenge is to make these computations tractable, either by a modification of the theory or by restricting the evidence domain in a reasonable way. By taking the latter approach, we have developed an algorithm for the implementation of the theory which merges a strict application of the D-S combining function with a simplifying approximation.

4.1 Simplifying the Evidence Domain to a Tree Structure

The key assumption underlying our proposed approach is that the experts who participate in the construction of large knowledge bases can define a strict hierarchy of hypotheses about which the reasoning system will gather evidence. In D-S terms, we are suggesting that, for a given domain, only some of the subsets in 2^{Θ} will be of semantic interest and that these can be selected to form a strict hierarchy. In medical diagnosis, for example, evidence often bears on certain disease categories as well as on specific disease entities. In the simplified case of cholestatic jaundice discussed earlier, for which $\Theta =$ {Hep, Cirr, Gall, Pan}, evidence available to the physician tends to support either intrahepatic cholestasis = {Hep, Cirr}, extrahepatic cholestasis = {Gall, Pan}, or the singleton hypotheses {Hep}, {Cirr}, {Gall}, and {Pan}. The other nodes of 2^{Θ} shown in Fig. 1 are not particulary meaningful notions in this context. The network of subsets in Fig. 1 could thus be pruned to that of Fig. 2, which summarizes the hierarchical relations of clinical interest. The hierarchy of Fig. 2 is a tree in the strict sense—each node below Θ has a unique parent. In the medical expert system known as MDX, the causes of jaundice have been usefully structured in precisely this way [4]. We believe, as do others [13], that such a structuring is characteristic of medical diagnostic tasks (as well as of many other problem-solving situations).

4.2 Evidence Combination Scheme for a Strict Hierarchy

We now propose a new three-step scheme for the implementation of the D-S theory in the situation in which the hypotheses of interest have been restricted by domain experts to subsets which form a strict hierarchy. It should be noted that, in general, the negations of hypotheses in the hierarchy (i.e., their set complements) will not be in the tree. For example, $\{\text{Hep}\}^{c} =$ {Cirr, Gall, Pan} does not occur in the hierarchy of Fig. 2. Thus, as did Barnett in his Step 1, we propose an approach in which disconfirming evidence is handled computationally by associating it directly with the disconfirmed hypothesis rather than by converting it to be manipulated as confirming evidence regarding the complement of the disconfirmed hypothesis. The first two steps in our approach are a strict application of the D-S theory, in which simple formulae can be derived due to the tree structure of the hypotheses of interest. In the first step all confirmatory evidence is combined for each node in the tree, and the same is done for all disconfirmatory evidence. This step is similar to the first step in Barnett's approach (Sect. 3.4) except that the hypotheses are not restricted to singletons. In the second step all confirmatory evidence is combined for the entire tree. The third step is an approximation for combining

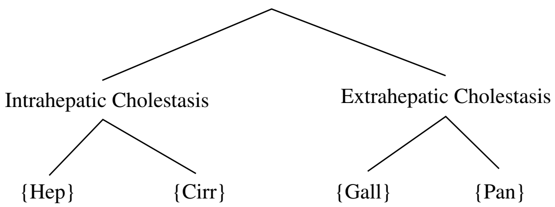


Fig. 2. The subsets of clinical interest in cholestatic jaundice

disconfirmatory evidence. Strict application of the D-S theory in this step may result in an exponential-time computation, whereas our approximation is computationally more efficient.

To illustrate these formulae, we use a slightly expanded version of the cholestatic-jaundice tree depicted in Fig. 2. Suppose we add to Θ a fifth cause of cholestatic jaundice, impaired liver function due to effects of oral contraceptives, denoted Orcon = {Orcon}. This addition will permit us to better demonstrate the properties of the technique we are proposing. Note that now Θ = cholestatic jaundice = {Hep, Cirr, Orcon, Gall, Pan} whereas intrahepatic cholestasis becomes the three-element subset {Hep, Cirr, Orcon} and has three direct descendents {Hep}, {Cirr}, and {Orcon}. This new tree is shown in Fig. 3 with only the first letter of each singleton hypothesis used, and commas and set brackets omitted for convenience of notation.

For the general case, we shall let T denote the set of all subsets (except for Θ itself) in the hierarchy of hypotheses that has been defined by the domain expert. Note that T is itself a subset of 2^{Θ}. However, it is convenient to think of T as simply the hypothesis tree without Θ. In our example, T is the set consisting of intrahepatic cholestasis, extrahepatic cholestasis, and the five single disease entities—i.e., {HCO, GP, H, C, O, G, P}. Let T' denote the set of all complements of subsets in T. T' is also a subset of 2^{Θ}, but the entities in T' will generally not be in T and hence are of interest only because they correspond to negations of pertinent hypotheses. In this example, T' is the set {HCO^{c}, GP^{c}, H^{c}, C^{c}, O^{c}, G^{c}, P^{c}}.

Step 1. Using the combining functions described in Step 1 of Barnett's evidence-combination scheme detailed in Sect. 3.4, for each subset X_i in T, combine all confirmatory evidence to obtain a *bpa*, m_{X_i}, and all disconfirmatory evidence to obtain another *bpa*, $m_{X_i^c}$.[8] Note that m_{X_i} can have a nonzero value on only X_i and Θ, $m_{X_i^c}$ on only X_i^c and Θ. Using our example, we would thus compute the following *bpa*'s: m_{HCO}, m_{GP}, m_{H}, m_{C}, m_{O}, m_{G}, m_{P},

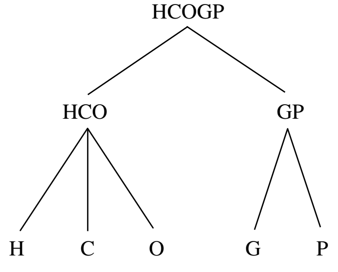


Fig. 3. The expanded tree of cholestatic jaundice

[8] Note that we have introduced a variation on the notation used up to this point: m_i has denoted the *bpa* associated with the ith piece of evidence, whereas m_{X_i} denotes the *bpa* associated with the set X_i after all evidence confirming X_i has been combined.

$m_{\mathrm{HCO}^{\mathrm{c}}}$, $m_{\mathrm{GP}^{\mathrm{c}}}$, $m_{\mathrm{H}^{\mathrm{c}}}$, $m_{\mathrm{C}^{\mathrm{c}}}$, $m_{\mathrm{O}^{\mathrm{c}}}$, $m_{\mathrm{G}^{\mathrm{c}}}$, $m_{\mathrm{P}^{\mathrm{c}}}$. Thus, $m_{\mathrm{HCO}}(\mathrm{HCO})$ is the belief in intrahepatic cholestasis (i.e., HCO) after all evidence confirmatory of this disease category has been combined. The remaining belief, $1-m_{\mathrm{HCO}}(\mathrm{HCO})$, is assigned to Θ. Similarly, $m_{\mathrm{HCO}^{\mathrm{c}}}(\mathrm{HCO}^{\mathrm{c}})$ is the total belief against intrahepatic cholestasis and $1 - m_{\mathrm{HCO}^{\mathrm{c}}}(\mathrm{HCO}^{\mathrm{c}})$ is assigned to Θ.

Our goal is to compute the single aggregate *bpa* that assigns net belief to all elements of T (by definition the only hypotheses of semantic interest for the domain) by blending in the disconfirming evidence associated with the sets in T'. This corresponds to the *bpa*

$$m_{Y_1} \oplus m_{Y_2} \oplus \cdots$$

where Y_i takes on the value of all subsets occurring in either T or T'. However, a strict application of the D-S theory in determining this *bpa* will assign nonzero values to many subsets that are in neither T nor T', precisely the event that we wish to avoid in order to prevent the enumeration of all sets in 2^{Θ}. The technique we propose combines in an organized fashion the *bpa*'s just computed in Step 1. Through a simple assumption defined below (see Step 3), we avoid the generation of new subsets.

We continue by observing that our aggregate final *bpa* can also be written as

$$m_T \oplus mT'$$

where

$$m_T = m_{x_1} \oplus m_{X_2} \oplus \cdots, X_i \in T$$

and

$$m_{T'} = m_{x_1^{\mathrm{c}}} \oplus m_{X_2^{\mathrm{c}}} \oplus \cdots, X_i^{\mathrm{c}} \in T'.$$

The *bpa*, m_T, has nonzero values on only Θ or subsets in T, i.e., on $T \cup \Theta$, since the intersection of any two subsets in T is either the empty set or in T (the smallest of the two subsets). This computation is therefore performed as Step 2.

Step 2. Combine all confirmatory evidence by computing the aggregate *bpa*, m_T, of the *bpa*'s in Step 1 of the form m_{x_i}, where

$$m_T = m_{x_1} \oplus m_{x_2} \oplus \cdots, X_i \in T.$$

Note that m_T has nonzero value only on $T \cup \Theta$. In our example,

$$m_T = m_{\mathrm{HCO}} \oplus m_{\mathrm{GP}} \oplus m_{\mathrm{H}} \oplus m_{\mathrm{C}} \oplus m_{\mathrm{O}} \oplus m_{\mathrm{G}} \oplus m_{\mathrm{P}}.$$

The quantity, $m_T(\mathrm{HCO})$, is the belief in HCO (intrahepatic cholestasis) after combining all evidence confirmatory of this disease category with all evidence confirmatory of every other disease category or entity in the tree.

Note that the calculation in Step 2 does not include evidence *disconfirmatory* of HCO or the other hypotheses in T. That task is left to Step 3, i.e., the remaining problem is to compute $m_T \oplus M_{T'}$. However, as mentioned earlier, if $m_{T'}$ is computed by a strict application of the D-S combining rule, it has nonzero value on many subsets that are in neither T nor T'. Even the aggregation of evidence disconfirmatory of a single subset in T (i.e., confirmatory of a single subset in T') with m_T leads to the generation of new subsets. For example, the combination of m_T with evidence disconfirmatory of hepatitis leads to a *bpa*, $m_T \oplus m_{\mathrm{H}^c}$, which assigns belief to the diagnosis of CO, i.e., the set {Cirr, Orcon}.[9] This set is not in the tree of Fig. 3 because it was not originally defined to be of diagnostic interest. If this *bpa* is then combined with that representing evidence disconfirmatory of cirrhosis, belief is assigned to the diagnosis of HO = {Hep, Orcon}. This set also is not in T. As more *bpa*'s are aggregated via the D-S combination rule, more subsets are generated which are not in T and thus not of diagnostic interest. Hence, we make the approximation described in Step 3.

Step 3. Combine disconfirmatory evidence by step-wise combination of the $m_{X^c{}_i}$'s in the following way. Choose any set X_1^c in T' and compute $m_T \ominus m_{X_1^c}$, which is an approximation to $m_T \oplus m_{X_1^c}$ with the property that $m_T \ominus m_{X_1^c}$ has nonzero value on only $T \cup \Theta$. Belief assigned to a subset A by $\oplus$ is instead assigned by $\ominus$ to the smallest superset of A in T if A itself is not in T. Now choose another set, X_2^c, in T', and compute $(m_T \ominus m_{X_1^c}) \ominus m_{X_2^c}$. Continue until all sets in T' have been chosen. The result is an aggregate *bpa* in which belief assigned to a set A in 2^Θ by the D-S function is sometimes assigned instead to an ancestor of A in $T \cup \Theta$. It may be shown (see Appendix A) that such an assignment is unique. Belief is thus displaced upward in the tree in order to avoid consideration of subsets not in T. Note that belief in A implies belief in B if B is a superset of A. The function, $\ominus$, is order-independent except in an easily identifiable case (see Appendix A).

To illustrate, belief assigned in the previous example to CO, a set not in the tree, is instead assigned to HCO, the smallest set in the tree containing it. Belief assigned to HO is also assigned to HCO. Note that disbelief in a singleton, which is represented as belief in its complement, is assigned by the approximation as belief in Θ (unless the complement happens to be in T).

As we have noted, the final *bpa* obtained by step-wise application of the function $\ominus$ in Step 3 differs from that obtained by the D-S function in that some belief assigned to a given subset by the latter is assigned to an ancestor of that subset by the former. Since belief in a subset of hypotheses implies belief in a superset of that subset, the upward displacement of belief in the hierarchy seems to be a reasonable exchange for the computational simplicity of our approximation method.

A final point is important to stress regarding the approach in Step 3. It should be clear that the scheme assigns *all* belief to subsets in T or to Θ. Thus, for A in *T, Bel(A)* can be computed by summing net belief in A with belief

[9] Note that $m_T \oplus m_{\mathrm{H}^c}$ assigns the quantity of belief, $m_T(\mathrm{HCO})m_{\mathrm{H}^c}(\mathrm{H}^c)$ to CO = HCO $\cap$ H^c.

assigned to all its descendents. However, it will not in general be possible to compute $Bel(A^{\mathrm{c}})$ since A^{c} will usually be in T' but not in T. Thus the notion of a belief interval, $[Bel(A),\ 1 - Bel(A^{\mathrm{c}})]$ is lost in the scheme we have proposed. Competing hypotheses would need to be compared based upon *Bel* alone without regard to the width of the plausibility interval (see Sect. 2.5).

In summary, the proposed evidence aggregation scheme is as follows.

Step 1. Calculate m_{X_i} for all X_i in T and $m_{X_i^{\mathrm{c}}}$ for all X_i^{c} in T'.
Step 2. Calculate $m_T = m_{X_1} \oplus m_{X_2} \oplus \cdots$ for all X_i in T.
Step 3. Calculate $m_T \ominus m_{X_1^{\mathrm{c}}}$, then $(m_T \ominus m_{X_1^{\mathrm{c}}}) \ominus m_{X_2^{\mathrm{c}}}$, etc. for all X_i^{c} in T'.

Recall that Step 1 is accomplished using the technique described in Sect. 3.4 and does not require the assumption of the tree structure of the domain or an approximation technique. Steps 2 and 3 do depend upon the assumption of the tree structure, however, and Step 3 requires the approximation outlined above. The formulae for the calculations in Steps 2 and 3 are given below, with their derivations provided in an Appendix A.

Step 2.

$$m_T\left(A\right) = \begin{cases} K m_A\left(A\right) \prod\limits_{\substack{X \in T \\ X \not\supseteq A}} m_X\left(\Theta\right) & \text{if } A \in T, \\ K \prod\limits_{X \in T} m_X\left(\Theta\right) & \text{if A} = \Theta, \end{cases}$$

where $K = 1/(1 - \kappa)$ and

$$1 - \kappa = \sum_{A \in T} \left[m_A\left(A\right) \prod_{\substack{X \in T \\ X \not\supseteq A}} m_X(\Theta) \right].$$

Step 3. There are different formulae in Step 3 depending upon which of three relationships hold between X and A : $X \subseteq A, X \cap A = \emptyset$, or $X \supset A$, where X is a subset of $T \cup \Theta$ and A is a subset of T. In all cases, $K = 1/(1 - \kappa)$ where

$$\kappa = m_{A^{\mathrm{c}}}\left(A^{\mathrm{c}}\right) \sum_{\substack{X \in T \\ X \subseteq A}} m_T\left(X\right).$$

Case 1. $X \subseteq A$:

$$m_T \ominus m_{A^{\mathrm{c}}}\left(X\right) = K m_T\left(X\right) m_{A^{\mathrm{c}}}\left(\Theta\right).$$

Case 2. $X \cap A = \emptyset$ (i.e., $X \cap A^{\mathrm{c}} = X$):

(i) If $X \cup A$ is a set in $T \cup \Theta$:

$$m_T \ominus m_{A^c}(X) = K[m_T(X) + m_T(X \cup A)\, m_{A^c}(A^c)].$$

(ii) If $X \cup A$ is not in $T \cup \Theta$:

$$m_T \ominus m_{A^c}(X) = K m_T(X)\,.$$

Case 3. $X \supset A$:

(i) If $X \cap A^c$ is not a set in T:

$$m_T \ominus m_{A^c}(X) = K m_T(X)\,.$$

(ii) If $X \cap A^c$ is in T:

$$m_T \ominus m_{A^c}(X) = K m_T(X)\, m_{A^c}(\Theta)\,.$$

5 Conclusion

A major drawback for practical implementation of the Dempster–Shafer theory of evidence in reasoning systems has been its computational complexity (and resulting inefficiency). Based on the observation that evidence used in diagnostic reasoning involves abstract categories that can often be naturally represented in a strict hierarchical structure, we have designed a method for evidence aggregation based on the D-S theory. Using combinatorial analysis, a strict application of the theory, and an approximation, we have presented an approach which is computationally tractable.

Some observers may question the *value* of using the D-S scheme rather than the CF model or some other *ad hoc* method for handling uncertainty when dealing only with singleton hypotheses. Systems like MYCIN and INTERNIST-1 have demonstrated expert-level performance using their current techniques for inexact reasoning [12, 18]. We have previously suggested, in fact, that the details of a model of evidential reasoning in an AI system may be relatively unimportant since the careful semantic structuring of a domain's knowledge seems to blunt the sensitivity of its inferences to the values of the numbers used.[10] Some have even suggested that evidential reasoning can be handled without the use of a numerical model at all [5]. As was emphasized in Sect. 4, however, it is the D-S theory's techniques for managing reasoning about hypotheses in hierarchic abstraction spaces that we have found particularly appealing. The failure of previous models to deal coherently with these

[10] See [3, Chap. 10] for a discussion of this point and an analysis of the sensitivity of MYCIN's conclusions to the CF values used in its rules. As is discussed there, MYCIN's performance can be shown to extremely insensitive to rather wide variations in the CF's assigned to its rules.

issues has led to unnatural knowledge representation schemes that require evidential associations among related concepts to be stated explicity rather than provided automatically by the hierarchic structure of pertinent domain concepts.

Directions for further work lie in the implementation and evaluation of our method in an actual reasoning system. Additional conventions will need to be defined before this can be done. For example, it is common for the evidence itself to be of an uncertain nature, and partially supported hypotheses in one frame of discernment may themselves be used as evidence to assign belief to hypotheses in another frame of discernment. This is a key feature of rule-chaining systems, for example, where belief in the premise conditions of rules may be less than certain. The *ad hoc* methods being used currently (e.g., the CF model's multiplicate convention [15]) may simply be borrowed for a D-S implementation. More interesting, perhaps, is the issue of how best to *use* the belief in the hypotheses after the proposed scheme has been applied. There is not likely to be a 'correct' approach to this problem because the nature of the actions based on evidence varies so greatly from one domain to another. Heuristics may be devised, however, for using thresholding or relative belief measures to determine what level of abstraction in the hypothesis hierarchy is most appropriately selected as the basis for a final conclusion or recommendation from an advice system.

The techniques described here will be neither necessary nor adequate for all expert system application domains. Some tasks are well managed by purely categorical inference techniques, and others do not lend themselves to hierarchical domain structuring and the evidence gathering model of problem solving. However, for diagnostic or classification tasks in settings where the hypothesis space is well suited to assumptions of mutual exclusivity and hierarchical organization, we believe that our adaptation of the Dempster–Shafer theory holds great appeal as a computationally tractable and coherent belief model.

Appendix A

We present here the details of Steps 2 and 3 in the proposed evidence combination scheme outlined in Sect. 4.

A Step 2: Aggregation of Confirmatory Evidence

The *bpa* m_T is the aggregate of all *bpa*'s of the form, m_X, where X is a subset in T. Each m_X has been obtained by combining all confirmatory evidence for X. For the following discussion we shall use A to refer to an arbitrary subset in $T \cup \Theta$. We now derive formulae for m_T by first computing the normalization constant, $K = 1/(1-\kappa)$, and then $m_T(A)$ for any subset A in $T \cup \Theta$.

A.1 The Normalization Constant of m_T

Recall that $1-\kappa$ is the sum of all beliefs not attributed to the empty set. Thus, $1-\kappa$ is

$$\sum \prod_{X\in T} m_X(Y_X),$$

where Y_X is either X (a subset in T) or Θ and the Y_X's intersect to give a non-empty subset. For example, in the cholestatic jaundice hierarchy of Fig. 3, two of the summands in $1-\kappa$ would be:

$$m_{\mathrm{H}}(\mathrm{H}) m_{\mathrm{C}}(\Theta)\, m_{\mathrm{O}}(\Theta)\, m_{\mathrm{G}}(\Theta)\, m_{\mathrm{p}}(\Theta)\, m_{\mathrm{GP}}(\Theta)\, m_{\mathrm{HCO}}(\mathrm{HCO}),$$
$$m_{\mathrm{H}}(\mathrm{H}) m_{\mathrm{C}}(\Theta)\, m_{\mathrm{O}}(\Theta)\, m_{\mathrm{G}}(\Theta)\, m_{\mathrm{p}}(\Theta)\, m_{\mathrm{GP}}(\Theta)\, m_{\mathrm{HCO}}(\Theta).$$

Note that once we choose $Y_A = A$ for a specific A, then, in order to avoid the empty set as the final intersection, we must choose all other $Y_X = \Theta$ except for descendents (subsets) or ancestors (supersets) of A. In the above example, once we chose $Y_{\mathrm{H}} = \mathrm{H}$, we had to choose $Y_X = \Theta$ for all other X except for $X = \mathrm{HCO}$, the one ancestor of H in T. For Y_{HCO}, we could choose Y_{HCO} as either HCO or Θ. Thus, we claim that

$$1-\kappa = \sum_{A\in T} \left[m_A(A) \prod_{\substack{X\in T\\ X\not\supseteq A}} m_X(\Theta) \prod_{\substack{X\in T\\ X\supset A}} \left[m_X(X) + m_X(\Theta)\right] \right].$$

Since m_X has nonzero value on only X and Θ, $m_X(X) + m_X(\Theta) = 1$ for all X in T. Thus,

$$\prod_{\substack{X\in T\\ X\supset A}} \left[m_X(X) + m_X(\Theta)\right] = 1$$

and the above simplifies to

$$1-\kappa = \sum_{A\in T} \left[m_A(A) \prod_{\substack{X\in T\\ X\not\supseteq A}} m_X(\Theta) \right].$$

The two products given above for the cholestatic jaundice example would be represented in this expression by the summand in the expression for $1-\kappa$ formed by choosing $A = \mathrm{H}$. Because $m_{\mathrm{HCO}}(\mathrm{HCO}) + m_{\mathrm{HCO}}(\Theta) = 1$, note that these two summands add to $m_{\mathrm{H}}(\mathrm{H}) m_{\mathrm{C}}(\Theta) m_{\mathrm{O}}(\Theta) m_{\mathrm{G}}(\Theta) m_{\mathrm{P}}(\Theta) m_{\mathrm{GP}}(\Theta)$.

A.2 Computation of $m_T(A)$

In order to derive a formula for $m_T(A)$, where A is any subset in $T \cup \Theta$, we need to enumerate all products of the form

$$\prod_{X \in T} m_X\left(Y_X\right),$$

where Y_X is either X (a subset in T) or Θ and the intersection of the Y_X's is A. For $A = \Theta$, we must choose $Y_X = \Theta$ for each X in T. Thus,

$$m_T\left(\Theta\right) = K \prod_{X \in T} m_X\left(\Theta\right).$$

For every A in T, we must choose $Y_A = A$ for the factor $m_A(Y_A)$ since $Y_A = \Theta$ will in general make it impossible to achieve a final intersection of A due to the tree structure of the subsets. If X is not an ancestor of A, then we must choose $Y_X = \Theta$ since $Y_X = X$ will yield an empty intersection for some subset of A. If X is an ancestor of A, then both $Y_X = X$ and $Y_X = \Theta$ will yield A as the intersection. Thus, we obtain

$$m_T\left(A\right) = K m_A\left(A\right) \prod_{\substack{X \in T \\ X \not\supseteq A}} m_x\left(\Theta\right) \prod_{\substack{X \in T \\ X \supset A}} \left[m_X\left(X\right) + m_X\left(\Theta\right)\right].$$

Once again, the indicated sum, and hence the last product, is 1 and the above simplifies to

$$m_T\left(A\right) = K m_A\left(A\right) \prod_{\substack{X \in T \\ X \not\supseteq A}} m_X\left(\Theta\right).$$

For example, in our model of cholestatic jaundice from Fig. 3, the effect of all confirmatory evidence on belief precisely in hepatitis is given by

$$m_T(\mathrm{H}) = K m_{\mathrm{H}}(\mathrm{H}) m_{\mathrm{C}}(\Theta) m_{\mathrm{O}}(\Theta) m_{\mathrm{G}}(\Theta) m_{\mathrm{P}}(\Theta) m_{\mathrm{GP}}(\Theta).$$

The effect on belief in intrahepatic cholestasis (i.e., HCO) is given by

$$\begin{aligned} m_T\left(\mathrm{HCO}\right) = & K m_{\mathrm{HCO}}\left(\mathrm{HCO}\right) m_{\mathrm{H}}\left(\Theta\right) m_{\mathrm{C}}\left(\Theta\right) m_{\mathrm{O}}\left(\Theta\right) m_{\mathrm{G}}\left(\Theta\right) m_{\mathrm{P}}\left(\Theta\right) \\ & m_{\mathrm{GP}}\left(\Theta\right). \end{aligned}$$

B Step 3: Aggregation of Disconfirmatory Evidence

As mentioned in Sect. 4, it is in this step that we first depart from a strict application of the D-S combining function in order to avoid the assignment of belief to subsets which are neither in T nor T'. Our solution to this difficulty is an approximation, $m_T \ominus m_{A^c}$, which assigns all belief to subsets in $T \cup \Theta$; i.e., the subsets on which m_T may have nonzero value. For example, in the hierarchy of Fig. 3, belief that would be assigned to CO is instead assigned to its smallest ancestor in T, HCO. This is a justifiable assignment because:

- the subset CO is, by the domain expert's definition of T, not of diagnostic interest and so should not be assigned belief;
- evidence confirming a subset also logically supports supersets of that subset;
- there is a unique smallest superset due to the strict tree structure of the hierarchy defined by the subsets in T, i.e., each subset in T has precisely one parent in T, except for those at the top of the hierarchy whose parent is Θ.

Thus, $m_T \ominus m_{A^c}$ assigns $m_T(X)m_{A^c}(A^c)$ to $X \cap A^c$ if $X \cap A^c$ lies in $T \cup \Theta$ and to X (which can be shown to be the unique smallest superset in $T \cup \Theta$ containing $X \cap A^c$) if not. We now derive formulae for $m_T \ominus m_{A^c}$.

B.1 Computation of Normalization Constant for the Modified Combining Function

This time we consider κ, the sum of beliefs assigned to $\emptyset$, instead of $1 - \kappa$ as we did in Step 2. Thus, we want a simplified expression for

$$\kappa = \sum_{X \in T} m_T\left(X\right) m_{A^c}\left(Y_{A^c}\right),$$

where $Y_{A^c} = A^c$ or Θ and $X \cap Y_{A^c} = \emptyset$. Clearly, X and Y_{A^c} are not disjoint if $Y_{A^c} = \Theta$. If $Y_{A^c} = A^c$, then we must choose $X = A$ or X a subset of A to yield $X \cap Y_{A^c} = \emptyset$. Thus,

$$\kappa = m_{A^c}\left(A^c\right) \sum_{\substack{X \in T \\ X \subseteq A}} m_T\left(X\right).$$

B.2 Formulae for the Modified Combining Function

We derive formulae for $m_T \ominus m_{A^c}(X)$ where X lies in $T \cup \Theta$ and therefore falls into one of three cases. We are looking in each case for all sets in $T \cup \Theta$ which intersect with either A^c or Θ to give X. For purposes of illustration, consider the hypothesis tree of Fig. 3 and the calculations necessary for combining evidence disconfirmatory of pancreatic cancer ($A = \mathrm{P}$).

Case 1. $X \subseteq A$. There is no subset in $T \cup \Theta$ that will intersect with A^c to give X so the only possibility is to choose X and Θ to yield $X \cap \Theta = X$. Thus,

$$m_T \ominus m_{A^c}\left(X\right) = K m_T\left(X\right) m_{A^c}\left(\Theta\right).$$

In our example with $A = \mathrm{P}$, the only set X in this case is $X = A = \mathrm{P}$. Thus,

$$m_T \ominus m_{\mathrm{P}^c}\left(\mathrm{P}\right) = K m_T\left(\mathrm{P}\right) m_{\mathrm{P}^c}\left(\Theta\right).$$

Case 2. $X \cap A = \emptyset$ (i.e., $X \cap A^{\mathrm{c}} = X$). Note that we may choose either X, A^{c} or X, Θ as pairs yielding an intersection equal to X. Two subcases should be distinguished: that in which $X \cup A$ is in $T \cup \Theta$ and that in which $X \cup A$ is not. For if $X \cup A$ is in $T \cup \Theta$, then we may also choose the pair $X \cup A,\ A^{\mathrm{c}}$ to yield X as the intersection. Thus, in the first subcase:

$$\begin{aligned} m_T \ominus m_{A^{\mathrm{c}}}(X) =& K[m_T(X)\, m_{A^{\mathrm{c}}}(A^{\mathrm{c}}) + m_T(X)\, m_{A^{\mathrm{c}}}(\Theta) \\ &+ m_T(X \cup A)\, m_{A^{\mathrm{c}}}(A^{\mathrm{c}})]. \end{aligned}$$

This expression simplifies to

$$m_T \ominus m_{A^{\mathrm{c}}}(X) = K\left[m_T(X) + m_T(X \cup A)\, m_{A^{\mathrm{c}}}(A^{\mathrm{c}})\right],$$

since $m_{A^{\mathrm{c}}}(A^{\mathrm{c}}) + m_{A^{\mathrm{c}}}(\Theta) = 1$.

In our example with $A = \mathrm{P}$, the set G falls into this subcase and

$$m_T \ominus m_{\mathrm{p^c}}(\mathrm{G}) = K[m_T(\mathrm{G}) + m_T(\mathrm{GP})\, m_{\mathrm{p^c}}(\mathrm{P^c})].$$

The second subcase applies for all X in T such that $X,\ A^{\mathrm{c}}$ and $X,\ \Theta$ are the only two pairs yielding an intersection equal to X:

$$m_T \ominus m_{A^{\mathrm{c}}}(X) = K[K m_T(X)\, m_{A^{\mathrm{c}}}(A^{\mathrm{c}}) + m_T(X)\, m_{A^{\mathrm{c}}}(\Theta)],$$

which simplifies to

$$m_T \ominus m_{A^{\mathrm{c}}}(X) = K m_T(X),$$

since $m_{A^{\mathrm{c}}}(A^{\mathrm{c}}) + m_{A^{\mathrm{c}}}(\Theta) = 1$.

In our example with $A = \mathrm{P}$, the subsets HCO, H, C, and O fall in this subcase and thus

$$\begin{aligned} m_T \ominus m_{\mathrm{p^c}}(\mathrm{HCO}) &= K m_T(\mathrm{HCO}), \\ m_T \ominus m_{\mathrm{p^c}}(\mathrm{H}) &= K m_T(\mathrm{H}), \\ m_T \ominus m_{\mathrm{p^c}}(\mathrm{C}) &= K m_T(\mathrm{C}), \\ m_T \ominus m_{\mathrm{p^c}}(\mathrm{O}) &= K m_T(\mathrm{O}). \end{aligned}$$

Case 3. $X \supset A$: In this case, the only pair yielding an intersection of X is $X,\ \Theta$. However, consider the pair $X,\ A^{\mathrm{c}}$ whose intersection may or may not lie in T. If $X \cap A^{\mathrm{c}}$ does not lie in T, it may be shown that X is the smallest superset of $X \cap A^{\mathrm{c}}$ containing $X \cap A^{\mathrm{c}}$ and we assign $m_T(X) m_{A^{\mathrm{c}}}(A^{\mathrm{c}})$ to X. Then,

$$m_T \ominus m_{A^{\mathrm{c}}}(X) = K[m_T(X) m_{A^{\mathrm{c}}}(A^{\mathrm{c}}) + m_T(X) m_{A^{\mathrm{c}}}(\Theta)] = K m_T(X).$$

In our example, $m_T \ominus m_{\mathrm{P^c}}(\Theta) = K m_{\mathrm{T}}(\Theta)$, since $m_T(\Theta) m_{\mathrm{P^c}}(\mathrm{P^c})$ is assigned to Θ, the smallest superset of $\mathrm{P^c}$ in $T \cup \Theta$.

If $X \cap A^{\mathrm{c}}$ does lie in T, then $m_T(X)m_{A^{\mathrm{c}}}(A^{\mathrm{c}})$ was assigned to $X \cap A^{\mathrm{c}}$ in Case 2. Clearly, if $X \cap A^{\mathrm{c}}$ is a subset in T, $X \cap A^{\mathrm{c}}$ falls into Case 2 since $(X \cap A^{\mathrm{c}}) \cap A = \emptyset$. Thus, for $X \supset A$ and $X \cap A^{\mathrm{c}} \in T$:

$$m_T \ominus m_{A^{\mathrm{c}}}\,(X) = Km_T\,(X)\,m_{A^{\mathrm{c}}}(\Theta).$$

In our example, $m_T \ominus m_{\mathrm{P}^{\mathrm{c}}}(\mathrm{GP}) = Km_T(\mathrm{GP})m_{\mathrm{P}^{\mathrm{c}}}(\Theta)$.

B.3 Optimal Ordering of Evidence Aggregation

It can be shown that the function $\ominus$ is order-independent except in the case of evidence involving a subset A where both A and its parent have exactly one sibling. In the hierarchy shown in Fig. 3, for example, the configuration of concern occurs when A is taken to be either G or P. In this situation, evidence involving the higher level subset GP should be combined before that involving G or P. A small portion of the belief that would be assigned to G or P by the D-S function is correctly assigned to G or P if disconfirming evidence $m_{X_i^{\mathrm{c}}}$ is aggregated first with the higher level subset and then with G and P. However, it is assigned to GP, the parent of G and P, if the disconfirming evidence is aggregated with the lower level subsets first.

Thus, a better approximation to the D-S function is obtained depending on the order for aggregation chosen in Step 3. However, this difference is insignificant in that the amount of belief involved is small and more importantly, it is only displaced upward by one level from a subset to its parent. Such upward displacement is a common result of the approximation function anyway. Combining evidence in a breadth-first fashion, from higher to lower levels, will result in an optimal approximation.

Acknowledgment

Some of the introductory material in this paper is based on a discussion of the Dempster–Shafer theory in [9]. We are indebted to several colleagues who participated in discussions of this work and provided comments on early drafts of the paper: Byron W. Brown, Bruce Buchanan, Benjamin Grosof, David Heckerman, Michael Higgins, Curtis Langlotz, Mark Tuttle, and Lotfi Zadeh. Our thanks also to Joan Differding, Barbara Elspas, and Janice Rohn who assisted with manuscript preparation.

References

1. Adams, J. B., A probability model of medical reasoning and the MYCIN model, *Math. Biosci.* **32** (1976) 177–186.
2. Barnett, J. A., Computational methods for a mathematical theory of evidence, in: *Proceedings Seventh International Joint Conference on Artificial Intelligence*, Vancouver, BC (1981) 868–875.

3. Buchanan, B. G. and Shortliffe, E. H., *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project* (Addison-Wesley, Reading, MA, 1984).
4. Chandrasekaran, B., Gomez, F., Mittal, S. and Smith, M., An approach to medical diagnosis based on conceptual schemes, in: *Proceedings Sixth International Joint Conference on Artificial Intelligence*, Tokyo, Japan (1979) 134–142.
5. Cohen, P. R., *Heuristic Reasoning about Uncertainty: An Artificial Intelligence Approach* (Pitman, London, 1984).
6. Friedman, L., Extended plausible inference, in: *Proceedings Seventh International Joint Conference on Artificial Intelligence*, Vancouver, BC (1981) 487–495.
7. Garvey, T. D., Lowrence, J. D. and Fischler, M. A., An inference technique for integrating knowledge from disparate sources, in: *Proceedings Seventh International Joint Conference on Artificial Intelligence*, Vancouver, BC (1979) 319–325.
8. Gomez, F. and Chandrasekaran, B., Knowledge organization and distribution for medical diagnosis, in: W. J. Clancey and E. H. Shortliffe (Eds.), *Readings in Medical Artificial Intelligence: The First Decade* (Addison-Wesley, Reading, MA, 1984) 320–338.
9. Gordon, J. and Shortliffe, E. H., The Dempster–Shafer theory of evidence, in: B. G. Buchanan and E. H. Shortliffe (Eds.), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project* (Addison-Wesley, Reading, MA, 1984) 272–292.
10. Gouvernet, J., Ayme, S., Sanchez, E., Mattei, J. F. and Giraud, F., Diagnosis assistance in medical genetics based on belief functions and a tree structured thesaurus: a conversational mode realization, in: *Proceedings of MEDINFO 80*, Tokyo, Japan (1980) 798.
11. Gouvernet, J., Apport des methods de classification en genetique medicale, Thesis, University of Marseilles, 1979.
12. Miller, R. A., Pople, H. E. and Myers, J. D., INTERNIST-1: An experimental computer-based diagnostic consultant for general internal medicine, *New England J. Medicine* **307**(8) (1982) 468–476.
13. Pople, H. E., Heuristic methods for imposing structure on ill-structed problems: the structuring of medical diagnostics, in: P. Szolovits (Ed.), *Artificial Intelligence in Medicine* (Westview Press, Boulder, CO, 1982) 119–190.
14. Shafer, G., *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, NJ, 1976).
15. Shortliffe, E. H. and Buchanan, B. G., A model of inexact reasoning in medicine, *Math. Biosci.* **23** (1975) 351–379.
16. Strat, T. M., Continuous belief functions for evidential reasoning, in: *Proceedings Fourth National Conference on Artificial Intelligence*, Austin, TX (1984) 308–313.
17. Szolovits, P. and Pauker, S., Categorical and probabilistic reasoning in medical diagnosis, *Artificial Intelligence* **11** (1978) 115–44.
18. Yu, V. L., Fagan, L. M., Wraith, S. M., Clancey, W. J., Scott, A. C., Hannigan, J. F., Blum, R. L., Buchanan, B. G. and Cohen, S. N., Antimicrobial selection by a computer: a blinded evaluation by infectious disease experts, *J. Amer. Med. Assoc.* **242** (1979) 1279–1282.
19. Zadeh, L. A., A mathematical theory of evidence (book review), *AI Magazine* **5**(3) (1984) 81–83.

13

Languages and Designs for Probability Judgment*

Glenn Shafer and Amos Tversky

Abstract. Theories of subjective probability are viewed as formal languages for analyzing evidence and expressing degrees of belief. This article focuses on two probability language, the Bayesian language and the language of belief functions [19]. We describe and compare the semantics (i.e., the meaning of the scale) and the syntax (i.e., the formal calculus) of these languages. We also investigate some of the designs for probability judgment afforded by the two languages.

Introduction

The weighing of evidence may be viewed as a mental experiment in which the human mind is used to assess probability much as a pan balance is used to measure weight. As in the measurement of physical quantities, the design of the experiment affects the quality of the result.

Often one design for a mental experiment is superior to another because the questions it asks can be answered with greater confidence and precision. Suppose we want to estimate, on the basis of evidence readily at hand, the number of eggs produced daily in the U.S. One design might ask us to guess the number of chickens in the U.S. and the average number of eggs laid by each chicken each day. Another design might ask us to guess the number of people in the U.S., the average number of eggs eaten by each person, and some inflation factor to cover waste and export. For most of us, the second design is manifestly superior, for we can make a reasonable effort to answer the questions it asks.

* From *Cognitive Science* **9**, 309–339 (1985). Glenn Shafer has corrected some typographical errors and updated a reference. The original research was supported in part by NSF grants MCS-800213 and 8301282 to the Glenn Shafer and ONR Grant NR197-058 to Amos Tversky. The article benefited from the comments of Jonathan Baron, Morris DeGroot, Persi Diaconis and David Krantz. Amos Tversky died on June 2, 1996.

As this example illustrates, the confidence and precision with which we can answer a question posed in a mental experiment depends on how our knowledge is organized and stored, first in our mind and secondarily in other sources of information available to us.

The quality of the design of a mental experiment also depends on how effectively the answers to the individual questions it asks can be combined to yield an accurate overall picture or accurate answers to questions of central interest. An analogy with surveying may be helpful. There are usually many different ways of making a land survey—many different angles and lengths we may measure. When we design the survey we consider not only the accuracy and precision with which these individual measurements can be made but also how they can be combined to give an accurate plot of the area surveyed [13]. Singer shows how a mental experiment may be designed to give a convincing estimate of the total value of property stolen by heroin addicts in New York City [24]. Other examples of effective designs for mental experiments are given by Raiffa [15].

One way to evaluate competing designs for physical measurement is to apply them to instances where the truth is known. But such empirical evaluation of final results is not always possible in the case of a mental experiment, especially when the experiment is designed to produce only probability judgments. It is true that probability judgments can be interpreted as frequencies. But as we argue below, this interpretation amounts only to a comparison with a repeatable physical experiment where frequencies are known. How the comparison is made—what kind of repetitions are envisaged—is itself one of the choices we make in designing a mental experiment. There may not be a single set of repetitions to which the design must be referred for empirical validation.

Since empirical validation of a design for probability judgment is problematic, the result of carrying out the mental experiment must be scrutinized in other ways. The result of the whole experiment must be regarded as an argument, which, like all other arguments, is open to criticism and counterarguments.

Understanding and evaluating a design for probability judgment is also complicated by problems of meaning. When we are simply guessing the answer to a question of fact, such as the number of eggs produced daily in the U.S., the meaning of the question seems to be independent of our design. But when we undertake to make probability judgments, we find that we need a theory of subjective probability to give meaning to these judgments.

In the first place, we need a numerical scale or at least a qualitative scale (practically certain, very probable, fairly probable, etc.) from which to choose degrees of probability. We also need canonical examples for each degree of probability on this scale—examples where it is agreed what degree of probability is appropriate. Finally, we need a calculus—a set of rules for combining simple judgments to obtain complex ones.

Using a theory of subjective probability means comparing the evidence in a problem with the theory's scale of canonical examples and picking out the canonical example that matches it best. Our design helps us make this comparison. It specifies how to break the problem into smaller problems that can be more easily compared with the scale of canonical examples and how to combine the judgments resulting from these separate comparisons.

Thought of in this way, a theory of subjective probability is very much like a formal language. It has a vocabulary—a scale of degrees of probability. Attached to this vocabulary is a semantics—a scale of canonical examples that show how the vocabulary is to be interpreted and psychological devices for making the interpretation effective. Elements of the vocabulary are combined according to a syntax—the theory's calculus.

Proponents of different theories of subjective probability have often debated which theory best describes human inductive competence. We believe that none of these theories provide an adequate account of people's intuitive judgments of probability. On the other hand, most of these theories can be learned and used effectively. Consequently, we regard these theories as formal languages for expressing probability judgments rather than as psychological models, however idealized.

The usefulness of one of these formal languages for a specific problem may depend both on the problem and on the skill of the user. There may not be a single probability language that is normative for all people and all problems. A person may find one language better for one problem and another language better for another. Furthermore, individual probability judgments made in one language may not be directly translatable into another.

This article studies the semantics and syntax of two probability languages, the traditional Bayesian language and the language of belief functions, and it uses these languages to analyze several concrete examples. This exercise can be regarded as a first step toward the general study of design for probability judgment. It illustrates the variety of designs that may be feasible for a given problem, and it yields a classification of Bayesian designs that clarifies the role of Bayesian conditioning. Our treatment is incomplete, however, because it does not provide formal criteria or lay out general empirical procedures for evaluating designs. The choice of design is left to the ingenuity of the user.

1 Examples

With the help of some simple examples we illustrate several designs for probability judgments. We will return to these examples in Sects. 3 and 4.

1.1 The Free-Style Race

We are watching one of the last men's swim meets of the season at Holsum University. We have followed the Holsum team for several seasons, so we watch with intense interest as Curt Langley, one of Holsum's leading free-stylers, gets off to a fast start in the 1650-yard race. As Curt completes his first

1000 yards, he is swimming at a much faster pace than we have seen him swim before. His time for the first 1000 yards is 9 min and 25 s. His best previous times for 1650 yards have been around 16 min and 25 s, a time that translates into about 9 min and 57 s at 1000 yards. The only swimmer within striking distance of him is a member of the visiting team named Cowan, whom we know only by name. Cowan is about half a lap (about 12 yards or 7 s) behind Curt.

Will Curt Win the Race? The first question we ask ourselves is whether he can keep up his pace. Curt is known to us as a very steady swimmer—one who knows what he is capable of and seldom, if ever, begins at a pace much faster than he can keep up through a race. It is true that his pace is much faster than we have seen before—much faster than he was swimming only a few weeks ago. It is possible that there has been no real improvement in his capacity to swim—that he has simply started fast and will slow down before the race is over. But our knowledge of Curt's character and situation encourages us to think that he must have trained hard and greatly increased his endurance. This is his senior year, and the championships are near. And he must have been provoked to go all out by Jones, the freshman on the team, who has lately overshadowed him in the long-distance races. We are inclined to think that Curt will keep up his pace.

If Curt does keep up his pace, then it seems very unlikely that Cowan could have enough energy in reserve to catch him. But what if Curt cannot keep up his pace? Here our vision becomes more murky. Has Curt deliberately put his best energy into the first part of the race? Or has he actually misjudged what pace he can keep up? In the first case, it seems likely he will soon slow down, but not to a disastrously slow pace; it seems to be a toss-up whether Cowan will catch him. On the other hand, if he has misjudged what pace he can keep up, then surely he has not misjudged it by far, and so we would expect him to keep it up almost to the end and, as usually happens in such cases, "collapse" with exhaustion to a very slow pace. There is no telling what would happen then—whether Cowan would be close enough or see the collapse soon enough to take advantage of the situation.

There are many different designs that we might use to assess numerically the probability of Curt's winning. There is even more than one possible Bayesian design. The Bayesian design suggested by our qualitative discussion assesses the probabilities that Curt will keep up the pace, slow down, or collapse and the conditional probabilities that he will win under each of these hypotheses and then combines these probabilities and conditional probabilities to obtain his overall probability of winning. We call this a total-evidence design because each probability and conditional probability is based on the total evidence. In sect. 3 we will formalize and carry out this total-evidence design. We will also carry out a somewhat different Bayesian total-evidence design for the problem. In sect. 4 we will carry out a belief-function design for the problem.

1.2 The Hominids of East Turkana

In the August, 1978, issue of *Scientific American*, Alan Walker and Richard E. T. Leakey [27] discuss the hominid fossils that have recently been discovered in the region east of Lake Turkana in Kenya. These fossils, between a million and two million years of age, show considerable variety, and Walker and Leakey are interested in deciding how many distinct species they represent.

In Walker and Leakey's judgment, the relatively complete cranium specimens discovered in the upper member of the Koobi Fora Formation in East Turkana are of three forms: (I) A "robust" form with large cheek teeth and massive jaws. These fossils show wide-fanning cheekbones, very large molar and premolar teeth, and smaller incisors and canines. The brain case has an average capacity of about 500 cubic centimeters, and there is often a bony crest running fore and aft across its top, which presumably provided greater area for the attachment of the cheek muscles. Fossils of this form have also been found in South Africa and East Asia, and it is generally agreed that they should all be classified as members of the species Australopithecus robustus. (II) A smaller and slenderer (more "gracile") form that lacks the wide-flaring cheekbones of I, but has similar cranial capacity and only slightly less massive molar and premolar teeth. (III) A large-brained (c. 850 cubic cm) and small-jawed form that can be confidently identified with the Homo erectus specimens found in Java and northern China.

The placement of the three forms in the geological strata in East Turkana shows that they were contemporaneous with each other. How many distinct species do they represent? Walker and Leakey admit five hypotheses:

1. I, II, and III are all forms of a single, extremely variable species.
2. There are two distinct species: one, *Australopithecus robustus*, has I as its male form and II as its female form; the other, *Homo erectus*, is represented by III.
3. There are two distinct species: one, *Australopithecus robustus*, is represented by I; the other has III, the so-called *Homo erectus* form, as its male form, and II as its female form.
4. There are two distinct species: one is represented by the gracile form II; the other, which is highly variable, consists of I and III.
5. The three forms represent three distinct species.

Here are the items of evidence, or arguments, that Walker and Leakey use in their qualitative assessment of the probabilities of these five hypotheses:

(i). Hypothesis 1 is supported by general theoretical arguments to the effect that distinct hominid species cannot coexist after one of them has acquired culture.

(ii). Hypotheses 1 and 4 are doubtful because they postulate extremely different adaptations within the same species: The brain seems to overwhelm the chewing apparatus in III, while the opposite is true in I.

(iii). There are difficulties in accepting the degree of sexual dimorphism postulated by hypotheses 2 and 3. Sexual dimorphism exists among living anthropoids, and there is evidence from elsewhere that hints that dental dimorphism of the magnitude postulated by hypothesis 2 might have existed in extinct hominids. The dimorphism postulated by hypothesis 3, which involves females having roughly half the cranial capacity of males, is less plausible.

(iv). Hypotheses 1 and 4 are also impugned by the fact that specimens of type I have not been found in Java and China, where specimens of type III are abundant.

(v). Hypotheses 1 and 3 are similarly impugned by the absence of specimens of type II in Java and China.

Before specimens of type III were found in the Koobi Fora Formation, Walker and Leakey thought it likely that the I and II specimens constituted a single species. Now on the basis of the total evidence, they consider hypothesis 5 the most probable.

What Bayesian design might we use to analyze this evidence? A total evidence design may be possible, but it is natural to consider instead a design in which some of the evidence is treated as an "observation" and used to "condition" probabilities based on the rest of the evidence. We might, for example, first construct a probability distribution that includes probabilities for whether specimens of Type I and II should occur in Java and China and then condition this distribution on their absence there. It is natural to call this a conditioning design. It is not a total-evidence design, because the initial (or "prior") probabilities for whether the specimens occur in Java and China will be based on only part of the evidence.

Later in Sect. 3, we will work this conditioning design out in detail. In Sect. 4 we will apply a belief-function design to the same problem.

2 Two Probability Languages

In order to make numerical probability judgments, we need a numerical scale. We need, in other words, a scale of canonical examples in which numerical degrees of belief are agreed upon. Where can we find such a scale?

The obvious place to look is in the picture of chance. In this picture, we imagine a game which can be played repeatedly and for which we know the chances. These chances, we imagine, are facts about the world: they are long-run frequencies, they can be thought of as propensities, and they also define fair betting rates—rates at which a bettor would break even in the long run.

There are several ways the picture of chance can be related to practical problems, and this means we can use the picture to construct different kinds of canonical examples and thus different theories or probability languages. In this essay, we shall consider two such languages: the Bayesian language, and the language of belief functions. The Bayesian language uses a scale of canonical examples in which the truth is generated by chance and our evidence consists of complete knowledge of the chances. The language of belief functions uses a scale of canonical examples in which our evidence consists of a message whose meaning depends on known chances.

We emphasize the Bayesian language because it is familiar to most readers. We study the language of belief functions as well in order to emphasize that our constructive view of probability, while not implying that all probability languages have equal normative claims, leaves open the possibility that no single language has a preemptively normative status.

2.1 The Bayesian Language

As we see it, a user of the Bayesian probability language makes probability judgments in a particular problem by comparing the problem to a scale of examples in which the truth is generated according to known chances and deciding which of these examples is most like the problem. The probability judgment $P(A) = p$, in this language, is a judgment that the evidence provides support for A comparable to what would be provided by knowledge that the truth is generated by a chance setup that produces a result in A exactly p of the time. This is not to say that one judges the evidence to be just like such knowledge in all respects, nor that the truth is, in fact, generated by chance. It is just that one is measuring the strength of the evidence by comparing it to a scale of chance setups.

The idea that Bayesian probability judgment involves comparisons with examples where the truth is generated by chance is hardly novel. It can be found, for example, in Bertrand [2] and in Box [3]. Box states that the adoption of given Bayesian probability distribution means that "current belief . . . would be calibrated with adequate approximation by a physical stimulation involving random sampling" (p. 385) from the distribution. The Bayesian literature has not, however, adequately addressed the question of how this comparison can be carried out. One reason for this neglect may be the emphasis that twentieth-century Bayesians have put on betting. When "personal probabilities" are defined in terms of a person's preferences among bets, we are tempted to think that the determination of probabilities is a matter of introspection rather than a matter of examining evidence, but see Diaconis and Zabell [4].

Bayesian Semantics. The task of Bayesian semantics is to render the comparison of our evidence to the Bayesian scale of canonical examples effective—to find ways of making the scale of chances and the affinity of our evidence to it vivid enough to our imagination that we can meaningfully locate the evidence on the scale.

By concentrating on different aspects of the rich imagery of games of chance, we can isolate different ways of making the Bayesian scale of chances vivid, and each of these ways can be thought of as a distinct semantics for the Bayesian probability language. Three such semantics come immediately to mind: a frequency semantics, a propensity semantics, and a betting semantics. The frequency semantics compares our evidence to the scale of chances by asking how often, in situations like the one at hand, the truth would turn out in various ways. The propensity semantics makes the comparison by first interpreting the evidence in terms of a causal model and then asking about the model's propensity to produce various results. The betting semantics makes the comparison by assessing our willingness to bet in light of the evidence: at what odds is our attitude towards a given bet most like our attitude towards a fair bet in a game of chance?

It is traditional, of course, to argue about whether probability should be given a frequency, a propensity, or a betting interpretation. But from our perspective these "interpretations" are merely devices to help us make what may ultimately be an imperfect fit of our evidence to a scale of chances. Which of these devices is most helpful will depend on the particular problem. We do not insist that there exists, prior to our deliberation, some particular frequency or numerical propensity in nature or some betting rate in our mind that should be called the probability of the proposition we are considering.

Which of these three Bayesian semantics tends to be most helpful in fitting our evidence to the scale of chances? We believe that the frequency and propensity semantics are central to the successful use of the Bayesian probability language, and that the betting semantics is less useful. Good Bayesian designs ask us to make probability judgments that can be translated into well-founded judgments about frequencies or about causal structures.

Since we readily think in terms of causal models, the propensity semantics often seems more attractive than the frequency semantics. But this attraction has its danger; the vividness of causal pictures can blind us to doubts about their validity. A simple design based on frequency semantics can sometimes be superior to a more complex design based on propensity semantics. We may, for example, obtain a better idea about how long it will take to complete a complex project by taking an "outside view" based on how long similar projects have taken in the past than by taking an "inside view" that attempts to assess the strength of the forces that could delay the completion of the project [10].

The betting semantics has a generality that the frequency and propensity semantics lack. We can always ask ourselves about our attitude towards a bet, quite irrespective of the structure of our evidence. But this lack of connection with the evidence is also a weakness of the betting semantics.

In evaluating the betting semantics, one must distinguish logical from psychological and practical considerations. Ramsey [16], Savage [18], and their followers have made an important contribution to the logical analysis of subjective probability by showing that it can be derived from coherent preferences between bets. This logical argument, however, does not imply psychological precedence. Introspection suggests that people typically act on the basis of their beliefs, rather than form beliefs on the basis of their acts. The gambler bets on Team A rather than on Team B because he believes that A is more likely to win. He does not usually infer such a belief from his betting preferences.

It is sometimes argued that the prospect of monetary loss tends to concentrate the mind and thus permits a more honest and acute assessment of the strength of evidence than that obtained by thinking about that evidence directly. There is very little empirical evidence to support this claim. Although incentives can sometimes reduce careless responses, monetary payoffs are neither necessary nor sufficient for careful judgment. In fact, there is evidence showing that people are sometimes willing to incur monetary losses in order to report what they believe [12]. Personally, we find that questions about betting do not help us think about the evidence; instead they divert our minds to extraneous questions: our attitudes towards the monetary and social consequences of winning or losing a bet, our assessment of the ability and knowledge of our opponent, etc.

Bayesian Syntax. It follows from our understanding of the canonical examples of the Bayesian language that this language's syntax is the traditional probability calculus. A proposition that a person knows to be false is assigned probability zero. A proposition that a person knows to be true is assigned probability one. And in general probabilities add: if A and B are incompatible propositions, then $P(A \text{ or } B) = P(A) + P(B)$.

The conditional probability of A given B is, by definition,

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}. \tag{1}$$

If $B_1, \ldots, B_2$ are incompatible propositions, one of which must be true, then the rule of total probability says that

$$P(A) = \sum_{j=1}^{n} P(B_j)P(A \mid B_j), \tag{2}$$

and Bayes's theorem says that

$$P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{\sum_{j=1}^{n} P(B_j)P(A \mid B_j)}. \tag{3}$$

As we shall see in Sect. 3, both total-evidence and conditioning designs can use the concept of conditional probability. Total-evidence designs often use (2), while conditioning designs use (I). Some conditioning designs can be described in terms of (3).

2.2 The Language of Belief Functions

The language of belief functions uses the calculus of mathematical probability, but in a different way than the Bayesian language does. Whereas the Bayesian language asks, in effect, that we think in terms of a chance model for the facts in which we are interested, the belief-function language asks that we think in terms of a chance model for, the reliability and meaning of our evidence.

This can be put more precisely by saying that the belief-function language compares evidence to canonical examples of the following sort. We know a chance experiment has been carried out. We know that the possible outcomes of the experiment are $o_1, \ldots, o_n$. and that the chance of o_i is p_i We are not told the actual outcome but we receive a message concerning another topic that can be fully interpreted only with knowledge of the actual outcome. For each i there is a proposition A_i, say, such that if we knew the actual outcome was o_i then we would see that the meaning of the message is A_i, We have no other evidence about the truth or falsehood of the A_i and so no reason to change the probabilities p_i.

What degrees of belief are called for in an example of this sort? How strongly should be believe a particular proposition of A?

For each proposition A, set $m(A) = \sum\{p_i \mid A_i = A\}$. This number is the total of the chances for outcomes that would show the message to mean A; we can think of it as the total chance that the message means A. Now let $Bel(A)$ denote the total chance that the message implies A; in symbols, $Bel(A) = \sum\{m(B) \mid B \text{ implies } A\}$. It is natural to call $Bel(A)$ our degree of belief in A.

We call a function Bel a belief function if it is given by the above equation for some choice of $m(A)$. By varying the p_i and the A_i in our story of the uncertain message, we can obtain any such values for the $m(A)$, and so the story provides canonical examples for every belief function.

We call the propositions A for which $m(A) > 0$ the *focal elements* of the belief function Bel. Often the most economical way of specifying a belief function is to specify its focal elements and their "m-values."

Semantics for Belief Functions. We have based our canonical examples for belief functions on a fairly vague story: We receive a message and we see, somehow, that if o_i were the true outcome of the random experiment, then the message would mean A_i. One task of semantics for belief functions is to flesh out the story in ways that help us compare real problems to it. Here we shall give three ways of fleshing out the story. The first leads to canonical examples for a small class of belief functions, called simple support functions. The second leads to canonical examples for a larger class, the consonant support functions. The third leads to canonical examples for arbitrary belief functions.

(i) A Sometimes Reliable Truth Machine. Imagine a machine that has two modes of operation. We know that in the first mode it broadcasts truths. But we are completely unable to predict what it will do when it is in the Second mode. We also know that the choice of which mode the machine will operate in on a particular occasion is made by chance: There is a chance s that it will operate in the first mode and a chance $1 - s$ that it will operate in the second mode.

It is natural to say of a message broadcast by such a machine on a particular occasion that it has a chance s of meaning what it says and a chance $1-s$ of meaning nothing at all. So if the machine broadcasts the message that E is true, then we are in the setting of our general story: The two modes of operation for the machine are the two outcomes o_1 and o_2 of a random experiment; their chances are $p_1 = s$ and $p_2 = 1 - s$; if o_1 happened then the message means $A_1 = E$, while if o_2 happened the message means nothing beyond what we already know, i.e., it means $A_2 = \varTheta$, where $\varTheta$ denotes the proposition that asserts the facts we already know. So we obtain a belief function with focal elements E and θ; $m(E) = s$ and $m(\varTheta) = 1 - s$.

We call such a belief function a simple support function. Notice its nonadditivity: the two complementary propositions E and not E have degrees of belief $Bel(E) = s < 1$ and $Bel(\text{not } E) = 0$.

It is natural to use simple support functions in cases where the message of the evidence is clear but where the reliability of this message is in question. The testimony of a witness, for example, may be unambiguous, and yet we may have some doubt about the witness's reliability. We can express this doubt by comparing the witness to a truth machine that is less than certain to operate correctly.

(ii) A Two-Stage Truth Machine. Consider a sometimes reliable truth machine that broadcasts two messages in succession and can slip into its untrustworthy mode before either message. It remains in the untrustworthy mode once it has slipped into it. As before, we know nothing about whether or how often it will be truthful when it is in this mode. We know the chances that it will slip into its untrustworthy mode: r_1 is the chance it will be in untrustworthy mode with the initial message, and r_2 is the chance it will slip into untrustworthy mode after the first message, given that it was in trustworthy mode then.

Suppose the messages received are E_1 and E_2. and suppose these messages are consistent with each other. Then there is a chance $(1-r_1)(1-r_2)$ that the message "E_1 and E_2" is reliable, a chance $(1-r_1)r_2$ that the message "E_1" alone is reliable, and a chance r_1 that neither of the messages is reliable. If we set

$$\begin{aligned} p_1 &= (1-r_1)(1-r_2), & A_1 &= E_1 \& E_2, \\ p_2 &= (1-r_1)r_2, & A_2 &= E_1, \\ p_3 &= r_1, & A_3 &= \varTheta, \end{aligned}$$

then we are in the setting of our general story: there is a chance p_i that the messages mean A_i.

Notice that A_1, A_2, and A_3 are "nested": A_1 implies A_2, and A_2 implies A_3. In general, we call a belief function with nested focal elements a *consonant support function.* It is natural to use consonant support functions in cases where our evidence consists of an argument with several steps; each step leads to a more specific conclusion but involves a new chance of error.

(iii) A Randomly Coded Message. Suppose someone chooses a code at random from a list of codes, uses the chosen code to encode a message, and then sends us the results. We know the list of codes and the chance of each code being chosen—say the list is $o_1, \ldots, o_n$, and the chance of o_i being chosen is p_i. We decode the message using each of the codes and we find that this always produces an intelligible message. Let A_i denote the message we get when we decode using o_i. Then we have the ingredients for a belief function: a message that has the chance p_i of meaning A_i.

Since the randomly coded message is more abstract than the sometimes reliable truth machine, it lends itself less readily to comparison with real evidence. But it provides a readily understandable canonical example for an arbitrary belief function. (For other scales of canonical examples for belief functions, see [11] and [28].)

Syntax for Belief Functions. Our task, when we assess evidence in the language of belief functions, is to compare that evidence to examples where the meaning of a message depends on chance and to single out from these examples the one that best matches it in weight and significance. How do we do this? In complicated problems we cannot simply look at our evidence holistically and write down the best values for the $m(A)$. The theory of belief functions provides, therefore, a set of rules for constructing complicated belief functions from more elementary judgments. These rules, which ultimately derive from the traditional probability calculus, constitute the syntax of the language of belief functions. They include rules for combination, conditioning, extension, conditional embedding, and discounting.

The most important of these rules is Dempster's rule of combination. This is a formal rule for combining a belief function constructed on the basis of one item of evidence with a belief function constructed on the basis of another, intuitively independent item of evidence so as to obtain a belief function representing the total evidence. It permits us to break down the task of judgment by decomposing the evidence.

Dempster's rule is obtained by thinking of the chances that affect the meaning or reliability of the messages provided by different sources of evidence as independent. Consider, for example, two independent witnesses who are compared to sometimes reliable truth machines with reliabilities s_1 and s_2 respectively. If the chances affecting their testimonies are independent, then there is a chance $s_1 s_2$ that both will give trustworthy testimony, and a chance $s_1 + s_2 - s_1 s_2$ that at least one will. If both testify to the truth of A, then we can take $s_1 + s_2 - s_1 s_2$ as our degree of belief in A. If, on the other hand, the first witness testifies for A and the second testifies against A, then we know that not both witnesses are trustworthy, and so we consider the conditional chance that the first witness is trustworthy given that not both are: $s_1(1 - s_2)/(1 - s_1 s_2)$, and we take this as our degree of belief in A. For further information on the rules for belief functions, see Shafer [19, 22].

3 Bayesian Design

We have already distinguished two kinds of Bayesian designs: *total-evidence* designs, in which all one's probability judgments are based on the total evidence, and *conditioning designs*, in which some of the evidence is taken into account by conditioning. In this Sect. we will study these broad categories and consider some other possibilities for Bayesian design.

3.1 Total-Evidence Designs

There are many kinds of probability judgments a total-evidence design might use, for there are many mathematical conditions that can help determine a probability distribution. We can specify quantities such as probabilities, conditional probabilities and expectations, and we can impose conditions such as independence, exchangeability, and partial exchangeability. Spetzler and Stael von Holstein [25], Alpert and Raiffa [1], and Goldstein [8] discuss total-evidence designs for the construction of probability distributions for unknown quantities. Here we discuss total-evidence designs for a few simple problems.

Two Total-Evidence Designs for the Free-Style Race. The Bayesian design for the free-style race suggested by our discussion in Sect. 1.2 above is an example of a total-evidence design based on a causal model. This design involves six possibilities:

$$\begin{aligned}
A_1 &= \text{Curt maintains the pace and wins.}\\
A_2 &= \text{Curt maintains the pace but loses.}\\
A_3 &= \text{Curt soon slows down but still wins.}\\
A_4 &= \text{Curt soon slows down and loses.}\\
A_5 &= \text{Curt collapses at the end but still wins.}\\
A_6 &= \text{Curt collapses at the end and loses.}
\end{aligned}$$

The person who made the analysis (the story was reconstructed from actual experience) was primarily interested in the proposition

$$A = \{A_1 \text{ or } A_3 \text{ or } A_5\} = \text{Curt wins},$$

but her insight into the matter was based on her understanding of the causal structure of the swim race. In order to make the probability judgment $P(A)$, she first made the judgments $P(B_i)$ and $P(A \mid B_i)$, where

$$\begin{aligned}
B_1 &= \{A_1 \text{ or } A_2\} = \text{Curt maintains his pace,}\\
B_2 &= \{A_3 \text{ or } A_4\} = \text{Curt soon slows down,}\\
B_3 &= \{A_5 \text{ or } A_6\} = \text{Curt collapses near the end,}
\end{aligned}$$

and she then calculated $P(A)$ using the rule of total probability—in this case, the formula

$$P(A) = P(B_1)P(A \mid B_1) + P(B_2)P(A \mid B_2) + P(B_3)P(A \mid B_3). \tag{4}$$

She did this qualitatively at the time, but she offers, in retrospect, the quantitative judgments indicated in Table 1. These numbers yield $P(A) = .87$ by (4).

This example brings out the fact that the value of a design depends on the experience and understanding of the person carrying out the mental experiment. For someone who lacked our analyst's experience in swimming and her familiarity with Curt Langley's record, the design (4) would be worthless. Such a person might find some other Bayesian design useful, or he/she might find all Bayesian designs difficult to apply.

Though it is correct to call the design we have just studied a total-evidence design, there is a sense in which its effectiveness does depend on the fact that it allows us to decompose our evidence. The question of what the next event in a causal sequence is likely to be is often relatively easy to answer precisely because only a small part of our evidence bears on it. When we try to decide whether Curt will still win if he slows down, i.e., when we assess $P(A \mid B_2)$—we are able to leave aside our evidence about Curt and focus on how likely Cowan is to maintain his own pace.

Here is another total-evidence design for the free-style race, one which combines the causal model with a more explicit judgment that Cowan's ability is independent of Curt's behavior and ability. We assess probabilities for whether Curt will (a) maintain his pace, (b) slow down, but less than 3%, (c) slow down more than 3%, or (d) collapse. (Whether Curt slows down 3% is significant because this is how much he would have to slow down for Cowan to catch him without speeding up.) We assess probabilities for whether Cowan (a) can speed up significantly, (b) can only maintain his pace, (c) cannot maintain his pace. We judge that these two questions are independent. And finally, we assess the probability that Curt will win under each of the $4 \times 3 = 12$ hypotheses about what Curt will do and what Cowan can do.

Table 2 shows the results of carrying out this design. The numbers in the vertical margin are our probability judgments about Curt, those in the horizontal margin are our probability judgments about Cowan, and those in the cells are our assessments of the conditional probability that Curt will win. These numbers lead to an overall probability of $(.85 \times .10 \times .5) + (.85 \times .70 \times 1.0) + \cdots \approx .88$ that Curt will win.

Our judgments about Cowan are based on our general knowledge about swimmers in the league. The numbers .10, .70, and .20 reflect our impression that perhaps 20% of these swimmers are forced to slow down in the second

Table 1. Component judgments for the first total-evidence design

$P(B_1) = .8$	$P(A \mid B_1) = .95$
$P(B_2) = .15$	$P(A \mid B_2) = .5$
$P(B_3) = .05$	$P(A \mid B_3) = .7$

Table 2. Component judgments for the second total-evidence design

		Cowan		
		Can speed up significantly	Can only maintain pace	Cannot maintain pace
		.10	.70	.20
Curt				
Maintains pace	.85	0.5	1.0	1.0
Slows less than 3%	.03	0.2	1.0	1.0
Slows 3% or more	.07	0.0	0.0	0.5
Collapses	.05	0.2	0.7	0.8

half of a 1650-yard race and that only 10% would have the reserves of energy needed to speed up. We are, in effect, thinking of Cowan as having been chosen at random from this population. We are also judging that Curt's training and strategy are independent of this random choice. Curt's training has probably been influenced mainly by the prospect of the championships. We doubt that Cowan's ability and personality are well enough known to Curt to have caused him to choose a fast start as a strategy in this particular race.

When we compare the design and analysis of Table 2 with the design we carried out earlier, we see that we have profited from the new design's focus on our evidence about Cowan. We feel that the force and significance of this evidence is now more clearly defined for us. On the other hand, we are less comfortable with the conditional probability judgments in the cells of Table 2; some of these seem to be pure speculation rather than assessments of evidence.

Total-Evidence Designs Based on Frequency Semantics. In the two designs we have just considered the breakdown into probabilities and conditional probabilities was partly determined by a causal model. In designs that depend more heavily on frequency semantics, this breakdown depends more on the way our knowledge of past instances is organized.

Consider, for example, the problem of deciding what is wrong when an automobile fails to start. If a mechanic were asked to consider the possible causes for this failure, he might first list the major systems that could be at fault (fuel system, ignition system, etc.), and then list more specific possible defects within each system. This would result in a "fault tree" that could be used to construct probabilities. The steps in the tree would not have a causal interpretation, but the tree would correspond, presumably, to the way the mechanic's memory of the frequencies of similar problems is organized. Fischhoff, Slovic, and Lichtenstein [7] have studied the problem of designing fault trees so as to make them as effective and unbiased as possible.

Here is another simple example based on an anecdote reported by Kahneman and Tversky [10]. An expert undertakes to estimate how long it will take to complete a certain project. He does this by comparing the project to similar past projects. And he organizes his effort to remember relevant information about these past projects into two steps: First he asks how often such projects were completed, and then he asks how long the ones that were completed tended to take. If he focuses on a particular probability judgment—"the probability that our project will be finished within 7 years" say—then he asks first how frequently such projects are completed and then how frequently projects that are completed take less than 7 years.

Why does the expert use this two-step design? Presumably because it facilitates his mental sampling of past instances. It is easier for the expert to thoroughly sample past projects he has been familiar with if he limits himself to asking as he goes only whether they were completed. He can then come back to the completed projects and attack the more difficult task of remembering how long they took.

The emphasis in this example is on personal memory. The lesson of the example applies, however, even when we are aided by written or electronic records. In any case, the excellence of a design depends in part on how the information accessible to us is organized.

3.2 Conditioning Designs

Bayesian conditioning designs can be divided into two classes: *observational* designs and *partitioning* designs. In observational designs, the evidence to be taken into account by conditioning is deliberately obtained after probabilities are constructed. In partitioning designs, we begin our process of probability judgment with all our evidence in hand, but we partition this evidence into "old evidence" and "new evidence," assess probabilities on the basis of the old evidence alone; and then condition on the new evidence.

It should be stressed that a conditioning design always involves two steps: constructing a probability distribution and conditioning it. The name "conditioning design" focuses our attention on the second step, but the first is more difficult. An essential part of any conditioning design is a subsidiary design specifying how the distribution to be conditioned is to be constructed. This subsidiary design may well be a total-evidence design.

Likelihood-Based Conditioning Designs. Bayesian authors often emphasize the use of Bayes's theorem. Bayes's theorem, we recall, says that if $B_1, ..., B_n$ are incompatible propositions, one of which must be true, then

$$P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{\sum_{j=i}^{n} P(B_j)P(A \mid B_j)}. \tag{5}$$

If A represents evidence we want to take into account, and if we are able to make the probability judgments on the right hand side of (5) while leaving this evidence out of account, then we can use (5) to calculate a probability for B_i.

When we use Bayes's theorem in this simple way, we are carrying out a conditioning design. Leaving aside the "new evidence" A, we use the "old evidence" to make probability judgments $P(B_i)$ and $P(A \mid B_i)$. Making these judgments amounts to constructing a probability distribution. We then condition this distribution on A. Formula (5) is simply a convenient way to calculate the resulting conditional probability of B_i.

This is a particular kind of conditioning design. The subsidiary design that we are using to construct the probability distribution to be conditioned is a total-evidence design that just happens to focus on the probabilities $P(B_i)$ and $P(A \mid B_i)$, where A is the new evidence and the B_i are the propositions whose final probabilities interest us. Since the conditional probabilities $P(A \mid B_i)$ are called "likelihoods," we may call this kind of conditioning design a likelihood-based conditioning design.

Both observational and partitioning designs may be likelihood-based. Bayesian theory has traditionally emphasized likelihood-based conditioning designs, and they will also be emphasized in this section. At the end of the section, however, we will give an example of a conditioning design that is not likelihood-based.

A Likelihood-Based Observational Design: The Search for Scorpion. The successful search for the remains of the submarine Scorpion, as reported by Richardson and Stone [17], provides an excellent sample of a likelihood-based observational design. The search was conducted from June to October, 1968, in an area about 20 miles square located 400 miles southwest of the Azores. The submarine was found on October 28.

Naval experts began their probability calculations by using a causal model to construct a probability distribution for the location of the lost submarine. They developed nine scenarios for the events attending the disaster and assigned probabilities to those scenarios. They then combined these probabilities with conditional probabilities representing uncertainties in the submarine's course, speed, and initial position to produce a probability distribution for its final location on the ocean floor. They did not attempt to construct this probability distribution for the final location in continuous form. Instead, they imposed a grid over the search area with cells about one square mile in size and used their probabilities and conditional probabilities in a Monte Carlo simulation to estimate the probability of Scorpion being in each of these approximately 400 cells. They then used these probabilities to plan the search: The cells with the greatest probability of containing Scorpion were to be searched first.

Searching a cell meant towing through the cell near the ocean bottom a platform upon which were mounted cameras, magnetometers, and sonars. The naval experts assessed the probability that this equipment would detect Scorpion if Scorpion were in the cell searched. So when they searched a cell and conditioned on the fact that Scorpion was not found there, they were, in effect, using a likelihood-based conditioning design to assess new probabilities for its location.

This example is typical of likelihood-based observational designs. The probabilities required by the design were subjective judgments, not known objective probabilities. (The assessed likelihood of detecting Scorpion when searching the cell where it was located turned out, for example, to be over optimistic.) But these judgments were made before the observation on which the experts conditioned was made. In fact, these judgments were the basis of deciding which of several possible observations to make, i.e., which cell to search.

A Likelihood-Based Partitioning Design: The Hominids of East Turkana. Let us now turn back to Walker and Leakey's discussion of the number of species of hominids in East Turkana one and a half million years ago. They begin, we recall, by taking for granted a classification of the hominids into three types: the "robust" type I, the "gracile" type II, and *Homo erectus*, type III. They were interested in five hypotheses as to how many distinct species these three types represent:

B_1 = One species.

B_2 = Two species, one composed of I (male) and II (female).

B_3 = Two species, one composed of III (male) and II (female).

B_4 = Two species, one composed of I and III.

B_5 = Three species.

We summarized the evidence they brought to bear on the problem under five headings:

(i). A theoretical argument for B_1.

(ii). Skepticism about such disparate types as I and III being variants of the same species.

(iii). Skepticism about the degree of sexual dimorphism postulated by B_2 and B_3.

(iv). Absence of type I specimens among the type III specimens in the Far East.

(v). Absence of type II specimens among the type III specimens in the Far East.

How might we assess this evidence in the Bayesian language?

Partitioning design seems to hold more promise in this problem than total-evidence design. Except for items (i) and possibly (ii), the evidence cannot be interpreted as an understanding of causes that generate the truth, and hence there is little prospect for a total-evidence design using propensity semantics. We also lack the experience with similar problems that would be required for a successful total-evidence design using frequency semantics. And since it is the diversity of the evidence that complicates probability judgments in the problem, a design that decomposes the evidence seems attractive.

Which of the items of evidence shall we classify as old evidence and which as new? The obvious move is to classify (i) as old evidence and to treat (ii)–(v), taken together, as our new evidence A. This means we will need to assess probabilities, $P(B_1), ..., P(B_5)$ and conditional probabilities, $P(A \mid B_1), ..., P(A \mid B_5)$ and calculate $P(B_i \mid A)$, by (5). The apparent complexity of (5) is lessened if we divide it by the corresponding expression for B_j, obtaining

$$\frac{P(B_i \mid A)}{P(B_j \mid A)} = \frac{P(B_i)}{P(B_j)} \frac{P(A \mid B_i)}{P(A \mid B_j)} \tag{6}$$

or

$$\frac{P(B_i \mid A)}{P(B_j \mid A)} = \frac{P(B_i)}{P(B_j)} L(A \mid B_i : B_j), \tag{7}$$

where $L(A \mid B_i : B_j) = P(A \mid B_i)/P(A \mid B_j)$ is called the *likelihood ratio* favoring B_i over B_j.

Expression (7) represents a real simplification of the design. Since the probabilities $P(B_1 \mid A), ..., P(B_5 \mid A)$ must add to one, they are completely determined by their ratios, $P(B_i \mid A)/P(B_j \mid A)$. Therefore, (7) tells us that it is not necessary to assess the likelihoods, $P(A \mid B_i)$ and $P(A \mid B_j)$. It is sufficient to assess their ratios, $L(A \mid B_i : B_j)$ [6].

One further elaboration of this design seems useful. Our new evidence A can be thought of as a conjunction: $A = A_1$ and A_2, where A_1 is the event that types I, II and III should be so disparate (items of evidence (ii) and (iii)) and A_2 is the event that specimens of types I and II should not be found along with the type III specimens in the Far East (items of evidence (iv) and (v)). The two events A_1 and A_2 seem to involve independent uncertainties, and this can be expressed in Bayesian terms by saying that they are independent events conditional on any one of the five hypotheses:

$$P(A \mid B_i) = P(A_1 \mid B_i) P(A_2 \mid B_i).$$

Substituting this into (6), we obtain

$$\frac{P(B_i \mid A)}{P(B_j \mid A)} = \frac{P(B_i)}{P(B_j)} \frac{P(A_1 \mid B_i) P(A_2 \mid B_i)}{P(A_1 \mid B_j) P(A_2 \mid B_j)} \quad \text{or}$$

$$\frac{P(B_i \mid A)}{P(B_j \mid A)} = \frac{P(B_i)}{P(B_j)} L(A_1 \mid B_i : B_j) L(A_2 \mid B_i : B_j).$$

We are not, of course, qualified to make the probability judgments called for by this design; it is a design for experts like Walker and Leakey, not a design for laymen. (If we ourselves had to make probability judgments about the validity of Walker and Leakey's opinions, we would need a design that analyzes our own evidence. This consists of their article itself, which provides internal evidence as to the integrity and the cogency of their thought, our knowledge of the standards of *Scientific American*, etc.) It will be instructive, nonetheless, to put ourselves in the shoes of Walker and Leakey and to carry out the design on the basis of the qualitative judgments they make in their article. As we shall see, there are several difficulties.

The first difficulty is in determining the prior probabilities $P(B_i)$ on the basis of the evidence (i) alone. This evidence is an argument for B_1 and so evaluation of it can justify a probability $P(B_1)$, say $P(B_1) = .75$. But how do we divide the remaining .25 among the other B_i? This is a typical problem in Bayesian design. In the absence of relevant evidence, we are forced to depend on symmetries, even though the available symmetries may seem artificial and conflicting. In this case, one symmetry suggests equal division among B_2, B_3, B_4, B_5 while another symmetry suggest equal division between the hypothesis of two species (B_2, B_3, B_4) and the hypothesis of three species (B_5). The $P(B_i)$ given in Table 3 represent a compromise.

Now consider A_1, the argument that the different types must represent three distinct species because of their diversity. Our design asks us, in effect, to assess how much less likely this diversity would be under the one-species hypothesis and under the various two-species hypotheses. Answers to these questions are given in the column of Table 3 labeled "$L(A_1 \mid B_i : B_5)$." These numbers reflect the great implausibility of the intraspecies diversity postulated by B_1 and B_4, the marginal acceptability of the degree of sexual dimorphism postulated by B_2, and the implausibility, especially in the putative ancestor of *Homo sapiens*, of the sexual dimorphism postulated by B_3. Notice how fortunate it is that we are required to assess only the likelihood ratios, $L(A_1 \mid B_i : B_5) = P(A_1 \mid B_i)/P(A_1 \mid B_5)$ and not, say, the absolute likelihood $P(A_1 \mid B_5)$. We can think about how much less likely the observed disparity among the three groups would be if they represented fewer than three species, but we would be totally at sea if asked to assess the unconditional chance of this degree of disparity among three extinct hominid species.

Finally, consider A_2, the absence of specimens of type I or II among the abundant specimens of type III in the Far East. This absence would seem much less likely if I or II were forms of the same species as III than if they were not, say 100 times less likely. This is the figure used in Table 3. Notice again that we are spared the well-nigh meaningless task of assessing absolute likelihoods.

As the last column of Table 3 shows, the total evidence gives a fairly high degree of support to B_5, the hypothesis that there are three distinct species. This is Walker and Leakey's conclusion.

How good an analysis is this? There seems to be two problems with it. First, we lack good grounds for some of the prior probability judgments.

Table 3. Component judgments for the likelihood-based partitioning design

	$P(B_j)$	$L(A_1 \mid B_j : B_5)$	$L(A_2 \mid B_j : B_5)$	$P(B_j \mid A)$
B_1	.70	.01	.01	.00060
B_2	.05	.50	1.00	.19983
B_3	.05	.05	.01	.00020
B_4	.05	.01	.01	.00004
B_5	.10	1.00	1.00	.79933

Second, the interpretation of the likelihoods seems strained. Are we really judging that the observed difference between I and III is 100 times more likely if they are separate species than if they are variants of the same species? Or are we getting this measure of the strength of this argument for separate species in some other way?

We should remark that it is a general feature of likelihood-based partitioning designs that only likelihood ratios need be assessed. In likelihood-based observational designs, on the other hand, we do usually need to assess absolute likelihoods. This is because in an observational design we must be prepared to condition on any of the possible observations. If, for example, the possible observations are A and not A, then we need to have in hand both $L(A \mid B_i : B_j) = P(A \mid B_i)/P(A \mid B_j)$ and $L(\text{not } A \mid B_i : B_j) = P(\text{not } A \mid B_i)/P(\text{not } A \mid B_j)$. Since $P(A \mid B_i) + P(\text{not } A \mid B_i) = P(A \mid B_j) + P(\text{not } A \mid B_j) = 1$, these likelihood ratios fully determine the absolute likelihoods $P(A \mid B_i)$ and $P(A \mid B_j)$.

The Choice of New Evidence. Traditionally, Bayesian statistical theory has been concerned with what we have called likelihood-based observational designs. This is because the theory has been based on the idea of a statistical experiment. It is assumed that one knows in advance an "observation space"—the set of possible outcomes of the experiment—and a "parameter space"—the set of possible answers to certain questions of substantive interest. One assesses in advance both prior probabilities for the parameters and likelihoods for the observations.

Many statistical problems do conform to this picture. The search for Scorpion, discussed earlier, is one example. But Bayesians and other statisticians have gradually extended their concerns from the realm of planned experiments, where parameter and observation spaces are clearly defined before observations are made, to the broader field of "data analysis." In data analysis, the examination of data often precedes the framing of hypotheses and "observations." This means that the Bayesian data analyst will often use partitioning designs rather than genuine observational designs.

We believe that Bayesian statistical theory will better meet the needs of statistical practice if it will go beyond observational designs and deal explicitly with partitioning designs. In particular, we need more discussion of principles for the selection of evidence that is to be treated as new evidence. In the example of the hominids, we treated certain arguments as new evidence because we could find better grounds for probability judgment when thinking of the likelihood of their arising than when thinking about them as conditions affecting the likelihood of other events. In other cases, we may single out evidence because its psychological salience can give it excessive weight in total-evidence judgments. By putting such salient evidence in the role of new evidence in a partitioning design, we gain an opportunity to make probability judgments based on the other evidence alone. (Cf. [25], p. 346 and [14], Chap. 3.) We need more discussion of such principles, and more examples.

A Partitioning Design that is not Likelihood-Based. Here is a problem that suggests a partitioning design that is not likelihood-based. Gracchus is accused of murdering Maevius. Maevius's death brought him a great and sorely needed financial gain, but it appears that Maevius and Gracchus were good friends, and our assessment of Gracchus's character suggests only a slight possibility that the prospect of gain would have been sufficient motive for him to murder Maevius. On the other hand, some evidence has come to light to suggest that beneath the apparent friendship Gracchus actually felt a simmering hatred for Maevius, and Gracchus is known to be capable of violent behavior towards people he feels have wronged him. The means to commit the murder is not at issue: Gracchus or anyone else could have easily committed it. But we think it very unlikely that anyone else had reason to kill Maevius.

Our partitioning design uses the fact of Maevius's murder as the new evidence. We consider the propositions:

$$\begin{aligned}
H &= \text{Gracchus hated Maevius,}\\
GI &= \text{Gracchus intended to kill Maevius,}\\
SI &= \text{Someone else intended to kill Maevius,}\\
GM &= \text{Gracchus murdered Maevius,}\\
SM &= \text{Someone else murdered Maevius,}\\
NM &= \text{No one murdered Maevius.}
\end{aligned}$$

Using the old evidence alone, we make the following probability judgments:

$$\begin{array}{lll}
P(H) = .2, & P(GI \mid H) = .2, & P(GI \mid \text{not } H) = .01;\\
P(SI) = .001, & & SI \text{ is independent of } GI;\\
P(GM|GI \,\&\, SI) = .4, & P(SM \mid GI \,\&\, SI) = .4, & P(NM|GI \,\&\, SI) = .2;\\
P(GM|GI \,\&\, \text{not } SI) = .8, & & P(NM|GI \,\&\, \text{not } SI) = .2;\\
P(SM|SI \,\&\, \text{not } GI) = .8, & & P(NM|SI \,\&\, \text{not } GI) = .2;\\
 & & P(NM| \text{ not } GI \,\&\, \text{not } SI) = 1.
\end{array}$$

Combining these judgments, we obtain

$$\begin{aligned}
P(GI) &= P(GI \mid H)P(H) + P(GI \mid \text{not } H)P(\text{not } H)\\
&= (.2)(.2) + (.8)(.01) = .048\\
P(GM) &= P(GM \mid \text{not } GI)P(\text{not } GI) + P(GM \mid GI \,\&\, SI)P(GI)P(SI)\\
&\quad + P(GM \mid GI \,\&\, \text{not } SI)P(GI)P(\text{not } SI)\\
&= (0)(.952) + (.4)(.048)(.001) + (.8)(.048)(.999) = .03838.
\end{aligned}$$

Similarly,

$$P(SM) = .00078 \quad \text{and} \quad P(NM) = .96084.$$

Finally we bring in the new evidence—the fact that Maevius was murdered. We find a probability

$$P(GM \mid \text{not } NM) = \frac{.03838}{.03838 + .00078} = .98$$

that Gracchus did it.

One interesting aspect of this example is the fact that the "new evidence"—the fact that Maevius was murdered—is actually obtained before much of the other evidence. Only after Maevius's death would we have gathered the evidence against Gracchus.

3.3 Other Bayesian Designs

What other Bayesian designs are possible in addition to total-evidence and conditioning design?

A large class of possible designs is suggested by the following general idea. Suppose one part of our evidence lends itself to a certain design d, while the remainder of our evidence does not fit this design, but seems instead relevant to some of the judgments specified by a different design d'. Then we might first construct a distribution P_o using d and considering only the first part of the evidence, and then switch to d', using the total evidence to make those judgments for which the second part of the evidence is relevant and obtaining the other judgments from P_o.

An interesting special case occurs when the total evidence is used only to construct probabilities $p_1, ..., p_n$ for a set of mutually incompatible and collectively exhaustive propositions $A_1, ..., A_n$, and the final distribution P is determined by setting $P(A_i) = p_i$ and $P(B \mid A_i) = P_o(B \mid A_i)$ for all B. Since such designs were considered by Jeffrey [9], we may call them *Jeffrey designs.*

Here is an example of a Jeffrey design. Gracchus is accused of murdering Maevius and the evidence against him is the same as in the preceding example, except that it is not certain that Maevius has been murdered. Perhaps Maevius has disappeared after having been seen walking along a sea cliff. We partition our evidence into two bodies of evidence—the evidence that was used in the probability analysis above, and the other evidence that suggests Maevius may have been murdered. We use the first body of evidence to make the analysis of the preceding section, obtaining the probabilities obtained there: a probability of .03838 that Gracchus murdered Maevius, a probability of .00078 that someone else did, and a probability of .96084 that no one did. We label this probability distribution P_o. Then we use the total evidence to assess directly whether we think Maevius has been murdered or not. Say we assess the probability of Maevius's having been murdered at .95. We then obtain a conditional probability from $P_o : P_o(\text{Graccus did it}|\text{Maevius was murdered}) \approx .98$. The final result is a probability of $.95 \times .98 \approx .93$ for the event that Gracchus murdered Maevius. For further examples of Jeffrey designs, see [21] and [4].

4 Belief-Function Design

Belief-function design differs from Bayesian design in that it puts more explicit emphasis on the decomposition of evidence. As we have seen, total-evidence designs are basic to the Bayesian language. (Even conditioning and Jeffrey designs must have subsidiary designs for the construction of initial distributions, and these subsidiary designs are usually total-evidence designs.) These total-evidence designs break down the task of judgment by asking us to answer several different questions. It is a contingent matter whether different items of evidence bear on these different questions, though this seems to be the case with the most effective total-evidence designs. The belief-function language, on the other hand, since it directly models the meaning and reliability of evidence, breaks down the task of judgment by considering different items of evidence. It is a contingent matter whether these different items of evidence bear on relatively separate and restricted aspects of the questions that interest us, but again, as we shall see, this seems to be the case with the most effective belief-function designs.

Here we shall explore the possibilities for belief-function design for Curt's swim race and Walker and Leakey's hominids. For further examples of belief-function design, see [20, 21, 23, 22].

4.1 The Free-Style Race

The second of the two Bayesian total-evidence designs that we gave for the free-style race (Sect. 3.1) was based on independent judgments about Curt and Cowan. We gave Curt an 85% chance of maintaining his pace, a 3% chance of slowing less than 3%, a 7% chance of slowing more than 3%, and a 5% chance of collapsing. And we gave Cowan a 10% chance of being able to speed up, a 70% chance of only being able to maintain his pace, and a 20% chance of being unable to maintain his pace. Since we were using the Bayesian language, we compared our evidence to knowledge that the evolution of the race actually was governed by these chances. It is equally convincing, however, to interpret these numbers within the language of belief functions. We compare our knowledge about Curt to a message that has an 85% chance of meaning that he will maintain his pace, etc., and we compare our knowledge about Cowan to a message that has a 70% chance of meaning that he can only maintain his pace, etc.

Formally, we have a belief function Bel_1 that assigns degrees of belief .85, .03, .07, and .05 to the four hypotheses about Curt, and a second belief function Bel_2 that assigns degrees of belief .10, .70, and .20 to the three hypotheses about Cowan. Judging that our evidence about Curt is independent of our evidence about Cowan, we combine these by Dempster's rule. If no further evidence is added to the analysis, then our resulting degree of belief that Curt will win will be our degree of belief that Curt will maintain his pace or slow less than 3% while Cowan is unable to speed up: $(.85 + .03)(.70 + .20) = .792$. And our degree of belief that Cowan will win will be our degree of belief that Curt will slow 3% or more and Cowan will be able to at least maintain his pace: $(.07)(.10 + .70) = .056$.

These conclusions are weaker than the conclusions of the Bayesian analysis. This is principally due to the fact that we are not claiming to have evidence about what will happen in the cases where our descriptions of Curt's and Cowan's behavior do not determine the outcome of the race. If we did feel we had such evidence, it could be introduced into the belief-function analysis.

We can also relax the additivity of the degrees of belief about Curt and Cowan that go into the belief-function analysis. Suppose, for example, that we feel our evidence about Curt justifies only an 85% degree of belief that he will maintain his pace, but we do not feel we have any positive reason to think he will slow down or collapse. In this case, we can replace the additive degrees of belief .85, .03, .07, and .05 with a simple support function that assigns only degree of belief .85 to the proposition that Curt will maintain his pace. If we retain the additive degrees of belief .10, .70, and .20 for Cowan's behavior, this leads to a degree of belief $(.85)(.70 + .20) = .765$ that Curt will win and a degree of belief zero that Cowan will win.

As this example illustrates, a belief-function design can be based on a causal structure like those used in Bayesian total-evidence designs. The belief-function design must, however, go beyond this causal structure to an explicit specification of the evidence that bears on its different parts.

4.2 The Hominids of East Turkana

Recall that Walker and Leakey considered five hypotheses:

B_1 = One species.
B_2 = Two species, one composed of I (male) and II (female).
B_3 = Two species, one composed of III (male) and II (female).
B_4 = Two species, one composed of I and III.
B_5 = Three species.

In our Bayesian analysis in Sect. 3.2, we partitioned the evidence into three intuitively independent arguments:

1. A theoretical argument for B_1.
2. An argument that the three types are too diverse not to be distinct species. This argument bears most strongly against B_1 and B_4, but also carries considerable weight against B_3 and some weight against B_2.
3. The fact that neither I nor II specimens have been found among the III specimens in the Far East. This provides evidence against hypotheses B_1, B_3, and B_4.

Let us represent each of these arguments by a belief function. Making roughly the same judgments as in the Bayesian analysis, we have

1. Bel_1, with $m_1(B_1) = .75$ and $m_1(\Theta) = .25$,
2. Bel_2, with $m_2(B_5) = .5$, $m_2(B_2 \text{ or } B_5) = .45$, $m_2(B_2 \text{ or } B_3 \text{ or } B_4) = .04$, and $m_2(\Theta) = .01$, and
3. Bel_3, with $m_3(B_2 \text{ or } B_5) = .99$ and $m_3(\Theta) = .01$.

Combining these by Dempster's rule, we obtain a belief function Bel with $m(B_5) = .4998$, $m(B_2 \text{ or } B_5) = .4994$, $m(B_2 \text{ or } B_4 \text{ or } B_5) = .0004$, $m(B_1) = .0003$, and $m(\Theta) = .0001$. This belief function gives fair support to B_5 and overwhelming support to B_2 or B_5: $Bel(B_5) = .4998$ and $Bel(B_2 \text{ or } B_5) = .9992$.

These belief-function results can be compared to the Bayesian results of Sect. 3.2, where we obtained $P(B_5) = .7993$ and $P(B_2 \text{ or } B_5) = .9992$. The different results for B_5 can be attributed to the different treatments of the first item of evidence, the argument against coexistence of hominid species. In the belief-function analysis, we treated this argument simply by giving B_1 a 75% degree of support. In the Bayesian analysis, we had to go farther and divide the remaining 25% among the other four hypotheses. The belief-function analysis, while it reaches basically the same conclusion as the Bayesian argument, can be regarded as a stronger argument, since it is based on slightly more modest assumptions.

5 The Nature of Probability Judgment

We have suggested that probability judgment is a kind of mental experiment. Sometimes it is like a statistician's thought experiment, as when we search, in our mind or on a bookshelf, for examples on which to base a frequency judgment. Sometimes it is more like a physicist's thought experiment, as when we try to trace the consequences of an imagined situation.

Probability judgment is a process of construction rather than elicitation. People may begin a task of probability judgment with some beliefs already formulated. But the process of judgment, when successful, gives greater content and structure to these beliefs and tends to render initial beliefs obsolete. It is useful, in this respect, to draw an analogy between probability and affective notions such as love and loyalty. A declaration of love is not simply a report on a person's emotions. It is also part of a process whereby an intellectual and emotional commitment is created; so too with probability.

A probability judgment depends not just on the evidence on which it is based, but also on the process of exploring that evidence. The act of designing a probability analysis usually involves reflection about what evidence is available and a sharpening of our definition of that evidence. And the implementation of a design involves many contingencies. The probability judgments we make may depend on just what examples we sampled from our memory or other records, or just what details we happen to focus on as we examine the possibility of various scenarios [26].

It may be helpful to point out that we do not use the word "evidence" as many philosophers do—to refer to a proposition in a formal language. Instead, we use it in a way that is much closer to ordinary English usage. We refer to "our evidence about Cowan's abilities," to "our memory as to how frequently similar projects are completed," or to "the argument that distinct hominid species cannot coexist." The references are, as it were, ostensive definitions of bodies of evidence. They point to the evidence in question without translating it into statements of fact in some language. This seems appropriate, for in all these cases the evidence involves arguments and claims that would fall short of being accepted as statements of fact.

Evidence, as we use the word, is the raw material from which judgments, both of probability and of fact, are made. Evidence can be distinguished in this respect from information. Information can be thought of as answers to questions already asked, and hence we can speak of the quantity of information, which is measured by the number of these questions that are answered. Evidence, in contrast, refers to a potential for answering questions. We can speak of the weight of evidence as it bears on a particular question, but it does not seem useful to speak of the quantity of evidence.

Though we have directed attention to the notion of mental experimentation, we want also to emphasize that when an individual undertakes to make a probability judgment that individual is not necessarily limited to the resources of memory and imagination. He or she may also use paper, pencils, books, files, and computers. And an individual need not necessarily limit his or her sampling experiments to haphazard search of memory and personal bookshelves. The individual may wish to extend sampling to a large-scale survey, conducted with the aid of randomization techniques.

There is sometimes a tendency to define human probability judgment narrowly—to focus on judgments people make without external aids. But it may not be sensible to try to draw a line between internal and external resources. Psychologists who wish to offer a comprehensible analysis of human judgment should, as Ward Edwards [5] has argued, take into account the fact that humans are tool-using creatures. Moreover, statisticians and other practical users of probability need to recognize the continuity between apparently subjective judgments and supposedly objective statistical techniques. The concept of design that we have developed in this paper is meant to apply both to probability analyses that use sophisticated technical aids and to those that are made wholly in our heads. We believe that the selection of a good design for a particular question is a researchable problem with both technical and judgmental aspects. The design and analysis of mental experiments for probability judgment therefore represents a challenge to both statisticians and psychologists.

References

[1] Marc Alpert and Howard Raiffa, 1982. A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, and A. Tversky, editors, *Judgment under Uncertainty: Heuristics and Biases.* Cambridge University Press, New York.

[2] Joseph Bertrand, 1907. *Calcul des Probabilités (2nd ed).* Gauthier-Villars.

[3] George E. P. Box, 1980. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society Series A*, 143: 383–430.

[4] Persi Diaconis and Sandy L. Zabell, 1982. Updating subjective probability. *Journal of the American Statistical Association*, 77:822–830.

[5] Ward Edwards, 1975. Comment on paper by Hogarth. *Journal of the American Statistical Association*, 70:291–293.

[6] Ward Edwards, Lawrence D. Phillips, William L. Hays, and Barbara C. Goodman, 1968. Probabilistic information processing systems: Design and evaluation. *IEEE Transactions on Systems Science and Cybernetics*, 4:248–265.

[7] Baruch Fischhoff, Paul Slovic, and Sarah Lichtenstein, 1978. Fault trees: Sensitivity of established failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance*, 4:330–344.

[8] Michael Goldstein, 1981. Revising prevision: A geometric interpretation. *Journal of Royal Statistical Society, Series B*, 43:105–130.

[9] Richard Jeffrey, 1965. *The Logic of Decision.* McGraw-Hill, New York.

[10] Daniel Kahneman and Amos Tversky, 1982. Variants of uncertainty. *Cognition*, 11:143–157.

[11] David Krantz and John Miyamoto, 1983. Priors and likelihood ratios as evidence. *Journal of the American Statistical Association*, 78:418–423.

[12] I. Lieblich and A. Lieblich, 1969. Effects of different pay-off matrices on arithmetic estimation tasks: An attempt to produce "rationality". *Perceptual and Motor Skills*, 29:467–473.

[13] Dennis V. Lindley, Amos Tversky, and Rex V. Brown, 1979. On the reconciliation of probability assessments. *Journal of the Royal Statistical Society*, 147:146–180.

[14] Richard Nisbett and Lee Ross, 1980. *Human Inference: Strategies and Shortcomings of Social Judgment.* Prentice Hall, Englewood Cliffs, NJ.

[15] Howard Raiffa, 1974. *Analysis of decision making. An audiographic, self-instructional course.* Encyclopedia Britanica Educational Corporation, Chicago.

[16] Frank P. Ramsey, 1931. Truth and probability. In R. G. Braithwaite, editor, *The Foundations of Mathematics and Other Logic Essays.* Routledge and Kegan Paul.

[17] Henry R. Richardson and Lawrence D. Stone, 1971. Operations analysis during the underwater search for scorpion. *Naval Research Logistics Quaterly*, 18: 141–157.

[18] Leonard J. Savage, 1954. *The Foundations of Statistics.* Wiley, New York, NY.

[19] Glenn Shafer, 1976. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ.

[20] Glenn Shafer, 1981. Constructive probability. *Synthese*, 448:1–60.

[21] Glenn Shafer, 1981. Jeffrey's rule of conditioning. *Philosophy of Science*, 48:337–362.

[22] Glenn Shafer, 1982. Belief functions and parametric models. *Journal of Royal Statistical Society, Series B*, 44:322–352.

[23] Glenn Shafer, 1982. Lindley's paradox. *Journal of the American Statistical Association*, 77:325–351.

[24] Max Singer, 1971. The vitality of mythical numbers. *The Public Interest*, 23:3–9.

[25] Carl S. Spetzler and Carl-Axel S. Staël von Holstein, 1975. Probability encoding in decision analysis. *Management Science*, 22:340–358.

[26] Amos Tversky and Daniel Kahneman, 1983. Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90:293–315.

[27] Alan Walker and Richard E. T. Leakey, 1978. The hominids of East Turkana. *Scientific American*, 238:54–66.

[28] Sławomir T. Wierzchoń, 1984. An inference rule based on Sugeno measure. Institute of Computer Science, Polish Academy of Science, Warsaw, Poland. Later appeared in J. C. Bezdek, editor, *Analysis of Fuzzy Information*, CRC Press, 1:85–96, 1987.

14

A Set-Theoretic View of Belief Functions

Logical Operations and Approximations by Fuzzy Sets*

Didier Dubois and Henri Prade

Abstract. A body of evidence in the sense of Shafer can be viewed as an extension of a probability measure, but as a generalized set as well. In this paper we adopt the second point of view and study the algebraic structure of bodies of evidence on a set, based on extended set union, intersection and complementation. Several notions of inclusion are exhibited and compared to each other. Inclusion is used to compare a body of evidence to the product of its projections. Lastly, approximations of a body of evidence under the form of fuzzy sets are derived, in order to squeeze plausibility values between two grades of possibility. Through all the paper, it is pointed out that a body of evidence can account for conjunctive as well as a disjunctive information, i.e. the focal elements can be viewed either as sets of actual values or as restrictions on the (unique) value of a variable.

Key words: Theory of evidence, Possibility measure, Fuzzy set, Knowledge representation

Introduction

The framework of plausibility and credibility (or belief) functions[24] or, equivalently that of the random sets[19] encompasses both probability theory and possibility theory[7, 38]. It is now acknowledged that fuzzy sets[35] viewed as possibility distributions, are, using Shafer's terminology, contour functions of consonant belief functions[4, 17] or in the terminology of random sets, one-point coverages of random sets[13, 22, 28]. In a recent paper[25] Shafer carefully examines the rules of calculation of fuzzy sets and possibility measures as opposed to their counterparts for belief functions. It turns out that the main difference lies in the use of Dempster rule for combining belief functions versus fuzzy set-intersection for combining possibility measures. Dempster rule applied to the combination of possibility measures does

* This paper is based on a presentation at the 1st IFSA Conference, held in Palma de Mallorca, Spain, July 1–6, 1985.

not yield a possibility measure while a fuzzy set-intersection does. This paper is a contribution to the debate between possibility measures and belief functions. First, combination rules for belief functions in the spirit of Dempster rule are described; they are counterparts of fuzzy set-theoretic union, complementation, products and projection. This set-theoretic view of belief functions points out the fundamental identity of both approaches to combining. Next, an extensive study of the concept of inclusion of bodies of evidence is carried out. Four definitions are proposed and compared. The existence of two antagonistic points of view on bodies of evidence is stressed and it brings some light to discriminate between definitions of inclusion. The following section is devoted to projections and products of belief functions, and the links between a body of evidence and the product of its projections. Lastly the problem of approximating belief functions by consonant bodies of evidence is considered, and best approximations, which squeeze a plausibility measure between two possibility measures, are calculated.

1 Shafer's Theory of Evidence Revisited

In this section, basic notions are introduced in a concise manner. It borrows from several already published works [4, 24, 25, 38] to which the reader is referred for proofs or detailed explanations. However some new issues are raised, especially the convexity of the set of belief functions and the difference between conjunctive and disjunctive items of information. This last point follows some early remarks by Zadeh[37] and a more elaborated discussion by Yager[31] in the framework of fuzzy sets and linguistic variables. Moreover, the allocation of a probability weight on the empty set is no longer forbidden.

1.1 Uncertainty Measures Induced by a Body of Evidence

According to Shafer[24], a body of evidence is modelled by a weighted set of logical statements, each referring to a subset A of a frame of discernment Ω. This frame of discernment corresponds to a point of view on a problem, and contains the possible values of some variable x. A body of evidence supplies information about the actual value of x (which is some element in Ω), with the following conventions, given here in a finite setting for simplicity. Let $\mathcal{F}$ be a family of subsets of Ω. A body of evidence is viewed as a pair $(\mathcal{F}, m)$ where m is a mapping from 2^Ω to the unit interval such that $m(A) > 0$ if and only if $A \in \mathcal{F}$. Any element of A of $\mathcal{F}$ is called a focal element, because part of the available information focuses on A. $m(A)$ is the relative weight of the statement "$x \in A$", and is viewed as the share of total belief committed to this statement exactly, and not to any other statement of the form "$x \in B \subset A$". m is called a basic assignment and satisfies the following requirement

$$\sum_{A \subseteq \Omega} m(A) = 1 \tag{1}$$

where 1 stands for the amount of total belief. The set of bodies of evidence on Ω is denoted as $\mathcal{B}(\Omega)$. In Shafer's book a basic assignment satisfies the additional condition

$$m(\varnothing) = 0 \tag{2}$$

which claims that no belief should be committed to the impossible event. (2) is a normalization condition which looks reasonable if the statement "$x \in \Omega$" is taken for granted. However in some instances one may be uncertain as to whether Ω is definitely exhaustive, or whether assigning a value to x is ever meaningful. For instance, if x is the age of cars belonging to some population where some individuals may have no cars (Zadeh[42]). Such situations can be conveniently handled by letting $m(\varnothing) > 0$. See also Dubois and Prade[4], Yager[29], Zadeh[41, 42] for further discussions. A body of evidence satisfying (2) is said to be *normal.*

Viewed as an allocation of probability over subsets of Ω, a body of evidence is also a random set[19]. However it can be equivalently represented by one of the following set-functions

$$\forall\ A \subseteq \Omega,\ \mathrm{Cr}(A) = \sum_{\varnothing \neq B \subseteq A} m(B), \tag{3}$$

$$\forall\ A \subseteq \Omega,\ \mathrm{Pl}(A) = \sum_{B \cap A \neq \varnothing} m(B), \tag{4}$$

$$\forall\ A \subseteq \Omega,\ Q(A) = \sum_{A \subseteq B} m(B). \tag{5}$$

Cr is called a belief function by Shafer[24], but we had rather call it a *credibility measure* since $\mathrm{Cr}(A)$ gathers the pieces of evidence which support A. Pl is called a plausibility measure since $\mathrm{Pl}(A)$ gathers the pieces of evidence which make the occurrence of A possible. Pl and Cr are related through the duality relation

$$\forall\ A,\ \mathrm{Pl}\,(A) + \mathrm{Cr}\left(\bar{A}\right) = 1 - m\,(\varnothing) \tag{6}$$

i.e. $\mathrm{Pl}(A)$ accounts for evidence which does not support the opposite event $\bar{A}$ nor events "outside Ω" (i.e. $\varnothing$). Q is called a *commonality function* by Shafer[24] and gathers pieces of evidence supported by event A. So far, its usefulness has been purely technical. Note that

$$\mathrm{Pl}\,(\varnothing) = \mathrm{Cr}\,(\varnothing) = 0; \qquad Q(\varnothing) = 1 \tag{7}$$

$$\mathrm{Pl}\,(\Omega) = \mathrm{Cr}\,(\Omega) = 1 - m\,(\varnothing)\,; \qquad Q\,(\Omega) = m\,(\Omega)\,. \tag{8}$$

When $m(\varnothing) = 0$, Shafer[24] has proved that Cr is order-n superadditive $\forall\ n \in\ \mathbb{N}$. This property still holds when $m(\varnothing) > 0$, for $\mathrm{Cr} + m(\varnothing)$, hence for Cr too. Then the basic assignment is still expressed in terms of the credibility measures as

$$\forall\ A, m\,(A) = \sum_{B \subseteq A} (-1)^{[A-B]}\,(\mathrm{Cr}(B) + m(\varnothing)) = \sum_{B \subseteq A} (-1)^{|A-B|}\mathrm{Cr}(B) \tag{9}$$

where $|A-B|$ is the cardinality of the set-difference $A-B$. See Shafer[24] for other inversion formulae (Pl in terms of Q, etc....). Pl and Cr are monotonic increasing with respect to set-inclusion, while Q is monotonic decreasing. As a consequence of (6), Pl is subadditive, which reads at order n:

$$\mathrm{Pl}\left(A_1 \cap A_2 \ldots \cap A_n\right) \leqq \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ I \neq \varnothing}} (-1)^{|I|+1}\,\mathrm{Pl}\left(\bigcup_{i\in I} A_i\right). \tag{10}$$

The set of plausibility measures on Ω is isomorphic to $\mathcal{B}(\Omega)$ and has an interesting structure. Namely it is a convex set since the convex combination $\sum_{i=1}^{n} \alpha_i \cdot \mathrm{Pl}_i$ of subadditive functions Pl_i is subadditive too. The coefficients α_i are such that $\sum_{i=1}^{n} \alpha_i = 1,\ \alpha_i \geqq 0,\ \forall\, i$. The plausibility measure $\mathrm{Pl} = \sum_{i=1}^{n} \alpha_i \cdot \mathrm{Pl}_i$ is called a *mixture*. Let $(\mathcal{F}_i, m)$ be the body of evidence associated with Pl_i. Then, that associated with Pl is $(\mathcal{F}, m)$ such that

$$\mathcal{F} = \bigcup_{i=1,n} \mathcal{F}_i;\ \forall\ A \subseteq \Omega, m\,(A) = \sum_{i=1}^{n} \alpha_i m_i\,(A)\,.$$

The same remark holds for credibility measures. Let $\mathcal{B}^{+}(\Omega)$ be the set of normal bodies of evidence. Clearly $\mathcal{B}^{+}(\Omega)$ is a convex subset of $\mathcal{B}(\Omega)$.

1.2 Possibility, Necessity, Probability

Two extreme cases of plausibility measures can be obtained by adding constraints on the set of focal elements.

a) *$\mathcal{F}$ contains only singletons*, i.e. $\forall\ A \in \mathcal{F},\ \exists\ \omega \in A,\ A = \{\omega\}$. This occurs if and only if Cr $=$ Pl and is a probability measure P. m is a probability assignment in the usual sense $(P(\{\omega\}) = m(\{\omega\}))$.

The set functions Pl and Cr can be viewed as upper and lower probabilities (Dempster[1]) since any probability measure P generated from $(\mathcal{F}, m)$ by the following allocation procedure

i) $\forall\ A \in \mathcal{F}$ choose $\omega_A \in A$
ii) set $P(\{\omega\}) = \sum_{\omega_A = \omega} m(A), \forall\ \omega \in \Omega$

satisfies the following inequalities:

$$\forall\ A, \mathrm{Cr}(A) \leqq P(A) \leqq \mathrm{Pl}\,(A) \tag{11}$$

when $(\mathcal{F}, m)$ is normal.

Dempster[1] has proved that the set of probability measures satisfying (11) is convex and is the convex closure of the set of probability measures obtained by the procedure (i)–(ii).

b) *$\mathcal{F}$ contains only a nested sequence of subsets $E_1 \subseteq E_2 \ldots \subseteq E_p$.* It occurs if and only if $\forall\ A, B \subseteq \Omega$

$$\mathrm{Cr}\left(A \cap B\right) = \min\left(\mathrm{Cr}\left(A\right), \mathrm{Cr}\left(B\right)\right) \tag{12}$$

$$\mathrm{Pl}\left(A \cup B\right) = \min\left(\mathrm{Pl}\left(A\right), \mathrm{Pl}\left(B\right)\right). \tag{13}$$

Cr is called a consonant belief function by Shafer[24] and Pl a possibility measure by Zadeh[38]. A possibility measure is denoted Π, and the duality relationship (6) justifies the name of "necessity measure"[2] for consonant belief functions. They are also called certainty measures by Zadeh[40], and shall be denoted N in the following.

The set $\pi(\Omega)$ of possibility measures is not convex. Indeed if $\mathcal{F}$ and $\mathcal{F}'$ both define nested sequences, then generally $\mathcal{F} \cup \mathcal{F}'$ does not, so that $\alpha\Pi + (1-\alpha)\Pi'$ is not always a possibility measure. A possibility measure Π such that $\forall\ A, \Pi(A) \in \{0,1\}$ is called a *crisp* possibility measure. Any crisp possibility measure derives from a unique focal element which is a subset E of Ω, i.e. ($\mathcal{F} = \{E\}$).

These two extreme cases of bodies of evidence correspond to precise but scattered pieces of uncertain information (Case (a)) and imprecise but consonant pieces of information (Case (b)). The nature of the relevant uncertainty measure (possibility or probability) is dictated by the structure of the available body of evidence. Generally a body of evidence is neither consonant nor precise. A body of evidence $(\mathcal{F}, m)$ is said to be *consistent* if and only if $\bigcap_{A \in \mathcal{F}} A \neq \varnothing$. This condition is weaker than the consonant constraint of nested focal elements, but still expresses some agreement between the various statements which form the body of evidence.

The following result indicates that in some sense probability measures and possibility measures are the basic concepts in the theory of evidence:

Proposition 1. *Any plausibility measure other than a possibility or a probability measure is a convex combination of a probability measure and possibility measures which are not Dirac functions.*

Proof. For any subset A of Ω, denote Π_A the possibility measure such that $\{A\}$ is its set of focal elements. Let Pl be a plausibility measure. Then (4) also reads

$$\mathrm{Pl}\left(A\right) = \sum_{B \subseteq \Omega} m\left(B\right) \cdot \Pi_B\left(A\right).$$

Now if B is a singleton, then Π_B is a Dirac function, so that the plausibility measure defined by

$$P\left(A\right) = \frac{\sum\limits_{|B|=1} m\left(B\right) \cdot \Pi_B\left(A\right)}{\sum\limits_{|B|=1} m\left(B\right)}$$

is a probability measure when it exists. Q.E.D.

As a consequence, if we identify the set of crisp possibility measures with 2^Ω, the set of subsets of Ω, through the bijection $A \mapsto \Pi_A$ such that $m(A) = 1$, the set $\mathcal{B}(\Omega)$ can be viewed as the convex hull of 2^Ω, while the set $\mathcal{P}(\Omega)$ of probability measures is the convex hull of the subset of singletons of Ω.

1.3 Possibility Measures as Fuzzy Sets

A possibility or a probability measure is entirely characterized by the set $\{\mathrm{Pl}(\{\omega\}) | \omega \in \Omega\}$; $\mathrm{Pl}(\{\omega\})$ is the one-point coverage function, in terms of random sets[13] and is called a contour function by Shafer[24]. In the case of probability measures, $\mathrm{Pl}(\{\omega\}) = P(\{\omega\})$ and $\forall\ A,\ \mathrm{Pl}(A) = \sum_{\omega \in A} \mathrm{Pl}(\{\omega\})$. In case of a possibility measure

$$\Pi\left(A\right) = \max_{\omega \in A} \Pi\left(\{\omega\}\right);\ N(A) = \min_{\omega \in \bar{A}}\ 1 - \Pi\left(\{\omega\}\right). \tag{14}$$

When $\Pi(\Omega) = 1$, we have $\max_{\omega \in \Omega} \Pi(\{\omega\}) = 1$.

In the following, $\Omega = \{\omega_1, \ldots, \omega_n\}$ has n elements, $P(\{\omega_i\})$ is denoted p_i, and $\Pi(\{\omega_i\})$ is denoted π_i, for the sake of simplicity. When Π has values only in {0,1}, the function $\mu_F : \Omega \mapsto [0,1]$ defined by

$$\mu_F\left(\omega_i\right) = \pi_i \tag{15}$$

is the characteristic function of a set. In the general case it is the membership function of a fuzzy set[35] F.

Let $F_\alpha = \{\omega | \mu_F(\omega) \geqq \alpha\}$ be the *α-cut* of F. When Ω is finite the set $\{F_\alpha | \alpha \in [0,1]\}$ of α-cuts is finite, and it is proved[4] that it is the set of focal elements of the possibility measure such that $\mu_F(\omega) = \Pi(\{\omega\})$. More specifically assume $\pi_1 = 1 \geqq \pi_2 \geqq \cdots \geqq \pi_n \geqq \pi_{n+1} = 0$ and let $A_i = \{\omega_1, \ldots, \omega_i\}$. Then the basic assignment m is defined in terms of the π_i's by:[4]

$$\begin{cases} m(A) = 0 \text{ if } \ \nexists\ i : A = A_i \\ m\left(A_i\right) = \pi_i - \pi_{i+1} \end{cases}. \tag{16}$$

In the general case, $\mathrm{Pl}(\{\omega\})$ may still be interpreted as the membership grade of ω in a fuzzy set F. However the knowledge of $\{\mathrm{Pl}(\{\omega\}) | \omega \in \Omega\}$ is not enough to recover the body of evidence $(\mathcal{F}, m)$. Moreover F is not always a normalized fuzzy set. Namely, even if $(\mathcal{F}, m)$ is normal,

$$\exists\ \omega : \mathrm{Pl}\left(\{\omega\}\right) = 1 \text{ if and only if } \ (\mathcal{F}, m) \ \text{ is consistent.}$$

Moreover when Pl is a probability measure, it rather corresponds to the idea of a fuzzy point,[16] since the grade of complete membership (1) is shared among the singletons in that case. The characterization of plausibility measures Pl such that $\forall\ \omega,\ \mathrm{Pl}(\{\omega\}) = \mu_F(\omega)$, given μ_F, is done by Goodman[13] in the setting of random sets.

1.4 Disjunctive versus Conjunctive Evidence

In the preceding paragraphs, a set is viewed as restricting the possible values of a variable x, and these values are supposedly mutually exclusive. Similarly fuzzy sets are viewed as fuzzy restrictions[36]. There is another view of sets, as containing values which are actually taken by x. This point of view is considered by Yager[31] in terms of linguistic variables and by Prade and Testemale[21] in the framework of fuzzy relational databases. In the first case variables are single-valued and the body of evidence is said to be *disjunctive.* In the second case, variables are multiple-valued, and the body of evidence is said to be *conjunctive.* The difference between conjunctive and disjunctive fuzzy sets has been pointed out by Zadeh[37].

Example 1. "John is *tall*" means that John's height is some number restricted by the fuzzy set "tall".

"John stayed in Paris from 1980 to 1984" means that {1980, 1981, 1982, 1983, 1984} is a set of years when John actually stayed in Paris.

In the case of conjunctive knowledge, $\forall\ B \subseteq A$, if "$x = A$" is true then "$x = B$" is also true, so that the entailment principle[39] works backwards (Yager[31]). As a consequence, the quantity $Q(A)$, i.e. the commonality number, defined by (5), is the actual grade of credibility of "$x = A$" in the case of a conjunctive body of evidence, instead of $\mathrm{Cr}(A)$, as pointed out by Zadeh. Notice that, for singletons the identity

$$Q\left(\{\omega\}\right) = \mathrm{Pl}\left(\{\omega\}\right) \tag{17}$$

holds, and moreover if $\mathcal{F}$ is consonant then, equivalently

$$\forall\ A, B, Q\left(A \cup B\right) = \min\left(Q\left(A\right), Q\left(B\right)\right). \tag{18}$$

In the consonant case, $\{Q(\{\omega\}) | \omega \in \Omega\}$ also characterizes the body of evidence and

$$\forall\ A, Q\left(A\right) = \min_{\omega \in A} Q\left(\{\omega\}\right). \tag{19}$$

In the conjunctive context, the membership function μ_F defined by (15) is no longer viewed as a possibility distribution, but what could be termed as a "necessity" or "certainty" distribution since $\mu_F(\omega_i) = Q(\{\omega_i\})$ is now the grade of certainty that ω_i is a value of x. The grade of possibility is then defined by

$$\phi\left(A\right) = 1 - Q\left(\bar{A}\right) = \max_{\omega \notin A} 1 - \mu_F\left(\omega\right). \tag{20}$$

Note that when F is such that there are at least two elements ω' and ω'' such that $\mu_F(\omega') = \mu_F(\omega'') = 0$, then $\forall\ \omega, \phi(\{\omega\}) = 1$. Indeed $\mu_F(\omega) = 0$ does not forbid ω as a *value* of x but only let this statement be contingent (total uncertainty). In other words "$x = A$" means that x takes *at least* all values in A. Lastly note that from (19) and (20)

$$\phi(A) = \Pi_{\bar{F}}\left(\bar{A}\right), Q(A) = 1 - \Pi_{\bar{F}}(A) = N_{\bar{F}}\left(\bar{A}\right) \tag{21}$$

where $\Pi_{\bar{F}}$ is the possibility measure where the underlying possibility distribution is the membership function of the complement $\bar{F}$ of F, i.e. $1 - \mu_F$.

The notion of conjunctive versus disjunctive types of information seems to be an important issue in knowledge representation, and is encountered in the next section, as a by-product.

2 Set-Theoretic Operations on Bodies of Evidence

Dempster[1] has introduced a rule of combination for two disjunctive normalized bodies of evidence $(\mathcal{F}_1, m_1)$, $(\mathcal{F}_2, m_2)$, consistently with Bayes rule of conditioning. It reads:

$$\forall\ A \subseteq \Omega, (m_1 \cap m_2)(A) = \sum_{B \cap C = A} m_1(B) \cdot m_2(C) \tag{22}$$

$$\forall\ A \subseteq \Omega, m(A) = \frac{(m_1 \cap m_2)(A)}{1 - (\{m_1 \cap m_2\})\varnothing}. \tag{23}$$

Equation (22) can be justified in statistical terms on the basis of the independence of the sources which provide $(\mathcal{F}_1, m_1)$ and $(\mathcal{F}_2, m_2)$. Equation (23) underlies a complete reliability of these sources, and is a normalization technique. The term $(m_1 \cap m_2)(\varnothing)$ reflects the amount of dissonance between the sources, and is eliminated. Equation (22) can be viewed as performing the intersection of independent random sets[4, 13].

Contrastedly, if Π_1 and Π_2 are two possibility measures, with possibility distributions $\pi_1 = \mu_{F_1}, \pi_2 = \mu_{F_2}$, a possibility measure Π_{12} can be obtained from the possibility distribution $\pi_{12} = \mu_{F_1 \cap F_2}$ where the fuzzy set-theoretic intersection is defined by a triangular norm[8, 23]:

$$\pi_{12} = \pi_1 * \pi_2. \tag{24}$$

The main candidates for $*$ are $a * b = \min(a, b)$; $a \cdot b$; $\max(0, a + b - 1)$.[2, 8] Assuming the complete reliability of the sources leads to normalize π_{12} into

$$\forall\ \omega, \pi(\omega) = \frac{\pi_1(\omega) * \pi_2(\omega)}{\max\limits_{\omega \in \Omega} \pi_1(\omega) * \pi_2(\omega)}. \tag{25}$$

(24) and (25) are possibilistic counterparts of (22) and (23) respectively.

It was pointed[4] that if Π_1 and Π_2 are combined via (22) what is obtained is generally not a possibility measure. This is because when $\mathcal{F}_1$ and $\mathcal{F}_2$ are consonant, the set $\mathcal{F}_{1\cap 2} = \{A \cap B | A \in \mathcal{F}_1, B \in \mathcal{F}_2\}$ is generally not consonant. Besides, using (22) yields a not necessarily normalized plausibility function Pl_{12} where[4]:

$$\mathrm{Pl}_{12}\left(\{\omega\}\right) = \pi_1\left(\omega\right) \cdot \pi_2\left(\omega\right) \tag{26}$$

which is a particular instance of (24) where $*$ is the product. Hence Dempster rule is closely related to a fuzzy set intersection. But generally

$$\mathrm{Pl}_{12}\left(A\right) \geqq \Pi_{12}\left(A\right) = \max_{\omega \in A} \pi_1\left(\omega\right) \cdot \pi_2\left(\omega\right) \tag{27}$$

i.e. Π_{12} is more informative than Pl_{12}.

The set of combination operations for fuzzy sets is richer than for bodies of evidence since all connectives of propositional logic can be extended to the combination of fuzzy sets and this extension is not unique. Strangely enough counterparts of set-union, set-complementation, etc.. . .have not been considered for bodies of evidence, but in the mathematical literature of random sets[12, 14].

In the following, these connectives are defined at an elementary level[1], thus casting Dempster rule in a set-theoretic framework, and enriching the set of combination rules. This view of plausibility measures reflects the standpoint of logic and contrasts with the measure-theoretic view which Dempster had when he introduced his concept of upper and lower probabilities and expectations.

2.1 The Union of Bodies of Evidence

The union of two bodies of evidence $(\mathcal{F}_1, m_1)$ and $(\mathcal{F}_2, m_2)$ on Ω is defined, in the spirit of (22)–(23) by the basic assignment $m_1 \cup m_2$ such that

$$\forall\ A \subseteq \Omega, \left(m_1 \cup m_2\right)\left(A\right) = \sum_{B \cup C = A} m_1\left(B\right) \cdot m_2\left(C\right). \tag{28}$$

Note that (28) is (22) where $\cap$ is changed into $\cup$. While the intersection of two bodies of evidence only keeps the items of information asserted by both sources, the union does not reject anything. Especially if $m_1(\varnothing) = m_2(\varnothing) = 0$, it is easy to check that $(m_1 \cup m_2)(\varnothing) = 0$, i.e. the union does not generate any conflict, and the normalization step (23) is useless here. The resulting set of focal elements is indeed $\mathcal{F}_{1\cup 2} = \{A \cup B \mid A \in \mathcal{F}_1, A \in \mathcal{F}_2\}$.

The union of two bodies of evidence is more easily performed via the credibility measure since:

Proposition 2. *Let* $\mathrm{Cr}_1 \cup \mathrm{Cr}_2$ *be the credibility measure associated with* $m_1 \cup m_2$*. Then* $\forall\ A \subseteq \Omega, (\mathrm{Cr}_1 \cup \mathrm{Cr}_2)(A) = \mathrm{Cr}_1(A) \cdot \mathrm{Cr}_2(A)$.

[1] During the course of the investigation whose results are reported here, we became aware of similar attempts by Yager[32] and Oblow[20].

Proof.

$$\begin{aligned}(\mathrm{Cr}_1 \cup \mathrm{Cr}_2)(A) &= \sum_{\varnothing \neq B \cup C \subseteq A} m_1(B) \cdot m_2(C) \\ &= \sum_{\varnothing \neq B \subseteq A} m_1(B) \left(\sum_{\varnothing \neq C \subseteq A} m_2(C) \right). \quad \text{Q.E.D.}\end{aligned}$$

Notice that in the case of intersection of bodies of evidence, the counterpart of Proposition 2 holds for the commonality numbers only, since, as noted by Shafer[24], (22) implies

$$(Q_1 \cap Q_2)(A) = Q_1(A) \cdot Q_2(A). \tag{29}$$

The notion of conjunctive and disjunctive knowledge can shed light on these properties, if we recall, following Yager[31], that in the presence of conjunctive information, "$x = A$ or $x = B$" translates into $x = A \cap B$ and "$x = A$ and $x = B$" translates into "$x = A \cup B$".

Example 2. John stayed in Paris from 1980 till 1982 and from 1982 till 1984 is equivalent to "John stayed in Paris from 1980 till 1984".

But if we happen to know from two sources that he stayed in Paris from 1980 till 1983 *or* from 1981 till 1984, then the only sure resulting item of information is that he stayed in Paris from 1981 till 1983.

Now, if we remember that the commonality numbers play, for a conjunctive body of evidence, the same role as the credibility degrees in a disjunctive body of evidence, it is clear that (29) is the mirror image of Proposition 2, and achieves an "or" of two conjunctive bodies of evidence.

As a consequence of (29) and Proposition 2, the union and intersection of bodies of evidence are commutative and associative. If we denote by Ω (resp.: $\varnothing$) the body of evidence such that $m(\Omega) = 1$ (resp.: $m(\varnothing) = 1$), that is, total ignorance (resp.: the null value "not applicable") for variable x in the disjunctive interpretation, we have

$$\forall\, m, m \cap \Omega = m; \qquad m \cup \varnothing = m. \tag{30}$$

Now, applying (22) and (28) on subsets A of Ω, i.e. $m(A) = 1$, we recover the usual set-intersection and union in 2^{Ω}. But these operations are not idempotent on $\mathcal{B}(\Omega)$. Indeed, Proposition 2 and (29) lead to

$$\forall\, \omega \in \Omega, (\mathrm{Pl}_1 \cap \mathrm{Pl}_2)(\{\omega\}) = \mathrm{Pl}_1(\{\omega\}) \cdot \mathrm{Pl}_2(\{\omega\}) \tag{31}$$

$$(\mathrm{Pl}_1 \cup \mathrm{Pl}_2)(\{\omega\}) = \mathrm{Pl}_1(\{\omega\}) + \mathrm{Pl}_2(\{\omega\}) - \mathrm{Pl}_1(\{\omega\}) \cdot \mathrm{Pl}_2(\{\omega\}). \tag{32}$$

Note that the intersection and the union in $\mathcal{B}(\Omega)$ are not stable on the subset of possibility measures. This is because if $\mathcal{F}_1$ and $\mathcal{F}_2$ are consonant, then generally neither $\mathcal{F}_{1\cap 2}$ nor $\mathcal{F}_{1\cup 2}$ are. But (32) as (31), correspond to well known fuzzy set-theoretic operations. The set of probability measures is not closed under the union operation, since $P_1 \cup P_2$ corresponds to a set of focal elements some of which are 2-element sets. Strictly speaking, the closure property does not hold for intersection since the intersection of two probability measures is no longer normalized (the intersection of singletons is generally empty!). The closure property is recovered through normalization (23) i.e. using Dempster rule as a whole.

Lastly the union of two consistent bodies of evidence is consistent while their intersection may no longer be so.

2.2 Complement of a Body of Evidence

The complement of a body of evidence $(\mathcal{F}, m)$ is $(\neg\mathcal{F}, \bar{m})$ defined by

$$\forall\ A \subseteq \Omega, \bar{m}\left(A\right) = m\left(\bar{A}\right) \tag{33}$$

so that $\neg\mathcal{F} = \{\bar{A} | A \in \mathcal{F}\}$. This complementation is formally involutive. Moreover the union and intersection satisfy De-Morgan laws since

$$\begin{aligned}\forall\ \overline{(m_1 \cup m_2)}\left(A\right) = \left(m_1 \cup m_2\right)\left(\bar{A}\right) &= \sum_{B \cup C = \bar{A}} m_1\left(B\right) \cdot m_2\left(C\right) \\ &= \sum_{\bar{B} \cap \bar{C} = A} \bar{m}_1\left(\bar{B}\right) \cdot \bar{m}_2\left(\bar{C}\right) = \left(\bar{m}_1 \cap \bar{m}_2\right)\left(A\right).\end{aligned}$$

It is easy to see that (33) reduces to usual set complementation when $m(A) = 1$. Moreover if $\mathcal{F}$ is consonant, then $\neg\mathcal{F}$ is also consonant, so that (33) also reduces to fuzzy set complementation when applied to a possibility measure, i.e. the set of possibility measures is closed under complementation. But the set of probability measures is not for $|\Omega| > 2$ since all focal elements in $\neg\mathcal{F}$ then contain $|\Omega| - 1$ elements.

$\mathcal{B}(\Omega)$ is *not* a Boolean algebra. Indeed union and intersection are not idempotent. Moreover the laws of contradiction and excluded middle are not valid, i.e.

$$\text{for } m \in \mathcal{B}\left(\Omega\right) - 2^{\Omega}, \qquad \text{generally} \quad m \cap \bar{m} \neq \varnothing; \qquad m \cup \bar{m} \neq \Omega.$$

Actually it can be checked that $(m \cap \bar{m})(\varnothing) > 0, (m \cup \bar{m})(\Omega) > 0$ which expresses that these laws *somewhat* hold. If $\mathcal{F} = \{A, B\}$ then $\neg\mathcal{F} = \{\bar{A}, \bar{B}\}$ and $(m \cap \bar{m})(A \cap \bar{B}) > 0, (m \cup \bar{m})(A \cup \bar{B}) > 0$, etc....

Hence $\mathcal{B}(\Omega)$ has the same algebraic structure as the set of fuzzy subsets of Ω, $[0,1]^{\Omega}$, under the product, probabilistic sum, and usual complementation of fuzzy sets, i.e.

$$\begin{aligned}\mu_{F \cap G}\left(\omega\right) &= \mu_F\left(\omega\right) \cdot \mu_G\left(\omega\right) \\ \mu_{F \cup G}\left(\omega\right) &= \mu_F\left(\omega\right) + \mu_G\left(\omega\right) - \mu_F\left(\omega\right)\mu_G\left(\omega\right) \\ \mu_{\bar{F}}\left(\omega\right) &= 1 - \mu_F\left(\omega\right).\end{aligned}$$

Moreover these fuzzy set-theoretic operations are consistent with set-theoretic operations in $\mathcal{B}(\Omega)$, under independence assumption, up to stability of $\pi(\Omega)$.

An interesting feature of complementation in $\mathcal{B}(\Omega)$ is that it turns a disjunctive body of evidence into a conjunctive one. To see it consider the simple case $\mathcal{F} = \{A\}$, and A restricts the possible values of x. Then $\bar{A}$ is a set of values which are forbidden for x. Let $\bar{x}$ be the variable which takes values which x does not take. It is clear that $\overline{A} \subseteq \overline{x}$ ($\overline{x}$ takes at least all values in $\overline{A}$) is equivalent to x is A (the value of x is restricted by A). In the general case, the same transformation occurs, and any focal element $\bar{A} \in \neg F$ is a set of values which x certainly does not take (with weight $m(A)$). This transformation in the nature of evidence provides some explanation of the following property.

Proposition 3. *Let $\bar{Q}$ be the commonality function associated with the complement $(\neg\mathcal{F}, \bar{m})$ of a disjunctive body of evidence $(\mathcal{F}, m)$. Then*

$$\forall\ A, \mathrm{Cr}\left(A\right) = \bar{Q}\left(\bar{A}\right) - m\left(\varnothing\right).$$

Proof.

$$\sum_{\varnothing \neq B \subseteq A} m\left(B\right) = \sum_{\bar{A} \subseteq \bar{B} \neq \Omega} \bar{m}\left(\bar{B}\right) = \bar{Q}\left(\bar{A}\right) - \bar{m}\left(\Omega\right). \qquad \text{Q.E.D.}$$

This result stresses that Cr and Q play the same role in each type of knowledge, disjunctive and conjunctive respectively.

An important remark is that, reciprocally, the complement of a conjunctive body of evidence is *not* a disjunctive body of evidence in the sense defined in this paper. To see it, consider the case of the conjunctive statement '$A \subseteq x$', then, defining $\bar{x}$ as above, all we know about $\bar{x}$ is that any subset of $\bar{A}$ is a possible conjunctive set of values for $\bar{x}$, so that the knowledge about $\bar{x}$ is a possibility distribution on $2^{\bar{A}}$, say π, such that $\forall\ B \subseteq \bar{A}$, $\pi(B) = 1$ means B is a possible set of values for $\bar{x}$ (i.e. x possibly does not take any value in B). Hence π defines disjunctive knowledge *over* $2^{\bar{A}}$. The usual disjunctive information is recovered as a particular case, setting $\pi(B) = 1$ if and only if B is a singleton in $\bar{A}$ and 0 otherwise. This type of higher-order disjunctive information is not a mere game of the mind; it is often encountered in data-bases with multiple-valued attributes, when one wishes to represent the possible sets of tongues spoken by an individual, for instance (see Prade and Testemale[21]).

These remarks weaken the apparent strength of the involution property of the complementation operation in $\mathcal{B}(\Omega)$.

2.3 Inclusions

Concepts of inclusion can also be introduced on $\mathcal{B}(\Omega)$. Given a normal body of evidence $(\mathcal{F}, m)$, the interval $[\mathrm{Cr}(A),\ \mathrm{Pl}(A)]$ can be viewed as the range of the probability of A induced by the lack of precision of the focal elements (see *1.2.*). In other words, the body of evidence $\mathcal{F}$ defines a (convex) set of probability measures on Ω, say $\mathcal{C}(\mathcal{F})$.

A normal body of evidence $(\mathcal{F}, m)$ can be viewed as included in $(\mathcal{F}', m')$ as soon as $\mathcal{C}(\mathcal{F}) \subseteq \mathcal{C}(\mathcal{F}')$. In terms of the plausibility and credibility measures (Pl, Cr) and $(\mathrm{Pl}', \mathrm{Cr}')$, this is equivalent to:

$$\forall\ A \in \Omega, [\mathrm{Cr}\,(A)\,, \mathrm{Pl}\,(A)] \subseteq \left[\mathrm{Cr}'\,(A)\,, \mathrm{Pl}'\,(A)\right]. \tag{34}$$

We shall write $(\mathcal{F}, m) \subseteq (\mathcal{F}', m')$ when (34) holds true. (34) reduces to (11) when Cr = Pl = a probability measure. Note that because $\mathrm{Cr}(A) = 1 - \mathrm{Pl}(\bar{A})$, any of the following inequalities is equivalent to (34):

$$\mathrm{Cr}\,(A) \geqq \mathrm{Cr}'\,(A)\,, \forall\ A \in \Omega, \tag{35}$$

$$\mathrm{Pl}\,(A) \leqq \mathrm{Pl}'\,(A) \forall\ A \in \Omega. \tag{36}$$

Disjunctive and Conjunctive Inclusions

The definition of inclusion can be extended from $\mathcal{B}^{+}(\Omega)$ to $\mathcal{B}(\Omega)$, taking (36) as the actual definition. Note that (35) is not equivalent to (36) for bodies of evidence which are not normal. Indeed, in the general case, (36) is equivalent to:

$$\mathrm{Cr}\,(A) + m\,(\varnothing) \geqq \mathrm{Cr}'\,(A) + m'\,(\varnothing)\,, \forall\ A \in \Omega$$

due to the definitions of Cr and Pl. (36) induces some relationships between the respective contents of $\mathcal{F}$ and $\mathcal{F}'$ such that $(\mathcal{F}, m) \subseteq (\mathcal{F}', m')$. In the following the *core* (resp.: *support*) of $(\mathcal{F}, m)$ is the intersection (resp.: union) of focal elements and denoted $C(\mathcal{F})$ (resp.: $S(\mathcal{F})$). The following necessary condition for inclusion relationship is noticeable:

Proposition 4. *If,* $(\mathcal{F}, m) \subseteq (\mathcal{F}', m')$ *then*

i) $S(\mathcal{F}) \subseteq S(\mathcal{F}')$; $C(\mathcal{F}) \subseteq C(\mathcal{F}')$,
ii) $\forall\ A' \subseteq \mathcal{F}',\ \exists A \in \mathcal{F},\ A \subseteq A'$.

Proof. $\forall\ \omega,\ \mathrm{Pl}(\{\omega\}) = 1$ if and only if $\omega \in C(\mathcal{F})$. From (36) if $\omega \in C(\mathcal{F})$ then $\mathrm{Pl}(\{\omega\}) = 1 = \mathrm{Pl}'(\{\omega\})$; hence $\omega \in C(\mathcal{F}')$. Besides $\forall\ \omega,\ \mathrm{Pl}(\{\omega\}) > 0 \Leftrightarrow \omega \in S(\mathcal{F})$. From (36) if $\omega \in S(\mathcal{F})$ then $0 < \mathrm{Pl}(\{\omega\}) \leqq \mathrm{Pl}'(\{\omega\})$; hence $\omega \in S(\mathcal{F}')$. To prove (ii), let $A' \in \mathcal{F}'$ contain no focal element in $\mathcal{F}$. Then

$$\mathrm{Pl}\left(\bar{A}'\right) = 1 > 1 - m'\,(A') \geqq \mathrm{Pl}'\left(\bar{A}'\right)$$

which contradicts (36). Q.E.D.

Conditions on the relative structure of $(\mathcal{F}, m)$ and $(\mathcal{F}', m')$ which would be necessary and sufficient to ensure $(\mathcal{F}, m) \subseteq (\mathcal{F}', m')$ seem to be difficult to produce. Inclusion $\subseteq$ has natural properties such as transitivity, and mutual inclusion implies equality (since Cr determines m). Notice also that

$$(\mathcal{F},m)\cap(\mathcal{F}',m')\subseteq(\mathcal{F},m)\subseteq(\mathcal{F},m)\cup(\mathcal{F}',m')\,. \tag{37}$$

For instance

$$\begin{aligned}\left(\mathrm{Pl}\cap\mathrm{Pl}'\right)(A) &= \sum_{B\cap B'\cap A\neq\varnothing} m\,(B)\cdot m'\,(B') \\ &= \sum_{B\cap A\neq\varnothing} m\,(B)\cdot\left(\sum_{B\cap B'\cap A\neq\varnothing} m'\,(B')\right)\leqq \mathrm{Pl}\,(A)\,.\end{aligned}$$

The other inclusion in (37) can be obtained in a similar way.

If $(\mathcal{F},m)$ and $(\mathcal{F}',m')$ both generate possibility measures with possibility distributions $\pi=\mu_F$ and $\pi'=\mu_{F'}$, then

$$(\mathcal{F},m)\subseteq(\mathcal{F}',m')\Leftrightarrow F\subseteq F'\quad(\text{i.e. }\mu_F\leqq{\mu_F}'). \tag{38}$$

That is, the inclusion of bodies of evidence is completely consistent with Zadeh's[35] inclusion of fuzzy sets, hence with the usual inclusion in 2^Ω. To see it just notice that if $(\mathcal{F},m)\subseteq(\mathcal{F}',m')$ then, as a particular case of (36), $\mathrm{Pl}(\{\omega\})=\mu_F(\omega)\leqq\mathrm{Pl}'(\{\omega\})=\mu_{F'}(\omega)$. Conversely if $F\subseteq F'$ then $\mathrm{Pl}(A)=\max\{\mu_F(\omega)|\omega\in A\}\leqq\mathrm{Pl}'(A)=\max\{\mu_{F'}(\omega)\mid\omega\in A\}$.

More surprising, and a disquieting fact at first glance, is that the complementation introduced in *2.2.* is *not* order-reversing for $\subseteq$. To see it first notice that due to Proposition 3

$$\forall\ A,\mathrm{Pl}\,(A)\leqq\mathrm{Pl}'\,(A)\Leftrightarrow\forall\ A,\bar{Q}\,(A)\geqq\bar{Q}'\,(A) \tag{39}$$

where $\bar{Q}$ and $\bar{Q}'$ are the commonality functions of the complementary bodies of evidence $(^\neg\mathcal{F},\bar{m})$, $(^\neg\mathcal{F}',\bar{m}')$ respectively. Moreover $(\mathcal{F},m)\subseteq(\mathcal{F}',m')$ does not imply any inequality between Q and Q', as proved by the following:

Counter-example 1 Let

$$\Omega=\{a,b,c\}\,,0<k<\frac{1}{2}.$$
$$\mathcal{F}=\{\{a\}\,,\Omega\}\,;m\,(\{a\})=1-k,m\,(\Omega)=k.$$
$$\mathcal{F}'=\{\{a,b\}\,,\{a,c\}\}\,;m'\,(\{a,b\})=k,m'\,(\{a,c\})=1-k.$$

Then the reader can check that $(\mathcal{F},m)\subseteq(\mathcal{F}',m')$; especially $\forall\ A\neq\{b\},\{c\},\varnothing$, $\mathrm{Pl}'(A)=1$ and $\mathrm{Pl}(\{b\})=\mathrm{Pl}'(\{b\})=k$, $\mathrm{Pl}(\{c\})=k<\mathrm{Pl}'(\{c\})=1-k$. But $Q(\{c\})=k<Q'(\{c\})=1-k$, $Q(\{b,c\})=k>Q'(\{b,c\})=0$. Q.E.D.

This lack of order-reversingness should not hurt our intuition because (36) is meaningul only for disjunctive evidence, but $(^\neg\mathcal{F},\bar{m})$ is conjunctive and the grade of credibility of A deduced from $(^\neg\mathcal{F},\bar{m})$ is $\bar{Q}(A)$. But from

(39) $(\neg\mathcal{F}', \bar{m}')$ is contained in $(\neg\mathcal{F}, \bar{m})$ (remember that Q is a decreasing set-function for set-inclusion), in the sense of a new kind of inclusion, which makes sense only for conjunctive evidence, namely $\overline{\subset}$ such that $(\mathcal{F}, m)\overline{\subset}(\mathcal{F}', m')$ if and only if

$$\forall\ A, Q\left(A\right) \leqq Q'\left(A\right). \tag{40}$$

$\overline{\subset}$ can be called ‘conjunctive inclusion’ while $\subseteq$ is called ‘disjunctive inclusion’, respectively abbreviated as c-inclusion and d-inclusion.

Note that c-inclusion is transitive, that mutual c-inclusion means equality (since Q determines m as well). Moreover

$$(\mathcal{F}, m) \cap (\mathcal{F}', m')\,\overline{\subset}\,(\mathcal{F}, m)\,\overline{\subset}\,(\mathcal{F}, m) \cup (\mathcal{F}', m') \tag{41}$$

which is simply (37) transformed by complementation. Similarly, c-inclusion applied to possibility measures is equivalent to Zadeh's[35] inclusion of fuzzy sets, i.e. a counterpart of (38) holds. A necessary condition to get (40) is given now:

Proposition 5. *If* $(\mathcal{F}, m)\overline{\subset}(\mathcal{F}', m')$ *then*

i) $S(\mathcal{F}) \subseteq S(\mathcal{F}'),\ C(\mathcal{F}) \subseteq C(\mathcal{F}')$,
ii) $\forall\ A \in \mathcal{F}, \exists\ A' \in \mathcal{F}',\ A \subseteq A'$.

Proof. (i) is easily seen due to $S(\neg\mathcal{F}) = \overline{C(\mathcal{F})},\ C(\neg\mathcal{F}) = \overline{S(\mathcal{F})}$ using complementation to turn $\overline{\subset}$ into $\overline{\supset}$. Now let $A \in \mathcal{F}$ be contained in no focal element in $\mathcal{F}'$ then

$$Q(A) \geqq m\left(A\right) > 0 = Q'\left(A\right)$$

which contradicts (40). Q.E.D.

At this point it is natural to define a third concept of inclusion which requires both (36) and (40) to hold:

Definition 1. $(\mathcal{F}, m)$ *is said to be included in* $(\mathcal{F}', m')$*, denoted* $(\mathcal{F}, m) \Subset (\mathcal{F}', m')$ *if and only if* $(\mathcal{F}, m)$ *is both c-included and d-included in* $(\mathcal{F}', m')$*.*

Inclusion is transitive, mutual inclusion is equality, (37) and (38) hold for $\Subset$. (Note that c-inclusion and d-inclusion are already equivalent for possibility measures). Moreover the complementation is order-reversing for $\Subset$.

Strong Inclusion

Yager[33] has introduced a fourth definition of inclusion in $\mathcal{B}(\Omega)$, which, for reasons to be clarified below, can be called *strong inclusion*, and will be denoted $\subset\!\!\!\subset\!\!\!\subset$. This concept can be presented as follows.

Definition 2. $(\mathcal{F}, m) \subset\!\!\!\subset\!\!\!\subset (\mathcal{F}', m')$ *if and only if the three following statements are valid:*

i) $\forall\ A_i \in \mathcal{F},\ \exists\ A'_j \in \mathcal{F}',\ A_i \subseteq A'_j$,
ii) $\forall\ A'_j \in \mathcal{F}',\ \exists\ A_i \in \mathcal{F},\ A_i \subseteq A'_j$,

iii) there exists a matrix W *with size* $m \times n$, $m = |\mathcal{F}|$, $n = |\mathcal{F}'|$, *whose entries are* $W_{ij} \in [0,1]$ *such that* $W_{ij} > 0 \Rightarrow A_i \subseteq A'_j$, $\sum_{ij} W_{ij} = 1$ *and the basic assignments* m *and* m' *can be expressed in terms of the* W_{ij}*'s as follows:*

$$\forall\ A_i \in \mathcal{F}, m\left(A_i\right) = \sum_{\substack{j \\ A_i \subseteq A'_j}} W_{ij}, \tag{42}$$

$$\forall\ A'_j \in \mathcal{F}', m'\left(A'_j\right) = \sum_{\substack{i \\ A_i \subseteq A'_j}} W_{ij}. \tag{43}$$

Note that (42) and (43) look like flow conservation equations in a flow network (Ford and Fulkerson[11]). This analogy is explained in the appendix and is useful to make Definition 2 work. The name 'strong inclusion' is justified by the following result:

Proposition 6. *Strong inclusion implies inclusion i.e.*

$$(\mathcal{F}, m) \subset\!\!\!\subset\!\!\!\subset (\mathcal{F}', m') \Rightarrow \forall\ A, Q\left(A\right) \leqq Q'\left(A\right),\ \mathrm{Pl}\left(A\right) \leqq \mathrm{Pl}'\left(A\right).$$

The converse does not hold.

Proof. Assume $(\mathcal{F}, m) \subset\!\!\!\subset\!\!\!\subset (\mathcal{F}', m')$.

$$\mathrm{Pl}'\left(B\right) = \sum_{A'_j \cap B \neq \varnothing} m'\left(A'_j\right) = \sum_{i,j} \left\{W_{ij} | A_i \subseteq A'_j; A'_j \cap B \neq \varnothing\right\}$$

but

$$\left\{\left(i,j\right) | A_i \subseteq A'_j; A'_j \cap B \neq \varnothing\right\} \supseteq \left\{\left(i,j\right) | A_i \subseteq A'_j; A_i \cap B \neq \varnothing\right\}$$

hence

$$\mathrm{Pl}'\left(B\right) \geqq \mathrm{Pl}\left(B\right) = \sum_{i,j} \left\{W_{ij} | A_i \subseteq A'_j; A_i \cap B \neq \varnothing\right\}.$$

A similar proof holds for the commonality function. Q.E.D.

That the converse does not hold is indicated by the following:

Counter-example 2 $\Omega = \{a, b, c, d, e\}$. Consider the two normal bodies of evidence:

$$(\mathcal{F}, m) = \left(\{a,b\}, 0.3\right); \left(\{a,c\}, 0.3\right); \left(\{c,d\}, 0.3\right); \left(\{e\}, 0.1\right)$$
$$(\mathcal{F}', m') = \left(\{a,b,c\}, 0.4\right); \left(\{a,b,d\}, 0.3\right); \left(\{a,c,d\}, 0.2\right); \left(\{c,d,e\}, 0.1\right).$$

To check that $(\mathcal{F}, m) \subseteq (\mathcal{F}', m')$ see on Table 1. To see that no matrix W satisfying (42)–(43) exists, it is enough to verify that the following system of equations has no solution in [0,1]:

$$\left[\begin{array}{l} 0.4 = 0.3m_{11} + 0.3m_{21}\,(= m'\left(\{a,b,c\}\right)) \\ 0.3 = 0.3\,(1 - m_{11})\,(= m'\left(\{a,b,d\}\right) \\ 0.2 = 0.3\,(1 - m_{21}) + 0.3m_{33}\,(= m'\left(\{a,c,d\}\right)) \\ 0.1 = 0.1 + 0.3\,(1 - m_{33})\,(= m'\left(\{c,d,e\}\right)) \,. \end{array}\right.$$

This system is equivalent to (42)–(43) where the W_{ij}'s have been changed into $W_{ij} = m(A_i)m_{ij}$, with $\sum_j m_{ij} = 1$, which eliminates (42). Deeper

Table 1. Counter-example 2

Events	$\mathrm{Cr}(A)$	$\mathrm{Cr}'(A)$	$Q(A)$	$Q'(A)$
$\{a\}$	0	0	0.6	0.9
$\{b\}$	0	0	0.3	0.7
$\{c\}$	0	0	0.6	0.7
$\{d\}$	0	0	0.3	0.6
$\{e\}$	0.1	0.	0.1	0.1
$\{a, b\}$	0.3	0	0.3	0.7
$\{a, c\}$	0.3	0	0.3	0.6
$\{a, d\}$	0	0	0	0.5
$\{a, e\}$	0.1	0	0	0
$\{b, c\}$	0	0	0	0.4
$\{b, d\}$	0.	0	0	0.3
$\{b, e\}$	0.1	0	0	0
$\{c, d\}$	0.3	0	0.3	0.3
$\{c, e\}$	0.1	0	0	0.1
$\{d, e\}$	0.1	0	0	0.1
$\{a,\ b,\ c\}$	0.6	0.4	0	0.4
$\{a,\ b,\ d\}$	0.3	0.3	0	0.3
$\{a,\ b,\ e\}$	0.4	0	0	0
$\{a,\ c,\ d\}$	0.6	0.2	0	0.2
$\{a,\ c,\ e\}$	0.4	0	0	0
$\{a,\ d,\ e\}$	0.1	0	0	0
$\{b,\ c,\ d\}$	0.3	0	0	0
$\{b,\ c,\ e\}$	0.1	0	0	0
$\{b,\ d,\ e\}$	0.1	0.	0	0
$\{c,\ d,\ e\}$	0.4	0.1	0	0.1
$\{a,\ b,\ c,\ d\}$	0.9	0.9	0	0
$\{a,\ b,\ c,\ e\}$	0.7	0.4	0	0
$\{a,\ b,\ d,\ e\}$	0.4	0.3	0	0
$\{a,\ c,\ d,\ e\}$	0.7	0.3	0	0
$\{b,\ c,\ d,\ e\}$	0.4	0.1	0	0

(The use of Cr or Pl to check the inclusion is indifferent because the bodies of evidence are normal).

understanding about the reasons why this system has no solution is gained in the appendix. Q.E.D.

The nice feature of Definition 2 is that it provides a construction method to build two bodies of evidence $(\mathcal{F}, m)$ and $(\mathcal{F}', m')$ such that one is strongly included in the other. It may act as a sufficient condition for having inclusion in the sense of Definition 1. Namely note that letting

$$\mathcal{F}'(A) = \{A' \in \mathcal{F}', A \subseteq A'\}; \qquad \mathcal{F}(A') = \{A \in \mathcal{F}, A \subseteq A'\}$$

then

$$\mathcal{F} = \bigcup_{A' \in \mathcal{F}'} \mathcal{F}(A'); \qquad \mathcal{F}' = \bigcup_{A \in \mathcal{F}} \mathcal{F}'(A). \tag{44}$$

Given $(\mathcal{F}', m')$, *all* bodies of evidence $(\mathcal{F}, m) \subset\!\!\!\subset\!\!\!\subset (\mathcal{F}', m')$ can be obtained by the following procedure:

Procedure a $\forall\ A'_j \in \mathcal{F}'$ dispatch the weight $m'(A'_j)$ among any family $\mathcal{F}(A'_j)$ of subsets of A'_j, letting W_{ij} be the share of $m'(A'_j)$ allocated to $A_i \in \mathcal{F}(A'_j)$.

Define $\mathcal{F}$ and m by (44) and (42) respectively.

Similarly, given $(\mathcal{F}, m)$ all bodies of evidence $(\mathcal{F}, m) \subset\!\!\!\subset\!\!\!\subset (\mathcal{F}', m')$ can be obtained by the dual procedure.

Procedure b $\forall\ A_i \in \mathcal{F}$, dispatch the weight $m(A_i)$ among any family $\mathcal{F}'(A_i)$ of supersets of $A_i(\mathcal{F}'(A_i) \subseteq \{A | A_i \subseteq A\})$ letting W_{ij} be the share of $m(A_i)$ allocated to $A'_j \in \mathcal{F}'(A_i)$.

Define $\mathcal{F}'$ and m' by (44) and (43) respectively.

Note that a particular case of Procedure (a) is obtained by forcing $\mathcal{F}(A'_j)$ to contain only singletons (provided that $(\mathcal{F}', m')$ is normal). We then recover Dempster's[1] procedure to generate the set $\mathcal{C}(\mathcal{F}')$ of probability measures satisfying (11) as recalled in *1.2*. Procedure (a) thus generalizes Dempster's procedure, but cannot produce *all* bodies of evidence $(\mathcal{F}, m)$ satisfying (34), as indicated in Proposition 6. Procedure (b) was first suggested by Yager[33] who gives it as the very definition of inclusion in $\mathcal{B}(\Omega)$.

Inclusion $\subset\!\!\!\subset\!\!\!\subset$ is transitive. To see it, rewrite (42), (43) under the form

$$\forall\ A'_j \in \mathcal{F}', m'(A'_j) = \sum_{i=1}^{m} m(A_i) \cdot m_{ij}$$

as done in the proof of Proposition 6. Let M be the matrix with coefficient m_{ij}, $\mathbf{m}$ and $\mathbf{m}'$ be the column vectors expressing the basic assignments. Then using matrix notation:

$$\mathbf{m}' = M\mathbf{m}. \tag{45}$$

Now $(\mathcal{F}, m) \subset\!\!\!\subset\!\!\!\subset (\mathcal{F}', m')$ and $(\mathcal{F}', m') \subset\!\!\!\subset\!\!\!\subset (\mathcal{F}'', m'')$ translate into $\mathbf{m}' = M\mathbf{m}, \mathbf{m}'' = M'\mathbf{m}'$, where M, M' belong to the class $\mathcal{M}$ of Markovian matrices, i.e. with positive entries summing to 1 on each row. Hence $\mathbf{m}'' = M'M\mathbf{m}$, and thus $(\mathcal{F}, m) \subset\!\!\!\subset\!\!\!\subset (\mathcal{F}'', m'')$ since $\mathcal{M}$ is closed under matrix product, and (44) holds between $\mathcal{F}$ and $\mathcal{F}''$ as is straightforwardly checked.

Of course, mutual strong inclusion of two bodies of evidence means their equality. Strong inclusion applied to possibility measures is consistent with Zadeh's inclusion of fuzzy set:

Proposition 7. *If $(\mathcal{F}, m)$ and $(\mathcal{F}', m')$ are consonant then*

$$(\mathcal{F}, m) \subset\!\!\!\subset\!\!\!\subset (\mathcal{F}', m') \textit{ if and only if } \mu_F \leqq \mu_{F'}$$

where μ_F and $\mu_{F'}$ are the contour functions of $(\mathcal{F}, m)$ and $(\mathcal{F}', m')$.

Proof. The difficult part is to prove that Zadeh's inclusion of fuzzy sets implies the existence of a matrix W satisfying (42) and (43). The proof is given through network flow theory arguments in the appendix, which gives a constructive procedure to build W. Q.E.D.

Lastly $\subset\!\!\!\subset\!\!\!\subset$ is order-reversing in $\mathcal{B}(\Omega)$ since Procedures (a) and (b) exchange via complementation. Inequalities (37) hold for the strong inclusion. Note that (i) and (ii) of Definition 2 hold between $\mathcal{F}$ and $\mathcal{F}'' = \{A \cup B' | A \in \mathcal{F}, B' \in \mathcal{F}'\}$. Moreover define $W_{ij} = m(A_i) \cdot m'(B'_j)$ as the share of $m(A_i)$ allocated to the focal element $A_i \cup B'_j$.

Properties of $\mathcal{B}(\Omega)$ under Inclusions

Any of the introduced inclusions equips $\mathcal{B}(\Omega)$ with a partial ordering structure (reflexive, transitive and weakly antisymmetric, that is xRy and yRx implies $x = y$). $\subset\!\!\!\subset\!\!\!\subset$ is able to compare less elements in $\mathcal{B}(\Omega)$ than $\subset\!\!\!\subset$, which in turn is able to compare less elements in $\mathcal{B}(\Omega)$ than any of $\subseteq$ and $\overline{\supset}$. However on $\pi(\Omega) = [0,1]^\Omega$, the set of possibility measures (or fuzzy sets), all four inclusions collapse into Zadeh's fuzzy set inclusion.

The greatest element in $\mathcal{B}(\Omega)$ in the sense of any inclusion is the total ignorance function ($m(\Omega) = 1$) and the least element is the empty body of evidence ($m(\varnothing) = 1$). The least elements in $\mathcal{B}^+(\Omega)$, i.e. normal bodies of evidence, are the probability measures. This is in the sense of disjunctive inclusion $\subseteq$. Indeed, because $\sum_\omega P(\{\omega\}) = 1$, probability measures are not comparable using (34) or (35). This is consistent with the idea that probability measures are sort of 'fuzzy points' (Höhle[16]) for which inclusion is meaningless (there is equality or disjointness!). Moreover given a normal body of evidence $(\mathcal{F}, m)$ any probability measure in $\mathcal{C}(\mathcal{F})$ is contained in $(\mathcal{F}, m)$ in the sense of disjunctive inclusion.

Probability measures are always interpreted in the disjunctive information framework (an event A occurs if and only if $\exists\, \omega \in A$ which is observed, and not only if *all* $\omega \in A$ are observed at the same time). Hence the commonality function Q is not interesting for probabilistic bodies of evidence ($Q(A) = 0$ as soon as $|A| > 1$). Hence probability measures have no interesting role in $(\mathcal{B}^+(\Omega), \overline{\subset})$. However they are still the least elements in $(\mathcal{B}^+(\Omega), \subset\!\!\!\subset\!\!\!\subset)$, because any probability measure in $\mathcal{C}(\mathcal{F})$ is strongly included in $(\mathcal{F}, m)$, from Dempster's[1] construction.

Lastly there is an interesting convexity property related to the inclusions:

Proposition 8. *The following subsets of* $\mathcal{B}(\Omega)$ *are convex:*

$$\{(\mathcal{F},m)|(\mathcal{F},m)R(\mathcal{F}',m')\}$$
$$\{(\mathcal{F}',m')|(\mathcal{F},m)R(\mathcal{F}',m')\}$$

with $R = \subseteq,\ \overline{\subset},\ \subset\!\subset,\ \subset\!\subset\!\subset$.

Proof. Using the definition of the convex combination of two bodies of evidence $(\mathcal{F},m)$ and $(\mathcal{G},n)$ i.e. $\alpha(\mathcal{F},m)+(1-\alpha)\ (\mathcal{G},n) = (\mathcal{F}\cup\mathcal{G}, \alpha m+(1-\alpha)n)$ with credibility measure $\mathrm{Cr} = \alpha\mathrm{Cr}_m + (1-\alpha)\mathrm{Cr}_n$, it is obvious that Proposition 8 holds for $R = \subseteq$. Now $Q = \alpha Q_m + (1-\alpha)Q_n$ as well, so that Proposition 8 holds for $R = \overline{\subset}$ and $\subset\!\subset$. Lastly if $(\mathcal{F},m)$ and $(\mathcal{G},n)$ are strongly included in $(\mathcal{F}',m')$ then conditions (i) and (ii) in Definition 2 hold for $\mathcal{F}\cup\mathcal{G}$ with respect to $\mathcal{F}'$. Moreover $\mathbf{m} = M\mathbf{m}'$ and $\mathbf{n} = N\mathbf{m}'$ implies $\alpha\mathbf{m} + (1-\alpha)\mathbf{n} = (\alpha M + (1-\alpha)N)\mathbf{m}'$ where $\alpha M + (1-\alpha)N$ is still a Markovian matrix consistent with the conditions (i) and (ii) in Definition 2. Hence Proposition 8 holds for $\subset\!\subset\!\subset$. Q.E.D.

2.4 Projections and Cartesian Product

In this section only normal bodies of evidence are considered.

Let $(\mathcal{F},m)$ be a body of evidence on a Cartesian product $\Omega = U\times V$. If S is a subset of Ω its projection on U (resp.: V) is denoted $U(S)$ (resp.: $V(S)$) and defined by

$$U(S) = \{u\in U|\exists\ v\in V, (u,v)\in S\}.$$

More generally the projection of $(\mathcal{F},m)$ on U is $(\mathcal{F}_U, m_U)$ such that (Shafer[25])

$$\forall\ A\subseteq U, m_U(A) = \sum_{S:A=U(S)} m(S). \tag{46}$$

It is easy to check that $(\mathcal{F}_U, m_U)$ induces a plausibility measure Pl_U on U such that $\mathrm{Pl}_U(A) = \mathrm{Pl}(A\times V)$, and a credibility measure Cr_U such that $\mathrm{Cr}_U(A) = \mathrm{Cr}(A\times V)$, which sounds consistent.

Proof.

$$\mathrm{Pl}(A\times V) = \sum_{(A\times V)\cap S\neq\varnothing} m(S) = \sum_{A\cap U(S)\neq\varnothing} m(S) \triangleq \mathrm{Pl}_U(A).$$

Now $\mathrm{Cr}(A\times V) = 1 - \mathrm{Pl}(\bar{A}\times V)$. Q.E.D.

As a consequence, if Pl is a possibility measure Π i.e. its contour function π is a fuzzy relation on $U\times V$, Pl_U is the possibility measure based on the projection of the fuzzy relation (in the sense of Zadeh[36]), since

$$\pi_U(u) = \mathrm{Pl}_U(\{u\}) = \Pi(\{u\} \times V\} = \sup_{v \in V} \pi(u, v).$$

Conversely, given two bodies of evidence $(\mathcal{F}_U,\ m_U)$ and $(\mathcal{F}_V,\ m_V)$ on U and V respectively, we can define their cylindrical extensions and define the product of these extensions via Dempster rule (Shafer[25]). Namely, the *cylindrical extension* of $(\mathcal{F}_U,\ m_U)$ is $(c\mathcal{F}_U,\ cm_U)$ such that

$$\forall\ B \subseteq, U, cm_U\,(B \times V) = m_U\,(B)$$

and

$$cm_U\,(A) = 0 \text{ for other } A \subseteq \Omega = U \times V.$$

From $(\mathcal{F}_U,\ m_U)$, $(\mathcal{F}_V,\ m_V)$ on U and V respectively, $(\hat{\mathcal{F}},\ \hat{m}) \triangleq (\mathcal{F}_U,\ m_U)$ $\times$ $(\mathcal{F}_V,\ m_V)$, denotes a Cartesian product of bodies of evidence. $\hat{m}$ is calculated by:

$$\begin{aligned} \forall\ A \subseteq \Omega, \hat{m}\,(A) &= m_U\,(B) \cdot m_V\,(C) \quad \text{if} \quad A = B \times C \\ &= 0 \text{ otherwise.} \end{aligned} \tag{47}$$

Note that $\{(B, C) | A = B \times C\} = \{(U(A),\ V(A))\}$ and $B \times C = \varnothing$ only if B or $C = \varnothing$ so that Dempster rule really boils down to (47), and $\hat{\mathcal{F}} = \{B \times C | B \in \mathcal{F}_U,\ C \in \mathcal{F}_V\}$. Note that $(\hat{\mathcal{F}},\ \hat{m})$ is always normal since $(\mathcal{F}_U,\ m_U)$ and $(\mathcal{F}_V,\ m_V)$ are supposed to be so.

If $(\mathcal{F}_U,\ m_U)$ and $(\mathcal{F}_V,\ m_V)$ reduce to sets B and C, then their products in the sense of (47) is their Cartesian products. (47) is however not in accordance with Zadeh's[36] definition of the Cartesian product of fuzzy sets since if $(\mathcal{F}_U,\ m_U)$ and $(\mathcal{F}_V,\ m_V)$ are possibility measures, with contour functions μ_F and μ_G respectively then the fuzzy Cartesian product is the possibility measure with contour function $\min(\mu_F,\ \mu_G)$. Rather, (47) implies that $(\hat{\mathcal{F}},\ \hat{m})$ is generally not a fuzzy Cartesian product since it is consistent with $\mu_F \cdot \mu_G$, an operation previously introduced by the authors[3]; moreover $(\hat{\mathcal{F}},\ \hat{m})$ defines no possibility measure, generally.

The natural thing to do is now to project $(\mathcal{F},\ m)$ on U and V and recombine their projections. One may expect some relationship between $(\mathcal{F},\ m)$ and $(\hat{\mathcal{F}},\ \hat{m})$ in terms of specificity, namely that $(\mathcal{F}, m)$ is included in $(\hat{\mathcal{F}},\ \hat{m})$; unfortunately this property does not hold as shown below.

Counter example 3 $\mathcal{F} = \{S_1,\ S_2\}$ with $S_1 \cap S_2 = \varnothing,\ U(S_1) \cap U(S_2) = \varnothing,\ V(S_1) \cap V(S_2) = \varnothing$. $\hat{\mathcal{F}} = \{U(S_i) \times V(S_j) | i = 1, 2;\ j = 1, 2\}$. $\hat{\mathcal{F}}$ is made of four disjoint focal elements.

Now since $S = U(S_1) \times V(S_2) \notin \mathcal{F},\ \mathrm{Cr}(S) = 0$ while $\hat{\mathrm{Cr}}(S) = m(S_1) \cdot m(S_2) > 0$. Moreover $\mathrm{Cr}(S_1 \cup S_2) = 1$ while $\hat{\mathrm{Cr}}(S_1 \cup S_2) \leqq m(S_1)^2 + m(S_2)^2$ < 1 (the equality holds if S_1 and S_2 are Cartesian products).

More particularly, if $(\mathcal{F},\ m)$ is a probabilistic body of evidence then $(\hat{\mathcal{F}},\ \hat{m})$ also generates a probability measure, and no inclusion must be expected, relating these two bodies of evidence.

However, if $(\mathcal{F},\ m)$ is consonant, this relationship might be expected to hold. The following result leaves no hope about it for the d-inclusion.

Proposition 9. *Even if $(\mathcal{F},\ m)$ is consonant, the property $(\mathcal{F},\ m) \subseteq (\hat{\mathcal{F}},\ \hat{m})$ does not hold.*

Counter example 4 $\mathcal{F} = \{S_1, S_2\},\ S_1 \subset S_2$, with $U(S_1) \neq U(S_2),\ V(S_1) \neq V(S_2)$. Let $\alpha = m(S_1)$. Hence, $\hat{m}(U(S_1) \times V(S_1)) = \alpha^2$, $\hat{m}(U(S_1) \times V(S_2)) = \hat{m}(U(S_2) \times V(S_1)) = \alpha(1-\alpha)$, $\hat{m}(U(S_2) \times V(S_2)) = (1-\alpha)^2$.

Now assume S is such that:

$$(U(S_1) \times V(S_2)) \cup (U(S_2) \times V(S_1)) \subset S \subset S_2$$

where the inclusions are strict. It is easy to figure out that such a set S may exist. Then we have:

$$\mathrm{Cr}(S) = \alpha < \hat{\mathrm{Cr}}(S) = \alpha(2-\alpha),\ \forall \alpha < 1. \quad \text{Q.E.D.}$$

Note that the c-inclusion does not hold either. Indeed assume that $S_i \neq U(S_i) \times V(S_i)$ for $i = 1, 2$ in the above counter example. Clearly,

$$Q(U(S_1) \times V(S_1)) = 1 - \alpha < \hat{Q}(U(S_1) \times V(S_1)) = 1$$
$$Q(S_2) = 1 - \alpha > \hat{Q}(S_2) = (1-\alpha)^2$$

since $S_2 \not\subset U(S_i) \times V(S_j),\ i \neq j$.

Proposition 9 contrasts with a well-known result in fuzzy set theory, due to Zadeh[36]. Namely, a fuzzy relation R on $U \times V$ is included in the Cartesian product of its projections. The inclusion turns into an equality if and only if $\mu_R(u,\ v)$ is of the form $\min(\mu_A(u)\mu_B(v))$ where A and B are fuzzy sets on U and V respectively. In the *possibilitistic case* it is interesting to specify conditions under which $(\mathcal{F},\ m) = (\hat{\mathcal{F}},\ \hat{m})$.

First any $S \in \mathcal{F}$ must be of the form $A \times B$. Then $\mathcal{F} \subseteq \hat{\mathcal{F}}$ is ensured. Now $\mathcal{F} = \{A_i \times B_i | i = 1, p\}$ and $\hat{\mathcal{F}} = \{A_i \times B_j | i = 1, p, j = 1, p\}$. Let $A_i \neq A_j$ and $B_i \neq B_j$, then one of $A_i \times B_j,\ A_j \times B_i$ should not be in $\hat{\mathcal{F}}$ since there is no inclusion relationship between them. So, to preserve a nested structure in $\hat{\mathcal{F}}$ we must have $\forall\ i,\ j,\ A_i = A_j$ or $\forall\ i,\ j,\ B_i = B_j$. Hence the following result, stated in the case when $\exists\ B_i \neq B_j$:

Proposition 10. *If $(\mathcal{F},\ m)$ is consonant, $(\hat{\mathcal{F}},\ \hat{m}) = (\mathcal{F},\ m)$ if and only if*

$$\exists\ A \subseteq U,\ B_1 \subset B_2 \subset \cdots \subset B_p \subset V$$

such that

$$\mathcal{F} = \{A \times B_i | i = 1, p\}.$$

This is equivalent to state that $m_U(A) = 1$, i.e. the projection of $(\mathcal{F},\ m)$ on U is a set. So that $m_V(B_i) = m(B_i)\ \forall\ i$.

In the general case, the concept of inclusion, even the weaker one proves too strong to be able to compare $(\mathcal{F}, m)$ and the product of its projections. Such a comparison can be however carried out using the measures of uncertainty and specificity respectively introduced by Higashi and Klir[15] and Yager[30]. Then some interesting inequalities can be obtained expressing that $(\hat{\mathcal{F}}, \hat{m})$ is not more specific than $(\mathcal{F}, m)$ (see Dubois and Prade[9]). Note that we may have $(\mathcal{F}, m) = (\hat{\mathcal{F}}, \hat{m})$ in the general case, since when $\mathcal{F}$ is not consonant the requirement $\mathcal{F} = \hat{\mathcal{F}}$ does not induce the same constraints on $\mathcal{F}$, as in the consonant case.

3 Consonant Approximation of a Body of Evidence

It is easier to deal with a possibility measure or a probability measure rather than with a general plausibility measure. The main reason is that in both cases, the body of evidence is completely characterized by its contour function, i.e. a probability allocation or a fuzzy set. The question of approximation of a body of evidence by either a probability measure or a possibility measure is thus worth considering.

3.1 The Approximation Problem

A body of evidence $(\mathcal{F}', m')$ can be viewed as a valid substitute of $(\mathcal{F}, m)$ as soon as $(\mathcal{F}, m) \subseteq (\mathcal{F}', m')$ (here we assume bodies of evidence are disjunctive). This is a generalized version of Zadeh's entailment principle[39], and it encompasses Yager's[33] proposal based on the strong inclusion. Moreover the knowledge of another body of evidence $(\mathcal{F}'', m'') \subseteq (\mathcal{F}, m)$ enables the plausibility measure associated with $(\mathcal{F}, m)$ to be located in an interval, i.e.

$$\forall\ A, \mathrm{Pl}''(A) \leqq \mathrm{Pl}(A) \leqq \mathrm{Pl}'(A). \tag{48}$$

A related inequality holds for the credibility function, of course. Whenever (48) holds the pair $(\mathrm{Pl}'',\ \mathrm{Pl}')$ is said to be an approximation of Pl. Pl'' is the lower approximation, Pl' the upper approximation.

The approximation problem[2] for bodies of evidence can then be stated as follows: Let $\mathcal{A}$ be a suitable subset of $\mathcal{B}(\Omega)$ containing 'simple' bodies of evidence, in the sense that it is easy to deal with them for some reason. Given any body of evidence $(\mathcal{F},\ m) \notin \mathcal{A}$, find two bodies of evidence $(\mathcal{F}^*,\ m^*)$ and $(\mathcal{F}_*,\ m_*)$ in $\mathcal{A}$, upper and lower approximations of $(\mathcal{F},\ m)$ i.e.

[2] An example of this approximation methodology can be found in the recent paper by J. Gordon and E. H. Shortliffe: "A method for managing evidential reasoning in a hierarchical hypothesis space," *Artificial Intelligence*, **26**, 1985, pp. 323–357. In this paper the authors are looking for an approximation of the result of the combination of several bodies of evidence by means of Dempster rule because the exact result would be too difficult to compute.

$$(\mathcal{F}_*, m_*) \subseteq (\mathcal{F}, m) \subseteq (\mathcal{F}^*, m^*) \,. \tag{49}$$

Moreover $(\mathcal{F}_*, m_*)$ and $(\mathcal{F}^*, m^*)$ should be best approximations in the following sense: denote $\mathcal{A}^+(\mathcal{F}, m)$ and $\mathcal{A}_-(\mathcal{F}, m)$ the sets

$$\mathcal{A}^+ (\mathcal{F}, m) = \{(\mathcal{F}', m') \,|\, (\mathcal{F}, m) \subseteq (\mathcal{F}', m')\} \cap \mathcal{A}$$
$$\mathcal{A}^- (\mathcal{F}, m) = \{(\mathcal{F}'', m'') \,|\, (\mathcal{F}'', m'') \subseteq (\mathcal{F}, m)\} \cap \mathcal{A}$$

and let $\mathcal{A}^+_*(\mathcal{F}, m)$ (resp.: $\mathcal{A}^*_-(\mathcal{F}, m)$) be the set of minimal (resp.: maximal) elements in $\mathcal{A}^+(\mathcal{F}, m)$ (resp.: $\mathcal{A}_-(\mathcal{F}, m)$). Then we should require $(\mathcal{F}_*, m_*) \in \mathcal{A}^*_-(\mathcal{F}, m)$ and $(\mathcal{F}^*, m^*) \in \mathcal{A}^+_*(\mathcal{F}, m)$.

Clearly it is meaningless to choose $\mathcal{A}$ as being the set $\mathcal{P}(\Omega)$ of probability measures because an upper approximation will never exist when $(\mathcal{F}, m)$ is normal (except if $(\mathcal{F}, m)$ generate a probability measure) and a lower approximation only exists if $(\mathcal{F}, m)$ is normal. In such a case $\mathcal{A}^*_- (\mathcal{F}, m) = \mathcal{A} - (\mathcal{F}, m) = \mathcal{C}(\mathcal{F})$ since probability measures do not compare with one another via $\subseteq$. So all probability measures are equally candidate as lower approximations.

A member of $\mathcal{C}(\mathcal{F})$ is especially interesting and has been suggested by the authors[4, 5] previously. It is obtained by equally sharing the weights $m(A)$ among elements of A; we then have

$$\forall\, \omega \in \Omega, P\left(\{\omega\}\right) = \sum_{\omega \in A} \frac{m\left(A\right)}{|A|}. \tag{50}$$

(50) is in accordance with Laplace's principle of modeling a lack of information by uniformly distributed probability allocations. When $(\mathcal{F}, m)$ is consonant (50) defines a bijection between probability measures and possibility measures on a finite set, and the converse mapping can be useful to derive a possibilistic interpretation of histograms as explained in Dubois and Prade[4, 5].

3.2 Possibilistic Approximations of Normal Bodies of Evidence

A more satisfactory approach is to consider the set $[0, 1]^\Omega$ of consonant bodies of evidence as the approximation set $\mathcal{A}$. In this section we derive best upper and lower approximations of $(\mathcal{F}, m)$ when $\mathcal{A} = [0, 1]^\Omega$. The best lower approximation Π_* is first derived. The following result was already obtained in Dubois and Prade[4].

Proposition 11. *The best lower approximation in* $[0, 1]^\Omega$ *of a body of evidence* $(\mathcal{F}, m)$ *is unique and is the possibility measure* Π_**, whose possibility distribution* π_* *is the contour function of* $(\mathcal{F}, m)$.

Proof. [4]

$$\forall\, A, \mathrm{Pl}\left(A\right) = \sum_{B \subseteq \Omega} m\left(B\right) \cdot \sup_{\omega \in A} \mu_B\left(\omega\right),$$

where μ_B is the characteristic function of B. Hence

$$\mathrm{Pl}\,(A) \geqq \sup_{\omega \in A} \sum_{B \in \Omega} \mu_B\,(\omega)\, m\,(B) \triangleq \sup_{\omega \in A} \mathrm{Pl}\,(\{\omega\})\,.$$

Let Π_* be the possibility measure such that $\Pi_*(\{\omega\}) = \mathrm{Pl}(\{\omega\})$, clearly $\Pi_* \in \mathcal{A}_{\text{-}}(\mathcal{F}, m)$. Let Π be a possibility measure such that $\Pi \leqq \mathrm{Pl}$. Then $\forall\ \omega \in \Omega,\ \pi(\omega) \leqq \mathrm{Pl}(\{\omega\}) = \pi_*(\omega)$. Hence $\Pi \leqq \Pi_*$. Q.E.D.

Π_* is defined for any $(\mathcal{F}, m) \in \mathcal{B}(\Omega)$. However if $(\mathcal{F}, m)$ is not consistent (i.e. the core $C(\mathcal{F})$ is empty), Π_* is not normal, while it is always normal otherwise, since $\mathrm{Pl}(\{\omega\}) = 1, \forall\, \omega \in C(\mathcal{F})$. As a consequence the lower approximation is completely meaningful for consistent bodies of evidence. Obviously, if $(\mathcal{F}, m)$ is consonant, then $\mathrm{Pl} = \Pi_*$. At the opposite if $(\mathcal{F}, m)$ defines a probability measure then $\pi_*(\omega) = P(\{\omega\})$, which is not very interesting.

The use of the contour function of $(\mathcal{F}, m)$ has been suggested by Zhang[43, 44] and Wang[27] to derive the membership of a fuzzy set from statistical data made of error intervals.

The set of focal elements of the lower approximation is $\mathcal{F}_*$ defined by

$$\mathcal{F}_* = \{\{\omega | \mathrm{Pl}\,(\{\omega\}) \geqq \alpha\}\, | \alpha \in]0, 1]\}$$

and letting $\alpha_1 = 1 > \alpha_2 \cdots > \alpha_p > 0$ be the elements of the set $\{\mathrm{Pl}(\{\omega\}) | \omega \in \Omega\} \cup \{1\}$. $\mathcal{F}_*$ contains p focal elements $A_1 \subset A_2 \subset \cdots \subset A_p$ with $A_i = \{\omega | \mathrm{Pl}(\{\omega\}) \geqq \alpha_i\}$. $A_1 \neq \varnothing$ if and only if $(\mathcal{F}, m)$ is consistent. Indeed it is easy to see that A_1 is the core of $(\mathcal{F}, m)$ i.e.

$$C\,(\mathcal{F}) = \{\omega, \forall\ A \in \mathcal{F}, \omega \in A\}$$

and A_p is the support of $(\mathcal{F}, m)$, i.e.

$$S\,(\mathcal{F}) = \{\omega, \exists\ A \in \mathcal{F}, \omega \in A\}\,.$$

Hence $(\mathcal{F}, m)$ and $(\mathcal{F}_*, m_*)$ have the same core and support. This remark enables a member of $\mathcal{A}^+(\mathcal{F}, m)$ to be constructed from the knowledge of $\mathcal{F}_*$; to do it we use a technique described in Dubois and Prade[6], which is an alternative way of deriving a membership function from a set of statistical data consisting of error intervals. This technique, which contrasts with Zhang and Wang's approach goes as follows.

i) Define a mapping $f : \mathcal{F} \to \mathcal{F}_*$ where $f(A)$ is the smallest A_i containing A, i.e.

$$f(A) = A_{\mathrm{i}} \quad \text{such that } A \subseteq A_{\mathrm{i}}, A \not\subset A_{\mathrm{i}-1}.$$

ii) Let $(\mathcal{F}^*, m^*)$ be such that $\mathcal{F}^* = f(\mathcal{F}) \subseteq \mathcal{F}_*$

$$\forall\ A_i, m^*\,(A_{\mathrm{i}}) = \sum_{A_{\mathrm{i}} = f(A)} m\,(A)\,.$$

Note that $f(A)$ is never empty since $\forall\ A,\ A \subseteq A_p$. Moreover f defines a partition of $\mathcal{F}$ through the equivalence relation $\sim$: $A \sim B \Leftrightarrow f(B) = f(A)$. Hence

$$\sum_{A_i \in \mathcal{F}_*} m^*(A_\mathrm{i}) = \sum_{A \in \mathcal{F}} m(A)\,.$$

It is easy to check that $(\mathcal{F}, m)$ is strongly included in $(\mathcal{F}^*, m^*)$ since the above technique is a particular case of Procedure (b) of 2.3.2. where the whole mass $m(A)$ is allocated to $f(A)$. In Dubois and Prade,[6] however, the sets $A_1 \dots A_p$ are given independently of $(\mathcal{F}, m)$ except that $A_1 = C(\mathcal{F}),\ A_p = S(\mathcal{F})$, and $A_1 \neq \varnothing$ i.e. the procedure was defined only for consistent bodies of evidence. Here we improve it by prescribing what are the focal elements A_i for $1 < i < p$. We now prove that $(\mathcal{F}^*, m^*)$ is a best approximation in some sense. Let π and π' be two possibility distributions on Ω, π and π' are said to be *order-equivalent* if and only if

$$\forall\ \omega, \omega',\ \pi(\omega) > \pi(\omega') \Leftrightarrow \pi'(\omega) > \pi'(\omega')\,. \tag{51}$$

Order-equivalence can be nicely characterized in terms of focal elements:

Lemma 1. *π and π' are order-equivalent, if and only if their associated sets of focal elements are equal.*

Proof. Let $\{\alpha_1, \dots, \alpha_p\} = \{\pi(\omega) > 0 | \omega \in \Omega\}$. It is well-known that $\mathcal{F} = \{A_1, \dots, A_p\}$ where $i = 1, p,\ A_i = \{\omega | \pi(\omega) \geqq \alpha_i\}$.[4, 7] Let $\omega_i \in A_i$ such that $\pi(\omega_i) = \alpha_i$. Now from order-equivalence $A_i = \{\omega | \pi(\omega) \geqq \pi(\omega_i)\} = \{\omega | \pi'(\omega) \geqq \pi'(\omega_i)\}$. Hence $A_i \in \mathcal{F}'$, the set of focal elements of π'. Hence $\mathcal{F} \subseteq \mathcal{F}'$ and $\mathcal{F}' \subseteq \mathcal{F}$ since π and π' play the same role. The converse proposition is obvious. Q.E.D.

Note that π_* and π^* are generally not order-equivalent but satisfy the weaker statement

$$\forall\ \omega, \omega', \pi_*(\omega) > \pi_*(\omega') \Rightarrow \pi^*(\omega) \geqq \pi^*(\omega')\,. \tag{52}$$

This is because generally $\mathcal{F}^* \subset \mathcal{F}_*$.

Proposition 12. *$(\mathcal{F}^*, m^*)$ is the best upper approximation of $(\mathcal{F}, m)$ among its order-equivalent consonant bodies of evidence.*

Proof. Let N^* and Π^* be the necessity and possibility measures induced by $(\mathcal{F}^*, m^*)$. First note that

$$\forall\ A_i^* \in \mathcal{F}^*, \mathrm{Cr}(A_i^*) = \sum_{A \subseteq A_i^*} m(A) = \sum_{j=1}^{i} m^*\left(A_\mathrm{j}^*\right) = N^*(A_i^*)\,.$$

Let $(\mathcal{F}', m')$ be an upper approximation of $(\mathcal{F}, m)$. Because π' and π^* are order-equivalent, the Lemma yields $\mathcal{F}' = \mathcal{F}^*$. Now the inequality $N'(A) \leqq \mathrm{Cr}(A),\ \forall\ A$ implies $\forall\ A_i^* \in \mathcal{F}^*,\ N'(A_i^*) \leqq N^*(A_i^*)$ which also reads

$$\forall\ i \neq p, \max\left\{\pi'\left(\omega\right)|\omega \notin A_i^*\right\} \geqq \max\left\{\pi^*\left(\omega\right)|\omega \notin A_i^*\right\}. \tag{53}$$

Now the maximum in both sides of (53) is reached by any element ω in $A_{i+1}^* \cap \bar{A}_i^*$ since A_i^* is a focal element in both $\mathcal{F}'$ and $\mathcal{F}^*$. Hence (53) translates into:

$$\forall\ \omega \notin A_1^*, \pi'\left(\omega\right) \geqq \pi^*\left(\omega\right).$$

Moreover $\forall\ \omega \in A_1^*,\ \pi'(\omega) = \pi^*(\omega) = 1$, since $C(\mathcal{F}') = C(\mathcal{F}^*)$. Q.E.D.

The condition of order-equivalence is a necessary one to get optimality. Indeed (53) implies only the existence of some ω' in $\bar{A}_i^*$ such that $\forall\ \omega \in A_{i+1}^* \cap \bar{A}_i^*,\ \pi'(\omega') \geqq \pi^*(\omega)$.

Counter example 5 $\Omega = \{a, b, c, d, e\}$

$$\mathcal{F} = \left\{\{c\}, \{c, d\}, \{b, c\}, \{c, d, e\}, \{a, b, c\}\right\}$$

with a uniformly distributed basic assignment $m(A) = \frac{1}{5}, \forall\ A \in \mathcal{F}$.

We have the following results, where π_* and π^* are calculated from m and π' is given:

	a	b	c	d	e
π_*	0.2	0.4	1	0.4	0.2
π_*	0.4	0.8	1	0.8	0.4
π'	0.4	0.6	1	0.8	0.4

Note that $\Pi^*(A) = \Pi'(A)$ except for $A = \{b\},\ \{a, b\},\ \{b, e\}$, for which $\Pi'(A) = 0.6 < \Pi^*(A) = 0.8$. But $\Pi' \geqq \mathrm{Pl}$ since $\mathrm{Pl}(\{a\}) = 0.2,\ \mathrm{Pl}(\{a, b\}) = 0.4$ and $\mathrm{Pl}(\{b, e\}) = \mathrm{Pl}(\{a, b, e\}) = 0.6$. But the distribution π' possesses a dissymmetry which does not look natural since neither Pl, nor π^* have such a dissymmetry. Assuming $\pi'(d) = \pi'(b) = 0.6$ is not possible since then (53) is violated. Q.E.D.

It is clear that if $(\mathcal{F}, m)$ is consonant then $(\mathcal{F}^*, m^*) = (\mathcal{F}, m)$ which shows a good behavior of $(\mathcal{F}^*, m^*)$.

The possibility distribution associated with $(\mathcal{F}^*, m^*)$ is π^* defined by

$$\forall\ \omega \notin A_p^*, \pi^*\left(\omega\right) = 0,$$

$$\forall\ \omega \in A_1^*, \pi^*\left(\omega\right) = 1,$$

$$\forall\ \omega \in A_i^* - A_{\mathrm{i}-1}^*, \pi^*\left(\omega\right) = 1 - N^*\left(A_{i-1}^*\right) = 1 - \sum_{A \subseteq A_{i-1}^*} m\left(A\right)$$

$$= \sum_{j=i}^{p} m^*\left(A_j^*\right).$$

When $(\mathcal{F}, m)$ generates a probability measure, the formula becomes

$$\forall\ \omega, \pi^*\left(\omega\right) = \sum\left\{P\left(\{\omega'\}\right)\right)|P\left(\{\omega'\}\right) \leqq P\left(\{\omega\}\right)\}$$

and if $p_1 \geqq p_2 \ldots \geqq p_n$ are the probability weights on $\omega_1, \ldots, \omega_n$ we get values $\pi_1^* \geqq \cdots \geqq \pi_n^*$ such that

$$\pi_i^* = \sum_{j=i}^{p} p_{j,} \forall\ i. \tag{54}$$

(54) defines a bijection between probability and possibility measures on Ω since it is equivalent to

$$p_i = \pi_i^* - \pi_{i+1}^*, \quad \forall\ i \tag{55}$$

with $\pi_{n+1}^* = 0$. Equation (54) provides the best possibilistic approximation of a probability measure in the sense of the consistency condition:

$$\forall\ A, N^*\left(A\right) \leqq P\left(A\right) \leqq \Pi^*\left(A\right)$$

and under order-equivalence assumption.

Particularly this transformation provides a more specific result than the converse of (50), proposed in a previous paper[5]. Note that in the case of $(\mathcal{F}, m)$ being a probability allocation, the result was already given in Dubois and Prade[4].

Conclusion

Shafer's theory of evidence seems to make measure theory and logic interfere with each other. Mathematical beings, living in $\mathcal{B}(\Omega)$ have a dual nature: they are kinds of sets (more precisely convex combinations of sets) and as such can combine via logical connectives such as union, intersection and complementation, and consequently *any* connective of classical logic can have an extension in $\mathcal{B}(\Omega)$. But they are also kinds of measures, and concepts of expectations can be defined from them as Dempster did[1]. However, because bodies of evidence first emerged as upper and lower probabilities, the possibility of constructing a logic calculus on them was not really pointed out by Dempster or Shafer, but by the people working in random set theory[12, 14].

Logical operations cannot be introduced in the setting of probability measures because they are generalized *points*, not sets. This is why, may be, logic and probability theory seem to ignore each other. Contrastedly, possibility theory, first discussed in terms of fuzzy sets, was naturally equipped with a logic calculus. The measure-theoretic point of view came afterwards when possibility measures were also viewed as upper probabilities[4].

Similarly, Shafer's book[24], due to a probabilistic background, always assume information is disjunctive, as it must be in probability theory. On the contrary a set is classically viewed as a conjunctive of values as often as a restriction on the value of a variable. The framework of credibility and plausibility measures enables the conjunctive point of view to enter the probabilistic arena, and this is very important for knowledge representation issues. The existence of logical connectives in $\mathcal{B}(\Omega)$ has been exploited by Yager[34] to define new patterns of reasoning which generalize the modus ponens. However, contrary to approximate reasoning, based on fuzzy sets, the choice of the implication connective is very much restricted by the unicity of basic operations such as the union and complementation. This unicity stems from the unicity of Dempster rule under decomposability conditions[10].

As noted by Zadeh, fuzzy set theory is not a particular case of Shafer's theory, although a possibility measure (i.e. a fuzzy set) is a special kind of a body of evidence, where focal element are consonant. The reason is that Shafer's theory needs Dempster rule of combination to perform intersection in $\mathcal{B}(\Omega)$ while fuzzy sets are conjunctively combined by means of triangular norms[8, 23]. Shafer's rationale for Dempster rule stems from probabilistic independence between two basic assignments m and m', viewed as probability allocations on 2^{Ω}. As a consequence the intersection of two consonant bodies of evidence is generally no longer consonant. On the contrary Zadeh's approach starts from the requirements that any logical combination of fuzzy sets should be a fuzzy set again. This requirement is linked to the fact that possibility distributions model the meaning of imprecise statements, and that the meaning of complex statements should be expressed as some combination of simpler statements that they involve.

Shafer theory is based on the re-interpretation of results by Dempster, results which were cast in a frequentist framework. And indeed Dempster rule has a frequentist flavor, and the development of a frequentist theory of upper and lower probabilities receives attention in the literature[26]. Such attempts combined with results of Sect. 3, can provide grounds for statistical estimation of membership functions[6, 27]. However the possibility of a frequentist interpretation of fuzzy set-theoretic operations seems to be very unlikely, while the connections between these operations and the theory of conjoint measurement[8, 18] are more promising. In other words fuzzy set theory seems to be closer to research in psychological measurement than to statistics, although possibility measures may have frequentist interpretation. The rules of combination of frequentist possibility measures will be dictated by independence-like arguments deriving from the study of statistical experiments, while the rules of combination of subjectivist possibility measures may turn out to be those of fuzzy set theory. The core of the debate is the relevance of subjective probability theory. If subjective probability theory is acknowledged as being too restrictive to model uncertainty judgments, then Shafer's subjectivist interpretation of upper and lower probabilities can be questioned on the same grounds. From a mathematical point of view, the theory of evidence is nothing but the rules of probability theory applied to imprecise statements, while classical probability theory leaves no room to imprecision. As a consequence the rules of combination of bodies of evidence are given by the rules of probability theory, and what is behind the problem of validating Shafer's theory as a theory of measurement of subjective uncertainty is the validity of the rules of (subjective) probability theory (and especially the rule of additivity). From this point of view fuzzy set theory seems to be far less normative than the theory of evidence, although both provide tools for modeling imprecision and uncertainty in a unique setting.

Acknowledgments

This paper has benefited from the help of several people. We wish to thank Prof. Zadeh for his remarks about conjunctive information; Ron Yager for discussions about inclusion of bodies of evidence; and also J. B. Cavaillé for information about network flows, which was useful getting some of the results.

References

1. A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping." *Annals of Mathematical Statistics*, **38**, 1967, pp. 325–339.
2. D. Dubois and H. Prade, *Fuzzy Sets and Systems: Theory and Applications.* Academic Press, New York, 1980.
3. D. Dubois and H. Prade, "Additions of interactive fuzzy numbers," *IEEE Trans. on Automatic Control*, **26**, 1981, 926–936.
4. D. Dubois and H. Prade, "On several representations of an uncertain body of evidence," *Fuzzy Information and Decision Processes*, edited by M. M. Gupta, E. Sanchez, North-Holland, Amsterdam, 1982, pp. 167–181.
5. D. Dubois and H. Prade, "Unfair coins and necessity measures. Towards a possibilistic interpretation of histograms." *Fuzzy Sets and Systems*, **10**, 1983, pp. 15–20.
6. D. Dubois and H. Prade, "Fuzzy sets and statistical data," *European Journal of Operational Research*, **25**(3), 1986, 345–356.
7. D. Dubois and H. Prade, *Théorie des Possibilités. Applications à la Représentation des Connaissances en Informatique*, Masson, Paris, 1985.
8. D. Dubois and H. Prade, "A review of fuzzy set aggregation connectives," *Information Sciences*, **36**, 1985, pp. 85–121.
9. D. Dubois and H. Prade, "Additivity and monotonicity of measures of information defined in the setting of Shafer's evidence theory." *BUSEFAL* (Universitė P. Sabatier, Toulouse), No. 24, 1985, pp. 64–76.
10. D. Dubois, R. Giles and H. Prade, "On the unicity of Dempster's rule of combination." *International Journal of Intelligent Systems*, **1**(2), 1986, 133–142.
11. L. R. Ford, Jr., D. R. Fulkerson, *Flows in Networks.* Princeton University Press, Princeton, N.J., 1962.
12. R. Fortet and M. Kambouzia, "Ensembles aléatoires et ensembles flous," *Publications Econométriques*, **IX**, fascicule 1, 1976, pp. 1–23.
13. I. R. Goodman, "Fuzzy sets as equivalence classes of random sets." In: *Fuzzy Set and Possibility Theory: Recent Developments*, edited by R. R. Yager, Pergamon Press, Oxford, 1982, pp. 327–342.
14. I. R. Goodman, "Characterization of n-ary fuzzy set operations which induce homomorphic random set operations," In: *Fuzzy Information and Decision Processes*, edited by M. M. Gupta and E. Sanchez, North-Holland, Amsterdam, 1982, pp. 203–212.
15. M. Higashi and G. Klir, "Measures of uncertainty and information based on possibility distributions." *International Journal of General Systems*, **9**, 1983, pp. 43–58.
16. U. Höhle, "Fuzzy filters: a generalization of credibility measures." *Proc. IF AC Symposium on Fuzzy Information, Knowledge Representation, and Decision Analysis*, Marseille, June 1983, pp. 111–114.

17. J. Kampé de Feriet, "Interpretation of membership functions of fuzzy sets in terms of plausibility and belief." In: *Fuzzy Information and Decision Processes*, edited by M. M. Gupta and E. Sanchez, North-Holland, Amsterdam, 1982, pp. 93–98.
18. D. H. Kranz, R. D. Luce, R. Suppes and A. Tversky, *Foundations of Measurement.* Vol. I, Academic Press, New York, 1971.
19. H. T. Nguyen, "On random sets and belief functions." *Journal of Mathematical Analysis and Applications*, **65**, 1978, pp. 531–542.
20. E. M. Oblow, "A hybrid uncertainty theory," *Proc. 5th International Workshop: "Expert Systems and their Applications,"* Avignon, May 13–15, 1985, pp. 1193–1201, published by ADI Paris.
21. H. Prade and C. Testemale, "Representation of soft constraints and fuzzy attribute values by means of possibility distributions in databases." In: *The Analysis of Fuzzy Information*, edited by J. C. Bezdek, CRC Press, Boca Raton, Fl., 1986, to appear, Vol. II.
22. T. Sales, "Fuzzy sets as set classes," *Stochastica* (Barcelona, Spain), **VI**, 1982, pp. 249–264.
23. B. Schweizer and A. Sklar, "Associative functions and abstract semi-groups." *Publicationes Mathematicae Debrecen*, **10**, 1963, pp. 69–81.
24. G. Shafer, *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, N.J., 1976.
25. G. Shafer, *Belief Functions and Possibility Measures*, Working Paper No. 163, School of Business, The University of Kansas, Lawrence, 1984.
26. P. Walley and T. Fine, "Towards a frequentist theory of upper and lower probability." *The Annals of Statistics*, **10**, 1982, pp. 741–761.
27. Wang Pei-Zhuang, "From the fuzzy statistics to the falling random subsets." In: *Advances in Fuzzy Sets, Possibility Theory and Applications*, edited by P. P. Wang, Plenum Press, New York, 1983, pp. 81–96.
28. Wang Pei-Zhuang and E. Sanchez, "Treating a fuzzy subset as a projectable random subset." In: *Fuzzy Information and Decision Processes*, edited by M. M. Gupta and E. Sanchez, North-Holland, Amsterdam, 1982, pp. 213–220.
29. R. R. Yager, "Hedging in the combination of evidence." *Journal of Information and Optimization Science*, **4**, No. 1, pp. 73–81, 1983.
30. R. R. Yager, "Entropy and specificity in a mathematical theory of evidence." *Int. Journal of General Systems*, **9**, 1983, pp. 249–260.
31. R. R. Yager, "On different classes of linguistic variables defined via fuzzy subsets," *Kybernetes*, **13**, 1984, pp. 103–110.
32. R. R. Yager, "Arithmetic and other operations on Dempster–Shafer structures," *Tech. Report* MII-508, Iona College, New Rochelle, N.Y., 1985.
33. R. R. Yager, The entailment principle for Dempster–Shafer granule," *Tech. Report* MII-512, Iona College, New Rochelle, N.Y., 1985.
34. R. R. Yager, Reasoning with uncertainty in expert systems." *Proc. 9th Int. Joint Conf. on Artificial Intelligence*, Los Angeles, 1985, pp. 1295–1297.
35. L. A. Zadeh, "Fuzzy sets," *Information and Control*, **8**, 1965, pp. 338–353.
36. L. A. Zadeh, "Calculus of fuzzy restrictions." In: *Fuzzy Sets and their Applications to Cognitive and Decision Processes*, edited by L. A. Zadeh, K. S. Fu, K. Tanaka, M. Shimura, Academic Press, New York, 1975, pp. 1–39.
37. L. A. Zadeh, "PRUF: A meaning representation language for natural languages." *Int. J. Man–Machine Studies*, **10**, 1978, pp. 395–460.

38. L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility." *Fuzzy Sets and Systems*, **1**, 1978, pp. 3–28.
39. L.A. Zadeh, "A theory of approximate reasoning." In: *Machine Intelligence*, Vol. 9, edited by J. Hayes, D. Michie and L. I. Mikulich, John Wiley, New York, 1979, pp. 149–194.
40. L. A. Zadeh, "Fuzzy sets and information granularity." In: *Advances in Fuzzy Set Theory and Applications*, edited by M. M. Gupta, R. K. Ragade and R. R. Yager, North-Holland, Amsterdam, 1979, pp. 3–18.
41. L. A. Zadeh, "On the validity of Dempster's rule of combination of evidence." *Memo UCB/ERL* No. 79/24, University of California, Berkeley, 1979.
42. L. A. Zadeh, "A simple view of the Dempster–Shafer theory of evidence," *Berkeley Cognitive Science Report* No. 27, University of California, Berkeley, 1984.
43. Zhang Nan-Lun, "A preliminary study of the theoretical basis of the fuzzy set." *BUSEFAL* (Univ. P. Sabatier, Toulouse), No. 19, 1984, pp. 58–67.
44. Zhang Nan-Lun, "The membership and probability characteristics of random appearance," *Journal of Wuhan Institute of Building Materials*, No. 1, 1981.

Appendix

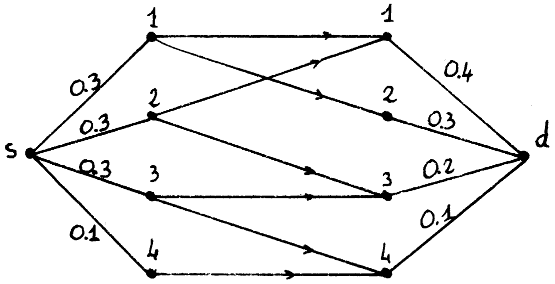
Flows in Networks and Inclusion

Let $(\mathcal{F}, m)$ and $(\mathcal{F}', m')$ be two bodies of evidence such that $(\mathcal{F}, m)$ is strongly included in $(\mathcal{F}', m')$ (cf. Definition 2). Let $A_1, \ldots, A_p$ (resp.: $A'_1, \ldots, A'_q$) be the elements of $\mathcal{F}$ (resp.: $\mathcal{F}'$). It is then possible to build a bipartite graph $(V, V', \mathcal{E})$ where V and V' are disjoint sets of nodes, and $\mathcal{E}$ is a set of arcs (v, v') where $(v \in V,\ v' \in V'$, defined as follows:

$$\text{each element of } V \text{ (resp. : } V'\text{) represents a focal element in } \mathcal{F} \text{ (resp. : } \mathcal{F}'\text{)},$$
$$\text{arc } (v_i, v'_j) \text{ exists if and only if } A_i \subseteq A'_j.$$

Note that $\forall\ v_i,\ \exists(v_i, v'_j) \in \mathcal{E}; \forall\ v'_j,\ \exists(v_i, v'_j) \in \mathcal{E}$ from strong inclusion.

Let s be a source node and d a sink node, which do not belong to $V \cup V'$. Build the arcs $(s, v_i) \forall\ v_i \in V$, with an associated capacity $a_i = m(A_i)$, and the arcs (v'_i, d) with an associated capacity $a'_i = m'(A_i)$. The graph corresponding to counter-example 2 is as follows:



Arcs in $\mathcal{E}$ are supposed to have infinite capacity. It is clear that the strong inclusion of $(\mathcal{F}, m)$ in $(\mathcal{F}', m')$ is equivalent to the existence of a flow of value 1 in the graph whose set of nodes is $N = V \cup V' \cup \{s, d\}$ and arcs, $E = \mathcal{E} \cup \{(s, v_i) | v_i \in V\} \cup \{(v'_j, d) | v'_j \in V'\}$. This fact is expressed by (42)–(43).

Now a *cut* in the graph is a partition $(X, \bar{X})$ of the nodes such that $s \in X$, $d \in \bar{X}$, and its capacity is the sum of the capacities of the arcs (i, j) such that $i \in X$, $j \in \bar{X}$. The max flow min-cut theorem[11] states that the maximal flow value from s to d is equal to the minimal cut capacity of all cuts separating s and d.

Obvious finite capacity cuts in (N, E) are obtained by stating $X = \{s\}$ or $\bar{X} = \{d\}$, and their capacity is 1. Hence the flow value through the graph is at most 1. Moreover if a cut involves an arc in $\mathcal{E}$, it has infinite capacity and is useless in the computation of the maximal flow. Hence interesting cuts are such that

if $S \subseteq V$ is a part of X then the set $\Gamma(S) \subseteq V'$ of successors of nodes in S is also in X,

if $T \subseteq V'$ is a part of $\bar{X}$ then the set $\Gamma^{-1}(T) \subseteq V$ of predecessors of nodes in T is also in $\bar{X}$.

Hence the set of cuts can be described as the set $\{(S, T) | S \subseteq V, T \subseteq V', \Gamma(S) \cap T = \varnothing, \Gamma^{-1}(T) \cap S = \varnothing\}$. The capacity of cut (S, T) is easily found as

$$C(S, T) = \sum_{v_i \in \bar{S}} a_i + \sum_{v_i \in \bar{T}} a'_i$$

since $X = \{s\} \cup S \cup \bar{T}$, $\bar{X} = \{t\} \cup T \cup \bar{S}$.

Now consider the cut $(\Gamma^{-1}(T), \Gamma(S))$. It is clear that

$$C(S, T) \geqq C\left(\Gamma^{-1}(T), \Gamma(S)\right).$$

So that the set of interesting cuts for the computation of the maximal flow is $\{(S, T) | S \subseteq V, T \subseteq V' \Gamma(S) = \bar{T}, \Gamma^{-1}(T) = \bar{S}\}$. A necessary and sufficient condition for the existence of a flow of value 1 through the network is thus

$$\forall\, S \subseteq V, \sum_{v_i \in \bar{S}} a_i + \sum_{v'_i \in \Gamma(S)} a'_i \geqq 1$$

$$\forall\, T \in V', \sum_{v_i \in \Gamma^{-1}(T)} a_i + \sum_{v'_i \in \bar{T}} a'_i \geqq 1$$

which also reads

$$\forall\, S \in V, \sum_{v_i \in S} a_i \leqq \sum_{v'_1 \in \Gamma(S)} a'_i \tag{I}$$

$$\forall\, T \in V', \sum_{v'_i \in T} a'_i \leqq \sum_{v_i \in \Gamma^{-1}(T)} a_i. \tag{II}$$

It is easy to check that the condition $Q(A) \leqq Q'(A)$ applied with $A = A_i$ gives (I) with $S = \{v_i\}$, and the condition $\mathrm{Cr}(A) \geqq \mathrm{Cr}'(A)$ applied with $A = A'_j$ gives (II) with $T = \{v'_j\}$. But generally it is possible to find $S \subseteq V$ such that

$$\nexists\ A, Q(A) = \sum_{u_i \in S} a_i$$

and T such that

$$\nexists\ A, \mathrm{Cr}'(A) = \sum_{u'_i \in T} a'_i.$$

This is why inclusion does not imply strong inclusion. In the above example if $S = \{v_3, v_4\}$ then $\Gamma(S) = \{v'_3, v'_4\}$ and (I) is violated i.e.

$$a_3 + a_4 = 0.4 > a'_3 + a'_4 = 0.3.$$

However, as the Table 1 shows, $Q(A) \leqq Q'(A),\ \mathrm{Cr}(A) \geqq \mathrm{Cr}'(A) \forall\ A$.

Now assume $(\mathcal{F}, m)$ and $(\mathcal{F}', m')$ are consonant. $\mathcal{F}$ and $\mathcal{F}'$ are ordered such that $A_1 \subset A_2 \cdots \subset A_p,\ A'_1 \subset A'_2 \cdots \subset A'_q$. The bipartite graph $(V, V', \mathcal{E})$ has a special structure since if $A_i \subseteq A'_j$ then $A_i \subseteq A'_k,\ \forall\ k \geqq j$.

We now prove that the flow equations always have a solution if the fuzzy set F associated to $(\mathcal{F}, m)$ is included in F' associated to $(\mathcal{F}', m')$.

$\forall\ v_i \in V$, let $\sigma(i)$ be the index such that

$$\sigma(i) = \min\left\{j | A_i \subseteq A'_j\right\}.$$

Similarly $\forall v'_j \in V'$, let $\tau(j)$ be the index such that

$$\tau(j) = \max\left\{i | A_i \subseteq A'_j\right\}.$$

Then the flow (42) and (43) reads

$$a_i = \sum_{j \geqq \sigma(i)} w_{ij} \qquad \forall\ i = 1, p \tag{III}$$

$$a'_j = \sum_{i \leqq \tau(j)} w_{ij} \qquad \forall\ j = 1, p. \tag{IV}$$

Let $n = |\sigma(V)|,\ k \in \sigma(V),\ \sigma^{-1}(k) = \{A_i | \sigma(i) = k\}$ and $i_k = \max\{i, A_i \in \sigma^{-1}(k)\}$. Because $(\mathcal{F}, m) \subseteq (\mathcal{F}',\ m'),\ A_1 \subseteq A'_1$ and $A_p \subseteq A'_q$. Moreover $\forall\ i \in \sigma^{-1}(k),\ A_i \subseteq A'_k$ but $A_i \not\subseteq A'_{k-1}$. Hence $\mathcal{F}$ can be partitioned into n groups of consecutive focal elements, and this partition creates a partition of $\mathcal{F}'$, also in n groups $\tau^{-1}(i_k),\ k \in \sigma(V)$ with

$$\tau^{-1}(i_k) = \{B_j | k \leqq j < \sigma(i_k + 1)\} \qquad \text{(see Fig. 1)}$$

and $\forall\ B_j \in \tau^{-1}(i_k),\ B_j \supseteq A_i$ for all $i \in \sigma^{-1}(\sigma(i_k))$, but $A_i \not\subset B_j,\ j < k$. Note that $\max_{k \in \sigma(V)} i_k = p$ so that $\sigma(p+1) = q+1$ by convention. It is clear that $i_k = \tau(k+1) - 1$ (see Fig. 1). Pairs $(\sigma^{-1}(k), \tau^{-1}(i_k))$ are ranked along increasing k's and can be renumbered as $\{(G_i,\ G'_i)\ i = 1,\ n\}$ as in Fig. 1. For all $(A_i,\ A'_j) \in G_k \times G'_l$ we define $x_{kl} = \Sigma\{w_{ij} | (A_i,\ A'_j) \in G_k \times G'_l\}$.

$$\bar{a}_k = \sum_{A_{\mathrm{i}} \in G_k} a_i, \qquad \bar{a}'_l = \sum_{A'_j \in G'_l} a'_j.$$

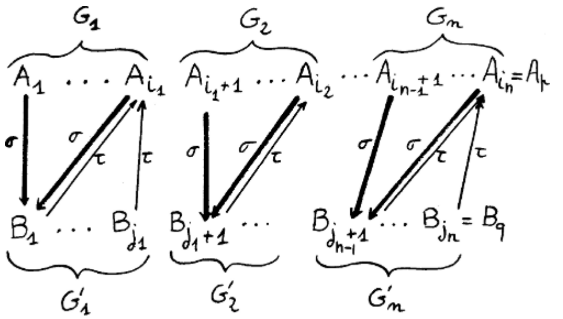

G_1
G_2
G_n
$A_1 \cdots A_{i_1}$
$A_{i_1+1} \cdots A_{i_2}$
$\cdots A_{i_{n-1}+1} \cdots A_{i_n} = A_p$
σ
τ
$B_1 \cdots B_{j_1}$
$B_{j_1+1} \cdots$
$B_{j_{n-1}+1} \cdots B_{j_n} = B_q$
G'_1
G'_2
G'_n

FIGURE 1

Now (III) and (IV) imply

$$\bar{a}_k = \sum_{\substack{l \\ l \geqq k}} x_{kl} \qquad k = 1, n, \tag{V}$$

$$\bar{a}'_l = \sum_{\substack{k \\ l \geqq k}} x_{kl} \qquad l = 1, n. \tag{VI}$$

Similarly, let Φ and Φ' be the fuzzy sets derived from F and F' as follows: Φ and Φ' have n α-cuts which are respectively the smallest set in each G_k and the greatest set in each G'_k, the mass allocated to the set from G_k (resp.: G'_k) being $\bar{a}_k$ (resp.: $\bar{a}'_k$). It is easy to check that $\Phi \subseteq F \subseteq F' \subseteq \Phi'$.

System (V) and (VI) always have solutions. Let

$$\mu_i = \sum_{k=i}^{n} \bar{a}_k, \qquad \mu'_i = \sum_{k=i}^{n} \bar{a}'_k.$$

$\Phi \subseteq \Phi'$ implies $\mu_i \leqq \mu'_i\ \forall\ i = 1,\ n$. Then let

$$\begin{aligned} x_{11} &= 1 - \mu'_2 \\ x_{ii} &= \mu_i - \mu'_{i+1} \qquad 1 < i < n \\ x_{i,i+1} &= \mu'_{i+1} - \mu_{i+1} \qquad 1 \leqq i < n \\ x_{nn} &= \mu_n \\ x_{ij} &= 0 \text{ otherwise.} \end{aligned}$$

This is a solution of (V–VI). Indeed, it is demonstrated as follows:

$$\begin{aligned}\bar{a}_k &= x_{kk} + x_{kk+1} = \mu_k - \mu_{k+1}\\ \bar{a}'_k &= x_{kk} + x_{k-1k} = \mu'_k - \mu'_{k+1}.\end{aligned}$$

From this solution, a solution to (III) and (IV) is easily deduced letting

$$w_{ij} = \frac{a_i \cdot a'_j}{\bar{a}_k \cdot \bar{a}'_l} x_{kl} \quad \text{whenever} \quad (A_i, A_j) \in G_{k.l}.$$

Hence if $F \subseteq F'$ then $(\mathcal{F}, m)$ is strongly included in $(\mathcal{F}', m')$.

$$\begin{aligned}\textit{Example}\quad \Omega &= \{a, b, c, d, e\}\\ F &= \{1/a, 0.5/b, 0.4/c, 0.2/d\}\\ F' &= \{1/a, 1/b, 0.5/c, 0.3/d, 0.3/e\}.\end{aligned}$$

Then

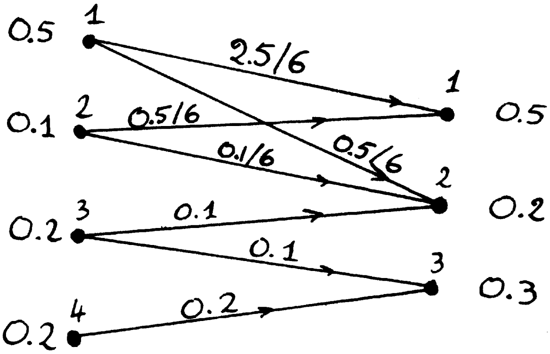
$$\begin{array}{ll|ll} A_1 = \{a\} & a_1 = 0.5 & A'_1 = \{a, b\} & a'_1 = 0.5\\ A_2 = \{a, b\} & a_2 = 0.1 & A'_2 = \{a, b, c\} & a'_2 = 0.2\\ A_3 = \{a, b, c\} & a_3 = 0.2 & A'_3 = \Omega & a'_3 = 0.3\\ A_4 = \{a, b, c, d\} & a_4 = 0.2 & & \end{array}$$

$$\begin{array}{ll|ll} G_1 = \{A_1, A_2\} & \bar{a}_1 = 0.6 & G'_1 = \{A'_1\} & \bar{a}'_1 = 0.5\\ G_2 = \{A_3\} & \bar{a}_2 = 0.2 & G'_2 = \{A'_2\} & \bar{a}'_2 = 0.2\\ G_3 = \{A_4\} & \bar{a}_3 = 0.2 & G'_3 = \{A'_3\} & \bar{a}'_3 = 0.3 \end{array}$$

$$x_{11} = 0.5$$

$$\begin{array}{l|l} x_{12} = 0.5 - 0.4 = 0.1 & \Rightarrow \begin{cases} w_{11} = 2.5/6\\ w_{21} = 0.5/6 \end{cases}\\ x_{22} = 0.1 & \quad w_{12} = 0.5/6\\ x_{23} = 0.1 & \quad w_{22} = 0.1/6 \text{ etc} \ldots.\\ x_{33} = 0.2 & \end{array}$$

Hence the flow



15

Weights of Evidence and Internal Conflict for Support Functions

Nevin L. Zhang

Abstract. Shafer [1] defined weights of evidence and the weight of internal conflict for separable support functions. He also formulated a conjecture, the weight-of-conflict conjecture, which implies that these definitions can be extended in a natural way to all support functions. In this paper I show that the extension to support functions can be carried out whether or not the weight-of-conflict conjecture is true.

1 Prerequisites

This section reviews basic concepts and results needed for the theorems in the next section. See Shafer [1] for details.

Let Θ be a finite set, called a frame of discernment. A function $\mathrm{Bel} : 2^{\Theta} \to [0,1]$ is called a belief function over Θ if

(1) $\mathrm{Bel}(\varnothing) = 0,\ \mathrm{Bel}(\Theta) = 1$, and
(2) for every integer n and arbitrary subsets $A_1, A_2, \ldots, A_n$ of Θ,

$$\mathrm{Bel}\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{k=1}^{n} (-1)^{k-1} \sum \left\{ \mathrm{Bel}\left(\bigcap_{i \in I} A_i\right) \middle| \, |I| = k, I \subseteq \{1,2,\ldots,n\} \right\}.$$

Given a belief function Bel over the frame Θ, there exists a unique map $m : 2^{\Theta} \to [0,1]$ (called the basic probability assignment for Bel) such that for each subset A of Θ,

$$\mathrm{Bel}\,(A) = \sum \left\{ m\,(B)\,|B \subseteq A \right\}.$$

The function $Q : 2^{\Theta} \to [0,1]$ defined by

$$Q\,(A) = \sum \left\{ m\,(B)\,|B \supseteq A \right\}$$

for each subset A of Θ is called the commonality function for Bel.

Suppose Bel_1 and Bel_2 are belief functions, with basic probability assignments m_1 and m_2, respectively. If the number

$$K = \sum \left\{ m_1(A_1) m_2(A_2) \,|A_1, A_2 \in 2^{\Theta}, A_1 \cap A_2 \neq \varnothing \right\}$$

is not zero, then we say that the orthogonal sum of Bel_1 and Bel_2 exists. We denote this orthogonal sum by $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$; by definition, it is the function over Θ whose basic probability assignment is given by

$$m(A) = \frac{1}{K} \sum \{ m_1(A_1) m_2(A_2) \,|A_1 \cap A_2 = A \} .$$

The number $-\log K$ is called the weight of conflict between Bel_1 and Bel_2. The weight of conflict and the orthogonal sum are defined similarly for more than two belief functions. They do not depend on the order of combination, and

$$\begin{aligned} \mathrm{Con}(\mathrm{Bel}_1, \ldots, \mathrm{Bel}_n) =& \mathrm{Con}(\mathrm{Bel}_1, \ldots, \mathrm{Bel}_{n-1}) \\ &+ \mathrm{Con}(\mathrm{Bel}_1 \oplus \cdots \oplus \mathrm{Bel}_{n-1}, \mathrm{Bel}_n), \end{aligned} \tag{1}$$

where $\mathrm{Con}(\mathrm{Bel}_1, \ldots, \mathrm{Bel}_n)$ stands for the weight of conflict among $\mathrm{Bel}_1, \ldots, \mathrm{Bel}_n$.

A subset of Θ to which the basic probability assignment assigns a positive number is called a focal element. If $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ exists, then its set of focal elements consists of all nonempty intersections of the form $A_1 \cap A_2$, where A_1 is a focal element of Bel_1 and A_2 is a focal element of Bel_2.

The belief function whose only focal element is Θ is called the vacuous belief function. If Bel_1 is vacuous, then $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2 = \mathrm{Bel}_2$. A belief function with at most one focal element other than Θ is called a simple support function; the focal element not equal to Θ is called the focus. A belief function which can be expressed as an orthogonal sum of simple support functions is called a separable support function.

Suppose S is a simple support function focused on A. Then

$$w = -\log[1 - S(A)]$$

is called the weight of evidence focused on A.

The union of the focal elements of a belief function is called its core. If A is the core of Bel, then $\mathrm{Bel}(B) = 1$ if and only if $B \supseteqq A$. If the core A of Bel is a proper subset of Θ, then it is sometimes convenient to replace the frame Θ by A or by some other set B such that $A \subset B \subset \Theta$. (This means that we work not with $\mathrm{Bel} : 2^{\Theta} \to [0,1]$ but with the restriction $\mathrm{Bel}|2^{B}$, which is a belief function over B whenever $\mathrm{Bel}(B) = 1$.)

Given a subset A of Θ, let Bel_A denote the belief function whose only focal element is A; this means that $\mathrm{Bel}_A(B) = 1$ whenever $B \supseteqq A$ and $\mathrm{Bel}_A(B) = 0$ otherwise. [The corresponding basic probability assignment m_A satisfies $m_A(A) = 1$ and $m_A(B) = 0$ when $B \neq A$.] If Bel is another belief function over Θ with $\mathrm{Bel}(\bar{A}) < 1$, then $\mathrm{Bel} \oplus \mathrm{Bel}_A$ exists and $(\mathrm{Bel} \oplus \mathrm{Bel}_A)(A) = 1$. The belief function $\mathrm{Bel} \oplus \mathrm{Bel}_A$ can be thought of as a belief function over Θ or as a belief function over A; in either case its values are given by

$$(\mathrm{Bel} \oplus \mathrm{Bel}_A)(B) = \frac{\mathrm{Bel}\left(B \cup \bar{A}\right) - \mathrm{Bel}\left(\bar{A}\right)}{1 - \mathrm{Bel}\left(\bar{A}\right)}. \qquad (2)$$

Changing Bel to $\mathrm{Bel} \oplus \mathrm{Bel}_A$ is called conditioning Bel on A.

The belief function Bel_A is idempotent with respect to the operation $\oplus : \mathrm{Bel}_A \oplus \mathrm{Bel}_A = \mathrm{Bel}_A$. This fact, together with the commutivity and associativity of $\oplus$, allows us to write

$$(\mathrm{Bel}_1 \oplus \cdots \oplus \mathrm{Bel}_n) \oplus \mathrm{Bel}_A = (\mathrm{Bel}_1 \oplus \mathrm{Bel}_A) \oplus \cdots \oplus (\mathrm{Bel}_n \oplus \mathrm{Bel}_A). \qquad (3)$$

In words: combining and then conditioning on A gives the same result as conditioning on A and then combining.

If we condition a simple support function on A, then the result, considered as a belief function over A, is again a simple support function. (Suppose S is a simple support function focused on B. Then S has at most two focal elements, B and Θ. Since Bel_A has only one focal element, A, the orthogonal sum $S \oplus \mathrm{Bel}_A$ has at most two focal elements, $B \cap A$ and $\Theta \cap A = A$. If $B \cap A = \varnothing$ or $B \cap A = A$, then A is $S \oplus \mathrm{Bel}_A$'s only focal element, and therefore $S \oplus \mathrm{Bel}_A$ is the vacuous belief function over A.) It follows from this and (3) that if we condition a separable support function on A, then the result, considered as a belief function over A, is again a separable support function.

Suppose S is a separable support function over the frame Θ; we assume, without loss of generality, that Θ is the core of S. In this case there exists a unique set $S_1, \ldots, S_n$ of nonvacuous simple support functions with distinct foci such that $S = S_1 \oplus \cdots \oplus S_n$. The weight of conflict among these S_i is called the weight of internal conflict in S and is denoted by W_S. If we denote the focus of S_i by A_i and denote the weight of evidence focused on A_i by w_i, then the function $V_s : 2^\Theta \rightarrow [0, \infty)$ defined by

$$V_S(A) = \sum \{w_i | A_i \not\supseteq A\}$$

is called the impingement function for $S; V(A)$ is the total weight of evidence impinging on A. It turns out that the commonality function Q_S for S satisfies

$$Q_S(A) = \exp\left[W_S - V_S(A)\right],$$

or

$$V_S(A) = W_S - \log Q_S(A), \qquad (4)$$

for every nonempty subset A of Θ.

Suppose M is a field of subsets of the frame Θ, and suppose A is a subset of Θ. Since Θ is finite, there is a smallest element of M containing A and a largest element of M contained in A. We denote these elements of M by A^+ and A^-, respectively:

$$A^{+} = \bigcup \{B | B \text{ is an atom of } M; B \cap A \neq \varnothing\}$$

or

$$A^{-} = \bigcup \{B | B \text{ is an atom of } M; B \subseteq A\} .$$

We say that Bel is carried by the field M if all the focal elements of Bel are in M. This is equivalent to the requirement that $\mathrm{Bel}(A) = \mathrm{Bel}(A^{-})$ for all $A \subseteq \Theta$.

2 Main Results

Theorem 1. *If* Bel_1 *and* Bel_2 *are both belief functions over* Θ*, and if* Bel_1 *and* Bel_2 *agree on the field* M *generated by the focal elements of* Bel_1*, then their commonality functions* Q_1 *and* Q_2 *satisfy* $Q_1 \geq Q_2$*.*

Proof. For any subset A of Θ,

$$\begin{aligned} \mathrm{Bel}_1(A) &= \mathrm{Bel}_1\left(A^{-}\right) = \mathrm{Bel}_2\left(A^{-}\right) \\ &= \sum \left\{m_2(B) \,|B \subseteq A^{-}\right\} \\ &= \sum \left\{m_2(B) \,|B^{+} \subseteq A\right\} , \end{aligned}$$

where m_2 is the basic probability assignment for Bel_2. This implies that the basic probability assignment for Bel_1 is given by

$$m_1(A) = \sum \left\{m_2(B) \,|B^{+} = A\right\} .$$

Therefore

$$\begin{aligned} Q_1(A) &= \sum \{m_1(B) \,|B \supseteqq A\} \\ &= \sum \left\{m_2(B) \,|B^{+} \supseteqq A\right\} \\ &\geq \sum \{m_2(B) \,|B \supseteqq A\} = Q_2(A) . \end{aligned}$$

Theorem 2. *If* S *and* T *are both separable support functions over* Θ*, and if* S *and* T *agree on the field* M *generated by the focal elements of* S*, then*

$$W_S \leq W_T \tag{5}$$

and

$$V_S \leq V_T . \tag{6}$$

Proof. We will assume, without loss of generality, that Θ is the core of S.

Let $S = S_1 \oplus \cdots \oplus S_n$ be the unique decomposition of S into nonvacuous simple support functions with distinct foci, and let $A_1, \ldots, A_n$ denote these foci. We will prove (5) by induction on n.

If $n = 1$, then (5) is immediate, because $W_S = 0$.

Suppose (5) is true for all $k < n$.

Consider the belief functions $S \oplus \mathrm{Bel}_{A_1}$ and $T \oplus \mathrm{Bel}_{A_1}$. Since A_1 is in M, it follows from (2) and from the agreement of S and T on M that $S \oplus \mathrm{Bel}_{A_1}$ and $T \oplus \mathrm{Bel}_{A_1}$ agree on M. In particular, they agree on

$$M' = \{A \cap A_1 | A \in M\} ,$$

which is a subset of M. When $S \oplus \mathrm{Bel}_{A_1}$ and $T \oplus \mathrm{Bel}_{A_1}$ are considered as belief functions over A_1, they are both separable support functions, and M' is the field of subsets generated by the focal elements of $S \oplus \mathrm{Bel}_{A_1}$. Moreover, the number of nonvacuous simple support functions with distinct foci in the decomposition of $S \oplus \mathrm{Bel}_{A_1}$ is less than n. To see this, use (3) to write

$$S \oplus \mathrm{Bel}_{A_1} = (S_1 \oplus \mathrm{Bel}_{A_1}) \oplus \cdots \oplus (S_n \oplus \mathrm{Bel}_{\mathrm{A}_1}) , \tag{7}$$

and recall that the $S_i \oplus \mathrm{Bel}_{A_1}$, considered as belief functions over A_1, are simple support functions. Since A_1 is the focus of S_1, $S_1 \oplus \mathrm{Bel}_{A_1}$ is vacuous, and others of the $S_i \oplus \mathrm{Bel}_{A_1}$ may also be vacuous. If we omit these from the right-hand side of (7), and if we then combine any of the $S_i \oplus \mathrm{Bel}_{A_1}$ that have a common focus to obtain a single simple support function with that focus, then we will have reduced (7) to the unique decomposition of $S \oplus \mathrm{Bel}_{A_1}$ into nonvacuous simple support functions with distinct foci, and the number of these simple support functions will be less than n.

It follows from the inductive hypothesis that

$$W_{S \oplus \mathrm{Bel}_{\mathrm{A}_1}} \leq W_{T \oplus \mathrm{Bel}_{A_1}} .$$

But by (1),

$$W_{S \oplus \mathrm{Bel}_{A_1}} = W_S + \mathrm{Con}\,(S, \mathrm{Bel}_{A_1})$$

and

$$W_{T \oplus \mathrm{Bel}_{A_1}} = W_T \oplus \mathrm{Con}\,(T, \mathrm{Bel}_{A_1}) .$$

And

$$\begin{aligned} \mathrm{Con}\,(S, \mathrm{Bel}_{A_1}) &= -\log\left[1 - S\left(\bar{A}_1\right)\right] \\ &= -\log\left[1 - T\left(\bar{A}_1\right)\right] \\ &= -\mathrm{Con}\,(T, \mathrm{Bel}_{A_1}) . \end{aligned}$$

So $W_S \leq W_T$.

From (5), (4), and Theorem 1, we immediately obtain (6).

3 Support Functions

In this section we show how the weight of internal conflict and the impingement function can be defined for support functions.

Suppose Θ and Ω are two frames of discernment. We call a function $\omega : 2^\Theta \rightarrow 2^\Omega$ a refining if $\{\omega(\theta)|\theta \in \Theta\}$ constitutes a disjoint partition of Ω, and $\omega(A) = \cup\{\omega(\theta)|\theta \in A\}$ for all subsets A of Θ. If $\omega : 2^\Theta \rightarrow 2^\Omega$ is a refining, then we say that Ω is a refinement of Θ and Θ is a coarsening of Ω.

If Bel is a belief function over Ω and $\omega : 2^\Theta \rightarrow 2^\Omega$ is a refining, then the function $\mathrm{Bel} \circ \omega$ is a belief function over Θ. If Bel_1 is a belief function over Ω, Bel_2 is a belief function over Θ, and $\mathrm{Bel}_2 = \mathrm{Bel}_1 \circ \omega$ for some refining ω, we say that Bel_1 is an extension of Bel_2.

If Bel is a belief function over Θ, m is the basic probability assignment for Bel, and $\omega : 2^\Theta \rightarrow 2^\Omega$ is a refining, then the belief function Bel^ω over Ω which is given by the basic probability assignment

$$m^\omega(A) = \begin{cases} m\,(B) & \text{if} \quad B \subseteq \Theta \text{ and } \omega\,(B) = A, \\ 0 & \text{if} \quad \text{there is no } B \subseteq \Theta \text{ such that } \omega\,(B) = A \end{cases} \tag{8}$$

is an extension of Bel to Ω. It is called the vacuous extension of Bel to Ω. It is obviously carried by the image $\omega(2^\Theta)$, which is a field of subsets of Ω.

As (8) makes clear, a belief function and its vacuous extension have the same structure, except that the vacuous extension is embedded in a finer frame. In general, any operation on belief functions on a given frame gives the same result when carried out on the vacuous extensions to a finer frame. For example,

$$\mathrm{Con}(\mathrm{Bel}_1, \ldots, \mathrm{Bel}_n) = \mathrm{Con}(\mathrm{Bel}_1^\omega, \ldots, \mathrm{Bel}_n^\omega, \tag{9}$$

and

$$(\mathrm{Bel}_1 \oplus \cdots \oplus \mathrm{Bel}_n)^\omega = \mathrm{Bel}_1^\omega \oplus \cdots \oplus \mathrm{Bel}_n^\omega. \tag{10}$$

A belief function is called a support function if it can be extended to a separable support function over some refinement. Given a support function S, we let $\mathcal{S}_s$ denote the set of all separable support functions which are extensions of S. We set

$$W_S' - \inf\{W_T | T \in \mathcal{S}_S\},$$

and we define a function V_s' on 2^Θ by

$$V_S'(A) = \inf\{V_T(\omega(A)) | T \in \mathcal{S}_S,\ S = T \circ \omega\}.$$

We would like to call W_s' and V_s' the weight of internal conflict in S and the impingement function for S, respectively. Doing so is justified by the following theorem.

Theorem 3. *If S is a separable support function over Θ, then $W_s' = W_s$, and $V_s' = V_s$.*

Proof. Consider an arbitrary extension T of S. Let $\omega : 2^\Theta \to 2^\Omega$ be the corresponding refining, and let S^ω denote the vacuous extension of S to Ω. Since $S = T \circ \omega = S^\omega \circ \omega$, T and S^ω agree on the field of subsets $\omega(2^\Theta)$. Since the focal elements of S^ω are all in $\omega(2^\Theta)$, it follows that T and S^ω agree on the field M generated by the focal elements of S^ω. Therefore, by Theorem 2, $W_{S^\omega} \leq W_T$ and $V_{S^\omega} \leq V_T$. Since T was an arbitrary element of $\mathcal{S}_S$, and since S^ω is in $\mathcal{S}_S$, it follows that $W'_S = W_{S^\omega}$ and $V'_S = V_{S^\omega}$. On the other hand, it is clear from (9) and (10) that $W_{S^\omega} = W_S$ and $V_{S^\omega} = V_S$.

4 The Weight-of-Conflict Conjecture

Shafer [1] was unable to prove Theorem 3 because he did not have Theorem 2 available. His attempt to prove Theorem 3 led him to formulate the weight-of-conflict conjecture: if the commonality functions Q_1 and Q_2 for two separable support functions S_1 and S_2 satisfy $Q_1 \geq Q_2$, then $W_{S_1} \leq W_{S_2}$. By reasoning equivalent to that in the proof of Theorem 1, he showed that this conjecture implied Theorem 3.

The results in this paper do not tell us whether Shafer's conjecture is true. They do show, however, that the conjecture is not needed for Shafer's purposes.

Acknowledgment

The author would like to thank Glenn Shafer for his help in revising this paper.

Reference

1. Glenn Shafer, *A Mathematical Theory of Evidence*, Princeton U.P.

16

A Framework for Evidential-Reasoning Systems*

John D. Lowrance, Thomas D. Garvey and Thomas M. Strat

Abstract. Evidential reasoning is a body of techniques that supports automated reasoning from evidence. It is based upon the Dempster-Shafer theory of belief functions. Both the formal basis and a framework for the implementation of automated reasoning systems based upon these techniques are presented. The formal and practical approaches are divided into four parts (1) specifying a set of distinct propositional spaces, each of which delimits a set of possible world situations (2) specifying the interrelationships among these propositional spaces (3) representing bodies of evidence as belief distributions over these propositional spaces and (4) establishing paths for the bodies of evidence to move through these propositional spaces by means of evidential operations, eventually converging on spaces where the target questions can be answered.

1 Introduction

For the past several years, we have been addressing perceptual problems that bridge the gap between low-level sensing and high-level reasoning [9, 5, 13, 12, 14, 18]. Problems that fall into this gap are often characterized by multiple evidential sources of real-time data, which must be properly integrated with general knowledge about the world to provide an understanding of the situation that is sufficiently rich to support high-level goals. In this paper, we describe a formal framework for reasoning with perceptual data that forms the basis for evidential-reasoning[1] systems.

The information required to understand the current state of the world comes from multiple sources: real-time sensor data, previously stored general

* This research was sponsored in part by the U.S. Navy Space and Naval Warfare Systems Command and the Defense Advanced Research Project Agency under contract N00039-83-K-0656 and by the U.S. Army Signal Warfare Center under contract DAAL02-85-C-0082.

[1] *Evidential reasoning* is a term coined by SRI International [11] to denote the body of techniques specifically designed for manipulating and reasoning from evidential information as characterized in this paper.

knowledge, and current contextual information. Sensors typically provide *evidence* in support of certain conclusions. Evidence is characteristically uncertain: it allows for multiple possible explanations; it is incomplete: the source rarely has a full view of the situation; and it may be completely or partially incorrect. The quality and the ease with which situational information may be extracted from a synthesis of current sensor data and prestored knowledge is a function both of how strongly the characteristics of the sensed data focus on appropriate intermediate conclusions and on the strength and effectiveness of the relations between those conclusions and situation events.

Given its characteristics, evidence is not readily represented either by logical formalisms or by classical probabilistic estimates. Because of this, developers of automated systems that must reason from evidence have frequently turned to informal, heuristic methods for handling uncertain information. The "probabilities" produced by these informal approaches often cause difficulties in interpretation. The lack of a formally consistent method can cause problems in extending the capabilities of such systems effectively. Our work in evidential reasoning was motivated by these shortcomings. Our theory is based on the Shafer-Dempster theory of evidence [3, 15, 16] and aims to overcome some of the difficulties in reasoning from evidence by providing a natural representation for evidential information, a formal basis for drawing conclusions from evidence, and a representation for belief.

In evidential reasoning, a *knowledge source* (KS) is allowed to express probabalistic opinions about the (partial) truth or falsity of statements composed of subsets of propositions from a space of distinct, exhaustive possibilities (called the *frame of discernment*). The theory allows a KS to assign belief to the individual propositions in the space or to disjunctions of these propositions or both. When it assigns belief to a disjunction, a KS is explicitly stating that it does not have enough information to distribute this belief more precisely. This condition has the attractive feature of enabling a KS to distribute its belief to statements whose granularity is appropriate to its state of knowledge. Also, the statements to which belief is assigned are not required to be distinct from one another. The distribution of beliefs over a frame of discernment is called a *body of evidence.*

Evidential reasoning provides a formal method, *Dempster's Rule of Combination*, for fusing (i.e., pooling) two bodies of evidence. The result is a new body of evidence representing the consensus of the two original bodies of evidence, which may in turn be combined with other evidence. Because belief may be associated directly with a disjunction of propositions, the probability in any selected proposition is typically underconstrained. This necessitates an interval measure of belief, because belief associated with a disjunction may, based upon additional information, devolve entirely upon any one of the disjuncts. Thus, an interval associated with a proposition implies that the true probability associated with that proposition must fall somewhere in the interval. A side-effect of applying Dempster's rule is a measure of *conflict* between the two bodies of evidence that provides a means for detecting possible gross errors in the information.

Current expert-systems technology is most effective when domain knowledge can be modeled as a set of loosely interconnected concepts (i.e., propositions) [2]; this loose interconnection justifies an *incremental* approach to updating beliefs. In most of our work, there is the potential for strong interconnectivity among beliefs in propositions. We, therefore, focus on a body of evidence as a primitive, meaningful collection of interrelated (dependent) beliefs; updating the belief in one proposition affects the entire body of evidence (other work has addressed the concept of a body of evidence in a production-rule formalism [6, 7] by creating special entities).

Evidential reasoning provides options for the representation of information: independent opinions are expressed by multiple (independent) bodies of evidence; dependent opinions (in which belief in one proposition depends on that of another) can either be expressed by a single body of evidence or by a network that describes the interrelationships among several bodies of evidence. These networks of bodies of evidence capture the geneology of each body (similar in spirit to those of [1]) and are used in a manner similar to data-flow models [17] updating interrelated beliefs (i.e., for belief revision [4]).

In this paper we assume some familiarity with the Dempster-Shafer theory of beliefs, although the appropriate equations from this theory are included. We begin with a discussion of the formal approach to the problem of reasoning from evidence and then progress to a description of the implementation approach, including an example. We close with a short description of the system that we have developed for applying evidential reasoning.

2 Formal Approach

2.1 Framing the Problem

The first step in applying evidential reasoning to a given problem is to delimit a propositional space of possible situations. Within the theory of belief functions, this propositional space is called the *frame of discernment.* It is so named because all bodies of evidence are expressed relative to this surrounding framework, and it is through this framework that the interaction of the evidence is discerned. A frame of discernment delimits a set of possible situations, exactly one of which is true at any one time. For example, the problem to be addressed is that of locating a ship. In this case, the frame of discernment consists of the set of all possible locations for that vessel. This might be represented by a set Θ_A in which each element a_i corresponds to a possible location:

$$\Theta_A = \{a_1, a_2, \ldots, a_n\} \quad .$$

Once a frame of discernment has been established, propositional statements can be represented by disjunctions of elements from the frame corresponding to those situations for which the statements are true. For example, the proposition A_i might correspond to the statement that the vessel is located in port, in which case A_i would be represented by the subset of elements from Θ_A that correspond to possible locations within port facilities:

$$\mathsf{A}_i \subseteq \Theta_A \quad .$$

Other propositions related to locating this vessel can be similarly represented as subsets of Θ_A (i.e., as elements of the power set of Θ_A, denoted 2^{Θ_A}). Once this has been accomplished, logical questions can be posed and resolved in terms of the frame. Given two propositions, A_i and A_j, the following logical operations and relation can be resolved through the associated set operations and relation:

$$\begin{aligned} \neg\mathsf{A}_i &\Longleftrightarrow \Theta_A - \mathsf{A}_i \\ \mathsf{A}_i \wedge \mathsf{A}_j &\Longleftrightarrow \mathsf{A}_i \cap \mathsf{A}_j \\ \mathsf{A}_i \vee \mathsf{A}_j &\Longleftrightarrow \mathsf{A}_i \cup \mathsf{A}_j \\ \mathsf{A}_i \Rightarrow \mathsf{A}_j &\Longleftrightarrow \mathsf{A}_i \subseteq \mathsf{A}_j \quad . \end{aligned}$$

If other aspects of ships are of interest besides their location, then additional frames of discernment might be defined. For example, the activities of these ships might be of interest. If so, an additional frame Θ_B might be defined to include elements corresponding to refueling, loading cargo, unloading cargo, being enroute, and the like. Propositional statements pertaining to a ship's activity can then be defined relative to this frame; e.g.,

$$\begin{aligned} &\Theta_B = \{b_1, b_2, \ldots, b_n\} \\ &\mathsf{B}_j \subseteq \Theta_B \quad . \end{aligned}$$

So far, propositional statements pertaining to a ship's location or pertaining to its activity can be addressed separately, but they cannot be jointly considered. To do this, one must first define a *compatibility relation* between the two frames. A compatibility relation simply describes which elements from the two frames can be true simultaneously. For example, a ship located at a loading dock might be loading or unloading cargo, but is not refueling, or enroute. In other words, being located at a loading dock is only compatible with one of two activities, loading or unloading. Thus, the compatibility relation between frames Θ_A and Θ_B is a subset of the cross product of the two frames. A pair (a_i, b_j) is included if and only if they can be true simultaneously. There is at least one pair (a_i, b_j) included for each a_i in Θ_A (the analogue is true for each b_j):

$$\varTheta_{A,B} \subseteq \varTheta_A \times \varTheta_B \quad .$$

Using the compatibility relation $\varTheta_{A,B}$ we can define a *compatibility mapping* $C_{A\mapsto B}$ for translating propositional statements expressed relative to $\varTheta_A$ to statements relative to $\varTheta_B$. If a statement A_k is true, then the statement $C_{A\mapsto B}(\mathsf{A}_k)$ is also true:

$$\begin{aligned} &C_{A\mapsto B} : 2^{\varTheta_A} \mapsto 2^{\varTheta_B} \\ &C_{A\mapsto B}(\mathsf{A}_k) = \{b_j | (a_i, b_j) \in \varTheta_{A,B}, a_i \in \mathsf{A}_k\} \quad . \end{aligned}$$

Instead of translating propositional statements between these two frames via $C_{A\mapsto B}$ and $C_{B\mapsto A}$, we might choose to translate these statements to a common frame that captures all of the information. This common frame is identical to the compatibility relation $\varTheta_{A,B}$. Frame $\varTheta_A$ (and analogously $\varTheta_B$) is trivially related to frame $\varTheta_{A,B}$ via the following compatibility relation and compatibility mappings:

$$\begin{aligned} \varTheta_{A,(A,B)} &= \{(a_i, (a_i, b_j)) | (a_i, b_j) \in \varTheta_{A,B}\} \\ C_{A\mapsto(A,B)}(\mathsf{A}_k) &= \{(a_i, b_j) | (a_i, (a_i, b_j)) \in \varTheta_{A,(A,B)}, a_i \in \mathsf{A}_k\} \\ &= \{(a_i, b_j) | (a_i, b_j) \in \varTheta_{A,B}, a_i \in \mathsf{A}_k\} \\ C_{(A,B)\mapsto A}(\mathsf{X}_k) &= \{a_i | (a_i, b_j) \in \varTheta_{A,B}, (a_i, b_j) \in \mathsf{X}_k\} \quad . \end{aligned}$$

Clearly, as more aspects of these ships become of interest, the number and complexity of the frames and compatibility mappings increases. However, there is a trade-off between the complexity of individual frames and the complexity of the network of compatibility mappings connecting them. We might define a single (complex) frame that encompasses all aspects of interest or, alternatively, define a (complex) network of frames that includes a distinct frame for each aspect of interest. Of course, these may not be equivalent. For example, consider the following frame:

$$\varTheta_{A,B,C} = \{(a_1, b_1, c_1), (a_2, b_1, c_2), (a_2, b_2, c_2)\} \quad .$$

If this frame properly captures the relationship among frames $\varTheta_A$, $\varTheta_B$, and $\varTheta_C$, then c_1 is the only element from $\varTheta_C$ compatible with a_1 from $\varTheta_A$. However, if we maintain these as three separate frames connected by compatibility mappings, $C_{A\mapsto B}, C_{B\mapsto A}, C_{B\mapsto C}$, and $C_{C\mapsto B}$, both c_1 and c_2 are compatible with a_1 because a_1 is compatible with b_1, and b_1 is compatible with both c_1 and c_2; i.e., $C_{B\mapsto C}(C_{A\mapsto B}(\{a_1\})) = \{c_1, c_2\}$. However, if a_1 is true, then it follows that either c_1 or c_2 is true. Thus, the reasoning based on a well-formed *gallery* of interconnected frames is sound but not necessarily complete. A gallery is well formed if there exists a single all encompassing frame whose answers are always included in the answers based upon the gallery.

In dynamic environments, compatibility relations can be used to reason over time. If Θ_{A1} represents the possible states of the world at time one and Θ_{A2} represents the possible states at time two, then a compatibility relation, $\Theta_{A1,A2}$, can capture the possible state transitions. For example, Θ_{A1} and Θ_{A2} might both represent the possible locations of a ship (i.e., they are identical to Θ_A as previously defined), then $\Theta_{A1,A2}$ could represent the constraints on that ship's movement. A pair of locations (a_i, a_j) would be included in $\Theta_{A1,A2}$ if a ship located at a_i on Day 1 (i.e., time) could reach a_j by Day 2. If we assume that the possible movements of a ship are constrained in the same way over any two day period, then the compatibility mapping associated with this compatibility relation can be reapplied as many times as necessary to constrain the possible locations of a ship across an arbitrary number of days.

2.2 Analyzing the Evidence

Once a gallery has been established, the available evidence can be analyzed. The goal of this analysis is to establish a line of reasoning, based upon both the possibilistic information in the gallery and the probabilistic information from the evidence that determines the most likely answers to some questions. The gallery delimits the space of possible situations, and the evidential information establishes the likelihoods of these possibilities. Within an analysis, bodies of evidence are expressed relative to frames in the gallery, and paths are established for the bodies of evidence to move through the frames via the compatibility mappings. An analysis also specifies if other evidential operations are to be performed, including whether multiple bodies of evidence are to be combined when they arrive at common frames. Finally, an analysis specifies which frame and ultimate bodies of evidence are to be used to answer each target question. Thus, an analysis specifies a means of arguing from multiple bodies of evidence towards a particular (probabilistic) conclusion. An analysis, in an evidential context, is the analogue of a proof tree in a logical context.

To begin, each body of evidence is expressed relative to a frame in the gallery. Each is represented as a mass distribution (e.g., m_A) over propositional statements discerned by a frame (e.g., Θ_A):

$$\begin{aligned} m_A : 2^{\Theta_A} &\mapsto [0,1] \\ \sum_{\mathsf{A}_i \subseteq \Theta_A} m_A(\mathsf{A}_i) &= 1 \\ m_A(\emptyset) &= 0 \quad . \end{aligned}$$

Intuitively, mass is attributed to the most precise propositions a body of evidence supports. If a portion of mass is attributed to a proposition A_i, it represents a minimal commitment to that proposition and all the propositions it implies. Additional mass attributed to a proposition A_j that is compatible with A_i, but does not imply it (i.e., $\emptyset \neq \mathsf{A}_i \cap \mathsf{A}_j \neq \mathsf{A}_j$), represents a potential commitment: mass that neither supports nor denies that proposition at present but might later move either way based upon additional information.

To *interpret* this body of evidence relative to the question A_j, we calculate its *support* and *plausibility* to derive its *evidential interval* as follows:

$$\begin{aligned} Spt(\mathsf{A}_j) &= \sum_{\mathsf{A}_i \subseteq \mathsf{A}_j} m_A(\mathsf{A}_i) \\ Pls(\mathsf{A}_j) &= 1 - Spt(\Theta_A - \mathsf{A}_j) \\ [Spt(\mathsf{A}_j), Pls(\mathsf{A}_j)] &\subseteq [0,1] \quad . \end{aligned}$$

The lower bound of an evidential interval indicates the degree to which the evidence supports the proposition, while the upper bound indicates the degree to which the evidence fails to refute the proposition, i.e., the degree to which it remains plausible. This evidential interval, for the most part, corresponds to bounds on the probability of A_j. Thus, complete ignorance is represented by an evidential interval of $[0.0, 1.0]$ and a precise probability assignment is represented by the "interval" collapsed about that point (e.g., $[0.7, 0.7]$). Other degrees of ignorance are captured by evidential intervals with widths other than 0 or 1 (e.g., $[0.6, 0.8], [0.0, 0.5], [0.9, 1.0]$).

If a body of evidence is to be interpreted relative to a question expressed over a different frame from the one over which the evidence is expressed, a path of compatibility relations connecting the two frames is required. The mass distribution expressing the body of evidence is then repeatedly *translated* from frame to frame, via compatibility mappings, until it reaches the ultimate frame of the question. In translating m_A from frame Θ_A to frame Θ_B via compatibility mapping $C_{A \mapsto B}$, the following computation is applied to derive the translated mass distribution m_B:

$$m_B(\mathsf{B}_j) = \sum_{C_{A \mapsto B}(\mathsf{A}_i) = \mathsf{B}_j} m_A(\mathsf{A}_i) \quad .$$

Intuitively, if we (partially) believe A_i, and A_i implies B_j, then we should have the same (partial) belief in B_j. This same method is applied to move mass distributions among frames that represent states of the world at different times. However, when this is the case, the operation is called *projection.*

Once two mass distributions m_A^1 and m_A^2 representing independent opinions are expressed relative to the same frame of discernment, they can be *fused* (i.e., combined) using *Dempster's Rule of Combination.* Dempster's rule pools mass distributions to produce a new mass distribution m_A^3 that represents the consensus of the original disparate opinions. That is, Dempster's rule produces a new mass distribution that leans towards points of agreement between the original opinions and away from points of disagreement. Dempster's rule is defined as follows:

$$m^3_A(\mathsf{A}_k) = (1-k)^{-1} \sum_{\mathsf{A}_i \cap \mathsf{A}_j = \mathsf{A}_k} m^1_A(\mathsf{A}_i) m^2_A(\mathsf{A}_j)$$

$$k = \sum_{\mathsf{A}_i \cap \mathsf{A}_j = \emptyset} m^1_A(\mathsf{A}_i) m^2_A(\mathsf{A}_j) \neq 1 \quad .$$

Since Dempster's rule is both commutative and associative, multiple (independent) bodies of evidence can be combined in any order without affecting the result. If the initial bodies of evidence are independent, then the derivative bodies of evidence are independent as long as they share no common ancestors. Thus, in the course of constructing an analysis, attention must be paid to the way that evidence is propagated and combined to guarantee the independence of the evidence at each combination.

Other evidential operations can also be included in an analysis. One frequently used operation is *discounting.* This operation adjusts a mass distribution to reflect its source's credibility (expressed as a discount rate $r \in [0,1]$). If a source is completely reliable ($r = 0$), discounting has no effect; if it is completely unreliable ($r = 1$), discounting strips away all apparent information content; otherwise, discounting lowers the apparent information content in proportion to the source's unreliability:

$$m^{\%}_A(\mathsf{A}_i) = \begin{cases} (1-r)m_A(\mathsf{A}_i), & \mathsf{A}_i \neq \Theta_A \\ r + (1-r)m_A(\Theta_A), & \text{otherwise} \end{cases} \quad .$$

Other evidential operations include *summarization* and *gisting* (among others). Summarization eliminates extraneous details from a mass distribution by collecting all of the extremely small amounts of mass attributed to propositions and attributing the sum to the disjunction of those propositions. Gisting produces the "central" Boolean-valued statement that captures the essence of a mass distribution. This is particularly useful when explaining lines of reasoning.

3 Implementation Approach

In implementing this formal approach, we have found that the gallery, frames, compatibility relations, and analyses can all be represented straightforwardly as graphs consisting of nodes connected by directed edges. This has led us to use **Grasper II**^{TM} [10, 8], a programming language extension to LISP that introduces graphs as a primitive data type. A graph in Grasper II consists of a set of labeled subgraphs. Each subgraph consists of a set of labeled nodes and a set of labeled, directed edges that connect pairs of nodes. Each node, edge, and subgraph have values that can be used as general repositories for information. Once the graphical representations have been established for the

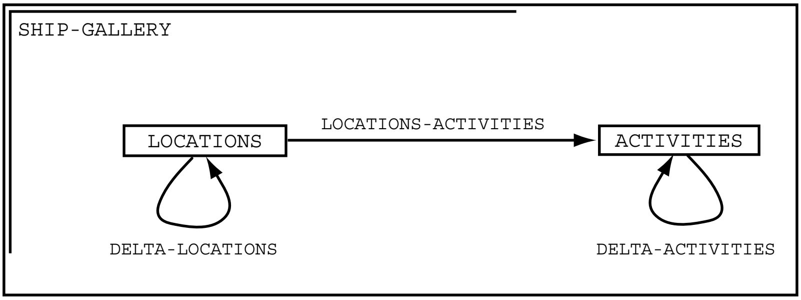


Fig. 1. SHIP-GALLERY gallery

gallery, frames, compatibility relations, and analyses, the remainder of the formal approach is easily implemented.

The first step is to define the gallery. If the problem is to reason about the locations and activities of ships, we might include two frames: a LOCATIONS frame and an ACTIVITIES frame. These are each represented as nodes in a subgraph called the SHIP-GALLERY (Fig. 1). In addition, the gallery might include three compatibility relations represented as edges. One compatibility relation, LOCATIONS-ACTIVITIES, relates locations to activities and is represented by an edge from LOCATIONS to ACTIVITIES. The two other compatibility relations, DELTA-LOCATIONS and DELTA-ACTIVITIES, describe how a ship's location and activity on one day are related to the next day's. Each of these is represented by an edge that begins and ends at the same node.

The next step is to define the frames in the gallery. Each of these is represented by a subgraph sharing the same name as a node from the gallery. Each such subgraph includes a node for each element of the frame and may include additional nodes representing aliases, i.e., named disjunctions of elements. Each of these additional nodes have edges pointing to elements of the frame (or other aliases) that make up the disjunction. The LOCATIONS frame (Fig. 2)

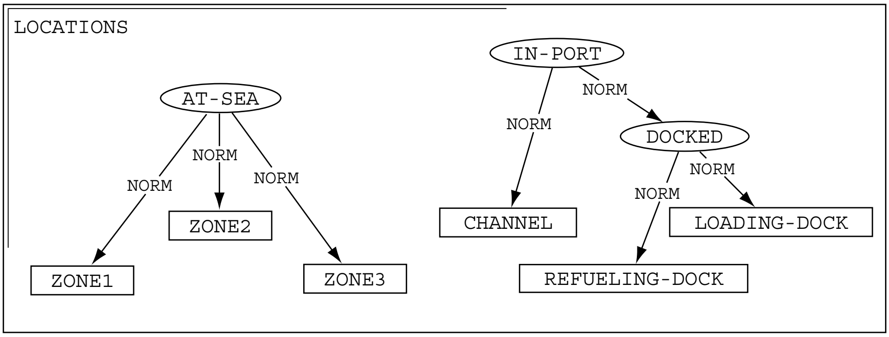


Fig. 2. LOCATIONS frame

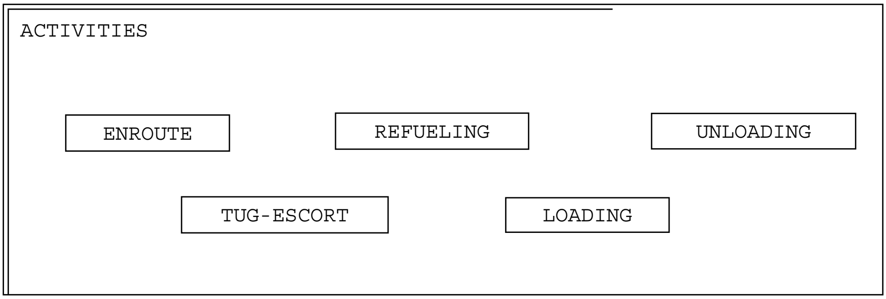


Fig. 3. ACTIVITIES frame

includes six elements (ZONE1, ZONE2, ZONE3, CHANNEL, LOADING-DOCK, REFUELING-DOCK) and three aliases (IN-PORT, DOCKED, AT-SEA). The ACTIVITIES frame (Fig. 3) includes five elements (ENROUTE, TUG-ESCORT, UNLOADING, LOADING, REFUELING).

Each compatibility relation in the gallery is represented as a subgraph that includes the nodes from the frames that they relate with edges connecting compatible elements. For example, in the LOCATIONS-ACTIVITIES compatibility relation (Fig. 4), ZONE1, ZONE2, and ZONE3 are all connected to ENROUTE (becuase these zones represent areas at sea), CHANNEL is connected to TUG-ESCORT (because a ship entering or leaving the port at the end of this channel would be under tugboat control), LOADING-DOCK is connected to both LOADING and UNLOADING (because either activity

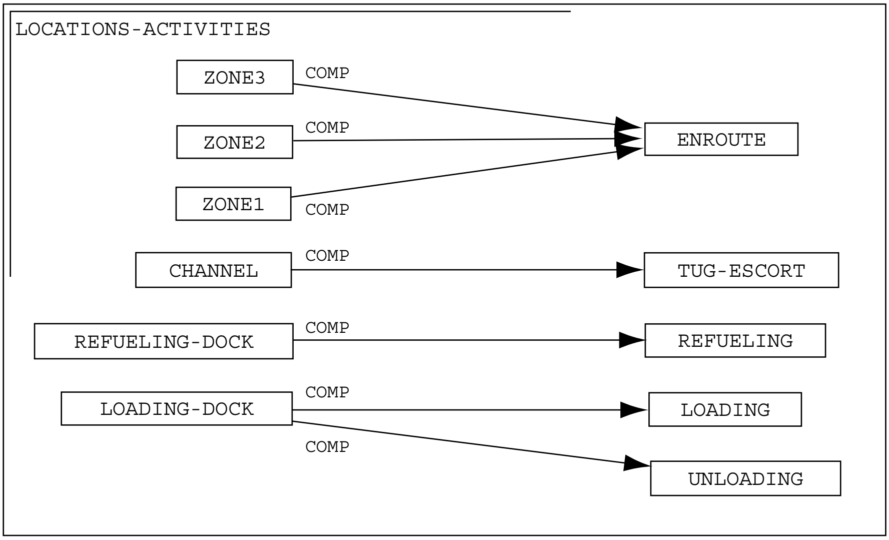


Fig. 4. LOCATIONS-ACTIVITIES compatibility relation

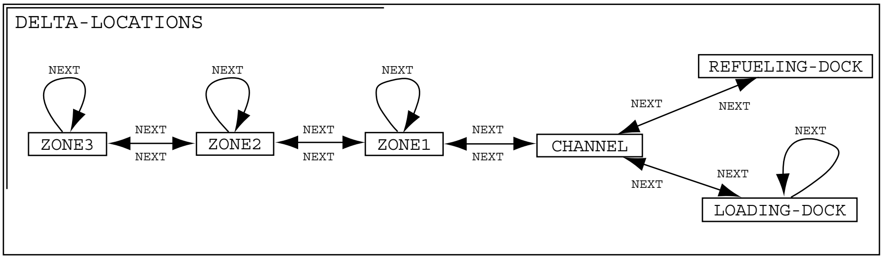


Fig. 5. DELTA-LOCATIONS compatibility relation

is consistent with being at that dock), and REFUELING-DOCK is connected to REFUELING. DELTA-LOCATIONS and DELTA-ACTIVITIES (Figs. 5 and 6) relate frames to themselves. They represent possible state transitions in their respective frames over any two day period. Edges connect compatible elements from one day to the next. DELTA-LOCATIONS indicates that the zones are linearly ordered and that a ship must pass through the channel to get to either the loading or refueling docks. It also indicates that a ship will only remain at the refueling dock or in the channel for one day at a time but may remain anywhere else for any number of days. In DELTA-ACTIVIES it can be seen that a ship must progress through TUG-ESCORT from ENROUTE before proceeding to REFUELING or UNLOADING and that REFUELING and TUG-ESCORT are one day activities. Further, a ship must go through LOADING after UNLOADING before returning to TUG-ESCORT.

After the gallery and its supporting frames and compatibility relations have been established, evidential analyses can be constructed. These analyses are represented as data-flow graphs where the data and the operations are evidential. Figure 7 is one such analysis. Here primitive bodies of evidence are represented by elliptical nodes and derivative bodies of evidence are represented by circular nodes. Diamond-shaped nodes represent interpretations of bodies of evidence. The values of these nodes are used as repositories for the information (i.e., data) that they represent (Fig. 7). For bodies of evidence this includes a frame of discernment (including the day to which the

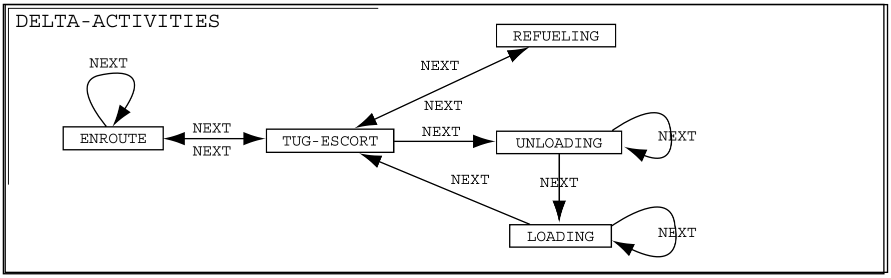


Fig. 6. DELTA-ACTIVITIES compatibility relation

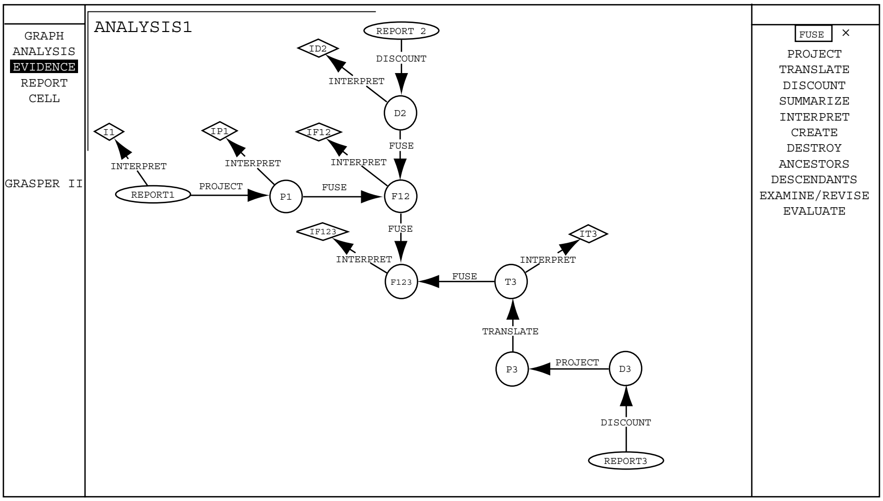


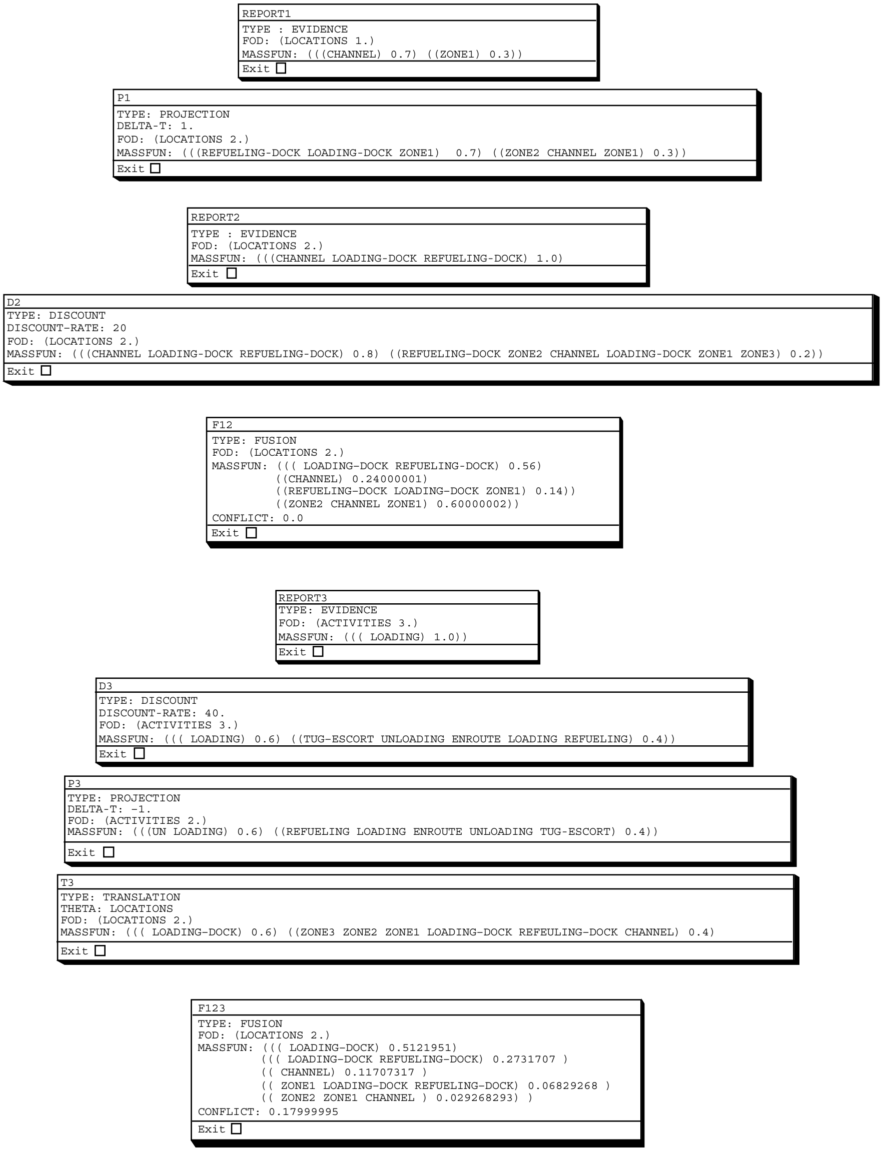
Fig. 7. ANALYSIS1 analysis

evidence pertains), a mass distribution, and other supporting information. Edges pointing to a derivative node are labeled with the evidential operation that is applied to the bodies of evidence, at the other ends of the edges, to derive the body of evidence represented by this node.

In the analysis of a ship in Fig. 8, there are three primitive bodies of evidence. REPORT1 locates the ship on Day 1 saying that there is a 70 percent chance that it can be found in the CHANNEL and a 30 percent chance that it is in ZONE1; REPORT2 says that the ship was IN-PORT on Day 2; and REPORT3 indicates that the ship was LOADING cargo on Day 3. REPORT1 is taken at face value, but REPORT2 and REPORT3 have been discounted by 20 percent and 40 percent, respectively, to derive D2 and D3, reflecting doubt in the credibility of these reports. REPORT1 has been projected forward by one day to derive P1 [2] and then has been fused with D2 to derive a consensus for Day 2, F12. D3 has been projected backwards in time by one day to derive P3 and then has been translated from the ACTIVITIES frame to the LOCATIONS frame. Finally, this result, T3, has been fused with F12 to derive a consensus, based on all three reports, about the ship's location on Day 2.

The interpretation nodes in this analysis track the evidential intervals for some key propositions. I1 is based soley on REPORT1 and indicates that there is precisely a 70 percent chance of the ship being IN-PORT$[0.7, 0.7]$ and no chance of it being DOCKED $[0.0, 0.0]$on Day 1. IP1 indicates that, based soley upon REPORT1, after one day has ellapsed, nothing is known about whether

[2] Note that the distribution at REPORT1 is a Bayesian distribution (i.e., a distribution over exclusive elements), but application of the projection operation results in a non-Bayesian distribution at P1.



```
REPORT1
TYPE : EVIDENCE
FOD: (LOCATIONS 1.)
MASSFUN: (((CHANNEL) 0.7) ((ZONE1) 0.3))
Exit □

P1
TYPE: PROJECTION
DELTA-T: 1.
FOD: (LOCATIONS 2.)
MASSFUN: (((REFUELING-DOCK LOADING-DOCK ZONE1)  0.7) ((ZONE2 CHANNEL ZONE1) 0.3))
Exit □

REPORT2
TYPE : EVIDENCE
FOD: (LOCATIONS 2.)
MASSFUN: (((CHANNEL LOADING-DOCK REFUELING-DOCK) 1.0)
Exit □

D2
TYPE: DISCOUNT
DISCOUNT-RATE: 20
FOD: (LOCATIONS 2.)
MASSFUN: (((CHANNEL LOADING-DOCK REFUELING-DOCK) 0.8) ((REFUELING-DOCK ZONE2 CHANNEL LOADING-DOCK ZONE1 ZONE3) 0.2))
Exit □

F12
TYPE: FUSION
FOD: (LOCATIONS 2.)
MASSFUN: ((( LOADING-DOCK REFUELING-DOCK) 0.56)
         ((CHANNEL) 0.24000001)
         ((REFUELING-DOCK LOADING-DOCK ZONE1) 0.14))
         ((ZONE2 CHANNEL ZONE1) 0.60000002))
CONFLICT: 0.0
Exit □

REPORT3
TYPE: EVIDENCE
FOD: (ACTIVITIES 3.)
MASSFUN: ((( LOADING) 1.0))
Exit □

D3
TYPE: DISCOUNT
DISCOUNT-RATE: 40.
FOD: (ACTIVITIES 3.)
MASSFUN: ((( LOADING) 0.6) ((TUG-ESCORT UNLOADING ENROUTE LOADING REFUELING) 0.4))
Exit □

P3
TYPE: PROJECTION
DELTA-T: -1.
FOD: (ACTIVITIES 2.)
MASSFUN: (((UN LOADING) 0.6) ((REFUELING LOADING ENROUTE UNLOADING TUG-ESCORT) 0.4))
Exit □

T3
TYPE: TRANSLATION
THETA: LOCATIONS
FOD: (LOCATIONS 2.)
MASSFUN: ((( LOADING-DOCK) 0.6) ((ZONE3 ZONE2 ZONE1 LOADING-DOCK REFEULING-DOCK CHANNEL) 0.4)
Exit □

F123
TYPE: FUSION
FOD: (LOCATIONS 2.)
MASSFUN: ((( LOADING-DOCK) 0.5121951)
         ((( LOADING-DOCK REFUELING-DOCK) 0.2731707 )
         (( CHANNEL) 0.11707317 )
         (( ZONE1 LOADING-DOCK REFUELING-DOCK) 0.06829268 )
         (( ZONE2 ZONE1 CHANNEL ) 0.029268293) )
CONFLICT: 0.17999995
Exit □
```

Fig. 8. Data from ANALYSIS1

the ship is IN-PORT $[0.0, 1.0]$, but that it may now be DOCKED $[0.0, .7.0]$. If REPORT2 is included after being discounted, IF12 indicates that there is strong reason to believe that the ship is IN-PORT $[0.8, 1.0]$, but there is conflicting information concerning whether or not it is DOCKED $[0.56, 0.7]$. IT3 indicates that based soley upon REPORT3, after having been discounted, projected backwards a day, and translated to the LOCATION frame, that there is 0.6 support and 1.0 plausibility for both IN-PORT and DOCKED. Finally, when all three reports are considered, IF123 indicates strong belief that the ship is IN-PORT $[0.9, 1.0]$ on Day 2 and a reasonably strong belief, though mixed, that it is also DOCKED $[0.78, 0.85]$.

4 Evidential-Reasoning Systems

To support the construction, modification, and interrogation of evidential analyses, we have developed **Gister**^{TM}. Gister supports an interactive, menu-driven, graphical interface that allows these structures to be easily manipulated. The user simply selects from a menu to add an evidential operation to an analysis, to modify operation parameters (e.g., discount rates), or to change any portion of a gallery including its frames and compatibility relations. In response, Gister updates the analyses.

All of the figures in this paper are actual screen images from Gister. Figure 3 includes the menus for working with analyses. On the left side of the screen is a menu of nouns. The user determines with what class of objects he wishes to work and selects the appropriate noun from the menu. Once a noun has been selected, a menu of verbs appears on the right side of the screen. A selection from this menu invokes the operation corresponding to the selected verb on the previously selected noun. The user then designates the appropriate nodes, edges, and the like for the selected operation.

Unlike other expert systems, Gister is designed as a tool for the domain expert. With this tool, an expert can quickly and flexibly develop a line of reasoning specific to a given domain situation. This differs markedly from other expert systems in which a single line of reasoning is developed by an expert and then is instantiated over different situations by nonexperts.

This approach has been successfully applied to Naval intelligence problems. New work is focusing on adapting this technology to multisource data fusion for the Army.

5 Summary

Evidential reasoning has already been successfully applied to problems in several domains. However, the addition of the compatability relation to the theory of beliefs, the formalization and development of new evidential operators, and the use of graphical representations have greatly improved the overall usefulness and accessibility of these techniques.

References

[1] Paul Cohen, *Heuristic reasoning about uncertainty: An artificial intelligence approach*, Pitman Publishing, Inc., 1985.

[2] R. Davis and J. J. King, *An overview of production systems*, Machine Intelligence 8 (E. Elcock and D. Michie, eds.), Ellis Horwood, Chichester, England, 1977, pp. 300–332.

[3] Arthur P. Dempster, *A generalization of Bayesian inference*, Journal of the Royal Statistical Society **30** (1968), 205–247.

[4] Jon Doyle, *A truth maintenance system*, Readings in Artificial Intelligence (Bonnie Lynn Webber and Nils J. Nilsson, eds.), Tioga Publishing Company, Palo Alto, CA, 1981, pp. 496–516.

[5] Thomas D. Garvey, John D. Lowrance, and Martin A. Fischler, *An inference technique for integrating knowledge from disparate sources*, Proceedings of the Seventh Joint Conference on Artificial Intelligence (Menlo Park, CA), American Association for Artificial Intelligence, August 1981, pp. 319–325.

[6] Kurt Konolige, *Bayesian methods for updating probabilities*, A Computer-Based Consultant for Mineral Exploration, Final Report, SRI Project 6415 (333 Ravenswood Avenue, Menlo Park, CA) (R. O. Duda, P. E. Hart, K. Konolige, and R. Reboh, eds.), 1979.

[7] John F. Lemmer and Stephen W. Barth, *Efficient minimum information updating for bayesian inferencing in expert systems*, Proceedings of the National Conference on Artificial Intelligence (Menlo Park, CA), American Association for Artificial Intelligence, August 1982, pp. 424–427.

[8] John D. Lowrance, *Grasper 1.0 reference manual*, COINS Technical Report 78–20, Department of Computer and Information Science, University of Massachusetts, Amherst, MA, December 1978.

[9] ———, *Dependency-graph models of evidential support*, Ph.D. thesis, Department of Computer and Information Science, University of Massachusetts, Amherst, MA, September 1982.

[10] ———, *Grasper ii reference manual*, Artificial Intelligence Center, SRI International, 333 Ravenswood Avenue, Menlo Park, CA, January 1987.

[11] John D. Lowrance and Thomas D. Garvey, *Evidential reasoning: A developing concept*, Proceedings of the Internation Conference on Cybernetics and Society, Institute of Electrical and Electronical Engineers, October 1982, pp. 6–9.

[12] ———, *Evidential reasoning: An approach to the simulation of a weapons operation center*, Technical report, Artificial Intelligence Center, SRI International, 333 Ravenswood Avenue, Menlo Park, CA, September 1983.

[13] ———, *Evidential reasoning: An implementation for multisensor integration*, Technical Report 307, Artificial Intelligence Center, SRI International, 333 Ravenswood Avenue, Menlo Park, CA, December 1983.

[14] John D. Lowrance, Thomas M. Strat, and Thomas D. Garvey, *Application of artificial intelligence techniques to naval intelligence analysis*, Final Report SRI Contract 6486, Artificial Intelligence Center, SRI International, 333 Ravenswood Avenue, Menlo Park, CA, June 1986.

[15] Glenn Shafer, *A mathematical theory of evidence*, Princeton University Press, Princeton, NJ, 1976.

[16] ———, *Belief functions and possibility measures*, The Analysis of Fuzzy Information **1** (1986), 51–84.

[17] W. W. Wadge and E. A. Ashcroft, *Lucid, the dataflow programming language*, Academic Press U. K., 1984.

[18] Leonard P. Wesley, *Evidential-based control in knowledge-based systems*, Ph.D. thesis, Department of Computer and Information Science, University of Massachusetts, Amherst, MA, 1988.

17

Epistemic Logics, Probability, and the Calculus of Evidence

Enrique H. Ruspini

Abstract. This paper, presents results of the application to epistemic logic structures of the method proposed by Carnap for the development of logical foundations of probability theory. These results, which provide firm conceptual bases for the Dempster-Shafer calculus of evidence, are derived by exclusively using basic concepts from probability and modal logic theories, without resorting to any other theoretical notions or structures.

A form of epistemic logic (equivalent in power to the modal system $S5$), is used to define a space of possible worlds or states of affairs. This space, called the epistemic universe, consists of all possible combined descriptions of the state of the real world and of the state of knowledge that certain rational agents have about it. These representations generalize those derived by Carnap, which were confined exclusively to descriptions of possible states of the real world.

Probabilities defined on certain classes of sets of this universe, representing different states of knowledge about the world, have the properties of the major functions of the Dempster-Shafer calculus of evidence: belief functions and mass assignments. The importance of these epistemic probabilities lies in their ability to represent the effect of uncertain evidence in the states of knowledge of rational agents. Furthermore, if an epistemic probability is extended to a probability function defined over subsets of the epistemic universe that represent true states of the real world, then any such extension must satisfy the well-known interval bounds derived from the Dempster-Shafer theory.

Application of this logic-based approach to problems of knowledge integration results in a general expression, called the additive combination formula, which can be applied to a wide variety of problems of integration of dependent and independent knowledge. Under assumptions of probabilistic independence this formula is equivalent to Dempster's rule of combination.

1 Introduction

The research work presented here was motivated by the need to improve the understanding of issues in the analysis and interpretation of evidence. In the context of this paper, the term evidence is used to describe the information usually imprecise and uncertain, that is conveyed by observations and measurements of real-world systems. We have sought to gain such an understanding by examining the basic concepts, structures, and ideas relevant to the characterization of imprecise and uncertain knowledge.

Our approach is strongly based on Carnap's methodology [1, 2] for the development of logical foundations of probability theory. In his formulation, Carnap developed an universe of possible worlds that encompasses all possible valid states of a real-world system. Information about that system, if precise and certain, identifies its actual state (e.g., a detailed diagnosis of a disease). If imprecise but certain, this information identifies a subset of possible system states (e.g., a number of possible diagnoses). If uncertain, then the information induces a probability distribution over system states (e.g., probability values for specific diagnoses).

It is important to note, however, that in Carnap's characterization no distinction is drawn between degrees of precision or detail when the information is uncertain. This representational shortcoming renders impossible the modeling of information that only assigns degrees of likelihood values to some subsets of possible states (i.e., instead of prescribing those values over all such subsets that are of relevance to the modeler). This type of information, providing some knowledge about the underlying probability distributions but not all the distribution values, is quite common in practical applications (e.g., in a medical diagnosis problem, tests and existing medical knowledge indicate that there is a 60% chance of liver disease but fail to provide any information about the likelihood of individual instances thereof).

Seeking to generalize Carnap's approach to allow for the treatment of this type of uncertain information, we directed our attention to epistemic logics–a form of modal logics developed to deal with problems of representation and manipulation of the states of knowledge of rational agents. Originally studied by Hintikka [6], their use in artificial intelligence problems was proposed by Moore [8]. Recently epistemic logics have also been applied to the design of intelligent robots [11].

In our extension of the Carnapian ideas the starting point is a generalization of Carnap's space of possible worlds, or universe. This generalization, obtained by considering representations of both the state of the world and the knowledge of rational agents, is called the *epistemic universe.* Described in the next section, the epistemic universe contains several interesting and important subset families. Two of these collections have as members truth sets and support sets, which are related, respectively, to different ontological and epistemological properties of possible worlds. Furthermore, these families have the properties of sigma algebras, i.e. the basic domain of definition of probability functions.

Again following Carnap's lead we define probabilities on these sigma algebras and consider their relationships. We differ from Carnap, however, in that we view evidence as generally providing information about the truth of some propositions while failing to give any indication about the truth of others. Evidence is further regarded as a potential modifier of our state of knowledge; accordingly, uncertain evidence is represented as a conventional probability function defined on the algebra of epistemic sets. This probability is then shown to have the structure of the basic functions of the Dempster-Shafer calculus of evidence [3, 14]. Furthermore, if such an epistemic probability is extended to the sigma algebra of the truth sets (representing probabilities of the truth of propositions that describe the world), then the extension must satisfy the bounds of the Dempster-Shafer theory. These bounds correspond to the well-known concepts of lower and upper probability functions and, in this particular regard, our results are in agreement with the characterization made by Suppes [15] of the role of uncertain information in determining the probability distribution values that underlie rational choices in decision problems.

Our approach is also related in several ways to the probabilistic logic approach of Nilsson [10]–the major differences being in the use of epistemic concepts and the derivation of global conditions for probability extension, in contrast to formulas derived from interval probability theory or from approximate-estimation techniques.

In addition, this work has similarities with that of Halpern and McAllester [5]–the dissimilarities in this case being in the methods used to model uncertainty. It is important to note, however, that Halpern and McAllester represent likelihood formally as the probability of knowledge (in the epistemic-logic sense) of propositional truth, using an interpretation that is similar to ours in several significant respects.

Section 4 deals with the problems associated with the combination of the knowledge of several mutually trusting agents. Under assumptions that guarantee that the integrated knowledge is solely the logical consequence of the states of knowledge of the agents, several results are presented, including a general formula for knowledge combination. This additive combination formula may be applied to several knowledge integration problems involving either dependent or independent evidential bodies. For the latter case, the corresponding result generalizes the Dempster's rule of combination.

It is important to emphasize that the results of Sects. 3 and 4, identifying the Dempster-Shafer calculus of evidence with the probability calculus in the epistemic universe, were derived by the direct application of conventional probability theory concepts without having to introduce other multivalued logic notions. The insight gained by using an epistemic model as the basic foundation of the Dempster-Shafer calculus of evidence has made possible the extension of this evidential formalism by the incorporation of new formulas for combining dependent evidence and for utilizing conditional knowledge.

In the exposition that follows, we have not included the proofs of any of the theoretical results obtained in the research being discussed, as such extensive discussion is well outside the scope of this paper. The reader interested in the actual details will find them discussed in a related work [13].

2 The Epistemic Universe

2.1 The Carnapian Universe

Carnap's logical approach to probability starts with the construction of a space of possible worlds that encompasses all valid states of a system of interest. First, all propositions (actually instantiated first-order-Iogic predicates in Carnap's formulation) of relevance to the system $p, q, r, s, ...$, are considered. All possible conjunctions of the type $p \wedge \neg q \wedge \neg r \wedge s \wedge ...$, where every proposition appears only once either as itself or as its negation, are then considered. After discarding logical impossibilities, the resulting set of logical expressions includes all possible system states that may be represented using the propositions $p, q, ...$.

Each such state corresponds to the truth of an atomic proposition about the system in question. These atomic propositions are equivalent to the elementary events introduced in most treatments of basic probability theory. Obviously, by construction, only one such proposition can truly describe the state of the world. The space of atomic propositions, or universe, is therefore a collection of all possible alternative states of the system.

Possible worlds can also be regarded as functions that map each relevant proposition into its truth-value (i.e. true or false) or, alternatively, as subsets of true propositions (i.e., those mapped into the true truth-value). If a possible world is viewed through a "conceptual microscope" as illustrated in Fig. 1, it

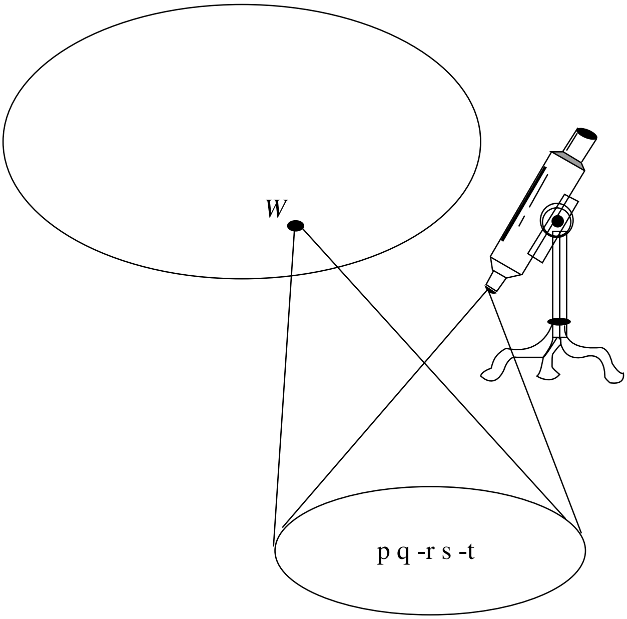


Fig. 1. The carnapian universe under the microscope

can be seen to contain all true propositions in that world, including the negations of those that are false; Two possible worlds will always be different since at least one proposition which is true in one of them will be false in the other.

The space of possible worlds (considered as a probabilistic space) is the basic structure used by Carnap to relate the values of probability functions of subsets associated with relevant propositions on the basis of the logical relationships between those propositions.

2.2 Epistemic Considerations

Carnap's logical approach, while enabling a clearer understanding of the relations between logical and probabilistic concepts, suffers from a major handicap: it assumes that observations of the real world always determine unambiguously probability values for every subset in the universe. This assumption leads inevitably to problems associated with the need to define probability values when the underlying information is not rich enough to furnish them.

If, for example, we have certain (i.e., sure) information that a guest to a party we are hosting is fond of French wine, we would ordinarily consider, in a nonprobabilistic setting, that this information constrains our spectrum of beverage choices (assuming, of course, that we aim to please our guest and are able to do so) without identifying what particular label or vintage he is likely to prefer. If, instead of being sure, our informant is uncertain and believes there is an 80% chance that our caller will like French wine and a 20% chance that he will opt for beer, it is unreasonable (simply because uncertainty has now entered the picture) to assume that this information can be used to assign probabilities for particular choices of wine or beer when before, in a world of certainty, we regarded similar information as being only capable of identifying a subset of possibilities.

These considerations have led to the development of schemes to represent uncertain information as constraints on the values of valid probability distributions. Interval probability theories [16], of which the Dempster-Shafer calculus is a particular case, are important examples of this technique.

The approach we have followed here, however, proceeds from a different logical foundation. Starting from the notion that certain information improves our knowledge by reducing the scope of possible valid states, it considers that uncertain information is associated with a probability function defined on some subsets (actually, a sigma algebra) of the universe, rather than on every subset of the universe. While in the case of certain information we say that we know that the system state is in a subset of possible states, in the case of uncertain information we similarly affirm, with some degree of likelihood, that state is in certain region of the universe. The corresponding probability values constrain the values of other probability functions defined over richer subset collections (i.e., probability extensions).

To identify a model that constitutes the basis for defining probabilities that take values over epistemic structure, we must look at abstract formalisms that allow proper differentiation between states of the world and states of knowledge. This framework is provided by epistemic logics.

2.3 Epistemic Logics and Epistemic Universes

The starting point for our generalization of the Carnapian universe is again a collection of propositions about the real world, denoted by p, q, r, s,... We consider, in addition, more complex propositions obtained therefrom by negation, conjunction, and disjunction. The resulting set of propositions is called a frame of discernment. Each of its members, describing a state of the world, is called an objective proposition or objective sentence.

In addition to objective sentences, we shall also deal with propositions that represent states of knowledge about the real world. When only one rational agent is concerned, the simplest of these epistemic propositions are denoted by $Kp, Kq, Kr, ...$, representing knowledge of their corresponding objective counterparts. We shall also consider expressions formed by combination of epistemic and objective propositions through disjunction, conjunction, implication, and negation, as in the examples $\neg Kr$, or $p \vee K(q \vee Ks)$. The set of all such propositions, which encompasses the frame of discernment as a subset, is called the sentence space, denoted by S. .

The next step in constructing an extension of the Carnapian universe is the generation of all possible states by the assignment of truth-values to propositions in the sentence space. In addition to compliance with the axioms of ordinary propositional logic, we shall also need the following axioms, which supply the unary operator K with the required epistemilogical semantics:

E1 If Kp is true, then p is true.
E2 If Kp is true, then KKp is true (positive introspection).
E3 If $K(p- > q)$ is true, then $Kp \rightarrow Kq$ is also true.
E4 If $\neg Kp$ is true, then $K\neg Kp$ is true (negative introspection).
E5 If p is an axiom, then Kp is an axiom.

This system is equivalent to the modal logic system $S5$ [7].

The space of possible worlds generated on the basis of the above schemata is called the epistemic universe and is denoted by $U(S)$. When seen through our imaginary conceptual microscope, as shown in Fig. 2, each possible world includes, as before, all objective propositions that are true in that world. Each possible world, however, includes also all true epistemic propositions representing knowledge of the truth (e.g., Kp) or falsehood ($K\neg p$) of propositions and, in addition, propositions describing ignorance regarding the truth or falsehood of certain propositions (e.g., $\neg Kp \wedge \neg K\neg p$).

It is important to note that, in the epistemic universe, possible worlds may share the same set of true objective propositions, even though the states of knowledge (i.e., true epistemic propositions) will be different in each case.

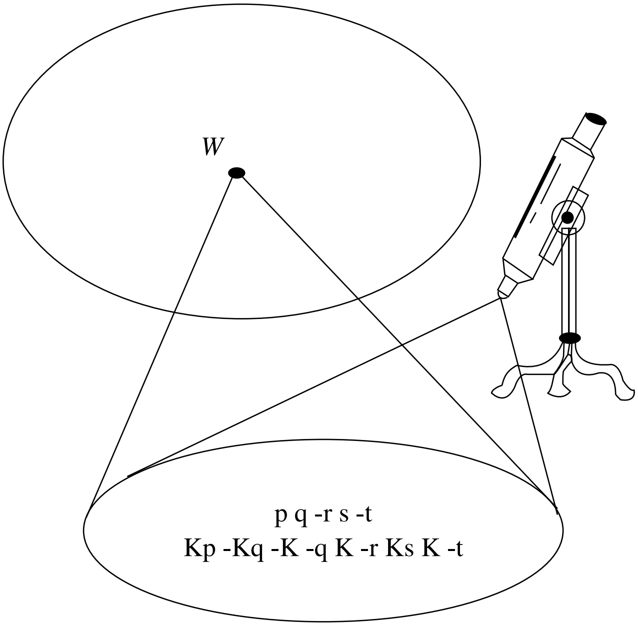


Fig. 2. The epistemic universe under the microscope

In the remainder of this work we will require to employ two important relations.

The first, called logical implication and denoted by $\Longrightarrow$, holds between propositions in sentence space. This relation, well known in modal logic, is used to indicate the fact that in any possible world the truth of some proposition implies that of another. In other words, if $p \Longrightarrow q$, then it is logically impossible for q to be false if p is true.

The second relation, called the accessibility relation and denoted by $\sim$, holds between possible worlds in the epistemic universe. Two possible worlds are related through the accessibility relation if the same epistemic propositions are true in both worlds. Clearly, such world pairs cannot be discriminated on the basis of the information (i.e. knowledge) available in each of them.

2.4 Special Sets in the Epistemic Universe

Several subsets of the epistemic universe are of importance in the definition of probability functions that adequately represent the effects of uncertain evidence in knowledge states.

The subset of all possible worlds where an objective proposition p is known to be true, i.e. in which the epistemic sentence Kp is true, is called the support set of p and is denoted by $k(p)$.

The epistemic set for an objective proposition p is the set of all possible worlds in which p is the most specific proposition that is known to be true (i.e., p is the conjunction of all objective propositions q such that Kq is true). The epistemic set $e(p)$ consists of possible worlds where Kq is true if and only if q is logically implied by p, i.e., $p \Longrightarrow q$. Pairs of possible worlds in the same epistemic set are always related by the accessibility relation $\sim$.

Epistemic sets and support sets are related by the set equation

$$k(p) = \cup_{q \Longrightarrow p} e(q) \tag{1}$$

which is of essential importance to establish the relationship between epistemic constructs and the Dempster-Shafer calculus. Epistemic sets corresponding to different propositions (i.e., those that are not logically equivalent, denoted simply by $\neq$ in this work) are disjoint. The above expression, therefore, represents the disjoint partition of support sets in terms of epistemic sets. Furthermore it can be proved that

$$e(p) = k(p) \cap \bigcup_{\substack{q \Longrightarrow p \\ q \neq p}} [\overline{k(q)}] \tag{2}$$

Finally, truth sets are important subsets of the epistemic universe that are directly related to the truth of objective, rather than epistemic, propositions. The truth set $t(p)$ for an objective proposition p is the collection of all possible worlds where the proposition p is true.

Since p is true in a possible world W whenever Kp is true in W, then it follows that the support set $k(p)$ is a subset of the truth set $t(p)$. It is also true that $k(p)$ is the largest support set contained in $t(p)$.

The inclusion relations between truth, support and epistemic sets are graphically illustrated in Fig. 3. This figure shows the truth set $t(p)$ for a proposition p; its corresponding support set $k(p)$; and the epistemic sets for several propositions which imply p (including the epistemic set for p itself). As noted before, epistemic sets $e(q)$ for propositions q that do not imply p are disjoint from the support set $k(p)$ and intersect the complement $\overline{t(p)}$ of the truth-set $t(p)$.

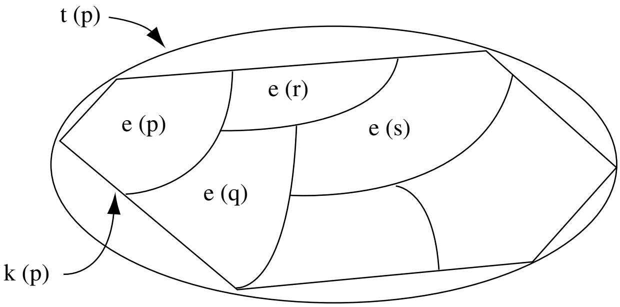


Fig. 3. Relations between epistemic, support, and truth sets

3 Epistemic Probabilities

3.1 Sigma Algebras

The collections of subsets defined in the previous section are of particular importance in a number of respects.

First, epistemic and support sets have a clear epistemological interpretation as representations of similar states of certain (i.e., sure) knowledge. Furthermore, the effect of uncertain information on states of knowledge can be represented by probability values assigned to these sets.

Truth sets, on the other hand, represent states of the world that share some ontological property. Probability values assigned to these sets represent the likelihood of certain events in the real world, namely, the truth of the proposition associated with the truth set. Because of the relations between knowledge and truth embodied in the axiom schema (E), these probability values can be expected to bear some relation, however, to probability values over support and epistemic sets. This relationship is discussed below.

Truth sets, on one hand, and epistemic and support sets, on the other, generate (by union, intersection, and complementation) sigma algebras of the epistemic universe, called the truth algebra and the epistemic algebra, respectively. Sigma algebras are the proper domain of definition for probability functions. This fact has often been ignored in the past when, usually for the sake of simplicity, probabilities have been assumed to be defined on every subset of some space. Consideration of the proper domain of definition for probabilities is, however, a most important issue in probability theory (e.g., when relating joint and marginal distributions).

3.2 Probabilities, Supports and Masses

A probability function defined over the sigma algebra of support and epistemic sets is called an epistemic probability. Epistemic probabilities represent the effect of uncertain evidence on a rational agent's state of knowledge. This effect can always be represented without ambiguity as the result of either previous experience or rational considerations. Under conditions of perfect probabilistic information (in conventional approaches this is assumed to be always available) the corresponding probability is defined for each atomic proposition. At the opposite end, the vacuous epistemic probability function assigns a probability of 1 to the epistemic set $e(U)$ and a probability of 0 to every other subset (i.e., the evidence does not convey any information).

Two functions, both defined in the frame of discernment, can be associated in a natural manner with an epistemic probability.

The first of these, called a mass function and denoted by m, is defined by the expression

$$m(p) = P(e(p)), \tag{3}$$

i.e., as the probability of the epistemic set associated with the objective proposition p.

The second function is called the support function and is denoted by S. It is defined by the expression

$$S(p) = P(k(p)). \tag{4}$$

Support functions and mass functions are related by the equation

$$S(p) = \sum_{q \Longrightarrow p} m(q), \tag{5}$$

which is valid for every objective proposition p in the frame of discernment. From this basic equation, by using results from combinatorial theory [4], it is possible to show that S and m are belief and mass functions, respectively, in the sense of Shafer [14].

In particular, it may be seen that m is expressed in terms of values of the support function S by the equation

$$m(p) = \sum_{q \Longrightarrow p} (-1)^{|p-q|} S(q), \tag{6}$$

where $|p - q|$ is the number of different (i.e., not logically equivalent) propositions r such that $q \Longrightarrow r \Longrightarrow p$, and where the sum is over all propositions q that imply p.

Furthermore, the following inequality, utilized by Shafer as an axiom for belief functions, can be derived as a necessary and sufficient condition characterizing support functions:

$$S(p_1 \vee \ldots \vee p_n) \geq \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ I \neq \phi}} (-1)^{|I|+1} S(\bigwedge_{i \in I} p_i) \tag{7}$$

where $|I|$ is the cardinality of the index subset I.

It is important to emphasize that the epistemic probability P associated with mass and support functions is a conventional probability defined on the epistemic algebra of the epistemic universe.

3.3 Lower and Upper Probabilities

Since both truth sets and epistemic sets are subsets of the epistemic universe, it is reasonable to ask what kind of relations exists between the probability values of members of either class. Answers to this question are obtained by considering the problems associated with the extension of an epistemic probability to a probability function defined over the truth algebra.

The problem of probability extension has received a great deal of attention in probability theory (see, for example, [9]). The standard procedure for its solution is to define lower and upper probabilities for sets not included in the domain of definition (i.e., sigma algebra) of the probability function being extended.

The lower probability of a set X is the probability of the largest subset of the sigma algebra (i.e., where the probability is actually defined) contained in X. Similarly, the upper probability of X is the probability of the smallest measurable subset that contains X.

If P_* and P^* denote the lower and upper probability functions, respectively, then well-known results of probability theory state that probability extensions P always exist and that the value $P(X)$ satisfies the inequality constraints

$$P_*(X) \leq P(X) \leq P^*(X) \tag{8}$$

In addition, the bounds provided by P_* and P^* may always be attained by some extension and are therefore the best possible.

If these basic theoretical results are applied to the epistemic universe, it can be seen that the value $P(t(p))$ of any epistemic probability extension P on the truth set $t(p)$ must satisfy the inequality

$$S(p) \leq P(t(p)) \leq Pl(p) \tag{9}$$

where Pl is the plausibility function of Shafer, defined by

$$Pl(p) = 1 - S(\neg p) = P(\overline{k(\neg p)}) \tag{10}$$

These basic results confirms the validity of the welI-known interval bounds of the Dempster-Shafer calculus.

Furthermore, lower and upper probabilities provide a general methodology to assess the impact of evidence upon understanding of the real-world state. The basic approach, according to these results, consists of representing knowledge as probabilities in an appropriate epistemic algebra, followed by estimation of the values of the lower and upper probabilities of truth sets.

4 Combination of Knowledge

This section briefly describes the results of investigations concerning the combination of the uncertain knowledge of several rational agents. For the sake of simplicity the results presented here are confined to problems involving the combination of the knowledge of two agents (Extensions to an arbitrary number of agents being straightforward).

Each of these two agents is assumed to have obtained information about the state of the world through observation devices that may possibly be dependent or correlated to some degree.

Construction of the epistemic universe that includes both the possible states of knowledge of the two agents, as well as the results of their integration requires the introduction of three unary operators: K_1 and K_2 representing the knowledge of each agent, and the unsubscripted operator K, describing results of knowledge combination. It is assumed that neither agent has information about the extent or nature of the information available to the other (i.e., propositions such as K_1K_2p are always false), and that each agent's domain of knowledge (i.e., the sentence spaces S_1 and S_2 and their related frames of discernment) may be different.

Since the operator K describes the results of integrating the knowledge of two agents, it is necessary to introduce an axiom that assures that the combined knowledge is solely a function of the states of knowledge being fused:

CK1 The proposition Kp is true if and only if there exist propositions p_1 and p_2 such that K_1p_1 and K_2p_2 are true and $p_1 \wedge p_2 \Longrightarrow p$

The epistemic universe constructed with this augmented framework is called a logical product universe. In this universe it is possible, as before, to define epistemic, support, and truth sets. However, since three epistemic operators are involved, these sets must be distinguished by subscripts that identify the respective knowledge sources.

If $e(p)$, $e_1(p)$, $e_2(p)$, denote the epistemic sets for the proposition p that are associated with the epistemic operators K, K_1 and K_2, respectively, then the basic set equation that relates these sets is

$$e(p) = \bigcup_{p_1 \wedge p_2 = p} [e_1(p_1) \cap e_2(p_2)] \tag{11}$$

where the union is over propositions p_1 and p_2 (in the respective domains of knowledge of K_1 and K_2) such that the conjunction $p_1 \wedge p_2$ is logically equivalent to p.

If P is an epistemic probability in the logical universe, the above set equation may be combined with basic probability results relating marginal and joint probability distributions to derive the following general expression for knowledge combination, called the additive combination formula:

$$m(p) = k \sum_{p_1 \wedge p_2 = p} P(e_1(p_1) \cap e_2(p_2)) \tag{12}$$

where k is a constant that makes $\sum m(p) = 1$.

Under assumptions of independence of the (marginal) epistemic algebras for K_1 and K_2, the above formula becomes a generalization of the Dempster's rule of combination:

$$m(p) = k \sum_{p_1 \wedge p_2} m_1(p_1) m_2(p_2). \tag{13}$$

Simple cases of combination of dependent evidence, such as those governed by compatibility relations, may also be derived directly from the additive combination formula, as we have discussed elsewhere [13].

In more general cases, the corresponding expressions must combine the knowledge of the two agents (expressed by the additive combination formula) with knowledge about the dependence relations between the two evidential bodies. The latter information is typically modeled as probabilities defined on a subalgebra of the epistemic algebra.

5 Conclusion

This paper has presented results that closely relate probability functions in epistemic universes to the concepts and constructs of the Dempster-Shafer calculus of evidence. The epistemic structures presented above also furnish important insight that is very useful to enhance the calculus of evidence by the development of expressions that allow for different types of dependent evidence to be combined. These expressions are the current object of our investigations, which focus particularly on the problems of combining multiple evidential bodies that share common information.

In addition, we are also concerned with problems related to the use of conditional evidence (i.e., evidence that is valid only when some proposition is true). This research expands upon and enhances our previous results in this area [12].

Our long term objectives include the treatment of problems involving combination of the knowledge of multiple agents that are aware, to different extents, of the information available to one another. The corresponding issues are of central importance in the design of distributed artificial intelligence systems with planning and counterplanning capabilities.

Acknowledgement

The author benefitted from advice and suggestions received from colleagues in the Artificial Intelligence Center. Bob Moore, Stan Rosenschein, and David Israel suggested investigation of the Carnapian viewpoint as a prospective avenue to gain a better understanding of the relationships among approaches to approximate reasoning. John Lowrance, Tom Garvey, Tom Strat, Len Wesley, and Stephen Lesh comprised an abundant source of insight and encouragement. I have also profited from conversations with Andy Hanson, Haim Gaifman, Pascal Fua, Leslie Kaelbling, Martin Fischler, Ray Perrault, Kurt Konolige, Piero Bonissone, Lotfi Zadeh, Terry Cronin, Ken Laws, Lorna Shinkle, Mike Georgeff, Martha Pollack, Ivan Leclerc, Todd Davies, and Oscar Firschein. To all of them, many thanks.

References

[1] R. Carnap, *Meaning and necessity*, University of Chicago Press, Chicago, Illinois, 1956.

[2] ———, *Logical foundations of probability*, University of Chicago Press, Chicago, Illinois, 1962.

[3] A. P. Dempster, *Upper and lower probabilities induced by a multivalued mapping*, Annals of Mathematical Statistics **38** (1967), 325–339.

[4] M. Jr Hall, *Combinatorial theory, second edition*, John Wiley and Sons, New York, 1986.

[5] J. Y. Halpern and D. A. McAllester, *Likelihood, probability, and knowledge*, Proceedings, 1984 AAAI Conference (AAAI, Menlo Park, California), 1984, pp. 137–141.

[6] J. Hintikka, *Knowledge and belief*, Cornell University Press, Ithaca, New York, 1962.

[7] G. E. Hughes and M. E. Creswell, *An introduction to modal logic*, Methuen, London, 1968.

[8] R. Moore, *Reasoning about knowledge and action*, Tech. report, SRI International, Menlon Park, California, 1980.

[9] J. Neveu, *Bases mathematiques du calcul des probabilities*, Masson, Paris, France, 1964.

[10] N. Nilsson, *Probabilistic logic*, Artificial Intelligence **28** (1986), 71–88.

[11] S. J. Rosenschein and L. P. Kaelbling, *The synthesis of digital machine with provable epistemic properties*, Proceedings of 1986 Conference on Theoretical Aspects of Reasoning about Knowledge (Los Altos, California), Kaufmann, 1986, pp. 83–98.

[12] E. H. Ruspini, *Approximate deduction in single evidential bodies*, Proceedings of 1986 AAAI Uncertainty Workshop (AAAI-86) (Menlo Park, California), 1986.

[13] ———, *The logical foundations of evidential reasoning*, Tech. report, SRI International, Menlo Park, California, 1986.

[14] G. Shafer, *A mathematical theory of evidence*, Princeton University Press, Princeton, NJ, 1976.

[15] P. Suppes, *The measurement of belief*, Journal of Royal Statistical Society Series B **36** (1974), 160–175.

[16] P. M. Williams, *Indeterminate probabilities*, Formal Methods in the Methodology of Empirical Sciences (K. Szaniawski M. Przelecki and R. Wojciki, eds.), Ossolineum and D. Reidel, 1976.

18

Implementing Dempster's Rule for Hierarchical Evidence

Glenn Shafer and Roger Logan

Abstract. This article gives an algorithm for the exact implementation of Dempster's rule in the case of hierarchical evidence. This algorithm is computationally efficient, and it makes the approximation suggested by Gordon and Shortliffe unnecessary. The algorithm itself is simple, but its derivation depends on a detailed understanding of the interaction of hierarchical evidence.

Introduction

Gordon and Shortliffe [4] propose an algorithm for approximating the results of Dempster's rule of combination for the case where the evidence being combined is evidence for and against hypotheses that can be arranged in a hierarchical or tree-like structure. This proposal is motivated by the computational complexity of Dempster's rule. In general, the amount of computation needed to implement the rule increases exponentially with the number of possible answers in a diagnostic problem. Gordon and Shortliffe's algorithm avoids this exponential explosion; the amount of computation it requires increases only linearly with the number of possible answers.

Gordon and Shortliffe's algorithm usually produces a good approximation. In the case of highly conflicting evidence, however, the approximation can be poor; an example is given in Sect. 2. Moreover, the algorithm does not give degrees of belief for all hypotheses (i.e., all subsets of the set of possible answers). It gives degrees of belief only for hypotheses in the tree.

In this article we show that it is not necessary to resort to Gordon and Shortliffe's approximation. We give an algorithm for exact implementation that is linear in its computational complexity. This algorithm works for slightly more general types of evidence than Gordon and Shortliffe's algorithm, and it gives degrees of belief for more hypotheses. In particular, it gives plausibilities as well as degrees of belief for hypotheses in the tree.

Dempster's rule is part of the theory of belief functions, sometimes called the Dempster–Shafer theory in the artificial intelligence community. A basic reference for the elementary aspects of this theory is Shafer [9]. A more recent exposition and an extensive bibliography are included in Shafer [10]. Expositions that discuss the theory's relevance to artificial intelligence include Garvey, Lowrence, and Fischler [2], Gordon and Shortliffe [3], and Shafer [11].

In the next section, we provide a reasonably self-contained discussion of those mathematical aspects of the theory of belief functions that are relevant to the algorithm presented in this article. Readers will need to turn to the references just cited for further details of the theory and for information on its intuitive interpretation.

In Sect. 2, we review the problem posed by Gordon and Shortliffe and describe the approximation they propose. In Sect. 3 we derive some mathematical facts about the problem, and in Sect. 4 we use these facts to derive our algorithm. In Sect. 5, we discuss generalizations.

1 The Mathematics of Belief Functions

Suppose Θ denotes a set of possible answers to some question, and assume that one and only one of these answers can be correct. We call Θ a *frame of discernment*. A function Bel that assigns a degree of belief Bel(A) to every subset A of Θ is called a *belief function* if it satisfies certain mathematical conditions.

Those familiar with the usual mathematical theory of probability can understand the mathematical structure of belief functions by thinking about random sets. A function Bel defined for every subset A of Θ qualifies as a belief function if and only if there is a random non-empty subset S of Θ such that

$$\mathrm{Bel}\,(A) = \Pr[S \subseteq A]$$

for all A. (It should be emphasized that this interpretation in terms of a random subset S provides insight only into the mathematical structure of belief functions. It does not provide insight into the interpretation of Bel(A) as a degree of belief based on evidence. See Shafer [9, 10] for explanations of the evidential interpretation.)

The information in a belief function Bel can also be expressed in terms of the *plausibility function* Pl, given by

$$\mathrm{Pl}\,(A) = 1 - \mathrm{Bel}\left(\bar{A}\right) = \Pr\left[S \cap A \neq \emptyset\right],$$

where $\bar{A}$ denotes the complement of A. In the evidential interpretation, Pl(A) is the plausibility of A in light of the evidence—a measure of the extent to which the evidence fails to refute A. To recover Bel from Pl, we use the relation $\mathrm{Bel}(A) = 1 - \mathrm{Pl}(\bar{A})$. Notice that $\mathrm{Bel}(A) \leq \mathrm{Pl}(A)$ for every subset A of Θ. Both Bel and Pl are monotone: $\mathrm{Bel}(A) \leq \mathrm{Bel}(B)$ and $\mathrm{Pl}(A) \leq \mathrm{Pl}(B)$ whenever $A \subseteq B$.

In this article we assume that the frame of discernment Θ is finite. In this case the information in Bel or Pl is also contained in the *commonality function* Q, defined by

$$Q\,(A) = \mathrm{Pr}\,[S \supseteq A]$$

for every subset A of Θ. Indeed, it is shown in [9, Chap. 2] that

$$Q\,(A) = \sum \left\{ (-1)^{|B|+1}\, \mathrm{Pl}\,(B)\, | \emptyset \neq B \subseteq A \right\} \tag{1}$$

and

$$\mathrm{Pl}\,(A) = \sum \left\{ (-1)^{|B|+1}\, Q\,(B)\, | \emptyset \neq B \subseteq A \right\} \tag{2}$$

for every non-empty subset A of Θ, where $|B|$ denotes the number of elements in the set B. (Formulas (1) and (2) do not give values for $Q(\emptyset)$ or $\mathrm{Pl}(\emptyset)$, but we know that $Q(\emptyset) = 1$ and $\mathrm{Pl}(\emptyset) = 0$ for any belief function.)

1.1 Dempster's rule

Consider two random non-empty subsets S_1 and S_2. Suppose S_1 and S_2 are probabilistically independent—i.e.,

$$\mathrm{Pr}[S_1 = A_1 \text{ and } S_2 = A_2] = \mathrm{Pr}[S_1 = A_1]\mathrm{Pr}\,[S_2 = A_2]\,.$$

And suppose $\mathrm{Pr}[S_1 \cap S_2 \neq \emptyset] > 0$. Let S be a random non-empty subset that has the probability distribution of $S_1 \cap S_2$ conditional on $S_1 \cap S_2 \neq \emptyset$—i.e.,

$$\mathrm{Pr}\,[S = A] = \frac{\mathrm{Pr}\,[S_1 \cap S_2 = A]}{\mathrm{Pr}\,[S_1 \cap S_2 \neq \emptyset]} \tag{3}$$

for every non-empty subset A of Θ.

If Bel_1 and Bel_2 are the belief functions corresponding to S_1 and S_2, then we denote the belief function corresponding to S by $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$, and we call $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ the *orthogonal sum* of Bel_1 and Bel_2. The rule for forming $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ is called *Dempster's rule of combination.* This rule corresponds, in the evidential interpretation, to the combination or pooling of independent bodies of evidence. (If $\mathrm{Pr}[S_1 \cap S_2 \neq \emptyset] = 0$, then the two belief functions contradict each other—i.e., there exists A such that $\mathrm{Bel}_1(A) = 1$ and $\mathrm{Bel}_2(\bar{A}) = 1$. It makes no sense to try to pool the evidence in this case.)

The formation of orthogonal sums by Dempster's rule corresponds to the multiplication of commonality functions. Indeed, if the commonality functions for Bel_1, Bel_2, and $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ are denoted by Q_1, Q_2, and Q, respectively, then

$$\begin{aligned} Q\left(A\right) &= \Pr[S \supseteq A] \\ &= K\ \Pr\left[S_1 \cap S_2 \supseteq A\right] = K\ \Pr\left[S_1 \supseteq A \text{ and } S_2 \supseteq A\right] \\ &= K\ \Pr\left[S_1 \supseteq A\right] \Pr\left[S_2 \supseteq A\right] = K\ Q_1\left(A\right) Q_2\left(A\right), \end{aligned}$$

where K does not depend on A;

$$K^{-1} = \Pr\left[S_1 \cap S_2 \neq \emptyset\right].$$

We can find K from Q_1 and Q_2 if we substitute $KQ_1(B)Q_2(B)$ for $Q(B)$ and Θ for A in (2). Since $\mathrm{Pl}(\Theta) = 1$, this gives

$$1 = \sum \left\{(-1)^{|B|+1} KQ_1\left(B\right) Q_2\left(B\right) | \emptyset \neq B \subseteq \Theta\right\},$$

or

$$K^{-1} = \sum \left\{(-1)^{|B|+1} Q_1\left(B\right) Q_2\left(B\right) | \emptyset \neq B \subseteq \Theta\right\}. \qquad (4)$$

We may summarize by saying that the multiplication of commonality functions gives a recipe for computing the plausibility function Pl for $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$. First we find the plausibility functions Pl_1 and Pl_2 using the relation

$$\mathrm{Pl}_i\left(A\right) = 1 - \mathrm{Bel}_i\left(\bar{A}\right).$$

Then we find the commonality functions Q_i using the relation

$$Q_i\left(A\right) = \sum \left\{(-1)^{|B|+1} \mathrm{Pl}_i\left(B\right) | \emptyset \neq B \subseteq A\right\}. \qquad (5)$$

Then we find Pl using the relation

$$\mathrm{Pl}\left(A\right) = K \sum \left\{(-1)^{|B|+1} Q_1\left(B\right) Q_2\left(B\right) | \emptyset \neq B \subseteq A\right\}, \qquad (6)$$

where K is given by (4). This recipe generalizes to the case where we wish to combine more than two belief functions; we merely put $Q_1(B) \cdots Q_n(B)$ in the place of $Q_1(B)Q_2(B)$ in (4) and (6).

Unfortunately, this recipe is computationally forbidding if Θ contains a large number of elements. The number of subsets of Θ increases exponentially with the number elements of Θ, and the sum in (4), for example, involves a term for each of these subsets.

This computational complexity seems to be intrinsic to Dempster's rule. There does not seem to be any general way of implementing the rule that will always involve fewer computations than are involved in (4), (5), and (6). There are, however, special cases where alternative methods involving less computation are possible.

1.2 Focal elements, simple support functions, and dichotomous belief functions

A subset S of Θ is called a *focal element* of Bel if $\Pr[\mathcal{S} = S]$ is positive.

The simplest belief function is the belief function whose only focal element is the whole frame Θ; in this case $\Pr[\mathcal{S} = \Theta] = 1$. This belief function is called the *vacuous belief function.* It is obvious that if Bel is the vacuous belief function, then $\text{Bel} \oplus \text{Bel}' = \text{Bel}'$ for any other belief function Bel'.

A belief function is called a *simple support function* if it has at most one focal element not equal to the whole frame Θ. If a simple support function does have a focal element not equal to Θ (i.e., if it is not vacuous), then this focal element is called the *focus* of the simple support function.

A belief function is called *dichotomous* with dichotomy $\{A, \bar{A}\}$ if it has no focal elements other than $A, \bar{A}$, and Θ.

In general, combination by Dempster's rule involves the intersection of focal elements. The focal elements for $\text{Bel}_1 \oplus \cdots \oplus \text{Bel}_n$ will consist of all non-empty intersections of the form $S_1 \cap \cdots \cap S_n$, where S_i is a focal element of Bel_i. Therefore, the orthogonal sum of simple support functions with a common focus will be another simple support function with that focus. Similarly, the orthogonal sum of dichotomous belief functions with a common dichotomy will be another dichotomous belief function with that dichotomy.

1.3 Bayesian belief functions

This theory of belief functions is a generalization of the more familiar Bayesian theory, which uses probability measures as expressions of subjective judgments and updates these measures by conditioning. A probability measure is a belief function, and conditioning is a special case of Dempster's rule.

Let us call a belief function a *Bayesian belief function* if it is a probability measure. A belief function is Bayesian if and only if its focal elements are all singletons. This is equivalent to saying that the corresponding random subset is always equal to a singleton. Since a singleton is contained in a subset A if and only if it has a non-empty intersection with A, a Bayesian belief function is equal to its plausibility function.

In the Bayesian theory, conditioning a belief function Bel_1 on knowledge that a subset B of Θ is true means changing one's degree of belief for each subset A from $\text{Bel}_1(A)$ to

$$\frac{\text{Bel}_1\,(A \cap B)}{\text{Bel}_1\,(B)}. \tag{7}$$

In the theory of belief functions, on the other hand, knowledge that a subset B of Θ is true is represented by a belief function, say Bel_2, that has B as its only focal element. And the way to change Bel_1 to take this knowledge into account is to combine Bel_1 with Bel_2 by Dempster's rule. In order to see that this application of Dempster's rule gives the same result as (7), let us return to (3) for a moment.

It is clear that from (3) that if S_1 is always a singleton, then S is also always a singleton, so $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ will indeed be Bayesian. Moreover, if we substitute $\{s\}$ for A in (3) and bear in mind that S_1 is always a singleton and S_2 is always equal to B, then we obtain

$$\begin{aligned}\Pr[S = \{s\}] &= \frac{\Pr[S_1 \cap B = \{s\}]}{\Pr[S_1 \cap B \neq \emptyset]} \\ &= \begin{cases} \dfrac{\mathrm{Bel}_1(\{s\})}{\mathrm{Bel}_1(B)}, & \text{if } s \in B, \\ 0, & \text{if } s \notin B. \end{cases}\end{aligned}$$

Adding $\Pr[S = \{s\}]$ for all s in A, we obtain (7) for our degree of belief in A.

1.4 Partitions

One case where the computational complexity of Dempster's rule can be reduced is the case where the belief functions being combined are carried by a partition $\mathcal{P}$ of the frame Θ. In this case, $\mathcal{P}$, which has fewer elements than Θ, can in effect be used in the place of Θ when the computations (5), (6) and (4) are carried out.

A *partition* of a frame of discernment Θ is a set of disjoint non-empty subsets of Θ whose union equals Θ. Such a partition $\mathcal{P}$ can itself be regarded as a frame of discernment; it is the set of possible answers to the question, "which element of $\mathcal{P}$ contains the correct answer to the question corresponding to Θ?" If $\mathcal{P}_1$ and $\mathcal{P}_2$ are partitions of Θ and for every element P_1 in $\mathcal{P}_1$ there is an element P_2 in $\mathcal{P}_2$ such that $P_1 \subseteq P_2$, then we say that $\mathcal{P}_1$ is a *refinement* of $\mathcal{P}_2$.

Given a partition $\mathcal{P}$ of Θ, we denote by $\mathcal{P}^*$ the set consisting of all unions of elements of $\mathcal{P}; \mathcal{P}^*$ is a field of subsets of Θ.

We say that a belief function Bel over Θ is *carried by* $\mathcal{P}$ if the random subset S corresponding to Bel satisfies

$$\Pr[S \in \mathcal{P}^*] = 1.$$

It is evident that if Bel_1 and Bel_2 are both carried by $\mathcal{P}$, then $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ will also be carried by $\mathcal{P}$; for if S_1 and S_2 are both in the field $\mathcal{P}^*$ with probability one, then $S_1 \cap S_2$ is as well.

For a given partition $\mathcal{P}$ of Θ and a given subset A of Θ, there is a largest element of $\mathcal{P}^*$ contained in A, namely

$$A_{\mathcal{P}} = \cup \{P | P \in \mathcal{P}, P \subseteq A\}.$$

There is also a smallest element of $\mathcal{P}^*$ containing A, namely

$$A^{\mathcal{P}} = \cup \{P | P \in \mathcal{P}, P \cap A \neq \emptyset\} .$$

When Bel is carried by $\mathcal{P}$, its values for elements of $\mathcal{P}^*$ determine its values for the other subsets of Θ. Indeed, since $S \in \mathcal{P}^*$, $S \subseteq A$ if and only if $S \subseteq A_{\mathcal{P}}$, and so

$$\begin{aligned}\mathrm{Bel}(A) &= \Pr [S \subseteq A] = \Pr [S \subseteq A_{\mathcal{P}}] \\ &= \mathrm{Bel} (A_{\mathcal{P}}) = \max \{\mathrm{Bel} (B) | B \subseteq A, B \in \mathcal{P}^*\} . \end{aligned} \tag{8}$$

Similarly,

$$\mathrm{Pl}(A) = \mathrm{Pl} \left(A^{\mathcal{P}}\right) = \min \{\mathrm{Pl} (B) | B \supseteq A, B \in \mathcal{P}^*\} . \tag{9}$$

It turns out that when Bel is carried by $\mathcal{P}$ we can replace (1) and (2) by analogous formulas that only involve elements of $\mathcal{P}^*$:

$$Q (A) = \sum \left\{(-1)^{|B|^{\mathcal{P}}+1} \mathrm{Pl} (B) | B \in \mathcal{P}^*, \emptyset \neq B \subseteq A\right\} \tag{10}$$

and

$$\mathrm{Pl} (A) = \sum \left\{(-1)^{|B|^{\mathcal{P}}+1} Q (B) | B \in \mathcal{P}^*, \emptyset \neq B \subseteq A\right\} \tag{11}$$

for every non-empty element A of $\mathcal{P}^*$, where $|B|^{\mathcal{P}}$ denotes the number of elements of $\mathcal{P}$ contained in B. It follows that if Bel_1 and Bel_2 are both carried by $\mathcal{P}$, we can compute the plausibility function Pl for $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ by first computing

$$\mathrm{Pl}_i (A) = 1 - \mathrm{Bel}_i \left(\bar{A}\right)$$

just for A in $\mathcal{P}^*$, then computing

$$Q_i (A) = \sum \left\{(-1)^{|B|^{\mathcal{P}}+1} \mathrm{Pl}_i (B) | B \in \mathcal{P}^*, \emptyset \neq B \subseteq A\right\} \tag{12}$$

just for A in $\mathcal{P}^*$, and then computing

$$\mathrm{Pl} (A) = K \sum \left\{(-1)^{|B|^{\mathcal{P}}+1} Q_1 (B) Q_2 (B) | B \in \mathcal{P}^*, \emptyset \neq B \subseteq A\right\} \tag{13}$$

for A in $\mathcal{P}^*$, where

$$K^{-1} = \sum \left\{(-1)^{|B|^{\mathcal{P}}+1} Q_1 (B) Q_2 (B) | B \in \mathcal{P}^*, \emptyset \neq B \subseteq A\right\} . \tag{14}$$

The values $\mathrm{Pl}(A)$ for A not in $\mathcal{P}^*$ can then be obtained, if they are desired, from (9).

Why do (10) and (11) hold for elements of $\mathcal{P}^*$? The easiest way to see that they do hold is to recognize that $\mathcal{P}^*$ is isomorphic to the set of all subsets of $\mathcal{P}$. And when we do this, we see that (10) and (11) are merely (1) and (2) with $\mathcal{P}$ in the place of Θ. When we use (12)–(14) we are treating our belief functions as if they were really belief functions on the simpler frame $\mathcal{P}$.

Formulas (13) and (14) generalize, of course, to the case where more than two belief functions carried by $\mathcal{P}$ are combined. As before, we simply replace $Q_1(B)Q_2(B)$ by $Q_1(B) \cdots Q_n(B)$.

1.5 Coarsenings

Given a random subset S and a partition $\mathcal{P}$, let $S^{\mathcal{P}}$ denote the random subset that is always equal to $A^{\mathcal{P}}$ when S is equal to A. If Bel is the belief function corresponding to S, then let $\mathrm{Bel}_{\mathcal{P}}$ denote the belief function corresponding to $S^{\mathcal{P}}$. Since $S^{\mathcal{P}}$ is always in $\mathcal{P}^*$, Bel is carried by $\mathcal{P}$. Since $S^{\mathcal{P}} \subseteq A$ if and only if $S \subseteq A_{\mathcal{P}}$,

$$\mathrm{Bel}_{\mathcal{P}}\left(A\right) = \Pr\left[S^{\mathcal{P}} \subseteq A\right] = \Pr[S \subseteq A_{\mathcal{P}}] = \mathrm{Bel}\left(A_{\mathcal{P}}\right).$$

This means in particular that $\mathrm{Bel}_{\mathcal{P}}(A) = \mathrm{Bel}(A)$ if $A \in \mathcal{P}^*$. Thus, $\mathrm{Bel}_{\mathcal{P}}$ is the unique belief function that agrees with Bel on $\mathcal{P}^*$ and is carried by $\mathcal{P}$.

Suppose we want to combine two belief functions Bel_1 and Bel_2. And suppose we are tempted to do so using (12), (13), and (14), even though Bel_1 and Bel_2 are not carried by the partition $\mathcal{P}$. We know that we will not get the right answer; we will get $\mathrm{Bel}_{1\mathcal{P}} \oplus \mathrm{Bel}_{2\mathcal{P}}$ instead of $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$. But suppose we are not interested in the whole belief function $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$. Suppose we are interested only in the values of $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ on $\mathcal{M}^*$ for some partition $\mathcal{M}$. We will get these values right if and only if

$$\left(\mathrm{Bel}_{1\mathcal{P}} \oplus \mathrm{Bel}_{2\mathcal{P}}\right)_{\mathcal{M}} = \left(\mathbf{Bel}_1 \oplus \mathrm{Bel}_2\right)_{\mathcal{M}}.$$

This is equivalent to

$$\left(S_1^{\mathcal{P}} \cap S_2^{\mathcal{P}}\right)^{\mathcal{M}} = \left(S_1 \cap S_2\right)^{\mathcal{M}}. \tag{15}$$

It is also equivalent to the condition that $M \cap P \neq \emptyset$, $S_1 \cap P \neq \emptyset$, and $S_2 \cap P \neq \emptyset$ together imply $S_1 \cap S_2 \cap M \neq \emptyset$ whenever $M \in \mathcal{M}$, $P \in \mathcal{P}$, S_1 is a focal element of Bel_1, and S_2 is a focal element of Bel_2. If this condition is satisfied, then we say that $\mathcal{P}$ *discerns the interaction* between Bel_1 and Bel_2 that is relevant to $\mathcal{M}$.

It is easy to see that if $\mathcal{P}$, $\mathcal{P}'$, $\mathcal{M}$, and $\mathcal{M}'$ are all partitions, $\mathcal{P}'$ is finer than $\mathcal{P}$, $\mathcal{M}'$ is coarser than $\mathcal{M}$, and $\mathcal{P}$ discerns the interaction relevant to $\mathcal{M}$, then $\mathcal{P}'$ discerns the interaction relevant to $\mathcal{M}'$.

We are most often interested in whether $\mathcal{P}$ discerns the interaction relevant to itself. In this case (15) becomes

$$S_1^{\mathcal{P}} \cap S_2^{\mathcal{P}} = \left(S_1 \cap S_2\right)^{\mathcal{P}},$$

and this is equivalent to the condition that $S_1 \cap P \neq \emptyset$ and $S_2 \cap P \neq \emptyset$ together imply $S_1 \cap S_2 \cap P \neq \emptyset$ whenever $P \in \mathcal{P}$, S_1 is a focal element of Bel_1, and S_2 is a focal element of Bel_2. Notice that if one of the pair Bel_1 and Bel_2 is carried by $\mathcal{P}$, then $\mathcal{P}$ will necessarily discern the interaction between Bel_1 and Bel_2 that is relevant to itself.

It might be thought that if $\mathcal{P}$ discerns the interaction relevant to itself and $\mathcal{P}'$ is finer than $\mathcal{P}$, then $\mathcal{P}'$ will also discern the interaction relevant to itself. But this is not necessarily true; $\mathcal{P}'$ will discern the interaction relevant to $\mathcal{P}$,

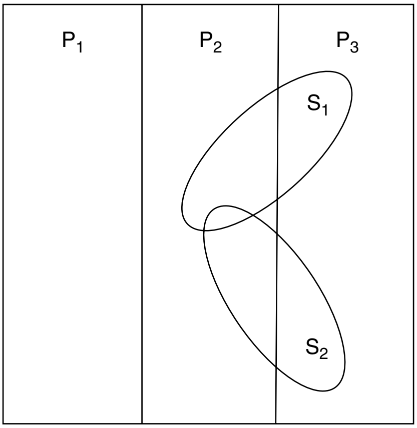


Fig. 1. A partition that does not discern the interaction relevant to itself

but it may not discern the interaction relevant to $\mathcal{P}'$. Figure 1 illustrates this point. If our two belief functions are simple support functions with foci S_1 and S_2, respectively, then the partition $\{P_1, P_2 \cup P_3\}$ discerns the interaction relevant to itself, but the partition $\{P_1, P_2, P_3\}$ does not. Figure 2 illustrates the opposite situation; $\{P_1, P_2, P_3\}$ discerns the interaction relevant to itself, but $\{P_1, P_2 \cup P_3\}$ does not.

The preceding discussion generalizes readily to the case where we have more than two belief functions. For example, $\mathcal{P}$ discerns the interaction among $\mathrm{Bel}_1, \ldots, \mathrm{Bel}_n$ that is relevant to itself if and only if

$$S_1^{\mathcal{P}} \cap \cdots \cap S_n^{\mathcal{P}} = (S_1 \cap \cdots \cap S_n)^{\mathcal{P}} ,$$

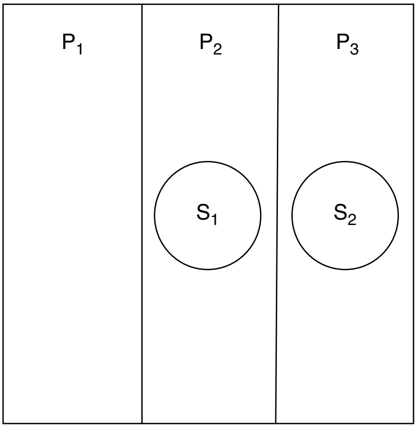


Fig. 2. A partition that does discern the interaction relevant to itself

and this is equivalent to the condition that $S_i \cap P \neq \emptyset$ for $i = 1, \ldots, n$ implies $S_1 \cap \ldots \cap S_n \cap P \neq \emptyset$ whenever $P \in \mathcal{P}$ and S_i is a focal element of Bel_i. Notice that if $\mathcal{P}$ discerns the interaction among $\mathrm{Bel}_1, \ldots, \mathrm{Bel}_n$ that is relevant to itself and $\mathrm{Bel}_{n+1}, \ldots, \mathrm{Bel}_{n+m}$ are carried by $\mathcal{P}$, then $\mathcal{P}$ discerns the interaction among $\mathrm{Bel}_1, \ldots, \mathrm{Bel}_{n+m}$ that is relevant to itself.

1.6 Barnett's technique

Barnett [1] has shown that Dempster's rule can be implemented in a number of computations that increases only linearly with the number of elements in Θ if the belief functions being combined are all simple support functions focused on singletons or their complements. Here we will explain Barnett's technique in terms of the commonality function.

Recall that a simple support function focused on S is a belief function whose only focal elements are S and Θ. If θ is an element of Θ, Bel_1 is a simple support function focused on the singleton $\{\theta\}$, and Bel_2 is a simple support function focused on its complement $\overline{\{\theta\}}$, then $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ is easy to calculate; it is dichotomous with the dichotomy $\{\{\theta\}, \overline{\{\theta\}}\}$. In describing Barnett's technique we may, therefore, assume that we begin with dichotomous belief functions of this form. In fact, we may assume, without loss of generality, that we have such a dichotomous belief function, Bel_θ say, for every element Θ of Θ; our task is to combine the Bel_θ. (If $\mathrm{Bel}_\theta(\{\theta\}) = \mathrm{Bel}_\theta(\overline{\{\theta\}}) = 0$, then Bel_θ is vacuous, and its presence in the combination makes no difference.)

For brevity, we denote $\mathrm{Bel}_\theta(\{\theta\})$ and $\mathrm{Bel}_\theta(\overline{\{\theta\}})$ by θ^+ and θ^- respectively. In order to avoid trivialities, we assume that both θ^+ and θ^- are less than one. Then the commonality function for Bel_θ is given by

$$Q_\theta(B) = \begin{cases} 1 - \theta^-, & \text{if } B = \{\theta\}, \\ 1 - \theta^+, & \text{if } \theta \notin B, \\ 1 - \theta^- - \theta^+, & \text{if } \theta \in B \text{ and } |B| > 1 \end{cases}$$

for all non-empty subsets B of Θ, and

$$\prod_{\theta \in \Theta} Q_\theta(B) = \begin{cases} \left(1 - \theta_0^-\right) \prod\limits_{\theta \neq \theta_0} \left(1 - \theta^+\right), & \text{if } B = \{\theta_0\} \\ \prod\limits_{\theta \in B} \left(1 - \theta^- - \theta^+\right) \prod\limits_{\theta \notin B} \left(1 - \theta^+\right), & \text{if } |B| > 1 \end{cases}$$

$$= \begin{cases} \prod\limits_{\theta \in \Theta} \left(1 - \theta^+)(1 - \theta_0^-\right) / (1 - \theta_0^+), & \text{if } B = \{\theta_0\}, \\ \prod\limits_{\theta \in \Theta} \left(1 - \theta^+\right) \prod\limits_{\theta \in B} \left(1 - \theta^- - \theta^+\right) / (1 - \theta^+), & \text{if } |B| > 1. \end{cases}$$

So implementation of the generalization of (6) involves calculating

$$\begin{aligned}
&\sum_{\substack{B \subseteq A \\ B \neq \theta}} (-1)^{|B|+1} \prod_{\theta\in\Theta} Q_\theta(B) \\
&= \prod_{\theta\in\Theta} \left(1-\theta^+\right) \left[\sum_{\theta\in A} \frac{1-\theta^-}{1-\theta^+} - \sum_{\substack{B \subseteq A \\ |B|>1}} (-1)^{|B|} \prod_{\theta\in B} \frac{1-\theta^- - \theta^+}{1-\theta^+} \right] \\
&= \prod_{\theta\in\Theta} \left(1-\theta^+\right) \\
&\quad \times \left[\sum_{\theta\in A} \frac{1-\theta^-}{1-\theta^+} + 1 - \sum_{\theta\in A} \frac{1-\theta^- - \theta^+}{1-\theta^+} \right. \\
&\quad \left. - \sum_{B\subseteq A} (-1)^{|B|} \prod_{\theta\in B} \frac{1-\theta^- - \theta^+}{1-\theta^+} \right] \\
&= \prod_{\theta\in\Theta} \left(1-\theta^+\right) \left[1 + \sum_{\theta\in A} \frac{\theta^+}{1-\theta^+} - \prod_{B \in A} \left(1 - \frac{1-\theta^- - \theta^+}{1-\theta^+} \right) \right] \\
&= \prod_{\theta\in\Theta} \left(1-\theta^+\right) \left[1 + \sum_{\theta\in A} \frac{\theta^+}{1-\theta^+} - \prod_{\theta \in A} \frac{\theta^-}{1-\theta^+} \right].
\end{aligned} \tag{16}$$

The next to last equality is the crucial step; it reduces the summation over subsets of A to a product over elements of A, which can be implemented in linear time.

Substituting (16) in the generalizations of (4) and (6) and omitting the common factor $\Pi_{\theta\in\Theta}(1-\theta^+)$, we obtain

$$\mathrm{Pl}\,(A) = K \left(1 + \sum_{\theta\in A} \frac{\theta^+}{1-\theta^+} - \prod_{\theta\in A} \frac{\theta^-}{1-\theta^+} \right), \tag{17}$$

where

$$K^{-1} = 1 + \sum_{\theta\in\Theta} \frac{\theta^+}{1-\theta^+} - \prod_{\theta\in\Theta} \frac{\theta^-}{1-\theta^+}. \tag{18}$$

The statement that (17) and (18) allow the implementation of Dempster's rule in linear time should be interpreted with caution. It is true that the number of computations required by (18) increases only linearly with the number of elements in Θ, and the same is true of any particular instance of (17). If, however, we wish to compute the whole belief function Bel, then we need to calculate Pl(A) for every subset A of Θ, and the number of such subsets increases exponentially with the size of Θ. In some problems this will cause no difficulty, for we will be able to identify a priori a few subsets A of Θ as the only ones for which we need to know Bel(A) or Pl(A). But in other problems we may be interested simply in finding the smallest subsets A that have high values of Bel(A), and if it is not feasible to calculate and look at Bel(A) for all A, then some search strategy may be needed.

If $\theta^+ + \theta^- = 1$ for all elements Θ in Θ, then it is easy to locate the subsets A of Θ that have the highest values of Bel(A). Indeed, in this situation Bel is a Bayesian belief function; Bel(A) = Pl(A) for all subsets A, and (17) and (18) become

$$\mathrm{Bel}(A) = \sum_{\theta \in A} f(\theta), \tag{19}$$

where

$$f(\theta) = \frac{\theta^+}{1-\theta^+} \Big/ \sum_{\theta' \in \Theta} \frac{\theta'^+}{1-\theta'^+}. \tag{20}$$

In this case, to locate subsets A with high values of Bel(A) we need only order the elements of Θ from largest to smallest in the value of $f(\theta)$, and consider subsets obtained by taking initial sequences from this list.

In general $\theta^+ + \theta^-$ will not equal one; in fact, $\theta^+ + \theta^-$ can approach one only as the weights of evidence for and against Θ become infinitely large (see [9, Chap. 9]). However, when there is a substantial amount of evidence both for and against most of the Θ, (19) and (20) may be nearly enough correct to help us identify subsets for which (17) should be computed.

Barnett's technique applies, of course, not only to the case where we begin with simple support functions for and against singletons but also to the case where we begin with simple support functions for and against elements of some coarser partition $\mathcal{P}$. Indeed, if Pl(A) is the plausibility function for the belief function $\oplus\{\mathrm{Bel}_P | P \in \mathcal{P}\}$, where Bel_P is dichotomous with dichotomy $\{P, \bar{P}\}$, and we write P^+ for $\mathrm{Bel}_P(P)$ and P^- for $\mathrm{Bel}_P(\bar{P})$, then

$$\mathrm{Pl}(A) = K \left(1 + \sum_{\substack{P \subseteq A \\ P \in \mathcal{P}}} \frac{P^+}{1-P^+} - \prod_{\substack{P \subseteq A \\ P \in \mathcal{P}}} \frac{P^-}{1-P^+} \right) \tag{21}$$

for every element A of $\mathcal{P}^*$, where

$$K^{-1} = 1 + \sum_{P \in \mathcal{P}} \frac{P^+}{1-P^+} - \prod_{P \in \mathcal{P}} \frac{P^-}{1-P^+}. \tag{22}$$

2 Gordon and Shortliffe's Problem

Gordon and Shortliffe [3, 4] discussed the problem of implementing Dempster's rule in the case where one begins with simple support functions focused for or against subsets of Θ that can be arranged hierarchically in a tree. They concluded that it is not feasible to compute Dempster's rule in such cases, and they proposed a simplification of the rule that can be computed easily.

Figure 3 shows a tree of the kind Gordon and Shortliffe considered. This tree represents the frame $\Theta = \{a,\ b,\ c,\ d,\ e,\ f\}$. We have labeled each node of the tree with a capital letter, which we will use to name both the node and the subset of Θ to which it corresponds. The terminal nodes of the tree correspond to singleton subsets; $A = \{a\},\ F = \{f\}$, etc. Each nonterminal node corresponds to the union of the terminal nodes below it; $G = \{a,\ b,\ c\},\ H = \{d,\ e\}$, and $I = \{a,\ b,\ c,\ f\}$. Notice that most subsets of Θ are not represented in the tree; there is no node, for example, that corresponds to the subset $\{d,\ f\}$.

In Gordon and Shortliffe's example, the elements of Θ are possible diseases, so that higher nodes in the tree correspond to classes of diseases. They suggested that diagnostic evidence tends either to support or refute particular diseases or natural classes of diseases that appear in the tree. Thus, they posed the problem of combining simple support functions focused on nodes of the tree and on the complements of these nodes.

Gordon and Shortliffe found that it is not difficult to combine simple support functions focused on nodes of the tree, because the intersection of two subsets corresponding to nodes will either be empty (because neither node lies below the other) or else equal to one of the two subsets (the one lying below the other). Combining negative evidence leads to computational difficulties, however, because the intersection of the complements of nodes may fail to correspond to a node or its complement. The intersection of $\bar{E}$ and $\bar{G}$

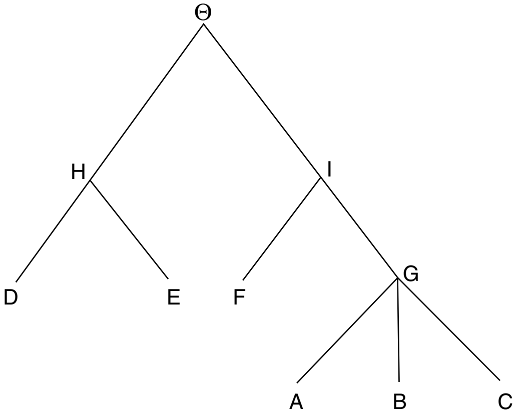


Fig. 3. A tree of diseases

in Fig. 3, for example, results in the subset $\{d, f\}$, and neither this subset nor its complement is represented by a node in the tree.

Gordon and Shortliffe suggested the following procedure. First we combine all the simple functions focused on nodes of the tree by Dempster's rule. Then we successively bring into the combination the simple support functions focused on the complements, working down the tree. But when we bring in one of the simple support functions focused on a complement, we modify Dempster's rule by replacing each intersection of focal elements by the smallest subset in the tree that contains it. The final result depends, in general, on the order in which the simple support functions focused on complements are brought in, but Gordon and Shortliffe conjectured that if we bring these simple support functions in as we work down the tree, then the result will approximate the result that we would get using Dempster's rule correctly.

We have found that Gordon and Shortliffe's approximation is usually very good when the degrees of support for the simple support functions are drawn at random from a uniform distribution. It is easy to construct examples, however, where the approximation is poor. Consider the tree in Fig. 4, and suppose that we have three items of evidence. One of these indicates fairly strongly that a patient's disease is in I, while the other two indicate very strongly that it is not f and not g. More precisely, we have three simple support functions to combine:

$$\begin{aligned}
&\mathrm{Bel}_1 \text{ focused on } I, \text{ with } \mathrm{Bel}_1(I) = 0.8,\\
&\mathrm{Bel}_2 \text{ focused on } \bar{F}, \text{ with } \mathrm{Bel}_2(\bar{F}) = 0.99,\\
&\mathrm{Bel}_3 \text{ focused on } \bar{G}, \text{ with } \mathrm{Bel}_3(\bar{G}) = 0.99.
\end{aligned}$$

Combining these by Dempster's rule, we obtain a belief function $\mathrm{Bel} = \mathrm{Bel}_1 \oplus \mathrm{Bel}_2 \oplus \mathrm{Bel}_3$, with $\mathrm{Bel}(H) \approx 0.91$, corresponding to the judgment that the

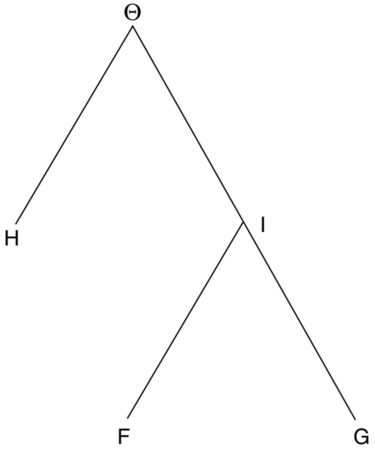


Fig. 4. A smaller tree of diseases

positive evidence for I represented by Bel_1 is overwhelmed by the negative evidence represented by Bel_2 and Bel_3. If, however, we combine using Gordon and Shortliffe's procedure, then we obtain, $\mathrm{Bel}(H) = 0$.

Another shortcoming of Gordon and Shortliffe's procedure is that it assigns degrees of belief only to the subsets of Θ that correspond to nodes in the tree. It does not assign degrees of belief to the complements of these nodes. Thus it does not allow us to assign plausibilities to the nodes. (Recall that the plausibility of A, $\mathrm{Pl}(A)$, is equal to $1 - \mathrm{Bel}(\bar{A})$.) Nor, for example, does it assign a degree of belief to the subset $\{d,\ f\}$ in Fig. 3. Since $\{d,\ f\}$ is not a natural class of diseases, it may be rare for evidence to support this class without supporting either d or f alone. But such a situation is conceivable; it would arise, for example, if one item of evidence weighed strongly against E and another weighed strongly against G. If this did happen, we would want it to come to our attention, so that we would know to look for further evidence that might help us decide which of these two diseases the patient really has.

Gordon and Shortliffe used the term "hierarchical hypothesis space" to emphasize that they were interested only in hypotheses corresponding to nodes of a tree. Since we think it is appropriate to be interested in degrees of belief for a broader class of hypotheses, we use instead the term "hierarchical evidence." This term reflects the assumption that the evidence bears directly only on hypotheses in the tree, but it leaves open the possibility that we might want to calculate degrees of belief for other hypotheses as well.

3 The Interaction of Hierarchical Evidence

In this section we derive some mathematical facts about the interaction of hierarchical evidence. In the next section we show how these facts enable us to implement Dempster's rule efficiently.

Here, as in the preceding section, we assume that we are working with a finite tree such as the one in Fig. 3. We denote by $\mathcal{A}$ the collection of all the nodes below Θ—i.e., all the nodes except Θ itself. If B is directly below A, we say that B is A's *daughter* and A is B's mother. In order to avoid trivialities, we assume that every node that is not a terminal node has more than one daughter. We call a set of nodes that consists of all the daughters of a given nonterminal node a *sib*. We denote by $\mathcal{S}_A$ the sib consisting of the daughters of A.

We suppose that for each node A in $\mathcal{A}$ we have one simple support function focused on A and another focused on the complement $\bar{A}$. Here, as in our discussion of Barnett's technique, we begin by combining these two simple support functions. Then for each node A in $\mathcal{A}$ we have a single dichotomous belief function Bel_A with the dichotomy $\{A,\ \bar{A}\}$. We assume that $\mathrm{Bel}_A(A)$ and $\mathrm{Bel}_A(\bar{A})$ are both strictly less than one, but we allow either or both to be zero.

For any node A in the tree, we denote by $\mathrm{Bel}_A^{\downarrow}$ the orthogonal sum of Bel_B for all nodes B that are strictly below A. In Fig. 3, for example, $\mathrm{Bel}_H^{\downarrow} = \mathrm{Bel}_D \oplus \mathrm{Bel}_E$, and

$$\begin{aligned}\mathrm{Bel}_I^{\downarrow} &= \mathrm{Bel}_F \oplus \mathrm{Bel}_G \oplus \mathrm{Bel}_G^{\downarrow} \\ &= \mathrm{Bel}_F \oplus \mathrm{Bel}_G \oplus \mathrm{Bel}_A \oplus \mathrm{Bel}_B \oplus \mathrm{Bel}_C.\end{aligned}$$

If A is a terminal node, then $\mathrm{Bel}_A^{\downarrow}$ is vacuous. Our purpose, of course, is to calculate values of $\mathrm{Bel}_{\Theta}^{\downarrow} = \oplus\{\mathrm{Bel}_A | A \in \mathcal{A}\}$.

For each node A in $\mathcal{A}$, we denote by $\mathrm{Bel}_A^{\diamond}$ the orthogonal sum of Bel_B for all nodes B in $\mathcal{A}$ that are neither below A nor equal to A. Thus

$$\mathrm{Bel}_{\Theta}^{\downarrow} = \mathrm{Bel}_A^{\downarrow} \oplus \mathrm{Bel}_A \oplus \mathrm{Bel}_A^{\diamond}. \tag{23}$$

Lemma 1. *Suppose $\mathcal{P}$ is a partition of Θ, and $P \in \mathcal{A} \cap \mathcal{P}$. Then* $(\mathrm{Bel}_P^{\downarrow})_{\mathcal{P}} = (\mathrm{Bel}_P^{\downarrow})_{\{P,\bar{P}\}}$.

Proof. The belief function Bel_A has only $A, \bar{A}$, and Θ as focal elements. If $A \subseteq P$, then each of these focal elements either contains $\bar{P}$ or else is contained in P. A focal element S of $\mathrm{Bel}_P^{\downarrow}$ is obtained by intersecting such focal elements and hence must also either contain $\bar{P}$ or else be contained in P. If S contains $\bar{P}$ but is not equal to $\bar{P}$, then $S^{\mathcal{P}} = S^{\{P,\bar{P}\}} = \Theta$. If S is equal to $\bar{P}$, then $S^{\mathcal{P}} = S^{\{P,\bar{P}\}} = \bar{P}$. If S is contained in P, then $S^{\mathcal{P}} = S^{\{P,\bar{P}\}} = P$. In any case, $S^{\mathcal{P}} = S^{\{P,\bar{P}\}}$. □

Lemma 2. *Suppose $\mathcal{P}$ is a partition of Θ, $A \in \mathcal{A}$, and $\bar{A} \in \mathcal{P}$. Then* $(\mathrm{Bel}_A^{\diamond})_{\mathcal{P}} = (\mathrm{Bel}_A^{\diamond})_{\{A,\bar{A}\}}$.

Proof. Again, Bel_B has only B, $\bar{B}$, and Θ as focal elements. If $B \in \mathcal{A}$ and $B \nsubseteq A$, then B is either disjoint from A or else contains A, and hence each focal element of Bel_B either contains A or else is contained in $\bar{A}$. Any focal element S of $\mathrm{Bel}_A^{\diamond}$ is the intersection of such focal elements and hence must also contain A or else be contained in $\bar{A}$. If S is equal to A, then $S^{\mathcal{P}} = S^{\{A,\bar{A}\}} = S$. If S contains A but is not equal to A, then $S^{\mathcal{P}} = S^{\{A,\bar{A}\}} = \Theta$. If A is contained in $\bar{A}$, then $S^{\mathcal{P}} = S^{\{A,\bar{A}\}} = \bar{A}$. In any case, $S^{\mathcal{P}} = S^{\{A,\bar{A}\}}$. □

Lemma 3. *Suppose $\mathcal{P}$ is a partition of Θ. Then $\mathcal{P}$ discerns the interaction relevant to itself among the belief functions in $\{\mathrm{Bel}_P^{\downarrow} | P \in \mathcal{A} \cap \mathcal{P}\}$.*

Proof. Suppose $\mathcal{A} \cap \mathcal{P} = \{P_1, \ldots, P_n\}$, and let S_i be a focal element of Bel_{p_i} for $i = 1, 2, \ldots, n$. Fix an element P of $\mathcal{P}$, and suppose $S_i \cap P \neq \emptyset$ for $i = 1, \ldots, n$. We must show that $S_1 \cap \cdots \cap S_n \cap P \neq \emptyset$.

By the proof of Lemma 1, S_i either contains $\bar{P}_i$ or else is contained in P_i. Since $\mathcal{P}$ is a partition, P_i, and P are either disjoint or equal. If they are disjoint, then since $S_i \cap P \neq \emptyset, S_i$ cannot be contained in P_i; instead it must contain $\bar{P}_i$, and hence it must contain P.

At most one of the P_i can equal P. If none equal P, then all the S_i contain P, and hence $S_1 \cap \cdots \cap S_n \cap P = P$. If one, say P_j, does equal P, then

$$\begin{aligned} S_1 \cap \cdots \cap S_n \cap P &= S_j \cap \left[\bigcap_{i \neq j} (S_i \cap P) \right] \\ &= S_j \cap P. \end{aligned}$$

In either case, $S_1 \cap \cdots \cap S_n \cap P \neq \emptyset$. □

Since the partition $\mathcal{P}$ carries Bel_P for each $P \in \mathcal{A} \cap \mathcal{P}$, we can strengthen Lemma 3 to the statement that $\mathcal{P}$ discerns the interaction relevant to itself among the belief functions in

$$\{\mathrm{Bel}_P | P \in \mathcal{A} \cap \mathcal{P}\} \cup \left\{\mathrm{Bel}_P^{\downarrow} | P \in \mathcal{A} \cap \mathcal{P}\right\}.$$

Consider, for example, the partition $\mathcal{S}_A \cup \{\bar{A}\}$, where A is a nonterminal node in $\mathcal{A}$. This partition discerns the interaction relevant to itself among

$$\{\mathrm{Bel}_B | B \in \mathcal{S}_A\} \cup \left\{\mathrm{Bel}_B^{\downarrow} | B \in \mathcal{S}_A\right\}.$$

Since $\mathrm{Bel}_A^{\downarrow}$ is the orthogonal sum of these belief functions, it follows that

$$\left(\mathrm{Bel}_A^{\downarrow}\right)_{\mathcal{S}_A \cup \{\bar{A}\}} = \oplus \left\{ (\mathrm{Bel}_B)_{\mathcal{S}_A \cup \{\bar{A}\}} \oplus \left(\mathrm{Bel}_B^{\downarrow}\right)_{\mathcal{S}_A \cup \{\bar{A}\}} | B \in \mathcal{S}_A \right\}.$$

This can be written more simply as

$$\begin{aligned} &\left(\mathrm{Bel}_A^{\downarrow}\right)_{\mathcal{S}_A \cup \{\bar{A}\}} = \oplus \left\{ \mathrm{Bel}_B \oplus \left(\mathrm{Bel}_B^{\downarrow}\right)_{\{B,\bar{B}\}} | B \in \mathcal{S}_A \right\}; \\ &(\mathrm{Bel}_B)_{\mathcal{S}_A \cup \{\bar{A}\}} = \mathrm{Bel}_B \end{aligned} \tag{24}$$

because Bel_B is carried by $\mathcal{S}_A \cup (\bar{A}\}$, and

$$\left(\mathrm{Bel}_B^{\downarrow}\right)_{\mathcal{S}_A \cup \{\bar{A}\}} = \left(\mathrm{Bel}_B^{\downarrow}\right)_{\{B,\bar{B}\}}$$

by Lemma 1. It should be borne in mind that if the element B of $\mathcal{S}_A$ is a terminal node, then $\mathrm{Bel}_B^{\downarrow}$ is vacuous, and the orthogonal sum $\mathrm{Bel}_B \oplus (\mathrm{Bel}_B^{\downarrow})_{\{B,\bar{B}\}}$ reduces to Bel_B.

The reasoning of the preceding paragraph applies to the case where A is the topmost node Θ, except that in this case the partition is simply $\mathcal{S}_\Theta$, not $\mathcal{S}_\Theta \cup \{\bar{\Theta}\}$. So

$$\left(\mathrm{Bel}_\Theta^{\downarrow}\right)_{\mathcal{S}_\Theta} = \oplus \left\{ \mathrm{Bel}_B \oplus \left(\mathrm{Bel}_B^{\downarrow}\right)_{\{B,\bar{B}\}} | B \in \mathcal{S}_\Theta \right\}. \tag{25}$$

Formulas (24) and (25) tell us that in order to find for A and her immediate daughters the degrees of belief resulting from all the evidence bearing on nodes below A, it is sufficient to consider each daughter separately. We find the degrees of belief for and against each daughter resulting from evidence bearing directly on it and on nodes below it, and then we combine the results for the different daughters.

Lemma 4. *Suppose A is a nonterminal element of $\mathcal{A}$. Then the partition $\mathcal{S}_A \cup \{\bar{A}\}$ discerns the interaction relevant to itself between* $\mathrm{Bel}^{\downarrow}_A$ *and* $\mathrm{Bel}^{\diamond}_A$*.*

Proof. Suppose S_1 is a focal element of $\mathrm{Bel}^{\downarrow}_A$, and S_2 is a focal element of $\mathrm{Bel}^{\diamond}_A$. Then S_1 either contains $\bar{A}$ or is contained in A, while S_2 either contains A or is contained in $\bar{A}$. Table 1 lists the four possibilities and shows what can happen when $S_1 \cap S_2$ is intersected with an element P of $\mathcal{S}_A \cup (\bar{A}\}$. Inspection of the table shows that if $S_1 \cap P \neq \emptyset$ and $S_2 \cap P \neq \emptyset$, then $S_1 \cap S_2 \cap P \neq \emptyset$. This establishes that $\mathcal{S}_A \cup (\bar{A}\}$ discerns the interaction relevant to itself between $\mathrm{Bel}^{\downarrow}_A$ and $\mathrm{Bel}^{\diamond}_A$. □

Since Bel_A is carried by $\mathcal{S}_A \cup \{\bar{A}\}$, Lemma 4 can be strengthened to the statement that $\mathcal{S}_A \cup \{\bar{A}\}$ discerns the interaction relevant to itself among $\mathrm{Bel}^{\downarrow}_A$, Bel_A, and $\mathrm{Bel}^{\diamond}_A$. So from (23) we can obtain

$$\left(\mathrm{Bel}^{\downarrow}_{\Theta}\right)_{\mathcal{S}_A \cup \{\bar{A}\}} = \left(\mathrm{Bel}^{\downarrow}_{A}\right)_{\mathcal{S}_A \cup \{\bar{A}\}} \oplus (\mathrm{Bel}_A)_{\mathcal{S}_A \cup \{\bar{A}\}} \oplus \left(\mathrm{Bel}^{\diamond}_{A}\right)_{\mathcal{S}_A \cup \{\bar{A}\}} . \qquad (26)$$

Since Bel_A is carried by $\{A, \bar{A}\}$, and since

$$\left(\mathrm{Bel}^{\diamond}_{A}\right)_{\mathcal{S}_A \cup \{\bar{A}\}} = \left(\mathrm{Bel}^{\diamond}_{A}\right)_{\{A, \bar{A}\}}$$

by Lemma 2, (26) reduces to

$$\left(\mathrm{Bel}^{\downarrow}_{\Theta}\right)_{\mathcal{S}_A \cup \{\bar{A}\}} = (\mathrm{Bel}^{\downarrow}_{A})_{\mathcal{S}_A \cup \{\bar{A}\}} \oplus \mathrm{Bel}_A \oplus (\mathrm{Bel}^{\diamond}_{A})_{\{A, \bar{A}\}} . \qquad (27)$$

This formula tells us that evidence from above A and down other branches affects our degrees of belief about the daughters of A only inasmuch as it affects our degrees of belief for and against A itself.

Table 1. Verifying the discernment

	$P = \bar{A}$	$P \in S_A$
$S_1 \supseteq \bar{A},\ S_2 \supseteq A$	$S_1 \cap S_2 \cap P = S_2 \cap P$	$S_1 \cap S_2 \cap P = S_1 \cap P$
$S_1 \subseteq A,\ S_2 \supseteq A$ $S_1 \subseteq A,\ S_2 \subseteq \bar{A}$	$S_1 \cap S_2 \cap P = S_1 \cap P = \emptyset$	
$S_1 \supseteq \bar{A},\ S_2 \subseteq \bar{A}$	$S_1 \cap S_2 \cap P = S_2$	$S_1 \cap S_2 \cap P = S_2 \cap P = \emptyset$

In the next section, we will have occasion to use two consequences of (27):

$$\left(\mathrm{Bel}_{\Theta}^{\downarrow}\right)_{\{A,\bar{A}\}} = \left(\mathrm{Bel}_{A}^{\downarrow}\right)_{\{A,\bar{A}\}} \oplus \mathrm{Bel}_A \oplus (\mathrm{Bel}_{A}^{\diamond})_{\{A,\bar{A}\}} \tag{28}$$

and

$$\left(\mathrm{Bel}_{\Theta}^{\downarrow}\right)_{\{B,A-B,\bar{A}\}} = \left(\mathrm{Bel}_{A}^{\downarrow}\right)_{\{B,A-B,\bar{A}\}} \oplus \mathrm{Bel}_A \oplus (\mathrm{Bel}_{A}^{\diamond})_{\{A,\bar{A}\}} \tag{29}$$

for every B in $\mathcal{S}_A$. These formulas follow from (27) because both the partition $\{A, \bar{A}\}$ and the partition $\{B, A - B, \bar{A}\}$ carry the belief function $\mathrm{Bel}_A \oplus (\mathrm{Bel}_A^{\diamond})_{\{A,\bar{A}\}}$. Whenever a partition carries a belief function, it discerns the interaction relevant to itself between that belief function and any other belief function.

4 Implementing Dempster's Rule

We now present our algorithm for calculating $\mathrm{Bel}_{\Theta}^{\downarrow}(A)$ for A in $\mathcal{A}$. We first present the algorithm in general terms and explain how it is justified by the results of the preceding section. We then give detailed formulas for the actual calculations. We conclude with a complexity analysis and a comparison of the complexity with that of Gordon and Shortliffe's algorithm.

The algorithm can be broken down into three states. In the first stage we begin with sibs of terminal nodes, combine the belief functions attached to them to find degrees of belief for and against their mothers, then do the same for the mothers' mothers, and so on, until we have a dichotomous belief function attached to each each daughter of Θ to obtain the values of $\mathrm{Bel}_{\Theta}^{\downarrow}$ for these daughters. In the third stage we use information stored as we moved up the tree to move back down, calculating further values of $\mathrm{Bel}_{\Theta}^{\downarrow}$ as we go.

4.1 First stage

Recall that we begin with a dichotomous belief function Bel_A attached to each node A of $\mathcal{A}$.

Choose a sib of terminal nodes, and let A denote its mother. According to (24),

$$\left(\mathrm{Bel}_{A}^{\downarrow}\right)_{\mathcal{S}_A \cup \{\bar{A}\}} = \oplus \left\{\mathrm{Bel}_B | B \in \mathcal{S}_A\right\}. \tag{30}$$

Since Bel_B is dichotomous with dichotomy $\{B, \bar{B}\}$, and since B is an atom of the partition $\mathcal{S}_A \cup \{\bar{A}\}$, Barnett's technique can be used to calculate values of the orthogonal sum in this formula. We use it to calculate $\mathrm{Bel}_A^{\downarrow}(A)$ and $\mathrm{Bel}_A^{\downarrow}(\bar{A})$ – i.e., to find $(\mathrm{Bel}_A^{\downarrow})_{\{A,\bar{A}\}}$.

We now compute $\mathrm{Bel}_A \oplus (\mathrm{Bel}_A^{\downarrow})_{\{A,\bar{A}\}}$. This is easy, since both $(\mathrm{Bel}_A^{\downarrow})_{\{A,\bar{A}\}}$ and Bel_A are dichotomous with dichotomy $\{A, \bar{A}\}$. We discard Bel_A and store in its place both $(\mathrm{Bel}_A^{\downarrow})_{\{A,\bar{A}\}}$ and $\mathrm{Bel}_A \oplus (\mathrm{Bel}_A^{\downarrow})_{\{A,\bar{A}\}}$. This means that we store four numbers at $A:$ $\mathrm{Bel}_A^{\downarrow}(A)$, $\mathrm{Bel}_A^{\downarrow}(\bar{A})$, $(\mathrm{Bel}_A \oplus \mathrm{Bel}_A^{\downarrow})(A)$, and $(\mathrm{Bel}_A \oplus \mathrm{Bel}_A^{\downarrow})(\bar{A})$.

After we have completed the procedure of the two preceding paragraphs for every sib of terminal nodes, we temporarily prune these terminal nodes from the tree, as it were, so that the mothers of the original sibs of terminal nodes are not themselves terminal nodes. We then repeat the procedure with the sibs of terminal nodes we now see, except that instead of (30), we now use (24),

$$\left(\mathrm{Bel}_A^{\downarrow}\right)_{\mathcal{S}_A \cup \{\bar{A}\}} = \oplus \left\{ \mathrm{Bel}_B \oplus \left(\mathrm{Bel}_B^{\downarrow}\right)_{\{B,\bar{B}\}} | B \in \mathcal{S}_A \right\}$$

to calculate $\mathrm{Bel}_A^{\downarrow}(A)$ and $\mathrm{Bel}_A^{\downarrow}(\bar{A})$ for the mother A of what are now terminal sibs. (Of course, we really used (24) in the first round, too. When we wrote Bel_B instead of $\mathrm{Bel}_B \oplus (\mathrm{Bel}_B^{\downarrow})_{\{B,\bar{B}\}}$ in (30) above, we were just taking advantage of the fact that $\mathrm{Bel}_B^{\downarrow}$ is vacuous when B is terminal.)

We continue this process until we have reached the daughters of the the topmost node Θ. We then have $(\mathrm{Bel}_A^{\downarrow})_{\{A,\bar{A}\}}$ and $\mathrm{Bel}_A \oplus (\mathrm{Bel}_A^{\downarrow})_{\{A,\bar{A}\}}$ stored at every node A in $\mathcal{A}$.

4.2 Second stage

Recall (25),

$$\left(\mathrm{Bel}_\Theta^{\downarrow}\right)_{\mathcal{S}_\Theta} = \oplus \left\{ \mathrm{Bel}_B \oplus \left(\mathrm{Bel}_B^{\downarrow}\right)_{\{B,\bar{B}\}} | B \in \mathcal{S}_\Theta \right\}.$$

We apply Barnett's technique to this formula to calculate $\mathrm{Bel}_\Theta^{\downarrow}(A)$ and $\mathrm{Bel}_\Theta^{\downarrow}(A)$ for each A in $\mathcal{S}_\Theta$. Knowing these two numbers amounts to knowing $(\mathrm{Bel}_\Theta^{\downarrow})_{\{A,\bar{A}\}}$. We store them at A, along side the four numbers already there.

4.3 Third stage

Now consider a particular daughter A of Θ. We want to calculate $\mathrm{Bel}_\Theta^{\downarrow}(B)$ and $\mathrm{Bel}_\Theta^{\downarrow}(\bar{B})$ for each daughter B of A. We can do this using (24), (28), and (29).

Consider first (28):

$$\left(\mathrm{Bel}_\Theta^{\downarrow}\right)_{\{A,\bar{A}\}} = \left(\mathrm{Bel}_{\mathrm{A}}^{\downarrow}\right)_{\{A,\bar{A}\}} \oplus \mathrm{Bel}_A \oplus (\mathrm{Bel}_A^{\diamond})_{\{A,\bar{A}\}}.$$

All the belief functions in this formula are dichotomous with dichotomy $\{A,\bar{A}\}$, and $(\mathrm{Bel}_\Theta^{\downarrow})_{\{A,\bar{A}\}}$ and $(\mathrm{Bel}_A^{\downarrow})_{\{A,\bar{A}\}}$ are stored at A. So we can easily find $\mathrm{Bel}_A \oplus (\mathrm{Bel}_A^{\diamond})_{\{A,\bar{A}\}}$ by division.

Now consider (24) again. We have already applied Barnett's technique to this formula to calculate $\mathrm{Bel}_A^{\downarrow}(A)$ and $\mathrm{Bel}_A^{\downarrow}(\bar{A})$. We now apply it again to calculate $\mathrm{Bel}_A^{\downarrow}(B)$, $\mathrm{Bel}_A^{\downarrow}(\bar{B})$, $\mathrm{Bel}_A^{\downarrow}(A-B)$, and $\mathrm{Bel}_A^{\downarrow}(B \cup \bar{A})$ for each B in $\mathcal{S}_A$. This gives us the belief function $(\mathrm{Bel}_A^{\downarrow})_{\{B,A-B,\bar{A}\}}$. (Actually, as we shall see in the next section, we do not need to calculate $\mathrm{Bel}_A^{\downarrow}(A-B)$.)

Now consider (29):

$$\left(\mathrm{Bel}_{\Theta}^{\downarrow}\right)_{\{B,A-B,\bar{A}\}} = \left(\mathrm{Bel}_{\mathrm{A}}^{\downarrow}\right)_{\{B,A-B,\bar{A}\}} \oplus \mathrm{Bel}_A \oplus (\mathrm{Bel}_A^{\diamond})_{\{A,\bar{A}\}} .$$

We have just found $\mathrm{Bel}_A \oplus (\mathrm{Bel}^{\diamond})_{\{A,\bar{A}\}}$ and $(\mathrm{Bel}_A^{\downarrow})_{\{B,A-B,\bar{A}\}}$. So we can use (29) to calculate $\mathrm{Bel}_{\Theta}^{\downarrow}(B)$ and $\mathrm{Bel}_{\Theta}^{\downarrow}(\bar{B})$. (Barnett's technique cannot be used here, since $(\mathrm{Bel}_A^{\downarrow})_{\{B,A-B,\bar{A}\}}$ is not dichotomous. But since the partition we are working with is only a trichotomy, a brute force application of Dempster's rule involves little computation.)

We have just seen how to go from $\mathrm{Bel}_{\Theta}^{\downarrow}(A)$ and $\mathrm{Bel}_{\Theta}^{\downarrow}(\bar{A})$ to $\mathrm{Bel}_{\Theta}^{\downarrow}(B)$ and $\mathrm{Bel}_{\Theta}^{\downarrow}(\bar{B})$ for the daughters B of A. This process can be repeated for the daughters of each B, and so on, until we have calculated $\mathrm{Bel}_{\Theta}^{\downarrow}(C)$ and $\mathrm{Bel}_{\Theta}^{\downarrow}(\bar{C})$ for every node C in the tree.

Usually, of course, we will not be interested in $\mathrm{Bel}_{\Theta}^{\downarrow}(C)$ and $\mathrm{Bel}_{\Theta}^{\downarrow}(\bar{C})$ for every node C in the tree. Once we have seen that $\mathrm{Bel}_{\Theta}^{\downarrow}(B)$ is very small, we know that $\mathrm{Bel}_{\Theta}^{\downarrow}(C)$ will be at least as small for every descendant C of B, and so we may not want to go to the trouble of finding these values. We may decide to look at descendants of B only if $\mathrm{Bel}_{\Theta}^{\downarrow}(B)$ is greater than 0.5, say. Since two disjoint sets cannot both have degree of belief greater than 0.5, this decision will result in our moving down the tree along just one path, which may stop before reaching a terminal node.

4.4 Details of the algorithm

The numerical calculations that our algorithm requires can be described by formulas, and we can group these formulas into six subroutines.

The following notation will allow us to write these formulas concisely. For each node A in $\mathcal{A}$, we set

$$\begin{array}{ll}
A_0^+ = \mathrm{Bel}_A\,(A)\,, & A_0^- = \mathrm{Bel}_A\left(\bar{A}\right), \\
A_{\downarrow}^+ = \mathrm{Bel}_A^{\downarrow}\,(A)\,, & A_{\downarrow}^- = \mathrm{Bel}_A^{\downarrow}\left(\bar{A}\right), \\
A^+ = (\mathrm{Bel}_A \oplus \mathrm{Bel}_A^{\downarrow})\,(A)\,, & A^- = (\mathrm{Bel}_A \oplus \mathrm{Bel}_A^{\downarrow})\left(\bar{A}\right), \\
A_{\diamond}^+ = (\mathrm{Bel}_A \oplus \mathrm{Bel}_A^{\diamond})\,(A)\,, & A_{\diamond}^- = (\mathrm{Bel}_A \oplus \mathrm{Bel}_A^{\diamond})\left(\bar{A}\right), \\
A_{\Theta}^+ = \mathrm{Bel}_{\Theta}^{\downarrow}\,(A)\,, & A_{\Theta}^- = \mathrm{Bel}_{\Theta}^{\downarrow}\left(\bar{A}\right),
\end{array}$$

(If A is a terminal node, then $\mathrm{Bel}_A^{\downarrow}$ is vacuous, and therefore $A_{\downarrow}^+ = A_{\downarrow}^- = 0$, $A^+ = A_0^+$, and $A^- = A_0^-$.) For each node B other than Θ and its daughters, we set

$$\begin{aligned} B_A^+ &= \mathrm{Bel}_A^{\downarrow}\left(B\right), \qquad B_A^- = \mathrm{Bel}_A^{\downarrow}\left(B\right), \\ B_A^* &= \mathrm{Bel}_A^{\downarrow}\left(B \cup \bar{A}\right), \end{aligned}$$

where A is B's mother.

Recall that the first stage of our algorithm begins with the computation of $(\mathrm{Bel}_A^{\downarrow})_{\{A,\bar{A}\}}$ for mothers of sibs of terminal nodes. Subroutine 1 specifies how this is done. This is followed by the calculation of $\mathrm{Bel}_A \oplus (\mathrm{Bel}_A^{\downarrow})_{\{A,A\}}$, by Subroutine 2. After these operations have been completed for every node A whose daughters are all terminal nodes, we pretend to prune all these terminal nodes from the tree, and we repeat the process with the new sibs of terminal nodes, and so on. Each round uses Subroutine 1 followed by Subroutine 2. We continue until we have calculated $(\mathrm{Bel}_A^{\downarrow})_{\{A,\bar{A}\}}$ for the the daughters A of Θ.

At the second stage we apply Barnett's technique to (25) to find $\mathrm{Bel}_{\Theta}^{\downarrow}(A)$ for $A \in \mathcal{S}_{\Theta}$. This is Subroutine 3.

In the second stage, we go back down the tree. When we go from A to its daughters, we first find $\mathrm{Bel}_A \oplus (\mathrm{Bel}_A^{\diamond})_{\{A,\bar{A}\}}$ using (28); this is Subroutine 4. Then we return to formula (24) and calculate $\mathrm{Bel}_A^{\downarrow}(B)$, $\mathrm{Bel}_A^{\downarrow}(\bar{B})$ and $\mathrm{Bel}_A^{\downarrow}(B \cup \bar{A})$ for each B in $\mathcal{S}_A$; this is Subroutine 5. Finally, we use (29) to calculate $\mathrm{Bel}_{\Theta}^{\downarrow}(B)$ and $\mathrm{Bel}_{\Theta}^{\downarrow}(\bar{B})$ for each B in $\mathcal{S}_A$; this is Subroutine 6. (Alternatively, to minimize storage, we may execute Subroutines 5 and 6 for a particular B in $\mathcal{S}_A$, then for another, and so on.)

In summary, we repeatedly cycle through Subroutines 4, 5 and 6 as we move up the tree, we execute Subroutine 3 once at the top of the tree, and then we repeatedly cycle through Subroutines 4, 5 and 6 as we move back down.

Subroutine 1. Calculating $A_{\downarrow}^+$ and $A_{\downarrow}^-$ from B^+ and B^- for B in $\mathcal{S}_A$;

$$A_{\downarrow}^+ = 1 - K, \qquad A_{\downarrow}^- = K \prod_{B \in \mathcal{S}_A} B^- / \left(1 - B^+\right),$$

where

$$K^{-1} = 1 + \sum_{B \in \mathcal{S}_A} B^+ / \left(1 - B^+\right).$$

Subroutine 2. Calculating A^+ and A^- from A_0^+, A_0^-, $A_{\downarrow}^+$, and $A_{\downarrow}^-$:

$$A^+ = 1 - K\left(1 - A_0^+\right)\left(1 - A_{\downarrow}^+\right), \qquad A^- = 1 - K\left(1 - A_0^-\right)\left(1 - A_{\downarrow}^-\right),$$

where

$$K^{-1} = 1 - A_0^+ A_{\downarrow}^- - A_0^- A_{\downarrow}^+.$$

Subroutine 3. Calculating A_{Θ}^+ and A_{Θ}^- for A in $\mathcal{S}_{\Theta}$ from A^+ and A^- for A in $\mathcal{S}_{\Theta}$:

$$A^+_\Theta = 1 - K\left(1 + \sum_{\substack{B \in \mathcal{S}_\Theta \\ B \neq A}} B^+ / \left(1 - B^+\right)\right) - \prod_{\substack{B \in \mathcal{S}_\Theta \\ B \neq A}} B^- / \left(1 - B^+\right),$$

$$A^-_\Theta = 1 - K\left(1 - A^-\right) / \left(1 - A^+\right).$$

where

$$K^{-1} = 1 + \sum_{B \in \mathcal{S}_\Theta} B^+ / \left(1 - B^+\right) - \prod_{B \in \mathcal{S}_\Theta} B^- / \left(1 - B^+\right).$$

Subroutine 4. Calculating $A^+_\diamond$ and $A^-_\diamond$ from $A^+_\Theta, A^-_\Theta,\ A^+_\downarrow$, and $A^-_\downarrow$;

$$A^+_\diamond = 1 - K\left(1 - A^+_\Theta\right) / (1 - A^+_\downarrow), \qquad A^-_\diamond = 1 - K\left(1 - A^-_\Theta\right) / \left(1 - A^-_\downarrow\right),$$

where

$$K^{-1} = \frac{1 - A^+_\Theta}{1 - A^+_\downarrow} + \frac{1 - A^-_\Theta}{1 - A^-_\downarrow} - \frac{1 - A^+_\Theta - A^-_\Theta}{1 - A^+_\downarrow - A^-_\downarrow}.$$

Subroutine 5. Calculating $B^+_A,\ B^-_A$, and B^*_A from C^+ and C^- for C in $\mathcal{S}_A$:

$$B^+_A = 1 - K\left(1 + \sum_{\substack{C \in \mathcal{S}_A \\ C \neq B}} C^+ / \left(1 - C^+\right)\right),$$

$$B^-_A = 1 - K\left(1 - B^-\right) / \left(1 - B^+\right),$$

$$B^*_A = 1 - K\left(1 + \sum_{\substack{C \in \mathcal{S}_A \\ C \neq B}} C^+ / \left(1 - C^+\right) - \prod_{\substack{C \in \mathcal{S}_A \\ C \neq B}} C^- / \left(1 - C^+\right)\right),$$

where

$$K^{-1} = 1 + \sum_{C \in \mathcal{S}_A} C^+ / \left(1 - C^+\right).$$

Subroutine 6. Calculating B^+_Θ and B^-_Θ from $A^+_\downarrow,\ A^-_\downarrow,\ A^+_\diamond,\ A^-_\diamond,\ B^+_A,\ B^-_A$ and B^*_A, where B is a daughter of A:

$$B^+_\Theta = K\left(A^+_\diamond\left(B^*_A - A^-_\downarrow\right) + \left(1 - A^+_\diamond - A^-_\diamond\right) B^+_A\right),$$

$$B^-_\Theta = 1 - K\left(1 - A^-_\diamond\right)\left(1 - B^-_A\right),$$

where

$$K^{-1} = 1 - A^+_\downarrow A^-_\diamond - A^-_\downarrow A^+_\diamond.$$

4.5 Miscellaneous comments

(1) The constant K in Subroutine 5 is the same as the constant K in Subroutine 1. Recognition of this fact will save computation on the way back down the tree, since we store $\mathrm{Bel}_A^{\downarrow}(A) = 1 - K$ on our way up the tree. It is probably most efficient, in fact, to store K^{-1} or

$$K^{-1} - 1 = \sum_{B \in \mathcal{S}_A} B^+ / \left(1 - B^+\right)$$

instead of $\mathrm{Bel}_A^{\downarrow}(A)$.

(2) Each sum or product in Subroutine 5 differs from the corresponding product in Subroutine 1 only by the omission of a single term or factor. So if we save the sums and products from Subroutine 1, we can obtain those in Subroutine 5 by subtraction and division. This may be advantageous when the sib sizes are large.

(3) In our description of the procedure for moving up the tree, we specified that $\mathrm{Bel}_A^{\downarrow}(A)$ and $\mathrm{Bel}_A^{\downarrow}(\bar{A})$ should be calculated first for those A whose daughters are all terminal, then for those A whose daughters are either terminal or else have only terminal daughters, and so on. In fact, however, we have more freedom of choice than this. In order to calculate $\mathrm{Bel}_A^{\downarrow}(A)$ and $\mathrm{Bel}_A^{\downarrow}(\bar{A})$ it is necessary only that these quantities should already have been calculated for each nonterminal daughter of A.

(4) We could move up the tree faster if we were to calculate $(\mathrm{Bel}_A^{\downarrow})_{\{A,\bar{A}\}}$ for disjoint A in parallel. A similar opportunity for parallelism occurs when we move back down the tree, provided we want to move down all the branches.

(5) We set out to calculate only $\mathrm{Bel}_{\Theta}^{\downarrow}(A)$ for all A in $\mathcal{A}$. As it turned out, we also calculated $\mathrm{Bel}_{\Theta}^{\downarrow}(\bar{A})$, since this was necessary for calculating the values of $\mathrm{Bel}_{\Theta}^{\downarrow}$ for A's daughters. (This means we can calculate the plausibility of A, $\mathrm{Pl}_{\Theta}^{\downarrow}(A) = 1 - \mathrm{Bel}_{\Theta}^{\downarrow}(\bar{A})$.) A glance at (24) and (27) makes it clear that we can also calculate $\mathrm{Bel}_{\Theta}^{\downarrow}(B)$ for any B that is in the field $(\mathcal{S}_A \cup \{\bar{A}\})^*$ for some node A. In general, however, there will remain many subsets B of subsets B of Θ for which our method is not helpful. It example, help us calculate $\mathrm{Bel}_{\Theta}^{\downarrow}(\{d,\ f\})$ in Fig. 3.

(6) We have used Barnett's technique in Subroutines 1, 3, and 5. (Subroutine 2 can also be regarded as an application of Barnett's technique, but there is really no distinction between Barnett's technique and brute-force calculation of an orthogonal sum when we are working with a single dichotomy.) However, we have used this technique only on the partitions $\mathcal{S}_A^- \cup (\bar{A}\}$. If the sibs $\mathcal{S}_A$ are all relatively small—if, say, no sib contains more than three or four daughters—then those calculations would be manageable even without Barnett's technique. Thus, the efficiency of our algorithm is mainly due not to Barnett's technique but to the fact that we are able to break the overall computation down into local computations.

4.6 Complexity analysis

It is clear from our description of the algorithm that the amount of arithmetic involving a particular node does not depend on the size of the tree. It depends only on the number of the node's daughters, and it increases linearly with the number of daughters. (Subroutine 1, for example has a product with a factor for each daughter and a sum with a term for each daughter.) It follows that the computational complexity of the algorithm is linear in the number of nodes in the tree.

We can make a closer complexity analysis if we assume that the number of daughters in a sib (the branching factor) is constant throughout the tree. Let f denote the branching factor. Let n denote the number of sibs or, equivalently, the number of nonterminal nodes. Then we can expect $a + bf$ arithmetic operations for each sib, or $n(a + bf)$ altogether, where a and b are positive constants.

This formula clarifies the role of Barnett's technique in our algorithm. Barnett's technique is responsible for the linearity with respect to the sib size f, while the localization of the computation is responsible for the linearity with respect to the number of sibs, n. If we did not use Barnett's technique, the computational complexity would be exponential in f but stll proportional to n. In place of $n(a + bf)$, we would have $n\exp(a + bf)$.

Instead of talking about the number of arithmetic operations per sib, we might wish to talk about the number per node. Since there are $nf + 1$ nodes altogether, this is

$$\frac{n\,(a+bf)}{nf+1} \approx \frac{n(a+bf)}{nf} = \frac{a}{f} + b. \tag{31}$$

Alternatively, we might wish to talk about the number of operations per terminal node, since the number of terminal nodes is the size of our frame. Since there are $nf - n + 1$ terminal nodes, the number of operations per terminal node is

$$\frac{n\,(a+bf)}{nf-n+1} \approx \frac{n\,(a+bf)}{n\,(f-1)} = \frac{a}{f-1} + \frac{bf}{f-1}. \tag{32}$$

Both (31) and (32) are greatest for binary trees ($f = 2$) and tend towards b as f increases. (There is no paradox here. When f is large, most nodes are terminal nodes.)

The formula $a + bf$ for the number of operations per sib can be verified empirically. We have verified it using a LISP implementation in a variety of trees, with f ranging up to 5 and n ranging up to 30,000. The fit was excellent, with 99.8% of the variance explained. The least squares estimates were $a = 158$ and $b = 44$. (Strictly speaking, the counts on which these estimates were based are counts of arguments in operations rather than counts of operations. Thus an addition of k terms counts as k, and the division of one number by another counts as 2.)

As we mentioned in Sect. 4.3, it is often possible to save computation by moving down only some of the branches in the third stage. Since Subroutine 5 involves the greatest computation, the savings can be substantial.

4.7 Comparison with Gordon and Shortliffe's algorithm

Gordon and Shortliffe [4] do not give details for the implementation of their algorithms. We have found, however, that it can also be implemented in linear time. The particular implementation we have used is analogous to the implementation of our own algorithm; it involves movements up and down the tree. We have found that this implementation of Gordon and Shortliffe's algorithm is comparable in complexity to our algorithm. In all the trees we checked it required fewer arithmetic operations than our algorithm, but never fewer than half as many.

The details of our implementation of Gordon and Shortliffe's algorithm are nearly as complicated as the details of our algorithm, and it is possible that a more efficient implementation might be found.

5 Generalizations

In this article, we have retained Gordon and Shortliffe's assumption that the belief functions being combined are simple support functions focused on nodes or their complements. The essence of our computational scheme can be retained, however, whenever each belief function is carried by a sib (more precisely, by a partition $\mathcal{S}_A \cup \{\bar{A}\}$ for some node A). Under this more general assumption, Barnett's technique is no longer available, and the amount of arithmetic is exponential in the sib size, but it remains proportional to the number of sibs. An interesting special case occurs when each belief function is conditionally Bayesian—i.e., when the belief function Bel_A carried by $\mathcal{S}_A \cup \{\bar{A}\}$ satisfies

$$\mathrm{Bel}_A\,(B|A) + \mathrm{Bel}_A\,\left(\bar{B}|A\right) = 1$$

and

$$\mathrm{Bel}_A\,\left(B|\bar{A}\right) + \mathrm{Bel}_A\,\left(\bar{B}|\bar{A}\right) = 1$$

for every element B of the field $(\mathcal{S}_A \cup \{\bar{A}\})^*$. In this case, the result of combining all the belief functions is Bayesian, and the computations can be simplified; the amount of arithmetic is again linear rather than exponential in the sib size. This case has been studied by Pearl [7].

A further generalization is to replace diagnostic trees with general trees of partitions or variables. We need only a "Markov" property: a given node in the tree should discern the interaction among the belief functions on the different branches of tree separated by the node. The problem of propagating belief functions in such Markov trees is discussed by Shenoy and Shafer [14] and by

Shafer, Shenoy, and Mellouli [13]. The Bayesian special case is discussed by Pearl [8].

The generalization to networks of variables has been studied by Kong [5]; see also Mellouli, Shafer, and Shenoy [6]. The last chapter of Kong [5] is of particular interest; it shows how the algorithm of this article can be generalized, without loss of computational efficiency, to the case where a patient may have more than one disease.

Acknowledgment

Research for this article has been partially supported by NSF grants IST-8405210 and IST-8610293 and by ONR grant N00014-85-K-0490. The article has been improved by the thoughtful criticism of a referee. Chuck Hand and Khaled Mellouli assisted with the complexity analysis in Sect. 4.6.

References

1. Barnett, J.A., Computational methods for a mathematical theory of evidence, in: *Proceedings IJCAI-81*, Vancouver, BC (1981) 868–875.
2. Garvey, T.D., Lowrance, J.D. and Fischler, M.A., An inference technique for integrating knowledge from disparate sources, in: *Proceedings IJCAI-81*, Vancouver, BC (1981) 319–325.
3. Gordon, J. and Shortliffe, E.H., The Dempster–Shafer theory of evidence, in: B.G. Buchanan and E.H Shortliffe (Eds.), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project* (Addison-Wesley, Reading, MA, 1985) 272–292.
4. Gordon, J. and Shortliffe, E.H., A method for managing evidential reasoning in a hierarchical hypothesis space, *Artificial Intelligence* **26** (1985) 323–357.
5. Kong, A., Multivariate belief functions and graphical models, Doctoral Dissertation, Department of Statistics, Harvard University, Cambridge, MA, 1986.
6. Mellouli, K., Shafer, G. and Shenoy, P., Qualitative Markov networks, in: *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Paris (1986) 31–35.
7. Pearl, J., On evidential reasoning in a hierarchy of hypotheses, *Artificial Intelligence* **28** (1986) 9–15.
8. Pearl, J., Fusion, propagation, and structuring in belief networks, *Artificial Intelligence* **29** (1986) 241–288.
9. Shafer, G., *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, NJ, 1976.
10. Shafer, G., Belief functions and possibility measures, in: J.C. Bizdek (Ed.), *The Analysis of Fuzzy Information* **2** (CRC Press, 1987).
11. Shafer, G., Probability judgment in artificial intelligence and expert systems, *Stat. Sci.* **2** (1987) 3–16.
12. Shafer, G., Hierarchical evidence, in: *Proceedings Second Conference on Artificial Intelligence Applications*, Miami, FL (1985) 16–21.

13. Shafer, G., Shenoy, P. and Mellouli, K., Propagating belief functions in qualitative Markov trees, Working Paper No. 190, School of Business, University of Kansas, Lawrence, KS, 1987.
14. Shenoy, P. and Shafer, G., Propagating belief functions with local computations, *IEEE Expert* **1**(3) (1986) 43–52.

19

Some Characterizations of Lower Probabilities and Other Monotone Capacities through the use of Möbius Inversion

Alain Chateauneuf and Jean-Yves Jaffray

Abstract. Monotone capacities (on finite sets) of finite or infinite order (lower probabilities) are characterized by properties of their Möbius inverses. A necessary property of probabilities dominating a given capacity is demonstrated through the use of Gale's theorem for the transshipment problem. This property is shown to be also sufficient if and only if the capacity is monotone of infinite order. A characterization of dominating probabilities specific to capacities of order 2 is also proved.

Key words: Decision theory; Lower probabilities; Belief functions; Capacities; Möbius inversion; Representation of uncertainty

1 Introduction

Dempster (1967) and Shafer (1976, 1981) have proposed a representation of uncertain environments which entails assigning a 'lower probability' (Dempster) or 'degree of belief' (Shafer) to every event – or proposition. Their model requires the lower probability (belief) function, which is, in general, not additive, to possess a weaker property: monotonicity of order K, for all K. This requirement is perfectly justified in some situations, such as the following example given by Dempster: suppose that there is a probability $\pi(x)$ of receiving a message 'B_x', $x \in X$, which informs one that event B_x obtains, in which case any event A such that $B_x \subset A$ also obtains; thus, although a given event A is not a probabilized event, one can assert that it is at least as likely to obtain as any event with probability $f(A) = \Sigma_{B \subset A} m(B)$, where $m(B) = \Sigma_{\{x \in X:\ B_x = B\}}\ \pi(x)$. The lower probability function f can indeed be shown to be monotone of order K for all K (for short: ∞-monotone).

However, it is easy to give other examples where partial information is best described by a function with weaker properties. In particular, consider the case where all probabilities of a given set $\mathcal{P}$ are compatible with the available data.

Function $f = \mathrm{Inf}_{P \in \mathcal{P}} P$ is in this case the natural 'lower probability' function. However f is not, in general, monotone of order K, for all K; f is nonetheless always monotone of order 1, and often monotone of superior orders (see Examples 6 and 7). Decision theorists, such as Kyburg (1974), Levi (1980), Walley and Fine (1979, 1982), Wolfenson and Fine (1982), and Papamarcou and Fine (1986) have studied this general form of lower probabilities.

Fortunately, many of the properties of ∞-monotone lower probabilities are shared by all functions which are monotone of order 2. This has been shown not only by decision theoreticians but also, independently, by other authors, who met with these functions in their own fields of research: pure mathematicians Choquet (1953), Revuz (1955), Dellacherie (1971) and Anger (1971, 1977); statisticians Huber (1973, 1976) and Huber and Strassen (1973); game theorist Shapley (1971); and specialists of matroïd theory Edmonds (1970) and Bixby et al. (1985).

Our aims, in this paper, are: (i) to determine which properties are specific to each category of monotone functions; (ii) to show that the use of Möbius inversion – a transformation applied by Dempster and Shafer – is not limited to belief functions, although inverses of other functions are not in that case non-negative; (iii) to characterize, in particular, members of each category of functions by properties of their Möbius inverses; (iv) to produce, using Möbius inversion and certain classic findings (such as Gale's theorem for the transshipment problem), both new findings and easier demonstrations of former ones. Potential applications of these results to decision making are discussed in §4.

Like Dempster and Shafer (and game or matroïd theorists), we shall only consider functions defined on 2^{Θ}, with Θ a finite set. The notations used are basically those of Shafer (1976), in particular: $|A|$ is the cardinal of set A; $A \backslash B = \{\theta \in \Theta : \ \theta \in A, \ \theta \notin B\}$; $\bar{A} = \Theta \backslash A$; $B \not\subset A$ means that $B \cap \bar{A} \neq \varnothing$; a summation such as '$\sum_{B \subset A} m(B)$' is short for '$\sum_{\{B \in 2^{\Theta} : B \subset A\}} m(B)$'; etc; by convention, $\sum_{\varnothing} \cdots = 0$.

We also use standard notations of probability theory: $f(X \geq x)$ for $f\{\theta \in \Theta : X(\theta) \geq x\}$, etc...

2 Elementary Properties of Monotone Capacities and their Möbius Inverses

Let Θ be a finite non-empty set and let $\mathcal{A} = 2^{\Theta}$ (set of all subsets of Θ), and $\mathcal{A}^* = \mathcal{A} \backslash \{\varnothing\}$. A mapping $f : \mathcal{A} \to \mathbb{R}$ (actually: $\mathcal{A} \to [0,1]$) is a (normalized) *capacity* whenever

$$f(\varnothing) = 0; f(\Theta) = 1; f(A_1) \leq f(A_2) \ \ \textit{for all} \ A_1, A_2 \in \mathcal{A} \text{ such that } A_1 \subset A_2. \tag{1}$$

The last property in (1) is monotonicity in the usual sense or 1-monotonicity; furthermore, given an integer $K \geq 2$, a mapping $f : \mathcal{A} \rightarrow \mathbb{R}$ is *K-monotone* (short for: monotone of order K) if and only if

$$f\left(\bigcup_{k=1}^{K} A_k\right) \geq \sum_{\substack{I \subset \{1,\ldots,K\} \\ I \neq \varnothing}} (-1)^{|I|+1} f\left(\bigcap_{k\in I} A_k\right), \text{ for all } A_k \in \mathcal{A},\ 1 \leq k \leq K. \tag{2}$$

By using the fact that the A_k's are not necessarily distinct, the first part of the following proposition can be easily proven; its second part is straightforward.

Proposition 1. (i) *If a mapping f is K-monotone for some $K \geq 2$, then f is also K'-monotone for $2 \leq K' \leq K$.*
(ii) *If, moreover, $f(\varnothing) = 0$ and $f(\{\theta\}) \geq 0$ for all $\theta \in \Theta$, it is also 1-monotone and $f \geq 0$.*

On the other hand, lower order monotonicities do not imply higher order monotonicities (see Example 3), and a mapping which is K-monotone for all $K \geq 2$ is said to be *∞-monotone* (short for: monotone of infinite order).

Note that probabilities are the particular instances of ∞-monotone capacities for which equality obtains in (2) for all K (Poincaré's equalities).

To any mapping $f : \mathcal{A} \rightarrow \mathbb{R}$ another mapping $m : \mathcal{A} \rightarrow \mathbb{R}$ can be associated by

$$m(A) = \sum_{B\subset A} (-1)^{|A\backslash B|} f(B) \text{ for all } A \in \mathcal{A}. \tag{3}$$

This correspondence proves to be one-to-one, since conversely,

$$f(A) = \sum_{B\subset A} m(B) \text{ for all } A \in \mathcal{A}. \tag{4}$$

The validity of (4) is proven by Shafer ([1976, Ch. 2, §7] (his proof is recalled in the Appendix), who calls the correspondence *Möbius inversion* (see Rota, 1964).

Capacities can then be characterized as follows:

Proposition 2. *f is a capacity if and only if its Möbius inverse satisfies*

$$m(\varnothing) = 0;\ \sum_{B\in\mathcal{A}} m(B) = 1;\ \sum_{\{\theta\}\subset B\subset A} m(B) \geq 0, \text{ for all } A \in \mathcal{A}, \text{all } \theta \in A. \tag{5}$$

In particular, it is necessary that $m(\{\theta\}) \geq 0$, for all $\theta \in \Theta$.

Proof. By (3) and (4), the two equalities are equivalent to $f(\varnothing) = 0$ and $f(\Theta) = 1$. For the inequalities, note that 1-monotonicity holds if and only if it holds for pairs $A_1 = A\backslash\{\theta\},\ A_2 = A$.

K-monotone mappings can be characterized as follows (this proof is a direct adaptation of the proof of Theorem 2.1 in Shafer (1976, p. 51)).

Proposition 3. *Let $f, m : \mathcal{A} \to \mathbb{R}$, and suppose that f and m are Möbius-inverse; then, f is K-monotone (K integer, $K \geq 2$) if and only if*

$$\sum_{\substack{B \subset \bigcup_{k=1}^{K} A_k \\ B \not\subset A_k, \text{all } k}} m(B) \geq 0 \text{ for } A_k \in \mathcal{A}, 1 \leq k \leq K. \tag{6}$$

Proof. To every $B \in \mathcal{A}$, let us associate

$$I(B) = \{k : 1 \leq k \leq K \text{ and } B \subset A_k\}.$$

$$\sum_{\substack{I \subset \{1,\ldots,K\} \\ I \neq \varnothing}} (-1)^{|I|+1} f\left(\bigcap_{k \in I} A_k\right) = \sum_{\substack{I \subset \{1,\ldots,K\} \\ I \neq \varnothing}} (-1)^{|I|+1} \sum_{\substack{B \subset \bigcap A_k \\ k \in I}} m(B)$$

$$= \sum_{I(B) \neq \varnothing} m(B) \sum_{\substack{I \subset I(B) \\ I \neq \varnothing}} (-1)^{|I|+1} = \sum_{I(B) \neq \varnothing} m(B) \left[1 - \sum_{I \subset I(B)} (-1)^{|l|}\right] = \sum_{I(B) \neq \varnothing} m(B),$$

by Lemma 2.1 in Shafer [1976, p. 47] (see Appendix). Further,

$$f\left(\bigcup_{k=1}^{K} A_k\right) = \sum_{\substack{B \subset \bigcup_{k=1}^{K} A_k \\ I(B) = \varnothing}} m(B) + \sum_{\substack{B \subset \bigcup_{k=1}^{K} A_k \\ I(B) \neq \varnothing}} m(B)$$

$$= \sum_{\substack{B \subset \bigcup_{k=1}^{K} A_k \\ B \not\subset A_k, \text{all } k}} m(B) + \sum_{I(B) \neq \varnothing} m(B).$$

It is now obvious that the proposition holds. □

From Proposition 3, it can be easily deduced that:

Corollary 1. *(i) If f is K-monotone and $2 \leq |A| \leq K$, then $m(A) \geq 0$.*
(ii) f satisfying $f(\varnothing) = 0$ is a non-negative ∞-monotone mapping if and only if m is itself non-negative.

Proof. For (i), use $A_k = A \backslash \{\theta_k\}$, where $A = \{\theta_1, \ldots, \theta_{K'}\}$, $K' \leq K$, in (6); for (ii) use (i) and Proposition 1. □

Note also that f is a probability if and only if $m(B) \geq 0$ when $|B| = 1$, $m(B) = 0$ otherwise, and $\Sigma_{B \in \Theta} m(B) = 1$.

A simpler characterization, which is similar to (5) for 1-monotonicity, can be given:

Proposition 4. *Let f,m: $\mathcal{A} \to \mathbb{R}$, and suppose that f and m are Möbius-inverse; then f is K-monotone (K integer, $K \geq 2$) if and only if*

$$\sum_{C \subset B \subset A} m(B) \geq 0, \text{ for all } A \in \mathcal{A} \text{ and } C \in \mathcal{A}, 2 \leq |C| \leq K. \tag{7}$$

Proof. (7) is obviously implied by (6), since one can always take $C = \{\theta_1, \ldots, \theta_1\}$ with $2 \leq l \leq K$, $A_1 = A \backslash \{\theta_1\}, A_2 = A \backslash \{\theta_2\}, \ldots, A_l = A \backslash \{\theta_l\}, A_{l+1} = \cdots = A_K = A_l$ so that $A = \bigcup_{k=1}^{K} A_k$, and, for $B \subset A$, $C \subset B$ is equivalent to $B \not\subset A_k$, $1 \leq k \leq K$.

Conversely, let us show that (7) implies (6). Let $A_1, A_k, \ldots, A_K \in \mathcal{A}$ and suppose that $B \subset A = \bigcup_{k=1}^{K} A_k$ and $B \not\subset A_k$ for $k = 1, \ldots, K$, which is equivalent to the existence of

$$\theta^{(1)} \in A \backslash A_1 = E_1,\ \theta^{(2)} \in A \backslash A_2 = E_{2,\ldots,} \theta^{(K)} \in A \backslash A_K = E_K$$

such that

$$\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(K)} \in B.$$

Let us choose enumeration orders

$$E_1 = \left\{\theta_1^{(1)}, \ldots, \theta_{L_1}^{(1)}\right\}, \ldots, E_K = \left\{\theta_1^{(k)}, \ldots, \theta_{L_K}^{(K)}\right\},$$

and endow $E = E_1 \times E_2 \times \cdots \times E_K$ with the lexicographical ordering ${}^L\leq$. If $(\theta_{l_1}^{(1)}, \ldots, \theta_{l_K}^{(K)})$ is the first element of E for ${}^L\leq$ satisfying

$$C = \bigcup_{k=1}^{K} \left\{\theta_{I_k}^{(k)}\right\} \subset B,$$

then B must be a subset of

$$A_{l_1,\ldots,l_K} = \left\{\theta_{l_1}^{(1)}, \ldots, \theta_{L_1}^{(1)}\right\} \cup \cdots \cup \left\{\theta_{I_K}^{(K)}, \ldots, \theta_{L_K}^{(K)}\right\} \cup A \Big\backslash \bigcup_{1}^{K} E_k;$$

further if

$$C' = \bigcup_{k=1}^{K} \left\{\theta_{l'_k}^{(k)}\right\} \subset B \subset A_{l'_1,\ldots,l'_k}$$

then, necessarily, $l'_1 = l_1, \ldots, l'_K = l_K$; therefore sets

$$\left\{B \in \mathcal{A} : \bigcup_{k=1}^{K} \left\{\theta_{l_k^{(k)}}\right\} \subset B \subset A_{l_1,\ldots,l_K}\right\},$$

$1 \leq l_k \leq L_k,\ 1 \leq k \leq K$, form a partition of

$$\left\{B \in \mathcal{A} : B \subset \bigcup_{k=1}^{K} A_k, B \not\subset A_k, k = 1, \ldots, K\right\};$$

thus the first member of (6) can be divided into a sum of non-negative terms.□

Note that if $K > 2$, condition (7) in Proposition 4 cannot be restricted to subsets C such that $|C| = K$, as shown by Example 1.

Example 1. $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$. Let the Möbius inverse, m, of capacity f satisfy: $m(\Theta) = \frac{1}{4}, m(\{\theta_1, \theta_2, \theta_3\}) = -\frac{1}{2}, m(\{\theta_1, \theta_2\}) = \frac{5}{4}$. Condition (7) is satisfied for every subset C such that $|C| = 4$, however f is not 4-monotone since $\Sigma_{\{\theta_1,\theta_2,\theta_3\} \subset B \subset \Theta} m(B) = -\frac{1}{4}$. □

For 2-monotone mappings, however, i.e. mappings f such that

$$f\left(A_1 \cup A_2\right) + f\left(A_1 \cap A_2\right) \geq f\left(A_1\right) + f\left(A_2\right) \text{ for all } A_1, A_2 \in \mathcal{A},$$

Propositions 3 and 4 have the following simple implication:

Corollary 2. *Let f, m: $\mathcal{A} \to \mathbb{R}$, and suppose that f and m are Möbius inverse. The following statements are equivalent:*

(i) $f \text{ is } 2-monotone;$

(ii) $$\sum_{\substack{B \subset A_1 \cup A_2 \\ B \not\subset A_1, B \not\subset A_2}} m\left(B\right) \geq 0 \; for \; all \; A_1, A_2 \in \mathcal{A}; \tag{8}$$

(iii) $$\sum_{\{\theta_1,\theta_2\} \subset B \subset A} m\left(B\right) \geq 0 \; for \; all \; A \in \mathcal{A} \; and \; all \; \theta_1, \theta_2 \in A, \theta_1 \neq \theta_2. \tag{9}$$

3 Characterization of the Probabilities Dominating Monotone Capacities

3.1 A Necessary Condition for Dominance

Any probability P (weakly) *dominating* a given capacity f, i.e., satisfying

$$P\left(A\right) \geq f\left(A\right) \quad \text{for all } A \in \mathcal{A}, \tag{10}$$

is simply related to f's Möbius inverse, as shown by the following result which generalizes a result of Dempster [1967, §2]:

Proposition 5. *Let f, m, P: $\mathcal{A} \to \mathbb{R}$, where f is a capacity; m, its Möbius inverse; and P, a probability dominating f. There exists then a weight function*

$$\lambda : \bigcup_{B \in \mathcal{A}^*} \{(B, \theta) : \theta \in B\} \to \mathbb{R} \; satisfying.$$

$$\lambda \geq 0 \; and \; \sum_{\theta \in B} \lambda\left(B, \theta\right) = 1 \; for \; all \; B \in \mathcal{A}^*, \tag{11}$$

such that P is identical to measure P_λ determined by

$$P_\lambda(\{\theta\}) = \sum_{B \supset \{\theta\}} \lambda(B,\theta)\, m(B) \quad \textit{for all } \theta \in \Theta, \tag{12}$$

and

$$P_\lambda(A) = \sum_{\theta \in A} P_\lambda(\{\theta\} \quad \textit{for all } A \in \mathcal{A}. \tag{13}$$

Proof. Let us show that the property to be demonstrated can simply be restated as a network-flow property, which results from Gale's (1960, p. 149) feasibility theorem for the transshipment problem.

Let us indeed consider the transshipment problem on a capacitated network consisting of: a set of sources $\mathcal{E} = \{e_B : B \in \mathcal{A}, m(B) > 0\}$ with supply $m(B)$ at $B \in \mathcal{E}$; a set of sinks $\Theta \cup \mathcal{S}$, where $\mathcal{S} = \{s_B : B \in \mathcal{A}, m(B) < 0\}$, with demand $P(\{\theta\})$ at $\theta \in \Theta$ and demand $-m(B)$ at $s_B \in \mathcal{S}$; arcs, with infinite capacities, joining a source $e_B \in \mathcal{E}$ to a sink $\theta \in \Theta$, or a sink $\theta \in \Theta$ to another sink $s_B \in \mathcal{S}$, if and only if $\theta \in B$ (see Figure 1).

Note that there is no excess supply, since

$$\sum_{\theta \in \Theta} P(\{\theta\}) = 1 = \sum_{B \in \mathcal{A}} m(B) = \sum_{e_B \in \mathcal{E}} m(B) - \sum_{s_B \in \mathcal{S}} (-m(B));$$

thus, a feasible flow, φ, has to saturate the supply and demand constraints:

$$\sum_{\theta \in B} \varphi(e_B, \theta) = m(B) \text{ for all } e_B \in \mathcal{E}, \tag{14}$$

$$\sum_{B \supset \{\theta\}, e_B \in \mathcal{E}} \varphi(e_B, \theta) = P(\{\theta\}) + \sum_{B \supset \{\theta\}, s_B \in \mathcal{S}} \varphi(\theta, s_B,) \text{ for all } \theta \in \Theta, \tag{15}$$

$$\sum_{\theta \in B} \varphi(\theta, s_B) = -m(B) \text{ forall } S_B \in \mathcal{S}. \tag{16}$$

To any feasible flow φ one can associate a function λ partially defined by:

$$\lambda(B,\theta) = \left\{ \begin{matrix} \frac{\varphi(e_B,\theta)}{m(B)} \text{ for } m(B) > 0 \\ \frac{\varphi(\theta,s_B)}{-m(B)} \text{ for } m(B) < 0 \end{matrix} \right\} \text{and } \theta \in B;$$

choosing, for $B \in \mathcal{A}^*$ and $m(B) = 0$, arbitrary $\lambda(B,\theta)$'s satisfying (11), λ shall, in fact, satisfy (11) for all $B \in \mathcal{A}^*$; moreover, it follows directly from (15) that $P(\{\theta\}) = P_\lambda(\{\theta\})$, given by (12).

Thus, all we need prove is the existence of a feasible flow in the network. According to Gale's theorem, this amount to checking that, for each partition $\{\mathcal{N}, \bar{\mathcal{N}}\}$ of the set of nodes, $k(\bar{\mathcal{N}}, \mathcal{N}) \geq d(\mathcal{N})$, where $k(\bar{\mathcal{N}}, \mathcal{N})$ is the sum of the capacities of the arcs joining a node in $\bar{\mathcal{N}}$ to a node in $\mathcal{N}$, and $d(\mathcal{N})$ is the net demand in $\mathcal{N}$ (i.e., the difference between the sum of the demands and the sum of the supplies at the various nodes in $\mathcal{N}$).

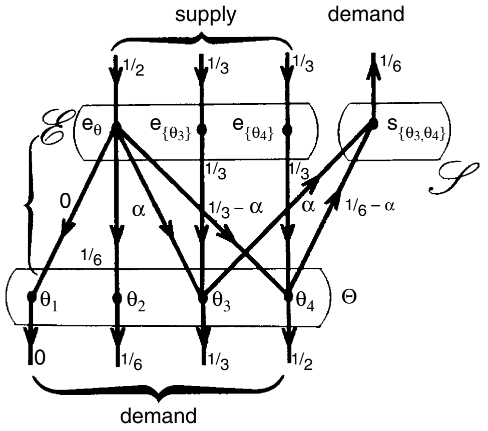


Fig. 1. An illustration of Proposition 5 based on Example 2

This inequality is obviously satisfied when $k(\bar{\mathcal{N}},\mathcal{N}) = +\infty$, i.e., when there exists $e_B \in \mathcal{E} \cap \bar{\mathcal{N}}$ and $\theta \in B \cap \mathcal{N}$, or $s_B \in \mathcal{S} \cap \mathcal{N}$ and $\theta \in B \cap \bar{\mathcal{N}}$.

In the alternative case, $k(\bar{\mathcal{N}},\mathcal{N}) = 0$, it can be noted: first, that $e_B \in \mathcal{E}$ and $B \cap \mathcal{N} \neq \varnothing$ imply that $e_B \in \mathcal{E} \cap \mathcal{N}$; secondly, that $s_B \in \mathcal{S} \cap \mathcal{N}$ implies that $B \subset \mathcal{N}$, hence that $B \cap \mathcal{N} \neq \varnothing$ (since necessarily $B \neq \varnothing$ when $s_B \in \mathcal{S}$); therefore,

$$\begin{aligned} d(\mathcal{N}) &= \sum_{s_B \in \mathcal{S} \cap \mathcal{N}} (-m(B)) + \sum_{\theta \in \Theta \cap \mathcal{N}} P(\{\theta\}) - \sum_{e_B \in \mathcal{S} \cap \mathcal{N}} m(B) \\ &\leq \sum_{\substack{s_B \in \mathcal{S} \\ B \cap \mathcal{N} \neq \varnothing}} (-m(B)) + P(\Theta \cap \mathcal{N}) - \sum_{\substack{e_B \in \mathcal{E} \\ B \cap \mathcal{N} \neq \varnothing}} m(B) \\ &= P(\Theta \cap \mathcal{N}) - \sum_{B \cap \mathcal{N} \neq \varnothing} m(B) = 1 - P\left(\Theta \cap \bar{\mathcal{N}}\right) - \left[1 - f\left(\Theta \cap \bar{\mathcal{N}}\right)\right] \leq 0. \end{aligned}$$

□

Example 2.

$$\Theta = \{\theta_1\theta_2,\theta_3,\theta_4\}\,; f\{\varnothing\} = 0; f(\{\theta_1\}) = f(\{\theta_2\}) = 0; f(\{\theta_3\}) = f(\{\theta_4\}) = \frac{1}{3};$$

$$f(\{\theta_1,\theta_2\}) = 0, f(\{\theta_3,\theta_4\}) = \frac{1}{2}, f(\{\theta_i,\theta_j\}) = \frac{1}{3} \text{ otherwise};$$

$$f(\{\theta_1,\theta_2,\theta_3\}) = f(\{\theta_1,\theta_2,\theta_4\}) = \frac{1}{3}; f(\{\theta_1,\theta_3,\theta_4\}) = f(\{\theta_2,\theta_3,\theta_4\}) = \frac{1}{2}; f(\Theta) = 1;$$

thus f is a capacity, but is not 2-monotone

$$(f(\{\theta_3,\theta_4\}) + f(\varnothing) < f(\{\theta_3\}) + f(\{\theta_4\}))\,.$$

Its Möbius inverse, m, satisfies

$$m(\{\theta_3\}) = m(\{\theta_4\}) = \frac{1}{3}, m(\{\theta_3, \theta_4\}) = -\frac{1}{6}, m(\Theta) = \frac{1}{2}, m(B) = 0 \text{ otherwise}$$

hence $\mathcal{E} = \{e_{\{\theta_3\}}, e_{\{\theta_4\}}, e_{\Theta}\}$ and $\mathcal{S} = \{s_{\{\theta_3, \theta_4\}}$.

Probability P, defined by $P(\{\theta_1\}) = 0,\ P(\{\theta_2\}) = \frac{1}{6},\ P(\{\theta_3\}) = \frac{1}{3}$ and $P(\{\theta_4\}) = \frac{1}{2}$, dominates f, and $P = P_\lambda$ for any weight function λ satisfying for some $\alpha \in [0, \frac{1}{6}] : \lambda(\{\theta_3\},\ \theta_3) = \lambda(\{\theta_4\},\ \theta_4) = 1$;

$$\lambda(\{\theta_3, \theta_4\}, \theta_3) = 6\alpha, \lambda(\{\theta_3, \theta_4\}, \theta_4) = 1 - 6\alpha; \lambda(\Theta, \theta_1) = 0,$$
$$\lambda(\Theta, \theta_2) = \frac{1}{3}, \lambda(\Theta, \theta_3) = 2\alpha \text{ and } \lambda(\Theta, \theta_4) = \frac{2}{3} - 2\alpha.$$

Note that, on the other hand, not all P_λ's defined by (11), (12) and (13) dominate f: for example, if $\lambda(\{\theta_3,\ \theta_4\}, \theta_3) = 1$ and $\lambda(\Theta, \theta_4) = 1$, then $P_\lambda(\{\theta_3\}) = \frac{1}{6} < \frac{1}{3} = f(\{\theta_3\})$. □

The set of probabilities dominating a capacity may even be empty:

Example 3.

$$\Theta = \{\theta_1, \theta_2, \theta_3\}; f(\varnothing) = 0; f(\{\theta_i\}) = 0, \text{ all } i; f(\{\theta_i, \theta_j\}) = \frac{3}{4}, \text{all } i, j, f(\Theta) = 1.$$

Dominance requires in particular $P(\{\theta_1\}) \leq \frac{1}{4},\ P(\{\theta_2\}) \leq \frac{1}{4}$ and $P(\{\theta_1, \theta_2\}) \geq \frac{3}{4}$; yet some P_λ's are probabilities, since f's Möbius inverse, m, satisfies $m(\{\theta_i, \theta_j\}) = \frac{3}{4}$, all i, j, and $m(\Theta) = -\frac{5}{4},\ m(B) = 0$ otherwise, hence P_λ is a probability for $\lambda(\{\theta_i,\ \theta_j\},\ \theta) = \frac{1}{2},\ \theta = \theta_i$ or θ_j, and $\lambda(\Theta,\ \theta) = \frac{1}{3}$, all θ. Note that f, which is not 2-monotone, is super-additive, i.e., $f(A \cup B) \geq f(A) + f(B)$ whenever $A \cap B = \varnothing$. □

Let: $\mathcal{P}_{\geq}$ be the set of all probabilities dominating capacity f, Λ be the set of all weight functions satisfying (11), and $\mathcal{M}_\Lambda$ be the set of measures P_λ associated to some λ by (12) and (13).

Proposition 5 merely asserts that $\mathcal{P}_{\geq} \subset \mathcal{M}_\Lambda$. We shall determine under what conditions the converse, $\mathcal{P}_{\geq} \supset \mathcal{M}_\Lambda$, is also true. Obviously, a first requirement is that measures P_λ be probabilities; let us examine this point.

Proposition 6. *Let f be a capacity; m its Möbius inverse; then, all measures in $\mathcal{M}_\Lambda$ (i.e., defined by (11), (12) and (13)), are probabilities if and only if*

$$m(\{\theta\}) + \sum_{\substack{B \supset \{\theta\} \\ B \neq \{\theta\}}} \text{Min}\{m(B), 0\} \geq 0, \quad \text{for all } \theta \in \Theta. \tag{17}$$

Proof. Since f is a capacity, m satisfies (5); thus, for any $P_\lambda \in \mathcal{M}_\Lambda$,

$$\sum_{\theta \in \Theta} P_\lambda(\{\theta\}) = \sum_{\theta \in \Theta} \sum_{B \supset \{\Theta\}} \lambda(B, \theta) m(B)$$
$$= \sum_{B \in \mathcal{A}^*} m(B) \sum_{\theta \in B} \lambda(B, \theta) = \sum_{B \in \mathcal{A}^*} m(B) = 1.$$

We need at this point only demonstrate that $P_\lambda \in \mathcal{M}_\Lambda$ implies $P_\lambda(\{\theta\}) \geq 0$ for all $\theta \in \Theta$ if and only if (17) is satisfied.

For any $P_\lambda \in \mathcal{M}_\Lambda$ and any $\theta \in \Theta$,

$$P_\lambda(\{\theta\}) = m(\theta) + \sum_{\substack{B \supset \{\theta\} \\ m(B) < 0}} \lambda(B,\theta)\, m(B) + \sum_{\substack{B \supset \{\theta\}, B \neq \{\theta\} \\ m(B) > 0}} \lambda(B,\theta)\, m(B);$$

hence, since $0 \leq \lambda(B,\theta) \leq 1$,

$$P_\lambda(\{\theta\}) \geq m(\{\theta\}) + \sum_{\substack{B \supset \{\theta\} \\ m(B) < 0}} m(B) = m(\{\theta\}) + \sum_{\substack{B \supset \{\theta\} \\ B \neq \{\theta\}}} \mathrm{Min}\{m(B), 0\};$$

It is thus straightforward that (17) is a sufficient condition for the nonnegativity of P_λ. The necessity of (17) results from the fact that the two members of the last inequality become equal if λ is chosen such that:

$$\begin{aligned} \lambda(B,\theta) &= 1 \quad \text{for } B \supset \{\theta\}, m(B) < 0, \\ \lambda(B,\theta) &= 0 \quad \text{for } B \supset \{\theta\}, B \neq \{\theta\}, m(B) > 0. \end{aligned}$$

□

It can be easily checked that (17) holds in Example 2 but does not hold in Example 3, nor in the following example, where f is 2-monotone:

Example 4.

$$\Theta\{\theta_1, \theta_2, \theta_3, \theta_4\}; f(\varnothing) = f(\{\theta_i\}) = 0, \text{ all } i : f(\{\theta_i; \theta_j\}) = \frac{1}{6}, \text{all } i, j;$$
$$f(\{\theta_i, \theta_j, \theta_k\}) = \frac{1}{3}, \text{ all } i, j, k; f(\Theta) = 1.$$

Its Möbius inverse, m, satisfies $m(\{\varnothing)\} = m(\{\theta_i\}) = 0$, all i; $m(\{\theta_i, \theta_j\}) = \frac{1}{6}$, all i, j; $m(\{\theta_i,\ \theta_j,\ \theta_k\}) = -\frac{1}{6}$, all *i,j,k*; $m(\Theta) = \frac{2}{3}$; it is easily checked, directly or by using (5) and (9), that f is a 2-monotone capacity; it is however not 3-monotone since triplets have negative masses. (17) does not hold, since

$$m(\{\theta_1\}) + \sum_{\substack{B \supset \{\theta_1\} \\ B \neq \{\theta_1\}}} \mathrm{Min}\{m(B), 0\} = -\frac{1}{2}.$$

Indeed, for λ_1 satisfying

$$\lambda_1(\{\theta_1 . \theta_j\}, \theta_1) = 0, \text{ all } j, \lambda_1(\{\theta_1, \theta_j, \theta_k,\}, \theta_1) = 1, \text{ all } j, k \text{ and } \lambda_1(\Theta, \theta_1) = 0,$$
$$P_{\lambda_1}(\{\ \theta_1\ \})\ = -\frac{1}{2}.$$

□

3.2 Characterization of the Probabilities Dominating an ∞-monotone Capacity

For capacities which are ∞-monotone, (17) is obviously satisfied since their Möbius inverses are non-negative; thus all measures in $\mathcal{M}_\Lambda$ are probabilities, which fact, however, can also be deduced from the following proposition (Dempster, 1967):

Proposition 7. *For ∞-monotone capacities, $\mathcal{P}_{\geq} \supset \mathcal{M}_\Lambda$; in other terms, if f is an ∞-monotone capacity, and m its Möbius inverse, every measure satisfying (11), (12) and (13) dominates f, and is a probability.*

Proof. Let $P_\lambda \in \mathcal{M}_\Lambda$ and $A \in \mathcal{A}$.

$$\begin{aligned} P_\lambda(A) &= \sum_{\theta\in A} P_\lambda(\{\theta\}) = \sum_{\theta\in A}\left(\sum_{B\ni\theta}\lambda(B,\theta)\,m(B)\right) \\ &= \sum_{B\in\mathcal{A}^*} m(B) \sum_{\theta\in A\cap B}\lambda(B,\theta) \\ &= \sum_{\substack{B\supset A \\ B\neq\varnothing}} m(B) + \sum_{B\cap\bar{A}\neq\varnothing} m(B) \sum_{\theta\in B\cap A}\lambda(B,\theta); \qquad (18) \end{aligned}$$

Since $m(\varnothing) = 0$, the first term to the right is equal to $f(A)$; since $m \geq 0$ for ∞-monotone mappings, the second term is non-negative; thus,

$$P_\lambda(A) \geq f(A) \geq 0.$$

Further since

$$P_\lambda(\Theta) = \sum_{B\subset\Theta} m(B) = 1,$$

P_λ is a probability.

In fact, the property in Proposition 7 characterizes capacities which are ∞-monotone, since:

Proposition 8. *Let f be a capacity, m its Möbius inverse; if $\mathcal{P}_{\geq} \supset \mathcal{M}_\Lambda$, i.e., if every measure satisfying (11), (12) and (13) dominates f, then f is ∞-monotone.*

Proof. Suppose that f is not ∞-monotone. By Corollary 1 (ii), there then exists $B_0 \in \mathcal{A}$ such that $m(B_0) < 0$; by Proposition 2, $|B_0| \geq 2$; let thus $\theta_1, \theta_2 \in B_0$. Take any $A \in \mathcal{A}$ such that $\theta_1 \in A$ and $\theta_2 \in \bar{A}$, and let $\lambda \in \Lambda$ satisfy $\lambda(B_0, \theta_1) = 1$ and, for every $B \neq B_0$ such that $B \cap \bar{A} \neq \varnothing$,

$$\sum_{\theta\in B\cap\bar{A}}\lambda(B,\theta) = 1 \quad \left(\text{hence, } \sum_{\theta\in B\cap A}\lambda(B,\theta) = 0\right).$$

By (18), $P_\lambda(A) = f(A)+m(B_0) < f(A)$, and, therefore, P_λ does not dominate f. □

Propositions 5, 7 and 8 together imply:

Corollary 3. *Let f be a capacity. Then $\mathcal{P}_\geq = \mathcal{M}_\Lambda$ if and only if f is ∞-monotone.*

Therefore, if f is a 2-monotone capacity, but is not ∞-monotone, $\mathcal{P}_\geq$ is a proper subset of $\mathcal{M}_\Lambda$, even when all members of $\mathcal{M}_\Lambda$ are probabilities as in the following example:

Example 5. Given Θ, f, and m of Example 4, and probability Q characterized by $Q(\{\theta_i\}) = \frac{1}{4}$, all i, consider the mapping $f' = \frac{2}{3}Q + \frac{1}{3}f$, i.e. $A \mapsto f'(A) = \frac{2}{3}Q(A) + \frac{1}{3}f(A)$. Its Möbius inverse, m', therefore satisfies $m' = \frac{2}{3}\mu + \frac{1}{3}m$, where μ is the inverse of Q : $m'(\{\theta_i\}) = \frac{2}{3} \cdot \frac{1}{4} = \frac{1}{6}$, all i, and $m'(B) = \frac{1}{3}m(B)$ for $|B| > 1$; hence, (17) holds for f', and $P_\lambda \geq 0$, for every $\lambda \in \Lambda$; however for λ_1 of Example 3, $P_{\lambda_1}(\{\theta_1\}) = 0 < f'(\{\theta_1\}) = \frac{1}{6}$. □

3.3 Characterization of the Probabilities Dominating a 2-monotone Capacity

Let us then try to characterize, for a 2-monotone capacity, subset $\mathcal{P}_\geq$ of $\mathcal{M}_\Lambda$, i.e., those members of $\mathcal{M}_\Lambda$ which are probabilities dominating that capacity.

To do so, let us first associate with any capacity f a certain family of probabilities, defined as follows: let Σ be the set of the permutations of all the elements in Θ; given a generic element of Σ,

$$S = (\theta_{i_1}, \ldots, \theta_{i_l}, \ldots, \theta_{i_L}), \ L = |\Theta|,$$

let us denote by S_l, $0 \leq l \leq L$ the subsets of Θ defined by $S_l = \{\theta_{i_l}, \ldots, \theta_{i_l}\}$ for $l \geq 1$, and $S_0 = \varnothing$. A measure P_S can be defined by

$$P_S(\{\theta_{i_l}\}) = f(S_l) - f(S_{l-1}), \quad \text{for } 1 \leq l \leq L. \tag{19}$$

and its (implied) additivity property. It is straightforward that, since f is a capacity, P_S is a probability and satisfies

$$P_S(S_l) = f(S_l), \quad \text{for } 1 \leq l \leq L. \tag{20}$$

We shall denote by $\mathcal{P}_\Sigma$ the set of probabilities equal to P_s for some $S \in \Sigma$. It is obvious that $\mathcal{P}_\Sigma$ cannot be empty, but is a singleton when f is a probability. The following property of $\mathcal{P}_\Sigma$ will later be useful:

Lemma 1. *Let f be a capacity. Given a decreasing sequence, $(A_n, \ 1 \leq n \leq N)$, of elements in $\mathcal{A}$, there exists a permutation $S \in \Sigma$, such that probability P_S satisfies*

$$P_S(A_n) = f(A_n), \text{ for } 1 \leq n \leq N. \tag{21}$$

Proof. There exists at least one enumeration S of Θ such that $A_n = S_{|A_n|}$, $1 \leq n \leq N$ (enumerate successively the elements of A_N, $A_{N-1}\backslash A_N, \ldots, A_1\backslash A_2$, $\Theta\backslash A_1$). (21) then follows from (20).

Proposition 9. *For 2-monotone capacities, $\mathcal{P}_\Sigma \subset \mathcal{P}_\geq \subset \mathcal{M}_\Lambda$. In other terms, if f is a 2-monotone capacity and m its Möbius inverse, then, for every $S \in \Sigma$, probability P_S, defined by (19), dominates f, and thus belongs to $\mathcal{M}_\Lambda$ (Proposition 5); more precisely $P_S = P_\lambda$, where P_λ satisfies (12) and (13), for λ (satisfying (11)) defined by: for every $B \in \mathcal{A}^*$, $\lambda(B, \theta_S(B)) = 1$, where $\theta_S(B)$ is the last element of B in permutation S.*

Proof. Let λ be defined as in the preceding statement. For every l, $1 \leq l \leq L$,

$$P_\lambda\left(\{\theta_{i_l}\}\right) = \sum_{\left\{B \supset \theta_{i_l}\right\}} \lambda\left(B, \theta_{i_l}\right) m\left(B\right);$$

hence,

$$P_\lambda\left(\{\theta_{i_l}\}\right) = \sum_{\theta_S(B)=\theta_{i_l}} m\left(B\right) = \sum_{\substack{B \subset S_l \\ B \not\subset S_{l-1}}} m\left(B\right) \quad = f\left(S_l\right) - f\left(S_{l-1}\right) = P_S\left(\{\theta_{i_l}\}\right);$$

thus $P_\lambda = P_S$.

Moreover, for any $A \in \mathcal{A}$,

$$\begin{aligned} P_S\left(A\right) = P_\lambda\left(A\right) &= \sum_{\theta_S(B)\in A} m\left(B\right) = \sum_{B \subset A} m\left(B\right) + \sum_{\substack{B \not\subset A \\ \theta_S\left(B\right) \in A}} m\left(B\right) \\ &= f\left(A\right) + \sum_{\substack{1 \leq l \leq L \\ \theta_l \in A}} \sum_{\substack{B \not\subset A \\ \theta_S\left(B\right) = \theta_l}} m\left(B\right) \\ &= f\left(A\right) + \sum_{\substack{1 \leq l \leq L \\ \theta_l \in A}} \sum_{\substack{B \subset \left(A \cap S_l\right) \cup S_{l-1} \\ B \not\subset S_{l-1}, B \not\subset A \cap S_l}} m\left(B\right) \geq f\left(A\right) \end{aligned}$$

since, according to (8), f being 2-monotone, the other terms are non-negative. Thus $P_S \in \mathcal{P}_\geq$. □

Remark 1. (i) Proposition 9 implies that $\mathcal{P}_\geq$ is not empty when f is 2-monotone.

(ii) The proof of Proposition 9 shows that, for all capacities, $\mathcal{P}_\Sigma \subset \mathcal{M}_\Lambda$.

The relations between $\mathcal{P}_\Sigma$ and $\mathcal{P}_\geq$ for 2-monotone capacities can be specified further; this however will require the preliminary proof of the following important result, due to Dempster (1967), and also derived, in a more general context, by Huber and Strassen (1973):

Proposition 10. *Let f be a 2-monotone capacity. For any mapping $X : \Theta \to \mathbb{R}$, with image*

$$X\left(\Theta\right) = \left\{x_1, \ldots, x_n, \ldots x_N\right\}, x_n \geq x_{n-1} \quad \text{for } 2 \leq n \leq N.$$

the infimum over $\mathcal{P}_{\geq}$ of the mathematical expectation

$$E_P\left(X\right) = \sum_{\theta \in \Theta} X\left(\theta\right) P\left(\{\theta\}\right)$$

is attained, for some $P_S \in \mathcal{P}_{\Sigma}$, and its value is

$$\operatorname*{Inf}_{P \in \mathcal{P}_{\geq}} E_P\left(X\right) = \sum_{n=2}^{N} \left(x_n - x_{n-1}\right) f\left(X \geq x_n\right) + x_1. \tag{22}$$

Proof. For every probability P,

$$E_P\left(X\right) = \sum_{\theta \in \Theta} X\left(\theta\right) P\left(\{\theta\}\right)$$

$$= \sum_{n=2}^{N} \left(x_n - x_{n-1}\right) P\left(X \geq x_n\right) + x_1.$$

The right side of equality (22) is obviously a lower bound of $E_P(X)$ for $P \in \mathcal{P}_{\geq}$, i.e., for P dominating f.

Moreover, since sets $A_n = \{\theta : X(\theta) \geq x_n\},\ 1 \leq n \leq N$, form a decreasing sequence, it results from Lemma 1 that there exists a probability $P_S \in \mathcal{P}_{\Sigma}$ and hence, by Proposition 9, $P_S \in \mathcal{P}_{\geq}$, such that

$$P_S\left(X \geq x_n\right) = f\left(X \geq x_n\right), \quad \text{for all } n, 1 \leq n \leq N.$$

□

From expression (22), another relation can be derived which involves, instead of f, its Möbius inverse m. Its validity for ∞-monotone capacities has already been proven by Shafer (1981).

Corollary 4. *Let f be a 2-monotone capacity and m its Möbius inverse. The infimum over $\mathcal{P}_{\geq}$ of the mathematical expectation of any mapping $X : \Theta \to \mathbb{R}$ is attained and its value is*

$$\operatorname*{Inf}_{P \in \mathcal{P}_{\geq}} E_P\left(X\right) = \sum_{B \in \mathcal{A}} x_{n(B)} m\left(B\right), \tag{23}$$

where $x_{n(B)} = \mathrm{Min}_{\theta \in B} X(\theta)$.

Proof. By (22),

$$\begin{aligned}\operatorname*{Inf}_{P\in\mathcal{P}_{\geq}} E_P\left(X\right) &= \sum_{n=2}^{N}\left(x_n - x_{n-1}\right) \sum_{B\subset\{\theta:X(\theta)\geq x_n\}} m\left(B\right) + x_1 \\ &= \sum_{B\in\mathcal{A}} m\left(B\right) \sum_{n=2}^{n(B)} \left(x_n - x_{n-1}\right) + x_1 = \sum_{B\in\mathcal{A}} m\left(B\right)\cdot x_{n(B)} - x_1 + x_1 .\end{aligned}$$

□

Remark 2. (i) In the particular case of ∞-monotone capacities, where $m \geq 0$ and $\mathcal{P}_{\geq} = \mathcal{M}_{\Lambda}$, (23) follows directly from

$$\begin{aligned}\operatorname*{Inf}_{P\in\mathcal{P}_{\geq}} E_P\left(X\right) &= \operatorname*{Inf}_{\lambda\in\Lambda}\left[\sum_{\theta\in\Theta} X\left(\theta\right) \sum_{B\supset\{\theta\}} \lambda\left(B,\theta\right) m\left(B\right)\right] \\ &= \sum_{B\in\mathcal{A}^*} m\left(B\right) \operatorname{Inf}\left\{\sum_{\theta\in B} \lambda\left(B,\theta\right)\cdot X\left(\theta\right) : \sum_{\theta\in B} \lambda\left(B,\theta\right) = 1; \lambda\left(B,\theta\right) \geq 0, \text{all } \theta\right\} .\end{aligned}$$

(ii) It results from the proofs of Propositions 10 and Corollary 4, that, for any capacity f, the right members of (22) and (23) are equal, and that their common value, $E_f^0(X)$, satisfies

$$\operatorname*{Inf}_{P\in\mathcal{P}_{\sum}} E_P\left(X\right) \leq E_f^0\left(X\right) \leq \operatorname*{Inf}_{P\in\mathcal{P}_{\geq}} E_P\left(X\right) .$$

Propositions 9 and 10 state properties which in fact characterize 2-monotone capacities, as shown by the proposition below.

Proposition 11. *Let f be a mapping: $\mathcal{A} \to \mathbb{R}$. The following statements are equivalent:*

(i) f is a 2-monotone capacity.
(ii) For every pair $A_1, A_2 \in \mathcal{A}$ such that $A_1 \supset A_2$, there exists a probability P, dominating f, and satisfying $P(A_1) = f(A_1)$ and $P(A_2) = f(A_2)$.
(iii) There exists a triplet $x_1, x_2, x_3 \in \mathbb{R}$, $x_1 < x_2 < x_3$, such that for every mapping $X : \Theta \to \{x_1, x_2, x_3\}$ there exists a probability P, dominating f, such that

$$E_P\left(X\right) = \left(x_3 - x_2\right) f\left(X \geq x_3\right) + \left(x_2 - x_1\right) f\left(X \geq x_2\right) + x_1 .$$

Proof. By Proposition 9 and Lemma 1, (i) implies (ii); by Proposition 10, it also implies (iii). Conversely, let us first note that (iii) implies (ii), since, to $A_1 \supset A_2$, one can associate X satisfying $X(A_2) = x_3, X(A_1 \backslash A_2) = x_2$ and $X(\bar{A}_1) = x_1$. Thus, the desired equalities follow from $P \geq f$ and

$$\begin{aligned}\left(x_3 - x_2\right) P\left(A_2\right) + \left(x_2 - x_1\right) P\left(A_1\right) + x_1 &= E_P\left(X\right) \\ &= \left(x_3 - x_2\right) f\left(A_2\right) \\ &\quad + \left(x_2 - x_1\right) f\left(A_1\right) + x_1 .\end{aligned}$$

Finally, let us show that (ii) implies (i): f is a capacity, since, when $A_1 \supset A_2$, $f(A_1) \geq f(A_2)$ results from $P(A_1) \geq P(A_2)$, and, moreover, pair $\varnothing$, Θ can be used to prove that $f(\varnothing) = 0$ and $f(\Theta) = 1$. And last, given $A_1, A_2 \in \mathcal{A}$, there exists $P \geq f$ such that

$$P\left(A_1 \cup A_2\right) = f\left(A_1 \cup A_2\right) \text{ and } P\left(A_1 \cap A_2\right) = f\left(A_1 \cap A_2\right);$$

hence,

$$\begin{aligned} f\left(A \cup A_2\right) + f\left(A_1 \cap A_2\right) &= P\left(A_1 \cup A_2\right) + P\left(A_1 \cap A_2\right) \\ &= P\left(A_1\right) + P\left(A_2\right) \geq f\left(A_1\right) + f\left(A_2\right) \end{aligned}$$

and therefore f is 2-monotone. □

Proposition 12. *A capacity f is 2-monotone if and only if $\mathcal{P}_\Sigma \subset \mathcal{P}_\geq$.*

Proof. The 'only if' statement is contained in Proposition 9. Conversely, given $A_1, A_2 \in \mathcal{A}$, $A_1 \supset A_2$, Lemma 1 asserts the existence of $S \in \Sigma$ such that $P_S(A_1) = f(A_1)$, and $P_S(A_2) = f(A_2)$; thus if $\mathcal{P}_\Sigma \subset \mathcal{P}_\geq, P_S$ has all the properties required by (ii) in Proposition 11; hence f is 2-monotone. □

The relation between $\mathcal{P}_\Sigma$ and $\mathcal{P}_\geq$ for 2-monotone capacities can be described more precisely, provided every probability P is identified with vector

$$\left(P\left(\{\theta_1\}\right), \ldots, P\left(\{\theta_l\}\right), \ldots, P\left(\{\theta_L\}\right)\right),$$

i.e. an element of the simplex of $\mathbb{R}^L$. With this identification, $\mathcal{P}_\geq$ as a subset of the simplex characterized by linear inequalities (10), becomes a bounded convex polyhedron; in particular, $\mathcal{P}_\geq$ is the convex closure of its profile, i.e. of the finite set of its extreme points (Krein-Milman theorem, see Berge, 1965, Chap. 8). We shall prove that:

Proposition 13. *If capacity f is 2-monotone, then $\mathcal{P}_\Sigma$ is the profile of $\mathcal{P}_\geq$.*

Proof. (i) Let us first show that, for every $S \in \Sigma$, P_S is an extreme point of $\mathcal{P}_\geq$. Suppose, on the contrary, that $P_S = \alpha P' + (1-\alpha)P''$ for some $\alpha \in (0,1)$, and $P', P'' \in \mathcal{P}_\geq, P' \neq P''$. For every S_l, $1 \leq l \leq L$, $P(S_l) = f(S_l), P'(S_l) \geq f(S_l)$ and $P''(S_l) \geq f(S_l)$, hence $P'(S_l) = P''(S_l) = f(S_l) = P(S_l)$, and therefore, for

$$1 \leq l \leq L,\ P'\left(\{\theta_{i_l}\}\right) = P''\left(\{\theta_{i_l}\}\right) = f\left(S_l\right) - f\left(S_{l-1}\right) = P\left(\{\theta_{i_l}\}\right);$$

thus $P' = P'' = P$, a contradiction.

(ii) Conversely, let us show that $\mathcal{P}_\Sigma$ contains every extreme point of $\mathcal{P}_\geq$. Suppose, on the contrary, that P is an extreme point of $\mathcal{P}_\geq$ and that $P \notin \mathcal{P}_\Sigma$. Let C be the convex closure of the finite set $\mathcal{P}_\Sigma$: C is closed and $C \subset \mathcal{P}_\geq$; since P is an extreme point of $\mathcal{P}_\geq$, and $P \notin \mathcal{P}_\Sigma$, necessarily $P \notin C$; hence, since C is a closed convex set, P can be strictly separated from C (Berge, 1965, second separation theorem, p. 171; Karlin, 1959, Vol. 1, Lemma 13.1.1., p. 397), i.e. there exists $a = (a_1, \ldots, a_l, \ldots, a_L) \neq 0$ such that

$$\sum_{l=1}^{L} a_1 P(\{\theta_l\}) < \sum_{l=1}^{L} a_l P'(\{\theta_l\}), \text{for all } P' \in C.$$

On the other hand, Proposition 10 implies that to X defined by $X(\theta_l) = a_l,\ 1 \leq l \leq L$, one can associate some $P_S \in \mathcal{P}_\Sigma$ such that

$$\sum_{l=1}^{L} a_l P_S(\{\theta_l\}) = E_{P_S}(X) \leq E_{P''}(X) = \sum_{l=1}^{L} a_l P''(\{\theta_l\}) \text{ for all } P'' \in \mathcal{P}_\geq,$$

contradicting the preceding inequality, since $P \in \mathcal{P}_\geq$ and $P_S \in C$. □

Proposition 12 is due to Ishiishi (1981), who completed the 'only if' statement of Shapley (1971). Proposition 13 was originally proved by Dempster (1967) for ∞-monotone capacities and extended by Shapley (1971) to 2-monotone capacities (in the language of game theory, a 2-monotone capacity is a convex game and $\mathcal{P}_\geq$ is the core of the game). A similar result is proved by Edmonds (1970) and Bixby et al. (1985) for matroïd polyhedra.

Remark 3. Möbius inversion, which has been our main tool in this paper, does not exist in the case where Θ is infinite. However, as noted by Shafer (1979), Choquet's (1953) theory of capacities provides the suitable tools for studying the infinite case, and, in actual fact, already contains generalizations of some of the preceding results.

3.4 Concave Extensions of 2-monotone Capacities

2-monotone capacities are also called convex set-functions. Shapley (1971) justifies this denomination by showing that some of their properties are similar to properties of convex functions. However, the natural extension of a 2-monotone capacity, f, to a function, F, defined on a vector space leads, in fact, as we shall see, to a concave function.

Let $\mathbb{X}$ be the vector space generated by the indicator functions I_B of events $B \in \mathcal{A}$; a generic element $X \in \mathbb{X}$ is defined by

$$\theta \rightarrowtail X(\theta) = \sum_{j \in J} \alpha_j I_{B_j}(\theta),$$

and denoted by $X = \Sigma_{j \in J} \alpha_j I_{B_j}$. However, the same X is the linear combination of diverse sets of indicator functions; in particular, every $X \in \mathbb{X}$ has a unique representation

$$X = \sum_{n=2}^{N} (x_n - x_{n-1}) I_{A_n} + x_1 I_\Theta,$$

where $x_n > x_{n-1}$ for all n, and events $A_1 = \Theta$ and $A_n, 2 \leq n \leq N$, form a decreasing sequence.

Define mapping $F : \mathbb{X} \to \mathbb{R}$ by

$$X \to F(X) = \sum_{n=2}^{N} (x_n - x_{n-1}) f(A_n) + x_1 .$$

Since $F(I_A) = f(A)$ for $A \in \mathcal{A}$, F becomes an extension of f, when $\mathcal{A}$ is identified with $\{I_A, A \in \mathcal{A}\}$, a subset of $\mathbb{X}$.

It is obvious that F is positively homogeneous of order 1, i.e., $F(\lambda X) = \lambda F(X)$ for $\lambda \geq 0$; thus, f is concave if and only if it is superadditive, i.e., $F(X' + X'') \geq F(X') + F(X'')$. Since, by (22), for any $X \in \mathbb{X}$, $F(X) = \mathrm{Inf}_{P \in \mathcal{P}} E_P(X)$, the last inequality results from an elementary property of infima.

4 Applications to Decision Making

Many statistical decision problems can be expressed as parametric problems, in which a sample distribution, P_ω, depends on a parameter ω only known to belong to some set Ω.

In other decision problems, in which the information is subjective rather than objective, decision makers often feel that they are only able to ascribe probability intervals to events, such as ‘the probability of event E is at least equal to p and at most equal to q’. In both cases, there exists a set $\mathcal{P}_0$ of probabilities compatible with the available information.

Let a decision be identified with a mapping $X : \Theta \to \mathbb{R}$, with $X(\theta)$ equal to the consequence (or its utility) resulting from that decision when θ obtains.

Decision criteria which are suited to the uncertainty situations just described include among others, Wald’s (1971) Maximin criterion, $\mathrm{Sup}_{X \in \mathbb{X}}$ $\mathrm{Inf}_{P \in \mathcal{P},}$ $E_P(X)$, and the more general criterion, $\mathrm{Sup}_{X \in \mathbb{X}} \upsilon(\mathrm{Inf}_{P \in \mathcal{P}_0}\ E_P(X)$, $\mathrm{Sup}_{P \in \mathcal{P}_0}\ E_P(X))$. This last criterion, which allows for varying degrees of pessimism, has been given an axiomatic justification in Cohen and Jaffray (1985). Since $\mathrm{Sup}_{P \in \mathcal{P}_0}\ E_P(X) = -\mathrm{Inf}_{P \in \mathcal{P}_0}\ E_P(-X)$, both criteria demand the same exact or approximate calculation, that of $\mathrm{Inf}_{P \in \mathcal{P}_0}\ E_P(X)$.

Let $f = \mathrm{Inf}_{P \in \mathcal{P}_0}\ P.f$ is obviously a capacity; let m be its Möbius inverse; then, according to Proposition 10, Corollary 4 and Remark 2,

$$\begin{aligned} E_f^0(X) \sum_{l=2}^{L} (x_l - x_{l-1}) f(X \geq x_l) + x_1 &= \sum_{B \in \mathcal{A}} x_{n(B)} \cdot m(B) \\ &\leq \operatorname*{Inf}_{P \in \mathcal{P}_{\geq}} E_P(X) \leq \operatorname*{Inf}_{P \in \mathcal{P}_0} E_P(X) . \end{aligned} \tag{24}$$

Thus expressions (22) and (23) provide us with an easy calculation of a lower bound for $\mathrm{Inf}_{P \in \mathcal{P}_0}\ E_P(X)$.

Unfortunately, it seems difficult to determine if this lower bound is attained or not: we know that the first inequality in (24) becomes an equality if and only if f is 2-monotone; however, the similar property for the second inequality does not depend on f's order of monotonicity as shown by the following examples.

Example 6.

$$\Theta = \{\theta_1, \theta_2, \theta_2\}\,; \mathcal{P}_0 = \{P_1, P_2\} \text{ with } P_1(\{\theta_1\}) = \frac{1}{4}, P_1(\{\theta_2\}) = \frac{1}{2},$$

$P_1(\{\theta_3\}) = \frac{1}{4}$ and $P_2(\{\theta_1\}) = \frac{1}{3}$, $P_2(\{\theta_2\}) = \frac{1}{6}$ and $P_2(\{\theta_3\}) = \frac{1}{2}$. To $f = \operatorname{Inf}_{P \in \mathcal{P}_0} P = \operatorname{Min}\{P_1, P_2\}$ corresponds a Möbius inverse m satisfying:

$$m(\{\theta_1\}) = \frac{1}{4}, m(\{\theta_2\}) = \frac{1}{6}, m(\{\theta_3\}) = \frac{1}{4}; m(\{\theta_1; \theta_2\}) = \frac{1}{12},$$
$$m(\{\theta_1, \theta_3\}) = 0,\ m(\{\theta_2, \theta_3\}) = \frac{1}{4},$$

and $m(\Theta) = 0$; thus f is ∞-monotone; however, for $X(\{\theta_1\}) = 2,\ X(\{\theta_2\}) = 1$ and $X(\{\theta_3\}) = 0$,

$$\operatorname*{Inf}_{P \in \mathcal{P}_{\geq}} E_P(X) = \frac{3}{4} < \frac{5}{6} = \operatorname*{Inf}_{P \in \mathcal{P}_0} E_P(X).$$

□

Example 7. Let $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ and let $\mathcal{P}_0 = \{P_1, P_2, P_3\}$, with

$$P_1(\{\theta_2\}) = P_1(\{\theta_3\}) = \frac{1}{2}, P_2(\{\theta_1\}) = \frac{1}{6}, P_2(\{\theta_2\}) = \frac{1}{4}, P_2(\{\theta_3\}) = \frac{1}{3},$$
$$P_2(\{\theta_4\}) = \frac{1}{4} P_3(\{\theta_1\}) = \frac{1}{6}, P_3(\{\theta_2\}) = \frac{1}{3}, P_3(\{\theta_3\}) = P_3(\{\theta_4\}) = \frac{1}{4}.$$

$f = \operatorname{Inf}_{P \in \mathcal{P}_0} P$ satisfies

$$f(\{\theta_1\}) = f(\{\theta_4\}) = 0, f(\{\theta_2\}) = f(\{\theta_3\}) = \frac{1}{4}; f(\{\theta_1, \theta_4\}) = 0,$$
$$f(\{\theta_2, \theta_4\}) = f(\{\theta_3, \theta_4\}) = \frac{1}{2}, f(\{\theta_1, \theta_2\}) = f(\{\theta_1, \theta_3\}) = \frac{5}{12},$$
$$f(\{\theta_2, \theta_3\}) = \frac{7}{12}; f(\{\theta_1, \theta_2, \theta_3\}) = \frac{3}{4}, f(\{\theta_2, \theta_3, \theta_4\}) = \frac{5}{6},$$
$$f(\{\theta_1, \theta_2, \theta_4\}) = f(\{\theta_1, \theta_3, \theta_4\}) = \frac{1}{2}.$$

f is not 2-monotone since

$$f(\{\theta_1, \theta_2, \theta_3\}) + f(\{\theta_1\}) = \frac{3}{4} < \frac{5}{6} = f(\{\theta_1, \theta_2\}) + f(\{\theta_1, \theta_3\}).$$

Indeed, for $S = (\theta_3, \theta_4, \theta_1, \theta_2),\ P_S \notin \mathcal{P}_{\geq}$, since, by (19),

$$P_S(\{\theta_1\}) = 0,\ P_S(\{\theta_2\}) = \frac{1}{2}, P_S(\{\theta_3\}) = P_S(\{\theta_4\}) = \frac{1}{4},$$

hence $P_S(\{\theta_1, \theta_3\}) < f(\{\theta_1, \theta_3\})$.

For X defined by $X(\theta_1) = 1,\ X(\theta_2) = 0,\ X(\theta_3) = 3$ and $X(\theta_4) = 2$,

$$E_{P_S}(X) = E_f^0(X) = \frac{15}{12},$$

whereas

$$\begin{aligned} &\operatorname*{Inf}_{P \in \mathcal{P}_0} E_P(X) = \frac{17}{12} = \operatorname*{Inf}_{P \in \mathcal{P}_\geq} E_P(X), \\ &(P \in \mathcal{P}_\geq \text{ implies } 3\ P(\{\theta_3\}) + 2P(\{\theta_4\}) + P(\{\theta_1\}) \\ &= 2P(\{\theta_3, \theta_4\}) + P(\{\theta_1, \theta_3\}) \geq 2f(\{\theta_3, \theta_4\}) + f(\{\theta_1, \theta_3\}) = \frac{17}{12}). \end{aligned}$$

□

Appendix

The following results and proofs are extracts from Shafer (1976), which are needed to make this paper self-contained.

Lemma 2. (p. 47). *If A is a finite set then*

$$\sum_{B \subset A} (-1)^{|B|} = \begin{cases} 1 & if\ A = \varnothing \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Obvious if $A = \varnothing$. When $A = \{\theta_1, \ldots, \theta_i, \ldots, \theta_n\} \neq \varnothing$,

$$\begin{aligned} \sum_{B \subset A} (-1)^{|B|} &= (-1)^{|\varnothing|} + \sum_i (-1)^{|\{\theta_i\}|} + \sum_{i<j} (-1)^{|\{\theta_i, \theta_j\}|} + \ldots + (-1)^{|A|} \\ &= \binom{n}{0} - \binom{n}{1} + \binom{n}{2} + \ldots + (-1)^n \binom{n}{n} = (1-1)^n = 0. \end{aligned}$$

□

Lemma 3. (p. 48). *If A is a finite set and $B \subset A$, then*

$$\sum_{B \subset C \subset A} (-1)^{|C|} = \begin{cases} (-1)^{|A|} & if\ A = B \\ 0 & otherwise. \end{cases}$$

Proof. Follows from Lemma 2, since

$$\sum_{B \subset C \subset A} (-1)^{|C|} = \sum_{D \subset A \backslash B} (-1)^{|B \cup D|} = (-1)^{|B|} \sum_{D \subset A \backslash B} (-1)^{|D|}.$$

□

Lemma 4. (p. 48). *Suppose Θ is a finite set and f and g are functions on 2^{Θ}. Then*

$$f(A) = \sum_{B \subset A} g(B) \ \ \textit{for all } A \subset \Theta \tag{25}$$

if and only if

$$g(A) = \sum_{B \subset A} (-1)^{|A \backslash B|} f(B) \ \ \textit{for all } A \subset \Theta \tag{26}$$

Proof. Both implications follow by simple calculations using Lemma 3.

(i) If (I) holds, then

$$\begin{aligned} \sum_{B \subset A} (-1)^{|A \backslash B|} f(B) &= (-1)^{|A|} \sum_{B \subset A} (-1)^{|B|} f(B) \\ &= (-1)^{|A|} \sum_{B \subset A} (-1)^{|B|} \sum_{C \subset B} g(C) \\ &= (-1)^{|A|} \sum_{C \subset A} g(C) \sum_{C \subset B \subset A} g(-1)^{|B|} \\ &= (-1)^{|A|} g(A) (-1)^{|A|} = g(A) . \end{aligned}$$

(ii) If (II) holds, then

$$\begin{aligned} \sum_{B \subset A} g(B) &= \sum_{B \subset A} \sum_{C \subset B} (-1)^{|B \backslash C|} f(C) \\ &= \sum_{C \subset A} (-1)^{|C|} f(C) \sum_{C \subset B \subset A} (-1)^{|B|} \\ &= (-1)^{A} f(A) (-1)^{|A|} f = f(A) . \end{aligned}$$

□

Acknowledgments

We wish to thank W. Batchelder and an anonymous referee for numerous helpful suggestions and comments.

References

B. Anger, Approximation of Capacities by Measures, in: Lecture Notes in Mathematics 226 (Springer, Berlin, 1971) pp. 152–170.

B. Anger, Representation of capacities, Math. Ann. 229 (1977) 245–258.

C. Berge, Espaces Topologiques, Fonctions Multivoques (Dunod, Paris, 1965).

R.E. Bixby, W.H. Cunningham and D.M. Tokpis, The partial order of a polymatroïd extreme point, Math. Oper. Res. 10 (1985) 367–378.

G. Choquet, Théorie des capacités, Ann. Inst. Fourier (Grenoble) (1953) V. 131–295.

M. Cohen and J.Y. Jaffray, Decision making in a case of mixed uncertainty: A normative model, J. Math. Psych. 29 (1985) 428–442.

C. Dellacherie, Quelques commentaires sur les prolongements de capacités, Lect. Notes Math. 191 (Sem. Prob. V) (1971) 77–81.

A.P. Dempster, Upper and lower probabilities induced by a multivalued mapping, Ann. Math. Statist. 38 (1967) 325–339.

J. Edmonds, Submodular functions, matroïds and certain polyhedra, Combinatorial structures and their applications (Proc. Calgary Int. Conf. 1969). R.L. Guy et al., eds (Gordon and Breach, New York, 1970) pp. 69–87.

D. Gale, The Theory of Linear Economic Models (Mc Graw Hill, New York, 1960).

P.J. Huber, The use of Choquet capacities in statistics, Bull. Int. Statist. Inst. XLV, Book 4 (1973) 181–188.

P.J. Huber, Kapazitäten statt Wahrscheinlichkeiten. Gedanken zur Grundlegung der Statistik, J. der Dt. Math. Verein. 78 (1976) 81–92.

P.J. Huber and V. Strassen, Minimax tests and the Neyman-Pearson lemma for capacities, Ann. Statist. 1 (1973) 251–263.

T. Ishiishi, Super-modularity: applications to convex games and to the greedy algorithm for LP, J. Econom. Theory 25 (1981) 283–286.

S. Karlin, Mathematical Methods and Theory in Games, Programming and Economics, Vol. 1 (Pergamon Press, London, Paris, 1959).

H. Kyburg, The Logical Foundations of Statistical Inference (Reidel, Dordrecht, 1974).

I. Levi, The Enterprise of Knowledge (MIT Press, Cambridge, 1980).

A. Papamarcou and T.L. Fine, A note on undominated lower probabilities, The Annals of Probab. 14 (1986) 710–723.

A. Revuz, Fonctions croissantes et mesures sur les espaces topologiques ordonnés, Ann. Instit. Fourier (Grenoble VI) (1955) 187–269.

G.C. Rota, Theory of Möbius functions, Z. fur Wahrscheinlichkeitstheorie und Verwandte Gebiete 2 (1964) 340–368.

G. Shafer, A Mathematical Theory of Evidence (Princeton University Press, Princeton, New Jersey, 1976).

G. Shafer, Allocations of probability, Ann. Prob. 7 (1979) 827–839.

G. Shafer, Constructive probability, Synthese 48 (1981) 1–59.

L.S. Shapley, Cores of convex games, Int. J. Game Theory 1 (1971) 11–26.

A. Wald, Statistical Decision Functions (Chelsea Publishing Company, Bronx, New York, 1971).

P. Walley and T.L. Fine, Towards a frequentist theory of upper and lower probability, Ann. Statist. 10 (1982) 741–761.

P. Walley and T.L. Fine, Varieties of modal (classificatory) and comparative probability, Synthese, 41 (1979) 321–374.

M. Wolfenson and T.L. Fine, Bayes-like decision making with upper and lower probabilities, J. Amer. Statist. Assoc. 77 (1982) 80–88.

20

Axioms for Probability and Belief-Function Propagation*

Prakash P. Shenoy and Glenn Shafer

Abstract. In this paper, we describe an abstract framework and axioms under which exact local computation of marginals is possible. The primitive objects of the framework are variables and valuations. The primitive operators of the framework are combination and marginalization. These operate on valuations. We state three axioms for these operators and we derive the possibility of local computation from the axioms. Next, we describe a propagation scheme for computing marginals of a valuation when we have a factorization of the valuation on a hypertree. Finally we show how the problem of computing marginals of joint probability distributions and joint belief functions fits the general framework.

1 Introduction

In this paper, we describe an abstract framework and present axioms for local computation of marginals in hypertrees. These axioms justify the use of local computation to find marginals for a probability distribution or belief function when the probability distribution or belief function is factored on a hypertree. The axioms are abstracted from the belief-function work of the authors (e.g., Shenoy and Shafer [30], Shenoy, Shafer and Mellouli [33], Shafer, Shenoy, and Mellouli [26]), but they apply to probabilities as well as to belief functions.

In the probability case, the factorization is usually a factorization of a joint probability distribution, perhaps into marginals and conditionals. Probability factorizations sometimes arise from causal models, which relate each variable to a relatively small number of immediate causes; see e.g., Pearl [20]. Probability factorizations can also arise from statistical models; see e.g., Darroch, Lauritzen and Speed [6]. Belief-function factorizations generally arise from the

* Research for this article has been partially supported by NSF grant IRI-8610293 and a Research Opportunities in Auditing grant 87-135 from the Peat Marwick Foundation. A condensed version of this paper previously appeared in the Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence in 1988.

decomposition of evidence into independent items, each involving only a few variables. We represent each item of evidence by a belief function and combine these belief functions by Dempster's rule [23].

It is shown in Shenoy [28] that Spohn's [35, 34] theory of epistemic beliefs also fits in the abstract framework described here. Furthermore, the axiomatic framework described here is extended in Shenoy and Shafer [32, 31] to include constraint propagation and optimization using local computation.

We first present our general axiomatic framework and then explain how it applies to probabilities and belief functions. Before we can present the axiomatic framework, we need to review some graph-theoretic concepts. We do this in Sect. 2. We present the framework in Sect. 3. We apply it to probabilities in Sect. 4 and to belief functions in Sect. 5.

2 Some Concepts from Graph Theory

Most of the concepts reviewed here have been studied extensively in the graph theory literature (see Berge [2], Golumbic [11], and Maier [18]). A number of the terms we use are new, however - among them, *hypertree*, *construction sequence*, *branch*, *twig*, *bud*, and *Markov tree*. A *hypertree* is what other authors have called an acyclic (Maier [18]) or decomposable hypergraph (Lauritzen, Speed and Vijayan [16]). A *construction sequence* is what other authors have called a sequence with the running intersection property (Lauritzen and Spiegelhalter [17]). A *Markov tree* is what authors in database theory have called a join tree (see Maier [18]). We have borrowed the term *Markov tree* from probability theory, where it means a tree of variables in which separation implies probabilistic conditional independence given the separating variables. For a fuller explanation of the concepts reviewed here, see Shafer and Shenoy [25].

As we shall see, hypertrees are closely related to Markov trees. The vertices of a Markov tree are always hyperedges of a hypertree, and the hyperedges of a hypertree can always be arranged in a Markov tree.

Hypergraphs and Hypertrees. We call a non-empty set $\mathbf{H}$ of non-empty subsets of a finite set $\boldsymbol{\chi}$ a *hypergraph* on $\boldsymbol{\chi}$. We call the elements of $\mathbf{H}$ *hyperedges.* We call the elements of $\boldsymbol{\chi}$ *vertices.*

Suppose t and b are distinct hyperedges in a hypergraph $\mathbf{H}$, $t \cap b \neq \phi$, and b contains every vertex of t that is contained in a hyperedge of $\mathbf{H}$ other than t; if $X \in t$ and $X \in h$, where $h \in \mathbf{H}$ and $h \neq t$, then $X \in b$. Then we call t a *twig* of $\mathbf{H}$, and we call b a *branch* for t. A twig may have more than one branch.

We call a hypergraph a *hypertree* if there is an ordering of its hyperedges, say $h_1 h_2 \dots h_n$, such that h_k is a twig in the hypergraph $\{h_1, h_2, \dots, h_k\}$ whenever $2 \leq k \leq n$. We call any such ordering of the hyperedges a *hypertree construction sequence* for the hypertree. We call the first hyperedge in a hypertree

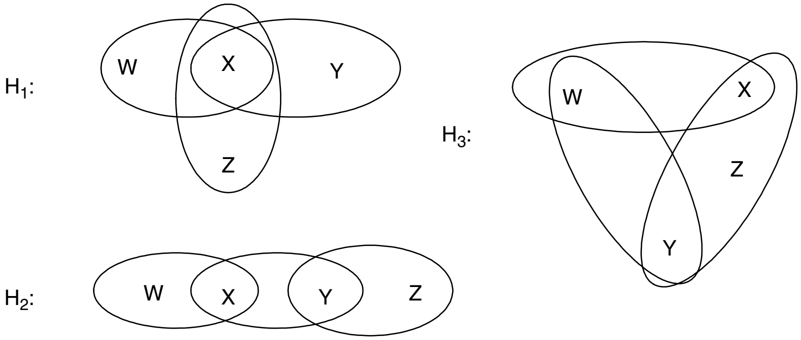


Fig. 1. Some hypergraphs on $\{W, X, Y, Z\}$. The hypergraph H_1 is a hyper tree, all of its hyperedges are twigs, and all six orderings of its hyperedges are hypertree construction sequences. The hypergraph H_2 is a hypertree, hyperedges $\{W, X\}$ and $\{Y, Z\}$ are twigs, and there are only four hypertree construction sequences: $\{W, X\}\{X, Y\}\{Y, Z\}$, $\{X, Y\}\{W, X\}\{Y, Z\}$, $\{X, Y\}\{Y, Z\}\{W, X\}$, and $\{Y, Z\}\{X, Y\}\{W, X\}$. The hypergraph H_3 is not a hypertree and it has no twigs

construction sequence the *root* of the hypertree construction sequence. Figure 1 illustrates hypergraphs, hypertrees, twigs and construction sequences.

Figure 1 illustrates hypergraphs, hypertrees, twigs and construction sequences.

If we construct a hypertree by adding hyperedges following a hypertree construction sequence, then each hyperedge we add is a twig when it is added, and it has at least one branch in the hypertree at that point. Suppose we choose such a branch, say $\beta(h)$, for each hyperedge h we add. By doing so, we define a mapping β from $\mathbf{H} - \{h_1\}$ to $\mathbf{H}$, where h_1 is the root of the hypertree construction sequence. We will call this function a *branching* for the hypertree construction sequence.

Since a twig may have more than one branch, a hypertree construction sequence may have more than one branching. In general, a hypertree will have many construction sequences. In fact, for each hyperedge of a hypertree, there is at least one construction sequence beginning with that hyperedge.

Hypertree Covers of Hypergraphs. We will justify local computation under two assumptions. The joint probability distribution function or the joint belief function with which we are working must factor into functions each involving a small set of variables. And these sets of variables must form a hypertree.

If the sets of variables form instead a hypergraph that is not a hypertree, then we must enlarge it until it is a hypertree. We can talk about this enlargement in two different ways. We can say we are adding larger hyperedges, keeping the hyperedges already there. Or, alternatively, we can say we are replacing the hyperedges already there with larger hyperedges. The choice between these two ways of talking matters little, because the presence of superfluous twigs (hyperedges contained in other hyperedges) does not affect whether a hypergraph is a hypertree, and because the computational cost of the procedures we will be describing depends primarily on the size of the largest hyperedges, not on the number of the smaller hyperedges (Kong [15], Mellouli [19]).

Formally, we will say that a hypergraph $\mathbf{H}^*$ covers a hypergraph $\mathbf{H}$ if for every h in $\mathbf{H}$ there is an element h^* of $\mathbf{H}^*$ such that $h^* \supseteq h$. We will say that $\mathbf{H}^*$ is a *hypertree cover* for $\mathbf{H}$ if $\mathbf{H}^*$ is a hypertree and it covers $\mathbf{H}$. Figure 2 shows a hypergraph that is not a hypertree and a hypertree cover for it.

Finding a hypertree cover is never difficult. The hypertree $\{\boldsymbol{\chi}\}$, which consists of the single hyperedge $\boldsymbol{\chi}$, is a hypertree cover for any hypergraph on $\boldsymbol{\chi}$. Finding a hypertree cover without large hyperedges, or finding a hypertree cover whose largest hyperedge is as small as possible, may be very difficult. How to do this best is the subject of a growing literature; see e.g., Rose [21], Bertele and Brioschi [3], Tarjan and Yannakakis [36], Kong [15], Arnborg, Corneil and Proskurowski [1], Mellouli [19], and Zhang [39].

Trees. A *graph* is a pair $(\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is a nonempty set and $\mathbf{E}$ is a set of two-element subsets of $\mathbf{V}$. We call the elements of $\mathbf{V}$ *vertices*, and we call the elements of $\mathbf{E}$ *edges*.

Suppose $(\mathbf{V}, \mathbf{E})$ is a graph. If $\{v, v'\}$ is an element of $\mathbf{E}$, then we say that v and v' are *neighbors*. We call a vertex of a graph a *leaf* if it is contained in only one edge, and we call the other vertex in that edge the *bud* for the leaf. If $v_1 v_2 \ldots v_n$ is a sequence of distinct vertices, where $n > 1$, and $\{v_k, v_{k+1}\} \in \mathbf{E}$ for $k = 1, 2, \ldots, n-1$, then we call $v_1 v_2 \ldots v_n$ a *path from* v_1 to v_n.

We call a graph a *tree* if there is an ordering of its vertices, say $v_1 v_2 \ldots v_n$ such that v_k is a leaf in the graph $(\{v_1, v_2, \ldots, v_k\}, \mathbf{E}_k)$ whenever $2 \leq k \leq n$, where $\mathbf{E}_k$ is the subset of $\mathbf{E}$ consisting of those edges that contain only vertices in $\{v_1, v_2, \ldots, v_k\}$. We call any such ordering of the vertices a *tree construction*

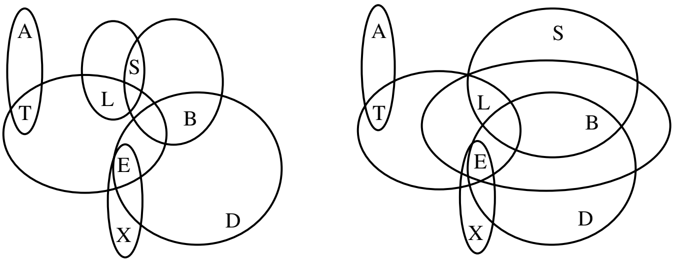


Fig. 2. Left: A hypergraph that is not a hypertree. **Right**: A hypertree cover for it obtained by adding hyperedges $\{S, L, B\}$ and $\{L, E, B\}$ and removing hyperedges $\{S, L\}$ and $\{S, B\}$

sequence for the tree. We call the first vertex in a tree construction sequence the *root* of the tree construction sequence. Note that in a tree, for any two distinct vertices v_i and v_j, there is a unique path from v_i to v_j.

If we construct a tree following a tree construction sequence starting with the root and adding vertices, then each vertex we add is a leaf when it is added, and it has a bud in the tree at that point. Given a tree construction sequence and a vertex v that is not the root, let $\beta(v)$ denote the bud for v as it is added. This defines a mapping β from $\mathbf{V} - \{v_1\}$ to $\mathbf{V}$, where v_1 is the root. We will call this mapping the *budding* for the tree construction sequence.

The budding for a tree construction sequence is analogous to the branching for a hypertree construction sequence, but there are significant differences. Whereas there may be many branchings for a given hypertree construction sequence, there is only one budding for a given tree construction sequence. In fact, there is only one budding with a given root.

Markov Trees. We have just defined a tree as a pair $(\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is the set of vertices, and $\mathbf{E}$ is the set of edges. In the case of a Markov tree, the vertices are themselves non-empty sets. In other words, the set $\mathbf{V}$ is a hypergraph. In fact, it turns out to be a hypertree.

Here is our full definition. We call a tree $(\mathbf{H}, \mathbf{E})$ a *Markov tree* if the following conditions are satisfied: (i) $\mathbf{H}$ is a hypergraph; (ii) If $\{h, h'\} \in \mathbf{E}$, then $h \cap h' \neq \phi$; and (iii) If h and h' are distinct vertices, and X is in both h and h', then X is in every vertex on the path from h to h'.

This definition does not state that $\mathbf{H}$ is a hypertree, but it implies that it is:

Proposition 1. (i) If $(\mathbf{H}, \mathbf{E})$ is a Markov tree, then $\mathbf{H}$ is a hypertree. Any leaf in $(\mathbf{H}, \mathbf{E})$ is a twig in $\mathbf{H}$. If $h_1h_2 \ldots h_n$ is a tree construction sequence for $(\mathbf{H}, \mathbf{E})$, with β as its budding, then $h_1h_2 \ldots h_n$ is also a hypertree construction sequence for $\mathbf{H}$, with β as a branching. (ii) If $\mathbf{H}$ is a hypertree, $h_1h_2 \ldots h_n$ is a hypertree construction sequence for $\mathbf{H}$, and β is a branching for $h_1h_2 \ldots h_n$, then $(\mathbf{H}, \mathbf{E})$ is a Markov tree, where $\mathbf{E}$ $= \{(h_2, \beta(h_2)), \ldots, (h_n, \beta(h_n))\}$; $h_1h_2 \ldots h_n$ is a tree construction sequence for $(\mathbf{H}, \mathbf{E})$, and β is its budding.

See Shafer and Shenoy [25] for a proof of Proposition 1. The key point here is the fact that a leaf in the Markov tree is a twig in the hypertree. This means that as we delete leaves from a Markov tree (a visually transparent operation), we are deleting twigs from the hypertree.

If $(\mathbf{H}, \mathbf{E})$ is a Markov tree, then we call $(\mathbf{H}, \mathbf{E})$ a *Markov tree representative* for the hypertree $\mathbf{H}$. As per Proposition 1, every hypertree has a Markov tree representative. Most hypertrees have more than one. Figure 3 shows three Markov tree representations for the hypertree in Fig. 2.

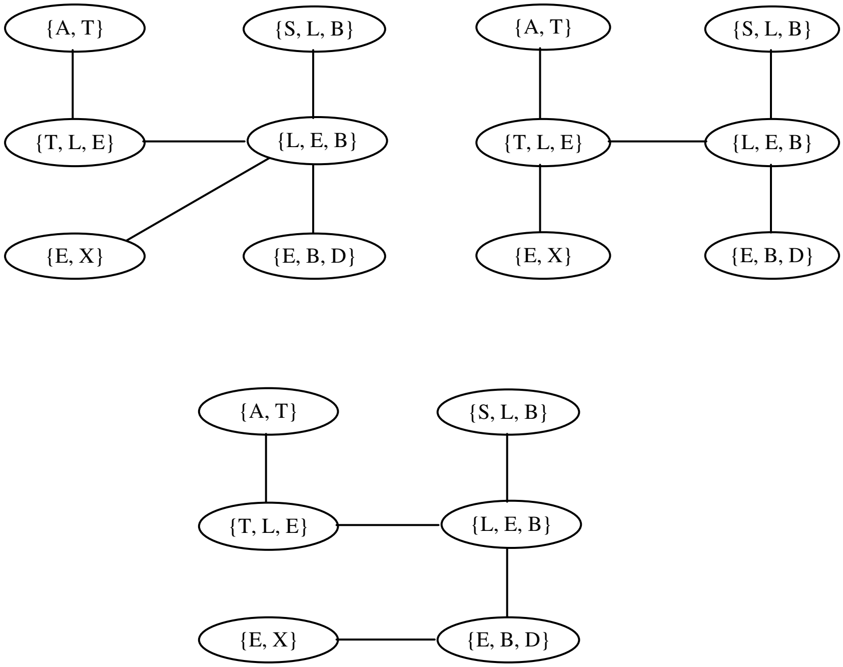


Fig. 3. If we choose $\{L, E, B\}$ as the root for the hypertree in Fig. 2, then $\{L, E, B\}$ must serve as the branch for $\{T, L, E\}$, $\{E, B, D\}$, and $\{S, L, B\}$, and $\{T, L, E\}$ must serve as the branch for $\{A, T\}$. This leaves only $\{E, X\}$, which can use $\{L, E, B\}$, $\{T, L, E\}$, or $\{E, B, D\}$ as its branch. It follows that the hypertree has exactly three Markov tree representations, which differ only in where the leaf $\{E, X\}$ is attached

3 An Axiomatic Framework for Local Computation

In this section, we describe a set of axioms under which exact local computation of marginals is possible.

In Sect. 3.1, we describe an axiomatic framework for local computation of marginals. The primitive objects of the framework are variables and valuations. The framework has two primitive operators, combination and marginalization. These operate on valuations. We state three axioms for these operators.

In Sect. 3.2, we show how local computation can be used to marginalize a factorization (of a valuation) on a hypergraph to the smaller hypergraph resulting from the deletion of a twig. Once we know how to delete a twig, we can reduce a hypertree to a single hyperedge by successively deleting twigs. When we have reduced a factorization on a hypertree to a factorization on a single hyperedge, it is no longer a factorization; it is simply the marginal for the hyperedge.

In Sect. 3.3, we shift our attention from a hypertree to the Markov tree determined by a branching for the hypertree. Using this Markov tree, we describe more graphically the process of marginalizing to a single hyperedge. Our description is based on the idea that each vertex in the tree is a processor, which can operate on valuations for the variables it represents and then send the result to a neighboring processor. In Sect. 3.4, we generalize this idea to a scheme of simultaneous computation and message passing that produces marginals for all the vertices in the Markov tree.

3.1 The Axiomatic Framework

The primitive objects of the framework are a finite set of variables, and a set of valuations. The framework has two primitive operators: combination and marginalization. These operate on valuations.

Variables and Valuations. Let χ be a finite set. The elements of χ are called *variables.* For each $h \subseteq \chi$, there is a set $\mathbf{V}_h$. The elements of $\mathbf{V}_h$ are called *valuations* on h. Let $\mathbf{V}$ denote $\cup\{\mathbf{V}_h \mid h \subseteq \chi\}$, the set of all valuations.

In the case of probabilities, a valuation on h will be a non-negative, real-valued function on the set of all configurations of h (a configuration of h is a vector of possible values of variables in h). In the belief-function case, a valuation is a non-negative, real-valued function on the set of all subsets of configurations of h.

Proper Valuations. For each $h \subseteq \chi$, there is a subset p_h of $\mathbf{V}_h$ whose elements will be called proper valuations on h. Let p denote $\cup\{p_h \mid h \subseteq \chi\}$, the set of all *proper valuations.* The notion of proper valuations is important as it will enable us to define combinability of valuations.

In the probability case, a valuation H on h is said to be proper if the values of the function H are not zero for all configurations of h. In the belief function case, a valuation H on h is said to proper if the values of the function H are not zero for all non-empty subsets of configurations of h.

Combination. We assume there is a mapping $\otimes : \mathbf{V} \times \mathbf{V} \to \mathbf{V}$, called *combination*, such that (i) If G and H are valuations on g and h respectively, then $G \otimes H$ is a valuation on $g \cup h$; (ii) If either G or H is not a proper valuation, then $G \otimes H$ is not a proper valuation; and (iii) If G and H are both proper valuations, then $G \otimes H$ may or may not be a proper valuation.

If $G \otimes H$ is not a proper valuation, then we shall say that G and H are *not combinable.* If $G \otimes H$ is a proper valuation, then we shall say that G and H are *combinable* and that $G \otimes H$ is the *combination of G and H.*

Intuitively, combination corresponds to aggregation. If G and H represent information about variables in g and h, respectively, then GH represents the aggregated information for variables in $g \cup h$. In the probability case, combination corresponds to pointwise multiplication. In the belief function case, combination corresponds to Dempster's rule.

Marginalization. We assume that for each $h \subseteq \chi$, there is a mapping $\downarrow h : \cup\{\mathbf{V}_g \mid g \supseteq h\} \to \mathbf{V}_h$, called *marginalization* to h, such that (i) If G

is a valuation on g and $h \subseteq g$, then $G^{\downarrow h}$ is a valuation on h; (ii) If G is a proper valuation, then $G^{\downarrow h}$ is a proper valuation; and (iii) If G is not a proper valuation, then $G^{\downarrow h}$ is not a proper valuation.

We will call $G^{\downarrow h}$ *marginal of* G for h.

Intuitively, marginalization corresponds to narrowing the focus of a valuation. If G is a valuation on g representing some information about variables in g, and $h \subseteq g$, then $G^{\downarrow h}$ represents the information for variables in h implied by G if we disregard variables in $g - h$. In both the probability and belief-function cases, marginalization corresponds to summation.

The Problem. We are now in a position to describe the problem. Suppose $\mathbf{H}$ is a hypergraph on $\boldsymbol{\chi}$. For each $h \in \mathbf{H}$, we have a proper valuation A_h on h. First, we need to determine if the proper valuations in the set $\{A_h \mid h \in \mathbf{H}\}$ are combinable. If the answer is in the affirmative then let A denote the proper valuation $\otimes\{A_h \mid h \in \mathbf{H}\}$. Second, we need to find the marginal of A for each $X \in \boldsymbol{\chi}$.

If $\boldsymbol{\chi}$ is a large set of variables, then computation of $A^{\downarrow\{X\}}$ by first computing the joint valuation A on $\boldsymbol{\chi}$ and then marginalizing A to $\{X\}$ will not be possible. For example, if we have 50 variables and each variable has 2 possible values, then we will have 2^{50} possible configurations of $\boldsymbol{\chi}$. Thus in the probability case, computing A will involve finding 2^{50} values. And in the belief function case, computing A will involve finding $2^{2^{50}}$ values. In either case, the task is infeasible. We will state axioms for combination and marginalization that make it possible to use local computation to determine if the given proper valuations are combinable and to compute $A^{\downarrow\{X\}}$ for each $X \in \boldsymbol{\chi}$ if they are.

We will assume that these two mappings satisfy three axioms.

Axiom A1 (*Commutativity and associativity of combination*): Suppose G, H, K are valuations on g, h, and k respectively. Then $G \otimes H = H \otimes G$, and $G \otimes (H \otimes K) = (G \otimes H) \otimes K$.

Axiom A2 (*Consonance of marginalization*): Suppose G is a valuation on g, and suppose $k \subseteq h \subseteq g$. Then $(G^{\downarrow h})^{\downarrow k} = G^{\downarrow k}$.

Axiom A3 (*Distributivity of marginalization over combination*): Suppose G and H are valuations on g and h, respectively. Then $(G \otimes H)^{\downarrow g} = G \otimes (H^{\downarrow g \cap h})$

One implication of Axiom A1 is that when we have multiple combinations of valuations, we can write it without using parenthesis. For example, $(\ldots((A_{h_1} \otimes A_{h_2}) \otimes A_{h_3}) \otimes \ldots \otimes A_{h_n})$ can be written simply as $\otimes\{A_{h_i} \mid i = 1, \ldots, n\}$ without indicating the order in which the combinations are carried out.

Factorization. Suppose A is a valuation on a finite set of variables $\boldsymbol{\chi}$, and suppose $\mathbf{H}$ is a hypergraph on $\boldsymbol{\chi}$. If A is equal to the combination of valuations on the hyperedges of h, say $A = \otimes\{A_h \mid h \in \mathbf{H}\}$, where A_h is a valuation on h, then we say that A *factorizes on* $\mathbf{H}$.

If we regard marginalization as a reduction of a valuation by deleting variables, then axiom A2 can be interpreted as saying that the order in which the variables are deleted does not matter.

Axiom A3 is the crucial axiom that makes local computation possible. Axiom A3 states that computation of $(G \otimes H)^{\downarrow g}$ can be accomplished without having to compute $G \otimes H$.

3.2 Marginalizing Factorizations

In this section, we learn how to adjust a factorization on a hypergraph to account for the deletion of a twig. This can be accomplished by local computations, computations involving only the valuations on the twig and a branch for the twig. This elimination of a twig by local computation is the key to the computation of marginals from a factorization on a hypertree, for by successively deleting twigs, we can reduce the hypertree to a single hyperedge.

Suppose $\mathbf{H}$ is a hypergraph on χ, t is a twig in $\mathbf{H}$, and b is a branch for t. The twig t may contain some vertices that are not contained in any other hyperedge in $\mathbf{H}$. These are the vertices in the set $t-b$. Deleting t from $\mathbf{H}$ means reducing $\mathbf{H}$ to the hypergraph $\mathbf{H}-\{t\}$ on the set $\chi' = \chi-(t-b) = \cup(\mathbf{H}-\{t\})$.

Suppose A is a valuation on χ, suppose A factors on $\mathbf{H}$, and suppose we have stored A in factored form. In other words, we have stored a valuation A_h for each h in $\mathbf{H}$, and we know that $A = \otimes\{A_h \mid h \in \mathbf{H}\}$. Adapting this factorization on A on $\mathbf{H}$ to the deletion of the twig t means reducing it to a factorization of $A^{\downarrow \chi'}$ on $\mathbf{H} - \{t\}$. Can we do this? Yes. The following proposition tells us that if A factors on $\mathbf{H}$, then $A^{\downarrow \chi'}$ factors on $\mathbf{H} - \{t\}$, and the second factorization can be obtained from the first by a local computation that involves only t and a branch.

Proposition 2. Under the assumptions of the preceding paragraph,

$$A^{\downarrow \chi'} = (A_b \otimes A_t{}^{\downarrow t \cap b}) \otimes (\otimes\{A_h \mid h \in \mathbf{H} - \{t, b\}\}) \tag{1}$$

where b is any branch for t. Thus the marginal $A^{\downarrow \chi'}$ factors on the hypergraph $\mathbf{H} - \{t\}$. The valuation on b is combined with $A_t{}^{\downarrow t \cap b}$, and the valuations on the other elements of $\mathbf{H} - \{t\}$ are unchanged.

Proposition 2 follows directly from axiom A3 by letting $G = \otimes\{A_h \mid h \in \mathbf{H} - \{t\}\}$ and $H = A_t$.

This result is especially interesting in the case of hypertrees, because in this case repeated application of (1) allows us to obtain A's marginal on any particular hyperedge of $\mathbf{H}$. If we want the marginal on a hyperedge h_1, we choose a construction sequence beginning with h_1, say $h_1 h_2 \ldots h_n$. Suppose χ_k denotes $h_1 \cup \ldots \cup h_k$ and $\mathbf{H}_k$ denotes $\{h_1, h_2, \ldots, h_k\}$ for $k = 1, \ldots, n-1$. We use (1) to delete the twig h_n, so that we have a factorization of $A^{\downarrow \chi_{n-1}}$ on the hypertree $\mathbf{H}_{n-1}$. Then we use (1) again to delete the twig h_{n-1}, so that we have a factorization of $A^{\downarrow \chi_{n-2}}$ on the hypertree $\mathbf{H}_{n-2}$. And so on, until we have deleted all the hyperedges except h_1, so that we have a factorization of $A^{\downarrow \chi_1}$ on the hypertree $\mathbf{H}_1$– i.e., we have the marginal $A^{\downarrow h_1}$. At each step, the computation is local, in the sense that it involves only a twig and a branch. Note that such a step-wise computation of the marginal of A for h_1 is allowed by axiom A2.

3.3 Computing Marginals in Markov Trees

As we learned in Sect. 2, the choice of a branching for a hypertree determines a Markov tree for the hypertree. We now look at our scheme for computing a marginal from the viewpoint of this Markov tree. This change in viewpoint does not necessarily affect the implementation of the computation, but it gives us a richer understanding. It gives us a picture in which message passing, instead of deletion, is the dominant metaphor, and in which we have great flexibility in how the message passing is controlled.

Why did we talk about deleting the hyperedge h_k as we marginalized h_k's valuation to the intersection with its branch $\beta(h_k)$? The point was simply to remove h_k from our attention. The "deletion" had no computational significance, but it helped make clear that h_k and the valuation on it were of no further use. What was of further use was the smaller hypertree that would remain were h_k deleted.

When we turn from the hypertree to the Markov tree, deletion of twigs translates into deletion of leaves. But a tree is easier to visualize than a hypertree. We can remove a leaf or a whole branch of a tree from our attention without leaning so heavily on metaphorical deletion. And a Markov tree also allows another, more useful, metaphor. We can imagine that each vertex of the tree is a processor, and we can imagine that the marginal is a message that one processor passes to another. Within this metaphor, vertices no longer relevant are kept out of our way by the rules guiding the message passing, not by deletion.

We cover a number of topics in this section. We begin by reviewing our marginalization scheme in the hypertree setting and seeing how its details translate into the Markov tree setting. We formulate precise descriptions of the operations that are carried out by each vertex and precise definitions of the messages that are passed from one vertex to another. Then we turn to questions of timing - whether a vertex uses a message as soon as it is received or waits for all its messages before it acts, how the order in which the vertices act are constrained, and whether the vertices act in serial or in parallel. We explain how the Markov tree can be expanded into an architecture for the parallel computation, with provision for storing messages as well as directing them. We explain how this architecture handles updating when inputs are changed. And finally, we explain how our computation can be directed by a simple forward-chaining production system.

Translating to the Markov Tree. We now translate our marginalization scheme from the hypertree to the Markov tree.

Recall the details in the hypertree setting. We have a valuation A on $\boldsymbol{\chi}$, in the form of a factorization on a hypertree $\mathbf{H}$. We want the marginal for the hyperedge h_1. We choose a hypertree construction sequence with h_1 as its root, say $h_1h_2 \ldots h_n$, and we choose a branching β for $h_1h_2 \ldots h_n$. On each hyperedge h_i, we have a valuation A_{h_i}. We repeatedly apply the following operation:

Operation H. Marginalize the valuation now on h_k to $\beta(h_k)$. Change the valuation now on $\beta(h_k)$ by combining it by this marginal.

We apply Operation H first for $k = n$, then for $k = n - 1$, and so on, down to $k = 2$. The valuation assigned to h_1 at the end of this process is the marginal on h_1.

We want now to redescribe Operation H, and the process of its repeated application, in terms of the actions of processors located at the vertices of the Markov tree $(\mathbf{H}, \mathbf{E})$ determined by the branching β.

The vertices of $(\mathbf{H}, \mathbf{E})$ are the hyperedges $h_1, h_2, \ldots, h_n$. We imagine that a processor is attached to each of the h_i. The processor attached to h_i can store a valuation defined on h_i, can compute the marginal of this valuation to h_j, where h_j is a neighboring vertex, can send the marginal to h_j as a message, can accept a valuation on h_i as a message from a neighbor, and can change the valuation it has stored by combining it by such an incoming message.

The edges of $(\mathbf{H}, \mathbf{E})$ are $\{h_n, \beta(h_n)\}, \{h_{n-1}, \beta(h_{n-1})\}, \ldots, \{h_3, \beta(h_3)\}$, $\{h_2, h_1\}$. When we move from h_n to $\beta(h_n)$, then from h_{n-1} to $\beta(h_{n-1})$, and so on, we are moving inwards in this Markov tree, from the outer leaves to the root h_1. The repeated application of Operation H by the processors located at the vertices follows this path.

In order to recast Operation H in terms of these processors, we need some more notation. Let Cur_h denote the valuation currently stored by the processor at vertex h of $(\mathbf{H}, \mathbf{E})$. In terms of the local processors and the Cur_h, Operation H becomes the following:

Operation M$_1$. Vertex h computes ${\mathrm{Cur}_h}^{\downarrow h \cap \beta(h)}$, the marginal of Cur_h to $\beta(h)$. It sends ${\mathrm{Cur}_h}^{\downarrow h \cap \beta(h)}$ as a message to vertex $\beta(h)$. Vertex $\beta(h)$ accepts the message ${\mathrm{Cur}_h}^{\downarrow h \cap \beta(h)}$ and changes $\mathrm{Cur}_{\beta(h)}$ by multiplying it by ${\mathrm{Cur}_h}^{\downarrow h \cap \beta(h)}$.

At the outset, $\mathrm{Cur}_h = A_h$ for every vertex h. Operation M$_1$ is executed first for $h = h_n$, then for $h = h_{n-1}$, and so on, down to $h = h_2$. At the end of this propagation process, the valuation Cur_{h_1}, the valuation stored at h_1, is the marginal of A on h_1.

An Alternative Operation. Operation M$_1$ prescribes actions by two processors, h and $\beta(h)$. We now give an alternative, Operation M$_2$, which is executed by a single processor. Since it is executed by a single processor, Operation M$_2$ will be easier for us to think about when we discuss alternative control regimes for the process of propagation.

Operation M$_2$ differs from Operation M$_1$ only in that it requires a processor to combine the messages it receives all at once, rather than incorporating them into the combination one by one as they arrive. Each time the Operation M$_1$ is executed for an h such that $\beta(h) = g$, the processor g must change the valuation it stores by combining it by the incoming message. But if processor g can store all its incoming messages, then it can delay the combination until it is its turn to marginalize. If we take this approach, then we can replace Operation M$_1$ with the following:

Operation $\mathbf{M}_{2a}$. Vertex h combines the valuation A_h with all the messages it has received, and it calls the result Cur_h. Then it computes ${\mathrm{Cur}_h}^{\downarrow h\cap\beta(h)}$, the marginal of Cur_h to $h \cap \beta(h)$. It sends ${\mathrm{Cur}_h}^{\downarrow h\cap\beta(h)}$ as a message to $\beta(h)$.

Operation M_{2a} involves action by only one processor, the processor h. When Operation M_{2a} is executed by h_n, there is no combination, because h_n, being a leaf in the Markov tree, has received no messages. The same is true for the other leaves in the Markov tree. But for vertices that are not leaves in the Markov tree, the operation will involve both combination and marginalization.

After Operation M_{2a} has been executed by h_n, h_{n-1}, and so on down to h_2, the root h_1 will have received a number of messages but will not yet have acted. To complete the process, h_1 must combine all its messages and its original valuation A_{h_1}, thus obtaining the marginal $A^{\downarrow h_1}$. We may call this Operation M_{2b}:

Operation $\mathbf{M}_{2b}$. Vertex h combines the valuation A_h with all the messages it has received, and it reports the result to the user of the system.

So Operation M_2 actually consists of two operations. Operation M_{2a} is executed successively by h_n, h_{n-1}, and so on down to h_2. Then Operation M_{2b} is executed by h_1.

Operation M_2 simplifies our thinking about control, or the flow of computation, because it allows us to think of control as moving with the computation in the Markov tree. In our marginalization scheme, control moves from one vertex to another, from the outer leaves inward towards the root. If we use Operation M_2, then a vertex is computing only when it has control.

Formulas for the Messages. We have described verbally how each vertex computes the message it sends to its branch. Now we will translate this verbal description into a formula that constitutes a recursive definition of the messages. The formula will not make much immediate contribution to our understanding, but it will serve as a useful reference in the next section, when we discuss how to extend our scheme for computing a single marginal to a scheme for computing all marginals.

Let $M^{h\to\beta(h)}$ denote the message sent by vertex h to its bud. Our description of Operation M_{2a} tells us that

$$M^{h\to\beta(h)} = {Cur_h}^{\downarrow h\cap\beta(h)}$$

where

$$Cur_h = A_h \otimes (\otimes\{M^{g\to\beta(g)} \mid g \in \mathbf{H} \;\&\; \beta(g) = h\})$$

Putting these two formulas together, we have

$$M^{h\to\beta(h)} = (A_h \otimes (\otimes\{M^{g\to\beta(g)} \mid g \in \mathbf{H} \;\&\; \beta(g) = h\}))^{\downarrow h\cap\beta(h)} \qquad (2)$$

If h is a leaf, then there is no $g \in \mathbf{H}$ such that $h = \beta(g)$, and so (2) reduces to

$$M^{h\rightarrow\beta(h)} = A_h^{\downarrow h\cap\beta(h)} \tag{3}$$

Formula (2) constitutes a recursive definition of $M^{h\rightarrow\beta(h)}$ for all h, excepting only the root h_1 of the budding β. The special case (3) defines $M^{h\rightarrow\beta(h)}$ for the leaves; a further application of (2) defines $M^{h\rightarrow\beta(h)}$ for vertices one step in towards the root from the leaves; a third application defines $M^{h\rightarrow\beta(h)}$ for vertices two steps in towards the root from the leaves; and so on.

We can also represent Operation M_{2b} by a formula:

$$A^{\downarrow h} = A_h \otimes (\otimes\{M^{g\rightarrow\beta(g)} \mid g \in \mathbf{H} \text{ and } \beta(g) = h\}) \tag{4}$$

Storing the Messages. If we want to think in terms of Operation M_2, then we must imagine that our processors have a way to store incoming messages.

Figure 4 depicts an architecture that provides for such storage. The figure shows a storage register at vertex g for each of g's neighbors. The registers for neighbors on the side of g away from the goal vertex are used to store incoming messages. The register for the neighbor in the direction of the goal vertex is used to store the vertex's outgoing message. The registers serve as communication links between neighbors; the outgoing register for one vertex being the incoming register for its neighbor in the direction of the goal vertex.

The message $M^{g\rightarrow\beta(g)}$, which vertex g stores in the register linking g to its bud, is a valuation on $g\cap\beta(g)$. It is the marginal for the bud of a valuation on g.

Flexibility of Control. Whether we use operation M_1 or M_2, it is not necessary to follow exactly the order h_n, h_{n-1}, and so on. The final result will be the same provided only that a processor never send a message until after it has received and absorbed all the messages it is supposed to receive.

This point is obvious when we look at a picture of the Markov tree. Consider, for example, a Markov tree with 15 vertices, as in Fig. 5. The vertices are numbered from 1 to 15 in this picture, indicating a construction sequence $h_1h_2\ldots h_{15}$. Since we want to find the marginal for vertex 1, all our messages

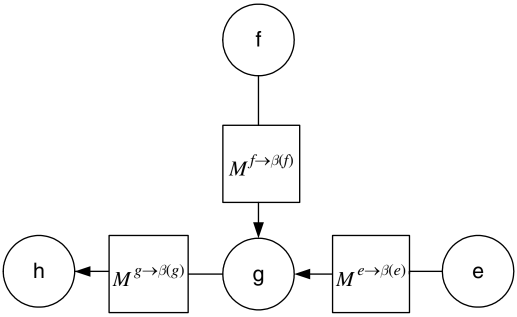

f
$M^{f\rightarrow\beta(f)}$
h
$M^{g\rightarrow\beta(g)}$
g
$M^{e\rightarrow\beta(e)}$
e

Fig. 4. A typical vertex processor g, with incoming messages from vertices f and e and outgoing message to h; here $g = \beta(f) = \beta(e)$ and $h = \beta(g)$

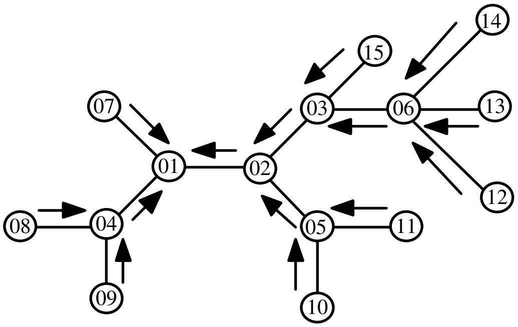


Fig. 5. A tree with 15 vertices

will be sent towards vertex 1, in the directions indicated by the arrows. Our scheme calls for a message from vertex 15 to vertex 3, then a message from vertex 14 to vertex 6, and so on. But we could just as well begin with messages from 10 and 11 to 5, follow with a message from 5 to 2, then messages from 12, 13, and 14 to 6, from 6 and 15 to 3, and so on.

Returning to the metaphor of deletion, where each vertex is deleted when it sends its message, we can say that the only constraint on the order in which the vertices act is that each vertex must be a leaf when it acts; all the vertices that used it as a branch must have sent their messages to it and then been deleted, leaving it a leaf.

The different orders of marginalization that obey this constraint correspond, of course, to the different tree construction sequences for $(\mathbf{H}, \mathbf{E})$ that use the branching β.

So far, we have been thinking about different sequences in which the vertices might act. This is most appropriate if we are really implementing the scheme on a serial computer. But if the different vertices really did have independent processors that could operate in parallel, then some of the vertices could act simultaneously. Figure 6 illustrates one way this might go for the Markov tree of Fig. 2. In step 1, all the leaf processors project to their branches. In step 2, vertices $4, 5$, and 6 (which would be leaves were the original leaves deleted) project. And so on.

If the different processors take different amounts of time to perform Operation M_2 on their inputs, then the lock-step timing of Fig. 6 may not provide the quickest way to find the marginal for h_1. It may be quicker to allow a processor to act as soon as it receives messages from its leaves, whether or not all the other processors that started along with these leaves have finished.

In general, the only constraint, in the parallel as in the serial case, is that action move inwards towards the root or goal, vertex h_1. Each vertex must receive and absorb all its messages from vertices farther away from h_1 before sending its own message on towards h_1. (In terms of Fig. 4, each processor must wait until all its incoming registers are filled before it can

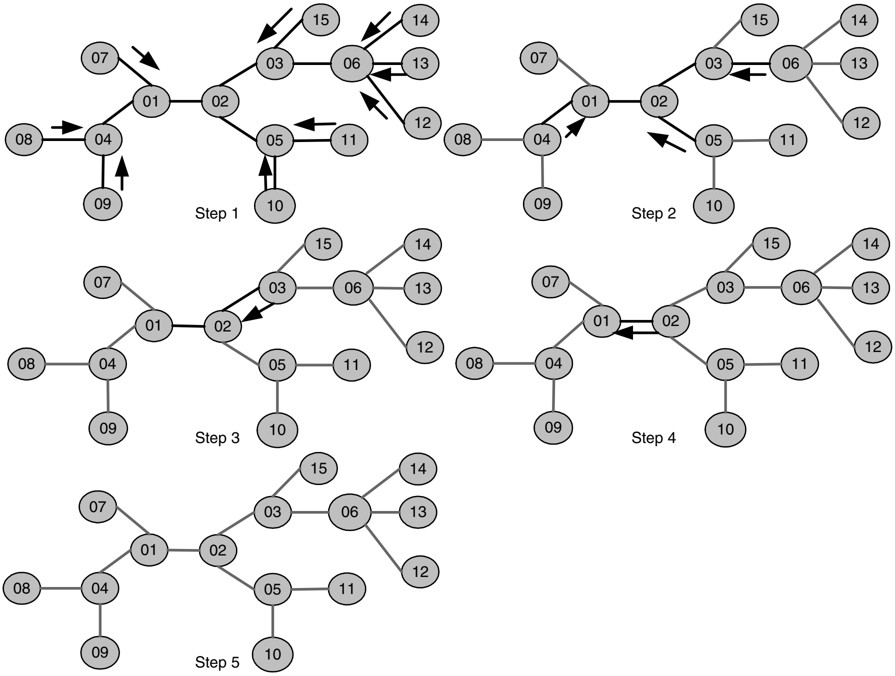


Fig. 6. An example of the message-passing scheme for computation of the marginal of vertex 1

compute a message to put in its outgoing register.) If we want to get the job done as quickly as possible, we will demand that each processor go to work as quickly as possible subject to this constraint. But the job will get done eventually provided only that all the processors act eventually. It will get done, for example, if each processor checks on its inputs periodically or at random times and acts if it has those inputs [20].

If we tell each processor who its neighbors are and which one of these neighbors lies on the path towards the goal, then no further global control or synchronization is needed. Each processor knows that it should send its outgoing message as soon as it can after receiving all its incoming messages. The leaf processors, which have no incoming messages, can act immediately. The others must wait their turn.

Updating Messages. Suppose we have completed the computation of $A^{\downarrow h_1}$, the marginal for our goal vertex. And suppose we now find reason to change A by changing one or more of our inputs, the A_h. If we have implemented the architecture just described, with storage registers between each of the vertices, then we may be able to update the marginal $A^{\downarrow h_1}$ without discarding all the work we have already done. If we leave some of the inputs unchanged, then some of the computations may not need to be repeated.

Unnecessary computation can be avoided without global control. We simply need a way of marking valuations, to indicate that they have received any needed updating. Suppose the processor at each vertex h can recognize the mark on any of its inputs (on A_h, our direct input, or on any message $M^{g \to \beta(g)}$ from a vertex g that has h as its bud), and can write the mark on its own output, the message $M^{h \to \beta(h)}$. When we wish to update the computation of $A^{\downarrow h_1}$, we put in the new values for those A_h we wish to change, and we mark all the A_h, both the ones we have changed, and the others, which we do not want to change. Then we run the system as before, except that a processor, instead of waiting for its incoming registers to be full before it acts, waits until all its inputs are marked. The processor can recognize when an input is marked without being changed, and in this case it simply marks its output instead of recomputing it.

Of course, updating can also be achieved with much less control. As Pearl [20] has emphasized, hardly any control at all is needed if we are indifferent to the possibility of wasted effort. If we do not care whether a processor repeats the same computations, we can forget about marking valuations and simply allow each processor to recompute its output from its inputs periodically or at random times. Under these circumstances, any change in one of the A_g will eventually be propagated through the system to change $A^{\downarrow h_1}$.

A Simple Production System. In reality, we will never have a parallel computer organized precisely to fit our problem. Our story about passing messages between independent processors should be thought of as metaphor, not as a guide to implementation. Implementations can take advantage, however, of the modularity the metaphor reveals.

One way to take advantage of this modularity, even on a serial computer, is to implement the computational scheme in a simple forward-chaining production system. A forward-chaining production system consists of a working memory and a rule-base, a set of rules for changing the contents of the memory. (See Brownston et al. [4] or Davis and King [7].)

A very simple production system is adequate for our problem. We need a working memory that initially contains A_h for each vertex h of $(\mathbf{H}, \mathbf{E})$, and a rule-base consisting of just two rules, corresponding to Operations M_{2a} and M_{2b}.

> **Rule 1**: If A_h is in working memory and $M^{g \to \beta(g)}$ is in working memory for every g such that $\beta(g) = h$, then use (3) to compute $M^{h \to \beta(h)}$, and place it in working memory.
>
> **Rule 2**: If A_{h_1} is in working memory and $M^{g \to \beta(g)}$ is in working memory for every g such that $\beta(g) = h_1$, then use (4) to compute $A^{\downarrow h_1}$, and print the result.

Initially, there will be no $M^{g \to \beta(g)}$ at all in working memory, so Rule 1 can fire only for h such that there is no g with $\beta(g) = h-$ i.e., only for h that are leaves. But eventually Rule 1 will fire for every vertex except the root h_1.

Then Rule 2 will fire, completing the computation. Altogether, there will be n firings, one for each vertex in the Markov tree.

Production systems are usually implemented so that a rule will fire only once for a given instantiation of its antecedent; this is called refraction [4]. If our simple production system is implemented with refraction, there will be no unnecessary firings of rules; only the n firings that are needed will occur. Even without refraction, however, the computation will eventually be completed.

Since refraction allows a rule to fire again for a given instantiation when the inputs for that instantiation are changed, this simple production system will also handle updating efficiently, performing only those recomputations that are necessary.

3.4 Simultaneous Propagation in Markov Trees

In the preceding section, we were concerned with the computation of the marginal on a single vertex of the Markov tree. In this section, we will be concerned with how to compute the marginals on all vertices simultaneously. As we will see, this can be done efficiently with only slight changes in architecture or rules.

Computing all the Marginals. If we can compute the marginal of A on one hyperedge in $\mathbf{H}$, then we can compute the marginals on all the hyperedges in $\mathbf{H}$. We simply compute them one after the other. It is obvious, however, that this will involve much duplication of effort. How can we avoid the duplication?

The first point to notice in answering this question is that we only need one Markov tree. Though there may be many Markov tree representatives for $\mathbf{H}$, any one of them can serve for the computation of all the marginals. Once we have chosen a Markov tree representative $(\mathbf{H}, \mathbf{E})$, then no matter which element h of $\mathbf{H}$ interests us, we can choose a tree construction sequence for $(\mathbf{H}, \mathbf{E})$ that begins with h, and since this sequence is also a hypertree construction sequence for $\mathbf{H}$, we can apply the method of Sect. 3.4 to it to compute $A^{\downarrow h}$.

The second point to notice is that the message passed from one vertex to another, say from f to g, will be the same no matter what marginal we are computing. If β is the budding that we use to compute $A^{\downarrow h}$, the marginal on h, and β' is the budding we use to compute $A^{\downarrow h'}$, and if $\beta(f) = \beta'(f) = g$, then the message $M^{f \to \beta(f)}$ that we send from f to g when computing $A^{\downarrow h}$ is the same as the message $M^{f \to \beta'(f)}$ that we send from f to g when computing $A^{\downarrow h'}$. Since the value of $M^{f \to \beta(f)}$ does not depend on the budding β, we may write $M^{f \to g}$ instead of $M^{f \to \beta(f)}$ when $\beta(f) = g$.

If we compute marginals for all the vertices, then we will eventually compute both $M^{f \to g}$ and $M^{g \to f}$ for every edge $\{f, g\}$. We will compute $M^{f \to g}$ when we compute the marginal on g or on any other vertex on the g side of the edge, and we will compute $M^{g \to f}$ when we compute the marginal on g or on any other vertex on the g side of the edge.

We can easily generalize the recursive definition of $M^{g\to\beta(g)}$ that we gave in Sect. 3.5 to a recursive definition of $M^{g\to h}$ for all neighbors g and h. To do so, we merely restate (2) in a way that replaces references to the budding β by references to neighbors and the direction of the message. We obtain

$$M^{g\to h} = (A_g \otimes (\otimes\{M^{f\to g} \mid f \in (\eta_g - \{h\})\}))^{\downarrow g\cap h} \tag{5}$$

where η_g is the set of all g's neighbors in $(\mathbf{H}, \mathbf{E})$. If g is a leaf vertex, then (5) reduces to $M^{g\to h} = A_g{}^{\downarrow g\cap h}$.

After we carry out the recursion to compute $M^{g\to h}$ for all pairs of neighbors g and h, we can compute the marginal of A on each h by

$$A^{\downarrow h} = A_h \otimes (\otimes\{M^{g\to h} \mid g \in \eta_h\}) \tag{6}$$

The General Architecture. A slight modification of the architecture shown in Fig. 4 will allow us to implement the simultaneous computation of the marginals on all the hyperedges. We simply put two storage registers between every pair of neighbors f and g, as in Fig. 7. One register stores the message from f to g; the other stores the message from g to f.

Figure 8 shows a more elaborate architecture for the simultaneous computation. In addition to the storage registers that communicate between vertices, this figure shows registers where the original valuations, the A_h, are put into the system and the marginals, the $A^{\downarrow h}$, are read out.

In the architecture of Fig. 4, computation is controlled by the simple requirement that a vertex g must have messages in all its incoming registers before it can compute a message to place in its outgoing register. In the architecture of Fig. 5, computation is controlled by the requirement that a vertex g must have messages in all its incoming registers except the one from h before it can compute a message to send to h.

This basic requirement leaves room for a variety of control regimes. Most of the comments we made about the flexibility of control for Fig. 4 carry over to Fig. 8.

In particular, updating can be handled efficiently if a method is provided for marking updated inputs and messages. If we change just one of the input, then efficient updating will save about half the work involved in simply reperforming the entire computation. To see that this is so, consider the effect of changing the input A_h in Fig. 8. This will change the message $M^{g\to f}$, but

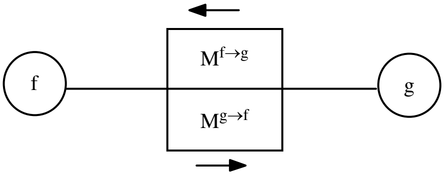


Fig. 7. The two storage registers between f and g

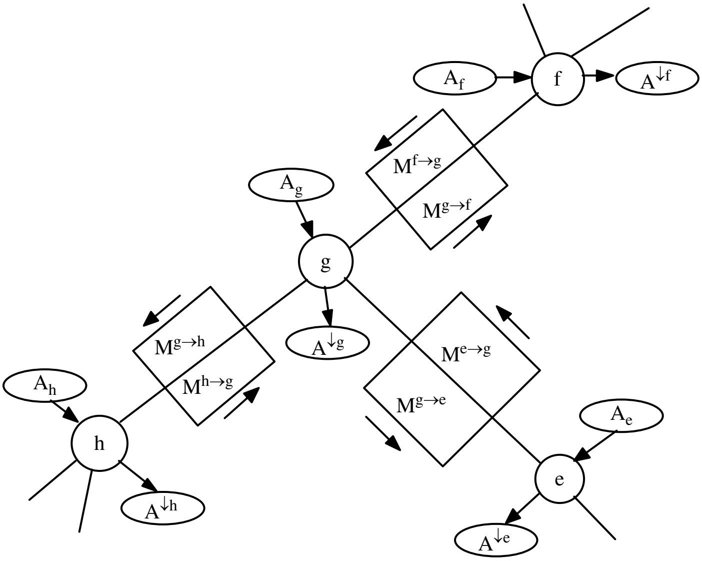


Fig. 8. Several vetices, with storage registers for communication between themselves and with user

not the message $M^{f \to g}$. The same will be true for every edge; one of the two messages will have to be recomputed, but not the other.

It may be enlightening to look at how the lock-step control we illustrated with Fig. 6 might generalize to simultaneous computation of the marginals for all vertices. Consider a lock-step regime where at each step, each vertex looks and sees what messages it has the information to compute, computes these messages, and sends them. After all the vertices working are done, they look again, see what other messages they now have the information to compute, compute these messages, and send them. And so on. Fig. 9 gives an example. At the first step, the only messages that can be computed are the messages from the leaves to their branches. At the second step, the computation moves inward. Finally, at step 3, it reaches vertex 2, which then has the information needed to compute its own marginal and messages for all its neighbors. Then the messages move back out towards the leaves, with each vertex along the way being able to compute its own marginal and messages for all its other neighbors as soon as it receives the message from its neighbor nearest vertex 2.

In the first phase, the inward phase, a vertex sends a message to only one of its neighbors, the neighbor towards the center. In the second phase, the outward phase, a vertex sends $k - 1$ messages, where k is the number of its neighbors. Yet the number of messages sent in the two phases is roughly the same, because the leaf vertices participate in the first phase and not in the second.

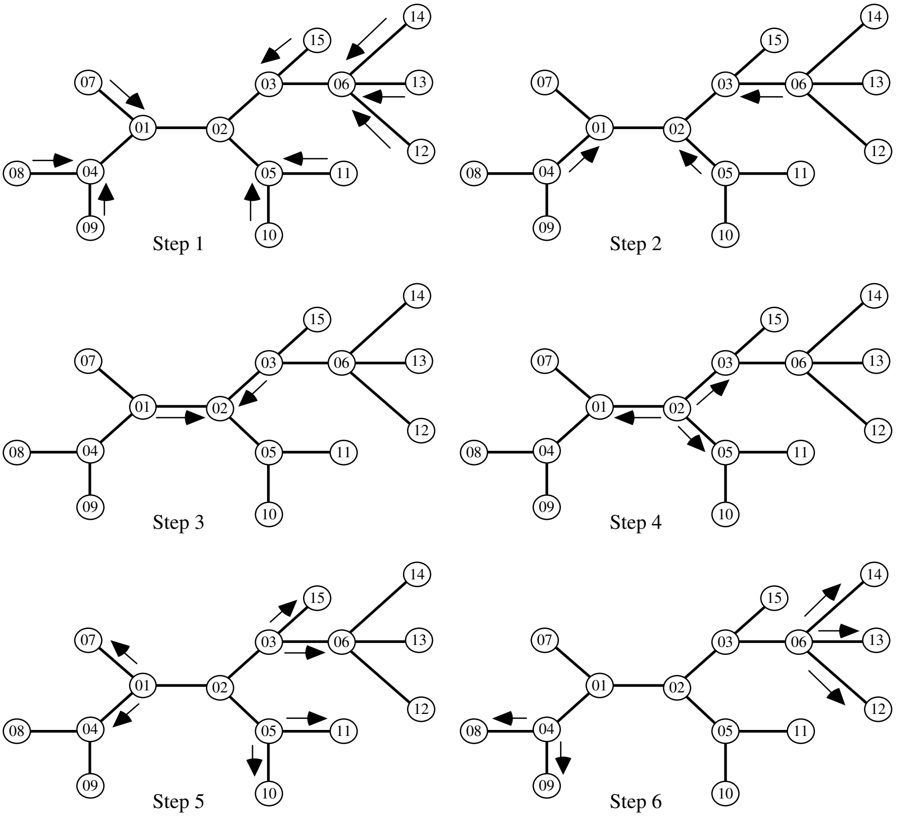


Fig. 9. An example of the message-passing scheme for simultaneous computation of all marginals

There are seven vertices in the longest path in the tree of Fig. 9. Whenever the number of vertices in the longest path is odd, the lock-step control regime will result in computation proceeding inwards to a central vertex and then proceeding back outwards to the leaves. Whenever this number is even, there will instead be two central vertices that send each other messages simultaneously, after which they both send messages back outwards towards the leaves.

If we really do have independent processors for each vertex, then we do not have to wait for all the computations that start together to finish before taking advantage of the ones that are finished to start new ones. We can allow a new computation to start whenever a processor is free and it has the information needed. On the other hand, we need not require that the work be done so promptly. We can assume that processors look for work to do only at random times. But no matter how we handle these issues, the computation will converge to some particular vertex or pair of neighboring vertices and then move back out from that vertex or pair of vertices.

There is exactly twice as much message passing in our scheme for simultaneous computation as there was in our scheme for computing a single marginal. Here every pair of neighbors exchange messages; there only one message was sent between every pair of neighbors. Notice also that we can make the computation of any given marginal the beginning of the simultaneous computation. We can single out any hyperedge h (even a leaf), and forbid it to send a message to any neighbor until it has received messages from all its neighbors. If we then let the system of Fig. 9 run, it will behave just like the system of Fig. 6 with h as the root, until h has received messages from all its neighbors. At that point, h can compute its marginal and can also send messages to all its neighbors; the second half of the message passing then proceeds, with messages moving back in the other direction.

The Corresponding Production System. Implementing simultaneous computation in a production system requires only slight changes in our two rules. The following will work:

> **Rule 1′**: If A_g is in working memory, and $M^{f \to g}$ is in working memory for every f in $\eta_g - \{h\}$, then use (5) to compute $M^{g \to h}$, and place it in working memory.
>
> **Rule 2′**: If A_h is in working memory, and $M^{g \to h}$ is in working memory for every g in η_h, then use (6) to compute $A^{\downarrow h}$, and print the result.

Initially, there will be no $M^{f \to g}$ at all in working memory, so Rule 1′ can fire only for g and h such that $\eta_g - \{h\}$ is empty - i.e., only when g is a leaf and h is its bud. But eventually Rule 1′ will fire in both directions for every edge $\{g, h\}$. Once Rule 1′ has fired for all the neighbors g of h, in the direction of h, Rule 2′ will fire for h. Altogether, there will be $3n - 2$ firings, two firings of Rule 1′ for each of the $n - 1$ edges, and one firing of Rule 2′ for each of the n vertices.

As the count of firings indicates, our scheme for simultaneous computation finds marginals for all the vertices with roughly the same effort that would be required to find marginals for three vertices if this were done by running the scheme of Sect. 3.5 three times.

4 Probability Propagation

In this section, we explain local computation for probability distributions. More precisely, we show how the problem of computing marginals of joint probability distributions fits the general framework described in the previous section.

For probability propagation, proper valuations will correspond to potentials.

Potentials. We use the symbol W_X for the set of possible values of a variable X, and we call W_X the *frame for* X. We will be concerned with a finite set $\boldsymbol{\chi}$ of variables, and we will assume that all the variables in $\boldsymbol{\chi}$ have finite frames. For each $h \subseteq \boldsymbol{\chi}$, we let W_h denote the Cartesian product of W_X for X in h; $W_h = \times\{W_X \mid X \in h\}$. We call W_h the *frame for* h. We will refer to elements of W_h as *configurations of* h. A *potential on* h is a real-valued function on W_h that has non-negative values that are not all zero. Intuitively, potentials are unnormalized probability distributions.

Projection of configurations. In order to develop a notation for the combination of potentials, we first need a notation for the projection of configurations of a set of variables to a smaller set of variables. Here projection simply means dropping extra coordinates; if (w, x, y, z) is a configuration of $\{W, X, Y, Z\}$, for example, then the projection of (w, x, y, z) to $\{W, X\}$ is simply (w, x), which is a configuration of $\{W, X\}$. If g and h are sets of variables, $h \subseteq g$, and $\mathbf{x}$ is a configuration of g, then we will let $\mathbf{x}^{\downarrow h}$ denote the projection of $\mathbf{x}$ to h.

Combination. For potentials, combination is simply pointwise multiplication. If G is a potential on g, H is a potential on h, and there exists an $\mathbf{x} \in W_{g \cup h}$ such that

$$G(\mathbf{x}^{\downarrow g})H(\mathbf{x}^{\downarrow h}) > 0 \tag{7}$$

then their *combination*, denoted simply by GH, is the potential on $g \cup h$ given by

$$(GH)(\mathbf{x}) = G(\mathbf{x}^{\downarrow g})H(\mathbf{x}^{\downarrow h}) \tag{8}$$

for all $\mathbf{x} \in W_{g \cup h}$. If there exists no $\mathbf{x} \in W_{g \cup h}$ such that $G(\mathbf{x}^{\downarrow g})H(\mathbf{x}^{\downarrow h}) > 0$, then we say that G and H are *not combinable.*

Intuitively, if the bodies of evidence on which G and H are based are independent, then $G \oplus H$ is supposed to represent the result of pooling these two bodies of evidence. Note that condition (7) ensures that GH defined in (8) is a potential. If condition (7) does not hold, this means that the two bodies of evidence corresponding to G and H contradict each other completely and it is not possible to combine such evidence.

It is clear from the definition of combination of potentials that it is commutative and associative (axiom A1).

Marginalization. Marginalization is familiar in probability theory; it means reducing a function on one set of variables to a function on a smaller set of variables by summing over the variables omitted.

Suppose g and h are sets of variables, $h \subseteq g$, and G is a potential on g. The *marginal of* G *for* h, denoted by $G^{\downarrow h}$, is the potential on h defined as follows: $\forall \mathbf{x} \in W_h$,

$$G^{\downarrow h}(\mathbf{x}) = \begin{cases} \Sigma\{G(\mathbf{x}, \mathbf{y}) \mid \mathbf{y} \in W_{g-h}\} & h \text{ is a proper subset of } g \\ G(\mathbf{x}) & h = g \end{cases}$$

It is obvious from the above definition that marginalization operation for potentials satisfies axiom A2.

Since multiplication distributes over addition, it is easy to show that combination and marginalization for potentials satisfy axiom A3. Thus all axioms are satisfied making local computation possible.

A number of authors who have studied local computation for probability, including Kelly and Barclay [14], Cannings, Thompson and Skolnick [5], Pearl [20], Shenoy and Shafer [30], and Lauritzen and Spiegelhalter [17], have described schemes that are variations on the basic scheme described in Sect. 2. Most of these authors, however, have justified their schemes by emphasizing conditional probability. We believe this emphasis is misplaced. What is essential to local computation is a factorization. It is not essential that this factorization be interpreted, at any stage, in terms of conditional probabilities. For more regarding this point, see Shafer and Shenoy [25].

We would like to make two important observations for the case of probability propagation. First note that it is sufficient, in order for a potential A to factor on $\mathbf{H}$, that A be proportional to a product of arrays on the hyperedges. Indeed, if

$$A \propto \Pi\{A_h \mid h \in \mathbf{H}\}$$

where A_h is a potential on h, then a representation of the form $A = \Pi\{A_h \mid h \in \mathbf{H}\}$ can be obtained simply by incorporating the constant of proportionality into one of the A_h. In practice, we will postpone finding the constant of proportionality until we have marginalized A to a hyperedge using the scheme described in Sect. 2.

The second observation relates to conditioning joint probability distributions. Suppose a probability distribution P represents our assessment of a given body of information, and we have been computing marginals of P from the factorization

$$P = \Pi\{A_h \mid h \in \mathbf{H}\} \tag{9}$$

where $\mathbf{H}$ is a hypertree on $\boldsymbol{\chi}$. Suppose we now observe the values of some of the variables in $\boldsymbol{\chi}$; say we observe $Y_1 = y_1$, $Y_2 = y_2$, and so on up to $Y_n = y_n$. We change our assessment from P to $P^{|f=y}$ where $f = \{Y_1, \ldots, Y_n\}$, $y = \{y_1, \ldots, y_n\}$, and $P^{|f=y}$ denotes the joint probability distribution conditioned on the observations. Can we adapt (9) to a factorization of $P^{|f=y}$? Yes, we can. More precisely, we can adapt (9) to a factorization of a potential proportional to $P^{|f=y}$, and this, as we noted in our first observation, is good enough. The adaptation is simple. It follows from the definition of conditional probability that

$$P^{|f=y} \propto B^{Y_1=y_1} \ldots B^{Y_n=y_n}\ \Pi\{A_h \mid h \in \mathbf{H}\}$$

where $B^{Y_i=y_i}$ is the *indicator potential for* $Y_i = y_i$ on $\{Y_i\}$ defined as follows: $\forall x \in W_{Y_i}$,

$$B^{Y_i=y_i}(x) = \begin{cases} 0 & x \neq y_i \\ 1 & x = y_i \end{cases} \quad .$$

We will now illustrate our propagation scheme using a simple example.

An Example. This example is adapted from Shachter and Heckerman [22]. Consider three variables D, B and G representing diabetes, blue toe, and glucose in urine, respectively. The frame for each variable has two configurations. $D = d$ will represent the proposition *diabetes is present* (in some patient) and $D = {\sim}d$ will represent the proposition *diabetes is not present*. Similarly for B and G. Let P denote the joint probability distribution for $\{D, B, G\}$. We will assume that diabetes causes blue toe and glucose in urine implying that variables B and G are conditionally independent (with respect to P) given D. Thus we can factor P as follows.

$$P = P^{D} P^{B|D} P^{G|D} \tag{10}$$

where P^D is the potential on $\{D\}$ representing the marginal of P for D, $P^{B|D}$ is the potential for $\{D, B\}$ representing the conditional distribution of B given D, and $P^{G|D}$ is the potential for $\{D, G\}$ representing the conditional distribution of G given D. For example, $P^{B|D}(d, b)$ represents the conditional probability of the proposition $B = b$ given that $D = d$. Thus P factors on the hypertree $\{\{D\}, \{D, B\}, \{D, G\}\}$. Since we would like to compute the marginals for B and G, we will enlarge the hypertree to include the hyperedges $\{B\}$ and $\{G\}$. It is easy to expand (10) so that we have a factorization of P on the enlarged hypertree - the potentials on these additional hyperedges consist of all ones. Suppose that the potentials P^D, $P^{B|D}$, and $P^{G|D}$ are as shown in Table 1.

The enlarged hypertree and a Markov tree representation are shown in Fig. 10.

Suppose we propagate the potentials using the scheme described in Sect. 2. The results are as shown in Fig. 11. For each vertex h, the input potentials are shown as I^h and the output potentials are shown as O^h. All the messages are also shown. Note that the output potentials have been normalized so that they represent marginal posterior probabilities.

Table 1. The potential tables

P^D		$P^{B\|D}$		$P^{G\|D}$	
d	.1	d,b	.014	$d,\ g$	.9
${\sim}d$	.9	$d,{\sim}b$	.986	$d,{\sim}g$	.1
		${\sim}d,b$	.006	${\sim}d,\ g$	.01
		${\sim}d,{\sim}b$	.994	${\sim}d,{\sim}g$	.99

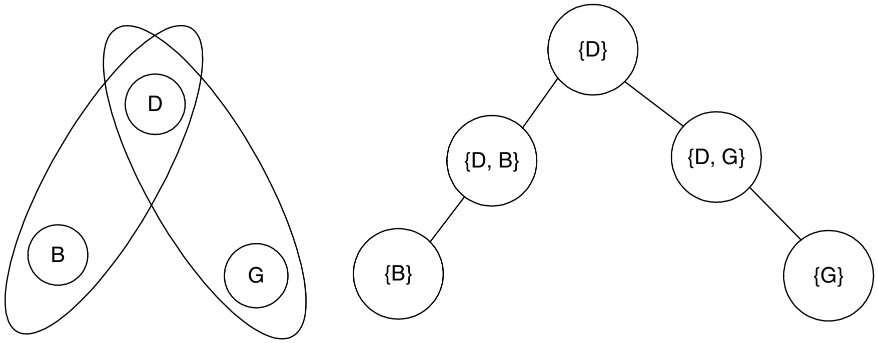


Fig. 10. The hypertree and a Markov tree representation

Now suppose we observe that the patient has blue toe. This is represented by the indicator potential for $B = b$. The other potentials are the same as before. If we propagate the potentials, the results are as shown in Fig. 12.

Note that the posterior probability of the presence of diabetes has increased (from .1 to .2059) and consequently the presence of glucose in urine has also increased (from .0990 to .1932). Now suppose that after the patient is tested for glucose in urine, the results indicate that there is an absence of glucose in urine. This information is represented by the indicator potential for $G = \sim g$. The other potentials are as before. If we propagate the potentials, the results are as shown in Fig. 13.

Note that the posterior probability of the presence of diabetes has decreased (from .2059 to .0255). This concludes our example.

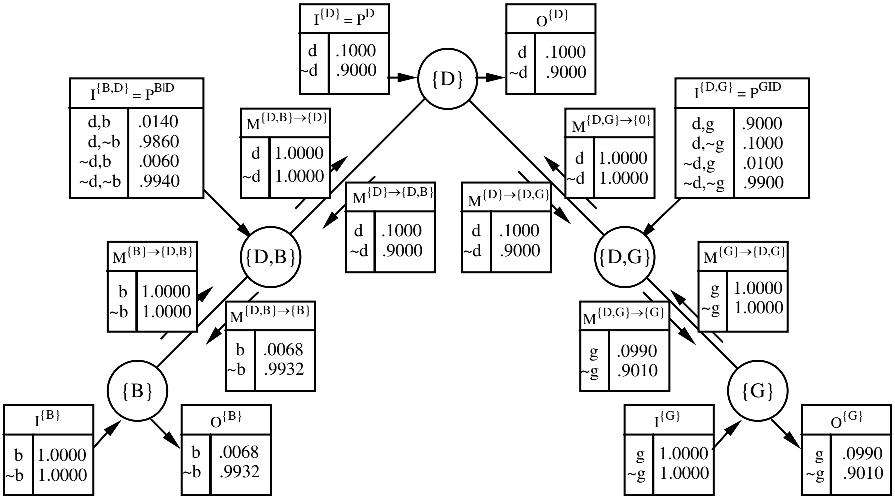


Fig. 11. The initial propagation of potentials

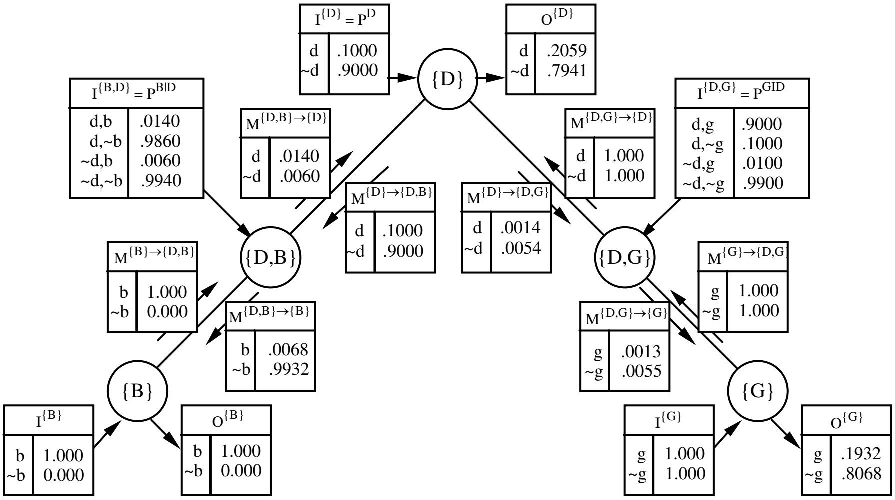


Fig. 12. The results of propagation after the presence of blue toe is observed

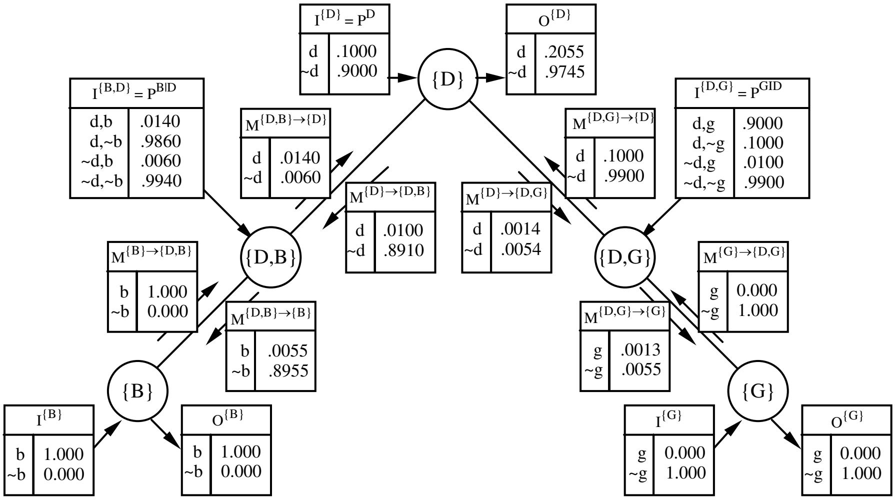


Fig. 13. The results of propagation after the observation that patient does not have glucose in urine

5 Belief-Function Propagation

In this section, we explain local computation for belief functions. More precisely, we show how the problem of computing marginals of a joint belief function fits the general framework described in Sect. 2.

For belief-function propagation, proper valuations correspond to either probability mass assignment functions, belief functions, plausibility functions or commonality functions. For simplicity of exposition, we will describe belief-function propagation in terms of superpotentials which are unnormalized basic probability assignment functions.

Basic Probability Assignment Functions. Suppose W_h is the frame for a subset h of variables. A basic probability assignment function (*bpa* function) for h is a non-negative, real-valued function m on the set of all subsets of W_h such that $m(\phi) = 0$ and $\Sigma\{m(\mathbf{a}) \mid \mathbf{a} \subseteq W_h\} = 1$. Intuitively, $m(\mathbf{a})$ represents the degree of belief assigned exactly to $\mathbf{a}$ (the proposition that the true configuration of h is in the set $\mathbf{a}$) and to nothing smaller. A *bpa* function is the belief function equivalent of a probability mass assignment function in probability theory. Whereas a probability mass function is restricted to assigning probability masses only to singleton configurations of variables, a *bpa* function is allowed to assign probability masses to sets of configurations without assigning any mass to the individual configurations contained in the sets.

Superpotentials. Suppose h is a subset of variables. A superpotential for h is a non-negative, real-valued function on the set of all subsets of W_h such that the values of non-empty subsets are not all zero. Given a superpotential H on h, we can construct a *bpa* function H' for h from H as follows: $H'(\phi) = 0$, and $H'(\mathbf{a}) = H(\mathbf{a})/\Sigma\{H(\mathbf{b}) \mid \mathbf{b} \subseteq W_h, \mathbf{b} \neq \phi\}$.Thus superpotentials can be thought of as unnormalized *bpa* functions. Superpotentials correspond to the notion of proper valuations in the general framework.

Projection and Extension of Subsets. Before we can define combination and marginalization for superpotentials, we need the concepts of projection and extension of subsets of configurations.

If g and h are sets of variables, $h \subseteq g$, and $\mathbf{g}$ is a non-empty subset of W_g, then the projection of $\mathbf{g}$ to h, denoted by $\mathbf{g}^{\downarrow h}$, is the subset of W_h given by $\mathbf{g}^{\downarrow h} = \{\mathbf{x}^{\downarrow h} \mid \mathbf{x} \in g\}$.

For example, If $\mathbf{a}$ is subset of $W_{\{W,X,Y,Z\}}$, then the marginal of $\mathbf{a}$ to $\{X, Y\}$ consists of the elements of $W_{\{X,Y\}}$ which can be obtained by projecting elements of $\mathbf{a}$ to $W_{\{X,Y\}}$.

By extension of a subset of a frame to a subset of a larger frame, we mean a cylinder set extension. If g and h are sets of variables, $h \subseteq g$, $h \neq g$, and $\mathbf{h}$ is a subset of W_h, then the extension of $\mathbf{h}$ to g is $\mathbf{h} \times W_{g-h}$. If $\mathbf{h}$ is a subset of W_h, then the extension of $\mathbf{h}$ to h is defined to be $\mathbf{h}$. We will let $\mathbf{h}^{\uparrow g}$ denote the extension of $\mathbf{h}$ to g.

For example, if $\mathbf{a}$ is a subset of $W_{\{W,X\}}$, for example, then the vacuous extension of $\mathbf{a}$ to $\{W, X, Y, Z\}$ is $\mathbf{a} \times W_{\{Y,Z\}}$.

Combination. For superpotentials, combination is called Dempster's rule [8, 9]. Consider two superpotentials G and H on g and h, respectively. If

$$\Sigma\{G(\mathbf{a})H(\mathbf{b}) \mid (\mathbf{a}^{\uparrow(g\cup h)}) \cap (\mathbf{b}^{\uparrow(g\cup h)}) \neq \phi\} \neq 0 \qquad (11)$$

then their combination, denoted by $G \oplus H$, is the superpotential on $g \cup h$ given by

$$G \oplus H(\mathbf{c}) = \Sigma\{G(\mathbf{a})H(\mathbf{b}) \mid (\mathbf{a}^{\uparrow(g\cup h)}) \cap (\mathbf{b}^{\uparrow(g\cup h)}) = \mathbf{c}\} \tag{12}$$

for all $\mathbf{c} \subseteq W_{g\cup h}$. If

$$\Sigma\{G(\mathbf{a})H(\mathbf{b}) \mid (\mathbf{a}^{\uparrow(g\cup h)}) \cap (\mathbf{b}^{\uparrow(g\cup h)}) \neq \phi\} = 0$$

then we say that G and H are not combinable.

Intuitively, if the bodies of evidence on which G and H are based are independent, then $G \oplus H$ is supposed to represent the result of pooling these two bodies of evidence. Note that (11) ensures that $G \oplus H$ defined in (12) is a superpotential. If (11) does not hold, this means that the two bodies of evidence corresponding to G and H contradict each other completely and it is not possible to combine such evidence.

It is shown in Shafer [23] that Dempster's rule of combination is commutative and associative. Thus combination for superpotentials satisfies axiom A1.

Marginalization. Like marginalization for potentials, marginalization for superpotentials corresponds to summation.

Suppose G is a superpotential for g and suppose $h \subseteq g$. Then the marginal of G for h is the superpotential $G^{\downarrow h}$ for h defined as follows: $\forall \mathbf{a} \subset W_h$,

$$G^{\downarrow h}(\mathbf{a}) = \Sigma\{G(\mathbf{b}) \mid \mathbf{b} \subseteq W_g \text{ such that } \mathbf{b}^{\downarrow h} = \mathbf{a}\}.$$

It is easy to see that marginalization for superpotentials satisfies axiom A2. In Shafer and Shenoy [25], it is shown that the above definitions of marginalization and combination for superpotentials satisfies axiom A3. Thus all axioms are satisfied making local computation possible.

Propagation of belief functions using local computation has been studied by Shafer and Logan [24], Shenoy and Shafer [30], Shenoy et al. [33], Kong [15], Dempster and Kong [10], Shafer et al. [24], Mellouli [19], and Shafer and Shenoy [25]. Shafer et al. [27], Shenoy [29], Zarley [37], Zarley et al. [38], Shenoy [28, 29], and Hsia and Shenoy [12, 13] discuss applications and implementations of these propagation schemes.

References

[1] Arnborg, S., Corneil, D. G., and Proskurowski, A. Complexity of finding embeddings in a k-tree. *SIAM Journal of Algebraic and Discrete Mathematics 8* (1987), 277–284.

[2] Berge, C. *Graphs and Hypergraphs.* translated from French by E. Minieka, North-Holland, 1973.

[3] Bertele, U., and Brioschi, F. *Nonserial Dynamic Programming.* Academic Press, New York, NY, 1972.

[4] Brownston, L. S., Farrell, R. G., Kant, E., and Martin, N. *Programming Expert Systems in OPS5: An Introduction to Rule-Based Programming.* Addison-Wesley, Reading, MA, 1985.

[5] Cannings, C., Thompson, E. A., and Skolnick, M. H. Probability functions on complex pedigrees. *Advances in Applied Probability 10* (1978).

[6] Darroch, J. N., Lauritzen, S. L., and Speed, T. P. Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics 8* (1980), 522–539.

[7] Davis, R., and King, J. J. The origin of rule-based systems in AI. In *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, B. G. Buchanan and E. H. Shortliffe, Eds. Addison-Wesley, Reading, MA, 1984, pp. 20–52.

[8] Dempster, A. P. New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics 37* (1966), 355–374.

[9] Dempster, A. P. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics 38* (1967), 325–339.

[10] Dempster, A. P. Uncertain evidence and artificial analysis. Tech. rep., Department of Statistics, Harvard University, Cambridge, MA, 1986.

[11] Golumbic, M. C. *Algorithmic Graph Theory and Perfect Graphs.* Academic Press, 1980.

[12] Hsia, Y., and Shenoy, P. P. An evidential language for expert systems. In *Methodologies for Intelligent Systems*, Z. Ras, Ed., vol. 4. North-Holland, 1989.

[13] Hsia, Y., and Shenoy, P. P. MacEvidence: A visual evidential language for knowledge-based systems. Tech. rep., School of Business, University of Kansas, Lawrence, KS, 1989.

[14] Kelly, C. W. I., and Barclay, S. A general Bayesian model for hierarchical inference. *Organizational Behavior and Human Performance 10* (1973), 388–403.

[15] Kong, A. *Multivariate Belief Functions and Graphical Models.* PhD thesis, Department of Statistics, Harvard University, Cambridge, MA, 1986.

[16] Lauritzen, S. L., Speed, T. P., and Vijayan, K. Decomposable graphs and hypergraphs. *Journal of Australian Mathematical Society 36, Series A* (1984), 12–29.

[17] Lauritzen, S. L., and Spiegelhalter, D. J. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society Series B 50* (1988), 157–224.

[18] Maier, D. *The Theory of Relational Databases.* Computer Science Press, Rockville, 1983.

[19] Mellouli, K. *On the Propagation of Beliefs in Networks using the Dempster-Shafer Theory of Evidence.* PhD thesis, School of Business, University of Kansas, Lawrence, KS, 1987.

[20] Pearl, J. Fusion, propagation and structuring in belief networks. *Artificial Intelligence 29* (1986), 241–288.

[21] Rose, D. J. Triangulated graphs and the elimination process. *Journal of Mathematical Analysis and Application 32* (1970), 597–609.

[22] Shachter, R. D., and Heckerman, D. A backwards view for assessment. *AI Magazine 8(3)* (1986), 55–61.

[23] Shafer, G. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ, 1976.

[24] Shafer, G., and Logan, R. Implementing Dempster's rule for hierarchical evidence. *Artificial Intelligence 33* (1987), 271–298.

[25] Shafer, G., and Shenoy, P. P. Local computation in hypertrees. Tech. rep., School of Business, University of Kansas, Lawrence, KS, 1988.

[26] Shafer, G., Shenoy, P. P., and Mellouli, K. Propagating belief functions in qualitative Markov trees. *International Journal of Approximate Reasoning 3* (1987), 383–411.

[27] Shafer, G., Shenoy, P. P., and Srivastava, R. P. Auditor's assistant: A knowledge engineering tool for audit decisions. In *Auditing Symposium IX: Proceedings of the 1988 Touche Ross University of Kansas Symposium on Auditing Problems* (1988), pp. 61–84.

[28] Shenoy, P. P. On Spohn's rule for revision of beliefs. Tech. rep., School of Business, University of Kansas, Lawrence, KS, 1989.

[29] Shenoy, P. P. A valuation-based language for expert systems. *International Journal for Approximate Reasoning 3*, 5 (1989), 383–411.

[30] Shenoy, P. P., and Shafer, G. Propagating belief functions using local computations. *IEEE Expert 1*, 3 (1986), 43–52.

[31] Shenoy, P. P., and Shafer, G. Axioms for discrete optimization using local computations, working paper no. 207. Tech. rep., School of Business, University of Kansas, Lawrence, KS, 1988.

[32] Shenoy, P. P., and Shafer, G. Constraint propagation. Tech. rep., School of Business, University of Kansas, Lawrence, KS, 1988.

[33] Shenoy, P. P., Shafer, G., and Mellouli, K. Propagation of belief functions: A distributed approach. *Uncertainty in Artificial Intelligence 2* (1988), 325–336.

[34] Spohn, W. A general non-prbabilistic theory of inductive reasoing. In *Uncertainty in Artificial Intellige. nce 4*, R. D. Shachter, T. S. Levitt, J. F. Lemmer, and L. N. Kanal, Eds. North-Holland, Amsterdam, 1988, pp. 149–158.

[35] Spohn, W. Ordinal conditional functions: A dynamic theory of epistemic states. In *Causation in Decision, Belief Change, and Statistics 2*, W. L. Harper and B. Skyrms, Eds. Reidel, Holland, 1988, pp. 105–134.

[36] Tarjan, R. E., and Yannakakis, M. Simple linear time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acylic hypergraphs. *SIAM Journal of Computing 13* (1984), 566–579.

[37] Zarley, D. K. An evidential reasoning system. Tech. rep., School of Business, University of Kansas, Lawrance, KS, 1988.

[38] Zarley, D. K., Hsia, Y., and Shafer, G. Evidential reasoning using delief. In *Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-88)* (Minneapolis, MN, 1988), vol. 1, pp. 205–209.

[39] Zhang, L. Studies on finding hypertree covers for hypergraphs. Tech. rep., School of Business, University of Kansas, Lawrance, KS, 1988.

21

Generalizing the Dempster–Shafer Theory to Fuzzy Sets

John Yen

Abstract. With the desire to manage imprecise and vague information in evidential reasoning, several attempts have been made to generalize the Dempster–Shafer (D–S) theory to deal with fuzzy sets. However, the important principle of the D–S theory, that the belief and plausibility functions are treated as lower and upper probabilities, is no longer preserved in these generalizations. A generalization of the D–S theory in which this principle is maintained is described. It is shown that computing the degree of belief in a hypothesis in the D–S theory can be formulated as an optimization problem. The extended belief function is thus obtained by generalizing the objective function and the constraints of the optimization problem. To combine bodies of evidence that may contain vague information, Dempster's rule is extended by 1) combining generalized compatibility relations based on the possibility theory, and 2) normalizing combination results to account for partially conflicting evidence. Our generalization not only extends the application of the D–S theory but also illustrates a way that probability theory and fuzzy set theory can be integrated in a sound manner in order to deal with different kinds of uncertain information in intelligent systems.

1 Introduction

EVIDENTIAL REASONING, which is the task of inferring the likelihood of some hypotheses by collecting and combining relevant evidence for or against these hypotheses, is central to many computer systems that help users in decisionmaking, diagnosis, pattern recognition, and speech understanding. The problem of evidential reasoning is complicated by information being conveyed

This article is partially based on the author's Ph.D. thesis at the University of California, Berkeley, which was supported by National Science Foundation Grant DCR-8513139. Part of the research described in the paper was done at USC/Information Sciences Institute and was supported by the Defense Advanced Research Projects Agency under Contract No. MDA903-86-C-0178.

by a piece of evidence is often not only uncertain, but also imprecise, incomplete, and vague. For example, a sensor's output may indicate that a flying object is about 50 miles from Los Angeles and that it belongs to a general class of missiles. But the sensor gives no further information about the specific type of the missile. Therefore, an evidential reasoning mechanism that can cope with all these different kinds of uncertainties in a sound manner is highly desirable.

Previous work on evidential reasoning has been largely based on three theoretical frameworks: the Bayesian probability theory, the Dempster–Shafer (D–S) theory of evidence, and the fuzzy set theory. These frameworks differ in their strengths and weaknesses. The Bayesian probability theory has a well-developed decision-making theory, but it requires precise probability judgments. Hence, it is weak in representing and managing imprecise information. To cope with this weakness, a Bayesian approach often needs to transform a piece of imprecise evidence into a precise one by using additional assumptions [1]. The D–S theory is based on probability theory, yet it allows probability judgments to capture the imprecise nature of the evidence. As a result, degrees of likelihood are measured by probability intervals, as opposed to point probabilities in the Bayesian approaches. One of the weaknesses of the D–S theory is that its decision theory is still a research topic [2]. The fuzzy set theory focuses on the issue of representing and managing vague information such as "the temperature is *high*" or "the missile is *about* 50 miles from Los Angeles." One of its strengths is its *possibility theory* as a foundation for dealing with imprecise data. Although the fuzzy set theory is still somewhat controversial at this point, it has been used successfully to solve many complex real-world problems. For example, Hitachi has used fuzzy control to develop an automatic train operation system for Sendai's municipal subway [3].

In this paper, we describe an approach that addresses the issue of managing imprecise and vague information in evidential reasoning by combining the D–S theory with the fuzzy set theory. Although several researchers have extended the D–S theory to deal with vague information [4]–[7], their extensions have not been able to preserve an important principle in the D–S theory: that the belief and the plausibility measures are lower and upper probabilities. Viewing this, we generalize the D–S theory in a way that preserves this principle. We achieve this by first generalizing the fundamental constructs of the theory and then deriving other extensions to the theory from these generalizations. The primitive constructs that have been generalized are 1) the compatibility relation, which relates the evidence to the hypotheses, and 2) the objective function and the constraints of the optimization problem, which compute the belief and the plausibility functions. From these generalized basic components, we derive the belief function, the plausibility function, and the rule of combination for the generalized theory of evidence. Finally, we discuss the relationship between Shafer's consonant support functions and the possibility distributions based on our generalized framework.

2 The Problem

The problem we want to solve in this paper can be described as follows. Suppose X and Y are two variables that take their possible values from two spaces, S and T, respectively. The space S is an evidence space that consists of a set of mutually exclusive and exhaustive evidential elements. The space T is a hypothesis space that is formed by a set of mutually exclusive and exhaustive hypotheses. A body of evidence for the hypothesis space T is constituted by (1) a set of rules that associate evidential elements to hypotheses in the form of

$$\text{if } X = s_i \text{ then } Y \text{ is } A_i$$

where s_i is an evidential element and A_i is a fuzzy subset of T, and (2) a probability distribution of the evidence space S. Our objective is to answer questions like "What is the likelihood that Y is B given a collection of bodies of evidence?" where B is a fuzzy subset of T.

To illustrate this, let us consider a computer system that infers the age of a person based on various information about the person. Such a system may contain two bodies of evidence, one regarding the boldness of the person, the other about whether he/she likes punk rock. The rules for these two bodies of evidence are listed below.

IF the person is bold, THEN his age is *NOT YOUNG.*
IF the person is not bold, THEN his age is *UNKNOWN.*
IF the person likes punk rock, THEN his age is *YOUNG.*
IF the person does not like punk rock, THEN his age is *UNKNOWN.*

where *not young* and *young* are fuzzy subsets of the interval [0, 100]. Suppose the system is given the following probability judgments about a person named John:

$$\begin{aligned} &P\,(\text{bold}) = 0.8, && P\,(\text{not bold}) = 0.2, \\ &P\,(\text{likes punk}) = 0.4, && P\,(\text{does not like punk}) = 0.6, \end{aligned}$$

The system is asked to determine how likely it is that John is a *middle-aged* person.

The important characteristic about the problem being considered here is that it contains both probabilistic information and vague information (e.g., *young, middle-aged*). The Dempster–Shafer theory has been shown to solve a special case of this problem where A_i and B are crisp sets [4]. Hence, we will briefly describe the basics of the D–S theory before we discuss previous work and our approach in generalizing the theory.

3 Basics of the Dempster–Shafer Theory

The Dempster–Shafer theory originated from the concept of lower and upper probability induced by a multivalued mapping [8]. Glenn Shafer further extended the theory in his book [9].

A multivalued mapping from space S to space T associates each element in S with a set of element in T, i.e., $\Gamma : S \to 2^T$. The image of an element s in S under the mapping is called the *granule* of s, denoted as $G(s)$. The multivalued mapping can also be viewed as a *compatibility relation* between the spaces S and T. A compatibility relation C between S and T characterizes the possibilistic relationship between their elements. An element s of S is compatible with an element t of T if it is possible that s is an answer to S and t is an answer to T at the same time [10] and the *granule* of s is the set of all elements in T that are compatible with s.

$$G\left(s\right) = \left\{t | t \in T, sCt\right\}.$$

Given a probability distribution of space S and a compatibility relation between S and T, a basic probability assignment (BPA) of space T, denoted by $m : 2^T \to\ [0, 1]$, is induced:[1]

$$m\left(A\right) = \frac{\sum\limits_{G(s_i)=A} p\left(s_i\right)}{1 - \sum\limits_{G(s_i)=\phi} p\left(s_i\right)} \tag{1}$$

where the subset A is also called a *focal element.*

The probability distribution of space T, which is referred to as *the frame of discernment*, is constrained by the basic probability assignment, but in general, it is not uniquely determined by the BPA. The belief measure and the plausibility measure of a set B are, respectively, the lower probability and the upper probability of the set subject to those constraints. These two quantities are obtained from the BPA as follows:

$$\text{Bel}\left(B\right) = \sum_{A \subset B} m\left(A\right) \tag{2}$$

$$\text{Pls}\left(B\right) = \sum_{A \subset B \neq \phi} m\left(A\right). \tag{3}$$

Hence, the belief interval [Bel(B), Pls(B)] is the range of B's probability.

An important advantage of the D–S theory is its ability to express degrees of ignorance. In the theory, the commitment of belief to a subset does not force the remaining belief to be committed to its complement, i.e., Bel(B) + Bel(B^c) $\leq =$ 1. The amount of belief committed to neither B nor B's complement is the degree of ignorance. Consequently, the theory provides a framework within which disbelief can be distinguished from a lack of evidence for belief.

[1] If we assume that T does not map any element of the space E to the empty set, the denominator (i.e., the normalization factor in 1) becomes one.

If m_1 and m_2 are two BPA's induced by two independent evidential sources, the combined BPA is calculated according to Dempster's rule of combination:

$$m_1 \oplus m_2 (C) = \frac{\sum\limits_{A_i \cap B_j = c} m_1 (A_i) m_2 (B_j)}{1 - \sum\limits_{A_i \cap B_j = \phi} m_1 (A_i) m_2 (B_j)}. \tag{4}$$

The basic combining steps that result in Dempster's rule are discussed in Sect. 5.6.

4 Previous Work

Zadeh was the first to generalize the Dempster-Shafer theory to fuzzy sets, based on his work on the concept of information granularity and the theory of possibility [4], [11]. A possibility distribution, denoted by Π, is a fuzzy restriction that acts as an elastic constraint on the values of a variable [12], [13]. Zadeh first generalized the granule of a D–S compatibility relation to a conditional possibility distribution. Then he defined the *expected certainty*, denoted by $EC(B)$, and the *expected possibility*, denoted by $E\Pi(B)$, as a generalization of D–S belief and plausibility functions:

$$E\Pi (B) = \sum_i m (A_i) \sup (B \cap A_i)$$
$$EC (B) = \sum_i m (A_i) \inf (A_i \Rightarrow B) = 1 - E\Pi (B^c)$$

where A_i denotes fuzzy focal elements induced from conditional possibility distributions, $\sup(B \cap A_i)$ measures the degree that B intersects with A_i, and $\inf(A \Rightarrow B)$ measures the degree to which A_i is included in B. It is easy to verify that the expected possibility and the expected certainty reduce to the D–S belief and plausibility measures when all A_i and B are crisp sets.

Following Zadeh's work, Ishizuka, Yager, and Ogawa have extended the D–S theory to fuzzy sets in slightly different ways [5]–[7]. They all extend D–S's belief function by defining a measure of inclusion $I(A \subset B)$, the degree to which set A is included in set B, and by using the following formula, similar to Zadeh's expected certainty $EC(B)$.

$$\mathrm{Bel}(B) = \sum_{A_i} I(A \subset B) m(A_i)$$

Their definitions of the measures of inclusion are listed as follows.

Ishizuka:

$$I_I (A \subset B) = \frac{\min_x [1, 1 + (\mu_B (x) - \mu_A (x))]}{\max_x \mu_A (x)}. \tag{5}$$

Yager:

$$I_Y\left(A \subset B\right) = \min_x \left[\mu_{\bar{A}}\left(x\right) \vee \mu_B\left(x\right)\right]. \tag{6}$$

Ogawa:

$$I_O\left(A \subset B\right) = \frac{\sum\limits_i \min\left[\mu_A\left(x_i\right), \mu_B\left(x_i\right)\right]}{\sum\limits_i \mu_B\left(x_i\right)}. \tag{7}$$

Based on Zadeh's expected certainty, Ishizuka and Yager arrive at different inclusion measures by using different implication operators in fuzzy set theory. Ogawa uses relative sigma count, which is analogous to conditional probability in spirit, to compute the degree of inclusion.

In order to combine two mass distributions with fuzzy focal elements, Ishizuka extended Dempster's rule by taking into account the degree of intersection of two sets, $J(A, B)$.

$$m_1 \oplus m_2\left(C\right) = \frac{\sum\limits_{A_i \cap B_j = c} J\left(A_i, B_j\right) m_1\left(A_i\right) m_2\left(B_j\right)}{1 - \sum\limits_{i,j}\left(1 - J\left(A_i B_j\right)\right) m_1\left(A_i\right) m_2\left(B_j\right)} \tag{8}$$

where

$$J\left(A, B\right) = \frac{\max_x \left[\mu_{A \cap B}\left(x\right)\right]}{\min\left[\max_x \mu_A\left(x\right), \max_x \mu_B\left(x\right)\right]}.$$

There are four problems with these extensions. First, the belief functions sometimes are not sensitive to significant changes in focal elements because degrees of inclusion are determined by certain "critical" points due to the use of "min" and "max" operators. Second, the definitions of "fuzzy intersection operator" and "fuzzy inclusion operator" are not unique. Consequently, it is difficult to choose the most appropriate definition for a given application. Third, although expected possibility and expected certainty (or, equivalently, expected necessity) degenerate to Dempster's lower and upper probabilities in the case of crisp sets, it is not clear that this is a "necessary" extension. Fourth, the generalized formula for combining evidence is not well justified.

5 Our Approach

Instead of directly modifying the formulas in the D–S theory, we generalize the primitive constructs of the theory and derive other extensions to the theory from these generalizations. We first generalize the compatibility relation in the D–S theory to a joint possibility distribution. Then, we formulate the linear programming problems that compute the belief measures and the plausibility measures. By extending the objective function and the constraints of the optimization problem, we obtain the formula for computing belief function in the generalized framework. We also extend Dempster's rule of combination by generalizing its steps in 1) combining the compatibility relations and 2) normalizing the combination result to account for the partial conflict between pieces of evidence. Finally, we achieve the commutativity of the extended Dempster rule by postponing its normalization step.

5.1 Generalizing the Compatibility Relation to a Possibility Distribution

In the Dempster–Shafer theory, the compatibility relation is limited to black-and-white answers. For example, given the question of whether s and t could be answers to S and T respectively, the compatibility relation may record only that the given situation is completely possible (i.e., (s, t) is in the relation C) or completely impossible (i.e., (s, t) is not in C). In general, however, the possibility that both s and t are answers to S and T is a matter of degree. To cope with this, we generalize Shafer's compatibility relation to a fuzzy relation that records joint possibility distribution of the spaces S and T.

Definition 1. *A generalized compatibility relation between the spaces S and T is a fuzzy relation $C : 2^{S\times T} \to [0, 1]$ that represents the joint possibility distribution of the two spaces, i.e.,*

$$C(s, t) = \Pi_{X,Y}(s, t)$$

where X and Y are variables that take values from the space S and the space T, respectively.

Shafer's compatibility relation is a special case of our fuzzy relation in which possibility measures are indicated by either zeros or ones.

In fuzzy set theory, if the relationship of two variables X and Y is characterized by a fuzzy relation R and the value of variable X is A, the value of variable Y can be induced using the *composition operation*, which is defined as:

$$\mu_{A\circ R}(y) = \max_x \{\min[\mu_A(x), \mu_R(x, y)]\}$$

So, we use the composition rule to generalize the definition of *granule*.

Definition 2. *Given a generalized compatibility relation $C : 2^{S\times T} \to [0, 1]$, the granule of an element s of S, denoted as $G(s)$, is defined to be the composition of the singleton $\{s\}$ and C, which turns out to be the possibility distribution conditioned on s, i.e.,*

$$G(s) = \{s\} \circ C = \Pi_{(Y|X=s)}.$$

Hence, we generalize granules to conditional possibility distributions just as Zadeh did; however, our approach is more general than Zadeh's approach because we go one step further to generalize the compatibility relation to a joint possibility distribution. As we will see in Sect. 5.6, the generalized compatibility relation is important for justifying our generalization of Dempster's rule.

Given a probability distribution of the space S and a joint possibility distribution between space S and space T such that the granules of S's elements are normal fuzzy subsets,[2] a basic probability assignment (BPA) m to T is induced using 1. Adopting the terminology of the D–S theory, we call a fuzzy subset of T with nonzero basic probability a *fuzzy focal element.* A fuzzy basic probability assignment (BPA) is a BPA that has at least one fuzzy focal element.

5.2 The Optimization Problem for Computing the Belief Function

As a basis for the following discussions, this section formulates the linear programming problems implicitly solved by the belief function. This serves as a foundation upon which we can generalize various basic components of the optimization problem (e.g., the objective function, the constraints) that correspond to basic concepts underlying the belief function.

Pls(B) and Bel(B) are the upper and lower probabilities of a set B under the constraints imposed by a basic probability assignment. Therefore, the belief function can be obtained by solving the following optimization problem:

$$\text{LP1---}\min \sum_{x_i \in B} \sum_{j} m\left(x_i : A_j\right)$$

subject to the following constraints:

$$m\left(x_i : A_j\right) \geq 0, \qquad i = 1, \cdots, n;\ j = 1, \cdots, 1 \tag{9}$$

$$m\left(x_i : A_j\right) = 0, \qquad \forall x_i \notin A_j \tag{10}$$

$$\sum_{i} m\left(x_i : A_j\right) = m\left(A_j\right) \qquad j = 1, \cdots, 1. \tag{11}$$

The variable $m(x_i : A_j)$ denotes the probability mass allocated to x_i from the basic probability of a focal element A_j. The objective function simply computes the total probability of the set B where the inner summation gives the probability of an element x_i. The inequality constraint, specified by (9), states the nonnegativity of probability masses. Equation (10) prohibits the basic probability of a focal from being assigned to any elements outside the focal. Equation (11) expresses that all the probability mass assigned by a focal should add up to its basic probability. It follows, from (9) and (11), that the upper bound on $m(x_i : A_j)$ is $m(A_j)$.

Since the distributions of focals' masses do not interact with one another, they can be optimized individually to reach a global optimal solution. Hence,

[2] A fuzzy subset A is normal if $\sup_x \mu_A(x) = 1$. The assumption that all focal elements are normal is further discussed in Sect. 5.6).

we partition the linear programming problem LP1 into subproblems, each one of which concerns the allocation of the mass of a focal element. The optimal value of the original problem LP1 is the sum of the subproblems' solutions. A subproblem for LP1 is formulated as follows:

$$\mathrm{LP1}_j\text{—}\min \sum_{x_i \in B} m\left(x_i : A_j\right)$$

subject to the following constraints:

$$\begin{aligned} m\left(x_i : A_j\right) &\geq 0 \\ m\left(x_i : A_j\right) &= 0, \qquad x_i \notin A_j \\ \sum_i m\left(x_i : A_j\right) &= m\left(A_j\right). \end{aligned}$$

The linear programming problem for computing the plausibility of set B differs only in the direction of optimization. It is formulated as LP2 as follows:

$$\mathrm{LP2}\text{—}\max \sum_{x_i \in B} \sum_j m\left(x_i : A_j\right) \text{subject to}(9) - (11).$$

Like LP1, the linear programming problem LP2 can be partitioned into l subproblems, each of which finds an optimal distribution of a focal's mass to make a maximum contribution to the belief in B.

The optimal solutions of the minimization subproblem $\mathrm{LP1}_j$ and the maximization subproblem $\mathrm{LP2}_j$ are denoted as $m_*(B : A_j)$ and $m^*(B{:}A_j)$ respectively. Adding the optimal solutions of subproblems, we get B's belief measure and plausibility measure as shown below.

$$\mathrm{Bel}\left(B\right) \sum_{A_j \subseteq T} m_*\left(B : A_j\right). \tag{12}$$

$$\mathrm{Pls}\left(B\right) \sum_{A_j \subseteq T} m_*\left(B : A_j\right). \tag{13}$$

It is easy to show that the optimal solutions of the subproblems are the following:

$$m_*\left(B : A_j\right) = \begin{cases} m\left(A_j\right) & \text{if } A_j \subset B \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

$$m^*\left(B : A_j\right) = \begin{cases} m\left(A_j\right) & \text{if } A_j \subset B \neq \phi \\ 0 & \text{otherwise} \end{cases}. \tag{15}$$

Equations (2) and (3), the formulas for calculating D–S belief and plausibility, thus follow directly from (12)–(15).

5.3 Generalizing Objective Functions

Philippe Smets has shown that the belief measure of a fuzzy set B, given a nonfuzzy basic probability assignment, can be obtained by computing the lower bound on the expected value of B's membership function [14]. Here we show that the same result can be obtained by modifying the objective functions of the optimization problems, discussed in Sect. 5.2, to account for the membership degree of the fuzzy set B.

The objective function of LP1 and LP2 computes the probability of a crisp set B. If B is a fuzzy subset of the frame of discernment, its probability is defined as

$$P\left(B\right)=\sum_{x_i} P\left(x_i\right)\times \mu_B\left(x_i\right)$$

in fuzzy set theory. We can thus generalize the objective function to

$$\sum_{x_i}\sum_{j} m\left(x_i : A_j\right)\times \mu_B\left(x_i\right).$$

Based on this generalization of the objective functions and the following theorem, we get the belief function of fuzzy sets for a nonfuzzy basic probability assignment.

Theorem 1. *Suppose A is a nonfuzzy focal element. The maximum and minimum probability masses that can be allocated to a fuzzy set B from A are*

$$m_*\left(B : A\right)=m\left(A\right)\times \inf_{x\in A}\mu_B\left(x\right) \tag{16}$$

$$m^*\left(B : A\right)=m\left(A\right)\times \sup_{x\in A}\mu_B\left(x\right). \tag{17}$$

Proof. $m^*(B : A)$ is the optimal solution to the following linear programming problem:

$$\min\sum_{x} m\left(x : A\right)\times \mu_B\left(x\right)$$

subject to the following constraints:

$$m\left(x : A\right)\geq 0$$
$$m\left(x : A\right)=0 \quad \forall x\notin A$$
$$\sum_{x} m\left(x : A\right)=m\left(A\right).$$

An optimal solution of this simple linear programming problem can be obtained by assigning all the mass of A to an element of A that has the lowest membership degree in B. Thus, we have $m_*(B : A) = m(A)\times \inf_{x\in A}\mu_B(x)$. Equation (17) can be proved in a similar way.

From (12), (13), (16), and (17), we obtain the following formula for computing the belief and plausibility of fuzzy sets from a crisp basic probability assignment:

$$\begin{aligned}\mathrm{Bel}\,(B) &= \sum_{A_j \subseteq T} m\,(A_j) \times \inf_{x \in A_j} \mu_B\,(x) \\ \mathrm{Pls}\,(B) &= \sum_{A_j \subseteq T} m\,(A_j) \times \sup_{x \in A_j} \mu_B\,(x)\,.\end{aligned}$$

Thus, we have shown that Smets' generalization of the D–S belief function is a result of generalizing the objective function of the optimization problem that the belief function is solving.

5.4 Representing the Probabilistic Constraints of Fuzzy Focal Elements Through Decomposition

To deal with fuzzy focal elements, we decompose them into nonfuzzy focal elements whose probabilistic constraints have been discussed in Sect. 5.2. A fuzzy focal element has two components: a fuzzy subset of the frame of discernment and the probability mass assigned to the subset. In this section, we first describe how a fuzzy set can be decomposed into nonfuzzy sets. Then we define the decomposition of a fuzzy focal element.

An α-level set of A, a fuzzy subset of T, is a crisp set denoted by A_α that comprises all elements of T whose grade of membership in A is greater than or equal to α:

$$A_\alpha = \{x | \mu_A\,(x) \geq \alpha\}$$

A fuzzy set A may be decomposed into its level-sets through the *resolution identity* [15]:

$$A = \sum_\alpha \alpha A_\alpha$$

where the summation denotes the set union operation and αA_α denotes a fuzzy set with a two-valued membership function defined by

$$\begin{aligned}\mu_{\alpha A_\alpha}\,(x) &= \alpha \quad \text{for } x \in A_\alpha \\ \mu_{\alpha A_n}\,(x) &= 0 \quad \text{elsewhere.}\end{aligned}$$

The importance of resolution identity is best described by Zadeh [15]: "The resolution identity provides a convenient way of generalizing various concepts associated with nonfuzzy sets to fuzzy sets" [15]. In fact, this is the underlying basis for many of the definitions of fuzzy set operations [16], [17].

In order to decompose a fuzzy focal element, we also need to decompose the focal's basic probability and distribute it among the focal's level-sets. Obviously, the decomposition has to satisfy two conditions:

1) The decomposed basic probabilities must add up to the basic probability assigned to the fuzzy focal.

$$\sum_{\alpha} m\left(A_{\alpha}\right) = m\left(A\right)$$

2) The decomposed basic probabilities must not be negative.

$$m\left(A_{\alpha}\right) \geqslant 0.$$

Using Dubois and Prade's observation on the relationship between possibility distribution (i.e., membership function of a fuzzy focal) and nonfuzzy consonant focals [18], we reach a decomposition of the fuzzy focal's basic probability that satisfies the two conditions stated above.

Dubois and Prade have shown that if a BPA is a set of nested focal elements, $A_1 \supset A_2 \cdots \supset A_n$, they can be related to the possibility distribution induced, denoted as Poss(x), as follows:[3]

$$m\left(A_i\right) = \pi_i - \pi_{i-1} \tag{18}$$

where $\pi_i = \inf_{x \in A_i}$ poss(x), $\pi_0 = 0$, and $\pi_n = 1$. This result can be directly applied to decompose a fuzzy focal element whose basic probability value is one (i.e., $m(A) = 1$) because the α-level sets of A form a set of nested focal elements. Since $\inf_{x \in A_{\alpha_i}}$poss$(x) = \alpha_i$, the π_i in (18) becomes the alpha value α_i of the level sets. Thus, we get

$$m\left(A_{\alpha_i}\right) = \alpha_i - \alpha_{i-1}. \tag{19}$$

We extend this idea to decompose fuzzy focal elements with arbitrary probability mass (i.e., $0 \leq m(A) \leq 1$) by multiplying the focal's mass with the right-hand side of (19). Formally, the decomposition of a fuzzy element is defined as follows.

Definition 3. *The decomposition of a fuzzy focal element A is a collection of nonfuzzy subsets such that 1) they are A's α-level sets that form a resolution identity, and 2) their basic probabilities are*

$$m\left(A_{\alpha_i}\right) = \left(\alpha_i - \alpha_{i-1}\right) \times m\left(A\right) \quad i = 1, 2, \cdots, n \tag{20}$$

where $\alpha_0 = 0$ and $\alpha_n = 1$.

When the focal element is a crisp set, its decomposition is the focal itself because the decomposition contains only one level set, which corresponds to the membership degree "one." The relationship between the decomposition of a fuzzy focal element and Shafer's consonant focals is discussed further in Sect. 6.1.

[3] We have paraphrased Dubois and Prade's results for the convenience of our discussion.

The probabilistic constraint of a fuzzy focal is defined to be that of its decomposition, which is a set of nonfuzzy focals. Since we already know how to deal with nonfuzzy focals, decomposing a fuzzy focal into nonfuzzy ones allows us to calculate the belief functions that are constrained by the fuzzy focals.

Definition 4. *The probability mass that a fuzzy focal A contributes to the belief (and plausibility) of a fuzzy subset B is the total contribution of A's decomposition to B's belief (and plausibility), i.e.,*

$$m^{*}\left(B:A\right)=\sum_{\alpha}m^{*}\left(B:A_{\alpha}\right) \tag{21}$$

$$m_{*}\left(B:A\right)=\sum_{\alpha}m_{*}\left(B:A_{\alpha}\right). \tag{22}$$

5.5 Computing the Belief Function

Based on generalizing the objective function and expressing the probabilistic constraints of fuzzy focal elements through their decompositions, we are able to derive the following formula for computing the belief function and the plausibility function.

$$\mathrm{Bel}\left(B\right)=\sum_{A}m\left(A\right)\sum_{\alpha_i}\left[\alpha_i-\alpha_{i-1}\right]\times\inf_{x\in A_{\alpha_i}}\mu_B\left(x\right) \tag{23}$$

$$\mathrm{Pls}\left(B\right)=\sum_{A}m\left(A\right)\sum_{\alpha_i}\left[\alpha_i-\alpha_{i-1}\right]\times\sup_{x\in A_{\alpha_i}}\mu_B\left(x\right). \tag{24}$$

It is also trivial to show that the derived formulas preserve the following important property of the D-S theory: The belief of a (fuzzy) set is the difference of one and the plausibility of the set's complement.

1) An Example: The following example illustrates how one applies the formula described in Sect. 5.5 for computing the belief function. Suppose the frame of discernment is the set of integers between 1 and 10. A fuzzy basic probability assignment consists of two focal elements A and C:

$$\begin{aligned}A=&\{0.25/1,0.5/2,0.75/3,1/4,1/5,\\&0.75/6,0.5/7,0.25/8\}\\C=&\{0.5/5,1.6,0.8/7,0.4/8\}\end{aligned}$$

where each member of the list is in the form of $\mu_A(x_i)/x_i$. We are interested in the degree of belief and the degree of plausibility of the fuzzy subset B:

$$B=\{0.5/2,1/3,1/4,1/5,0.9/6,0.6/7,0.3/8\}.$$

The decomposition of fuzzy focal A consists of four nonfuzzy focals:

$$\begin{aligned}A_{0.25} &= \{1,2,\cdots,8\} \text{ with mass } 0.25\times m\,(A)\\ A_{0.5} &= \{2,3,\cdots,7\} \text{ with mass } 0.25\times m\,(A)\\ A_{0.75} &= \{3,4,\cdots,6\} \text{ with mass } 0.25\times m\,(A)\\ A_{1} &= \{4,5\} \text{ with mass } 0.25\times m\,(A)\end{aligned}$$

and the decomposition of fuzzy focal C also consists of four nonfuzzy focals:

$$\begin{aligned}C_{0.4} &= \{5,6,7,8\} \text{ with mass } 0.4\times m\,(C)\\ C_{0.5} &= \{5,6,7\} \text{ with mass } 0.1\times m\,(C)\\ C_{0.8} &= \{6,7\} \text{ with mass } 0.3\times m\,(C)\\ C_{1} &= \{6\} \text{ with mass } 0.2\times m\,(C)\end{aligned}$$

Let us denote $\inf_{x\in A_{\alpha_i}}\mu_B(x)$ as $f_{B,A}(\alpha_i)$. So, we have

$$\begin{aligned}&m_*\,(B:\ A)\\ &\quad = m\,(A)\times[0.25\times f_{B,A}\,(0.25)\,0.25\times f_{B,A}\,(0.5)\\ &\quad + 0.25\times f_{B,A}\,(0.75) + 0.25\times f_{B,A}\,(1)]\\ &\quad = m\,(A)\times[0.25\times 0 + 0.25\times 0.5 + 0.25\times 0.9 + 0.25\times 1]\\ &\quad = 0.6\times m\,(A)\\ &m_*\,(B:\ C)\\ &\quad = m\,(C)\times[0.4\times f_{B,C}\,(0.4)\,0.1\times f_{B,C}\,(0.5)\\ &\quad + 0.3\times f_{B,C}\,(0.8) + 0.2\times f_{B,C}\,(1)]\\ &\quad = m\,(C)\times[0.4\times 0.3 + 0.1\times 0.6 + 0.3\times 0.6 + 0.2\times 0.9]\\ &\quad = 0.54\times m\,(C)\end{aligned}$$

Thus, we have

$$\mathrm{Bel}\,(B) = 0.6\ m\,(A) + 0.54\ m\,(C)$$

Similarly, we can calculate the plausibility of B:

$$\mathrm{Pls}\,(B) = m\,(A) + 0.86\ m\,(C)$$

2) A Comparison with Alternative Approaches: In this section, we will use the example discussed in Sect. 5.5-1 to compare our approach with the alternative fuzzy evidential reasoning methods discussed in Sect. 4. The degrees of belief in the fuzzy set B computed using these methods are listed as follows:

$$\begin{aligned}\text{Ishizuka}:\ \ \mathrm{bal}(B) =&0.75\ m\,(A) + 0.8\ m\,(C)\,.\\ \text{Yager}:\ \ \mathrm{bal}(B) =&0.5\ m\,(A) + 0.6\ m\,(C)\,.\\ \text{Ogawa}:\ \ \mathrm{bal}(B) =&0.8962\ m\,(A) + 0.434\ m\,(C)\,.\end{aligned}$$

We will compare how these results are changed in response to a change of fuzzy focal element. More specifically, we change the membership function of the fuzzy focal element A in three different ways. First, we increase the gradient of $\mu_A(x)$ for $1 \leq x \leq 3$ while keeping $\mu_A(2)$ unchanged. The modified focal element, denoted as A', is

$$A' = \{0.166/1, 0.5/2, 0.833/3,\ 1/4, 1/5, 0.75/6, 0.5/7, 0.25/8\}.$$

Second, we modify A into A'' by increasing the gradient of $\mu_A(x)$ for $1 \leq x \leq 3$ while preserving the membership value $\mu_A(1)$:

$$A' = \{0.25/1, 0.75/2, 1/3, 1/4, 1/5, 0.75/6, 0.5/7, 0.25/8\}.$$

Finally, we get A''' by decreasing the membership value $\mu_A(1)$ while maintaining the membership values of other points:

$$A'' = \{0/1, 0.5/2, 0.75/3, 1/4, 1/5, 0.75/6, 0.5/7, 0.25/8\}.$$

Since only the focal element A has been changed, we can analyze the impact to the belief function by comparing the contributions of the focal element A and its variations to the degree of belief in B. Table 1 lists the portion of each modified focal's mass that contributes to B's belief measure (i.e., the ratio $m_*(B : A)/m(A)$) for each fuzzy evidential reasoning method.

Table 2 shows how Bel(B) computed by different methods change as the focal element A changes in three ways. As shown in the table, Yager's method is insensitive to any of the three changes in the focal's membership function; Ishizuka's method is insensitive to a change from A to A''; and Ogawa's approach is insensitive to a change from A to A'''. Our approach is sensitive to all three kinds of changes in the focal's membership function.

This comparison indicates that previous approaches to generalizing the Dempster–Shafer model to fuzzy sets are not always responsive to a change of the focal element. In general, Ishizuka's belief function and Yager's belief function are insensitive to a focal element's change unless it results in a change of the "critical point," a point whose membership value is the minimal value in (5) and (6) for computing the inclusion measure, i.e.,

$$\mu_A(x_I) = \min_x [1, 1 + (\mu_B(x) - \mu_A(x))]$$
$$\mu_A(x_Y) = \min_x [\mu_{\bar{A}}(x) \vee \mu_B(x)]$$

Table 1. The Contribution to Bel(B) from the Focal Element A and its Variations

Focal Elements	Yager	Ishizuk	Ogawa	Yen
A	0.5	0.75	0.8962	0.6
A'	0.5	0.834	0.9119	0.6252
A''	0.5	0.75	0.9434	0.5
A'''	0.5	1	0.8962	0.675

Table 2. Changes to Bel(B) Due to Changes in the Focal Element A

Changes of Focal Element A	Yager	Ishizuk	Ogawa	Yen
$A \rightarrow A'$	unchanged	increased	increased	increased
$A \rightarrow A''$	unchanged	unchanged	increased	decreased
$A \rightarrow A'''$	unchanged	increased	unchanged	increased

where x_I and x_Y denote the critical points for Ishizuka's inclusion measure and Yager's inclusion measure respectively. In our example, the critical points for Yager's inclusion measure $I_Y(A \subset B)$ and Ishizuka's inclusion measure $I_I(A \subset B)$ are $x_Y = 2$ and $x_I = 1$ respectively. As the focal element A changes to A', the critical point for Yager's inclusion measure remains the same. As a result, Yager's belief measure of the fuzzy subset B remains unchanged. Similarly, a change from A to A'' does not change the critical point x_I. Hence, Ishizuka's belief measure of B remains the same in this case.

Ogawa's belief measure of a fuzzy subset B is not responsive to a change in the focal element's membership function unless the intersection between the focal and the fuzzy subset B is different. Since the intersection $A \cap B$ is the same as $A''' \cap B$, Ogawa's belief measure of B remains unchanged when the focal A changes to A'''.

A surprising result of this comparison is that a change from A to A'' increases Ogawa's belief measure, but decreases ours. This can be explained as follows. Ogawa's measure of inclusion is based on the sigma count of $A \cap B$ relative to the sigma count of B. Since the intersection of A'' and B is a fuzzy superset of the intersection of A and B, Ogawa's measure of inclusion increases as the focal change from A to A''. However, our belief measure in B decreases because the level set of A'' at membership degree 0.75 contributes less to the belief measure Bel(B) than A's level-set at 0.75 does (i.e., $f_{B,A''}(0.75) = 0.5$ is less than $f_{B,A}(0.75) = 0.9$) while the contributions of all other level sets remain the same.

In summary, the comparison above indicates that our method for computing the belief function of fuzzy sets is more responsive to any change to a focal element's membership function than previous approaches are. Moreover, a change in our belief measure can always be explained in terms of a change in the underlying probabilistic constraints imposed by the focal elements.

5.6 Generalizing Dempster's Rule of Combination

Dempster's rule combines the effects of two independent evidential sources, denoted as R and S, on the probability distribution of a hypothesis space, denoted as T. The rule can be viewed as a result of three steps.

1) *Combine the compatibility relations.* A combined compatibility relation between the product space $R \times S$ and T can be constructed from the compatibility relation between R and T and the one between S and T using the following principle:

$$rCt \text{ and } sCt \Rightarrow [r,s]\,Ct$$

where r, s, t, and $[r,s]$ denote elements of R, S, T, and $R \times S$ respectively. As a result, the granule of $[r,s]$ under the combined multivalued mapping is the intersection of the granule of r and the granule of s, i.e.,

$$G\left([r,s]\right) = G\left(r\right) \cap G\left(s\right). \tag{25}$$

This explains why focal elements of different evidential sources are intersected in Dempster's rule.

2) *Compute joint probability distributions of the combined evidential source.* Since R and S are assumed to be independent, the joint probability distribution of the space $R \times S$ can be computed from the probability distribution of each individual space:

$$P\left([r,s]\right) = P\left(r\right) \times P\left(s\right)$$

3) *Normalize the combined basic probability assignment.* Having obtained the probability distribution of $R \times S$ and the compatibility relation between $R \times S$ and T from the two previous steps, Dempster's rule follows directly from (1), which includes a normalization process to discard probability mass assigned to the empty set.

Two generalizations must be made to Dempster's rule before it can be used to combine fuzzy BPA's in our generalized framework: 1) the first step above has to be extended to allow the combination of fuzzy compatibility relations; and 2) the normalization step needs to consider subnormal fuzzy focal elements that result from combining fuzzy compatibility relations.

1) *Combination of Fuzzy Compatibility Relations*: By employing the noninteractiveness assumption in possibility theory, we generalize (25) in order to perform fuzzy intersection to obtain granules of the combined compatibility relation. A compatibility relation in our generalized D-S framework, as discussed in Sect. 5.1, is a joint possibility distribution. Thus, we have

$$C\left(r,t\right) = \prod_{X,Z}\left(r,t\right) \quad \text{and } C\left(s,t\right) = \prod_{Y,Z}\left(s,t\right) \tag{26}$$

where X, Y, and Z are variables that take values from the spaces R, S, and T, respectively. Let W be a variable that takes values from the space $R \times S$. The combined fuzzy compatibility relation can be expressed as

$$C\left([r,s],t\right) = \prod_{W,Z}\left([r,s],t\right) = \prod_{X,Y,Z}\left(r,s,t\right).$$

Marginal possibility distributions $\Pi_{X,Z}$ and $\Pi_{Y,Z}$ are the projection of joint possibility distribution on Y and X respectively, [12] i.e.,

$$\prod_{Y,Z}(s,t) = \min_{r} \prod_{X,Y,Z}(r,s,t)$$
$$\prod_{X,Z}(r,t) = \min_{s} \prod_{X,Y,Z}(r,s,t)\,.$$

Hence, the joint possibility distribution is bounded by the marginal possibility distributions:

$$\prod_{X,Y,Z}(r,s,t) \le \prod_{Y,Z}(s,t) \wedge \prod_{X,Z}(r,t)$$

where $\wedge$ denotes the minimum operator. By employing the assumption that the variables $Y,\ Z$ and $X,\ Z$ are noninteractive, a concept analogous to the independence of random variables, we obtain the following joint possibility distribution:

$$\prod_{X,Y,Z}(r,s,t) = \prod_{Y,Z}(s,t) \wedge \prod_{X,Z}(r,t)\,.$$

Thus, the combined fuzzy compatibility relation is

$$C\left(\left[r,s\right],t\right) = C\left(r,t\right) \wedge C\left(s,t\right). \tag{27}$$

For a fixed pair of r and s, applying (27) to all possible elements in T gives us the following relationship between conditional possibility distributions:

$$\prod_{(Z|W=[r,s])} = \prod_{(Z|X=r)} \cap \prod_{(Z|Y=s)}$$

where $\cap$ denotes the fuzzy intersection operator. Equivalently, the granule of the pair $[r,s]$ under the combined compatibility relation defined in (27) is the fuzzy intersection of $G(r)$ and $G(s)$:

$$G\left(\left[r,s\right]\right) = G\left(r\right) \cap G\left(s\right)$$

2) *Normalizing Subnormal Fuzzy Focal Elements*: An important assumption of our work is that all focal elements are normal. We avoid subnormal fuzzy focal elements because they assign probability mass to the empty set. For example, suppose A is a fuzzy subset of the frame of discernment $\{x0, x1, x2, x3, x4\}$, characterized by the membership function

$$A\left\{0/x0, 0.1/x1, 0.2/x2, 0.1/x3, 0/x4\right\}.$$

Let the basic probability value of the set A be "a". The decomposition of this focal element A is:

$$\begin{aligned} A_{0.1} &= \{x1, x2, x3\} \text{ with mass } 0.1 \times a \\ A_{0.2} &= \{x2\} \text{ withmass } 0.1 \times a \\ A_1 &= \phi \text{ with mass } 0.8 \times a \end{aligned}$$

In general, the probability mass assigned to the empty set by a subnormal fuzzy focal A is the basic probability assigned to the decomposed focal of A that is constructed from A's α-level set at the degree of membership one:

$$\left[1 - \max_x \mu\left(x\right)\right] \times m\left(A\right).$$

Although we have assumed that the focal elements of fuzzy BPA's are all normal, the intersections of focals may be subnormal. Hence, the combination of fuzzy BPA's should deal with the normalization of subnormal fuzzy focal elements. To do this, we need to normalize the two components of a fuzzy focal element: the focal itself, which is a subnormal fuzzy set, and the probability mass assigned to the focal.

It is straightforward to normalize the focal. Suppose A is a subnormal fuzzy set characterized by the membership function $\mu_A(x)$. A's normalized set, denoted as $\bar{A}$, is characterized by the following membership function.

$$\mu_{\bar{A}}\left(x\right) = \frac{\mu_A\left(x\right)}{\max\limits_x \mu_A\left(x\right)} = k \times \mu_A\left(x\right)$$

where k is the normalization factor

$$k = 1 / \max_x \mu_A\left(x\right).$$

The criterion for normalizing the probability mass of a subnormal focal is that the probabilistic constraints imposed by the subnormal focal should be preserved after the normalization. Since we use the decomposition of a focal to represent its probabilistic constraint, this means that the probability mass assigned to a decomposed focal should not be changed by the normalization process. Since the α_i cut of the subnormal focal becomes the $k\alpha_i$ cut of the normalized focal, the probability mass assigned to them should be the same:

$$m\left(A_{\alpha_i}\right) = m\left(\bar{A}_{k\alpha_i}\right). \tag{28}$$

From this condition, we can derive the relationship between $m(\bar{A})$ and $m(A)$ as follows. The left-hand side of (28) can be rewritten as

$$m\left(A_{\alpha_i}\right) = m\left(A\right)\left(\alpha_i - \alpha_{i-1}\right).$$

The right-hand side of (28) can be rewritten as

$$m\left(\bar{A}_{k\alpha_i}\right) = m\left(\bar{A}\right)\left(k\alpha_i - k\alpha_{i-1}\right) = km\left(\bar{A}\right)\left(\alpha_i - \alpha_{i-1}\right).$$

It follows from the three equations above that the mass of the normalized focal is reduced by a factor reciprocal to the ratio by which its membership function is scaled up:

$$m\left(\bar{A}\right) = m\left(A\right)/k.$$

The remaining mass $(1 - 1/k)m(A)$ is the amount assigned to the empty set by the subnormal fuzzy focal and, hence, should be part of the normalization factor in the generalized Dempster's rule.

We summarize our approach to normalize a subnormal focal element into three steps:

a) Scale up the membership function so that its peak (i.e., highest membership degree) is one.
b) Reduce the basic probability using a ratio reciprocal to the scaling factor of the first step.
c) Assign the basic probability lost during the second step to the empty set.

3) A Generalized Rule of Combination: Commutativity is an important requirement for any evidence combination rule, because it is highly desirable to have the effect of the aggregated evidence independent of the order of combination. It is well known that Dempster's rule is commutative [9, p. 62]. Our normalization step discussed in Sect. 5.6-2 is not commutative because it modifies the membership functions of the focal elements' subnormal intersections. To solve this problem, we first show that the normalization process in Dempster's rule can be postponed without changing the combination result. Then, we describe our generalized combining rule where the normalization process is postponed to achieve commutativity.

Normalization in Dempster's rule does not have to apply after each combining operation. It can be postponed to a later point without changing the result. More specifically, several BPA's in the D–S theory can be combined without normalization, and the normalized combined bpa can be obtained by applying the normalization process to the unnormalized combined BPA at the end. In the following discussion, we use the symbol $\otimes$ to denote Dempster's rule without normalization (i.e., the denominator in (4) is one), the letter "N" to denote the normalization process, and the primed letter m' to denote the unnormalized BPA. Fig. 1 and Fig. 2 show two ways to apply Dempster's rule: combine BPA's with immediate normalization, or combine BPA's with postponed normalization. To show that they obtain the same result, we consider three BPA's of a frame of discernment: m_1, m_2, and m_3. We want to show that applying normalization after the three BPA's are combined without normalization yields the same result as using Dempster's rule in the conventional way to combine them, i.e.,

$$(m_1 \oplus m_2) \oplus m_3 = N\left[(m_1 \oplus m_2) \oplus m_3\right] \qquad (29)$$

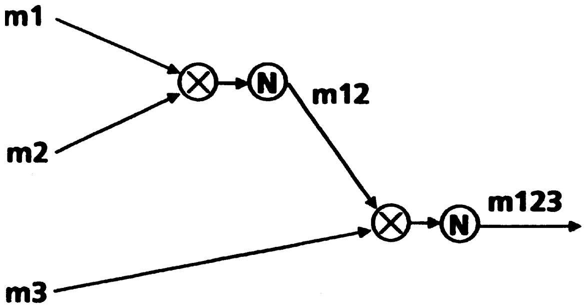


Fig. 1. Combination of evidence with immediate normalization

We first expand the result of combining the first two BPA's using Dempster's rule.

$$m_1 \oplus m_2 (C) = \frac{m'_{12} (C)}{1 - k_{12}}$$

where

$$m'_{12} (C) = \sum_{A \cap B = C} m_1 (A) m_2 (B) \tag{30}$$

and

$$k_{12} = \sum_{A \cap B = \phi} m_1 (A) m_2 (B) \tag{31}$$

The left-hand side of (29) thus becomes

$$(m_1 \oplus m_2) \oplus m_3 (E) = \frac{\frac{1}{1-k_{12}} \sum\limits_{C \cap D = E} m'_{12} (C) m_3 (D)}{1 - \frac{1}{1-k_{12}} \sum\limits_{C \cap D = \phi} m'_{12} (C) m_3 (D)}$$

Substituting $m'_{12}(C)$ with the right-hand side of (30), we get

$$= \frac{\frac{1}{1-k_{12}} \sum\limits_{A \cap B \cap D = E} m_1 (A) m_2 (B) m_3 (D)}{1 - \frac{1}{1-k_{12}} \sum\limits_{A \cap B \cap D = \phi, A \cap B \neq \phi} m_1 (A) m_2 (B) m_3 (D)}$$

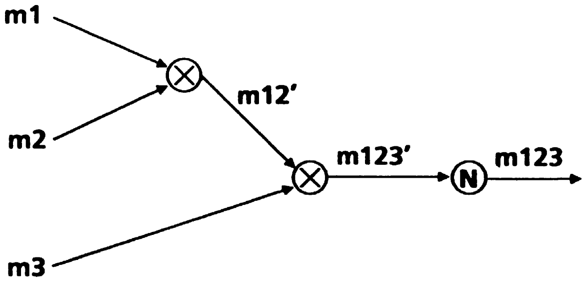


Fig. 2. Combination of evidence with postponed normalization

Multiplying both the numerator and the denominator by $1 - k_{12}$, we have

$$= \frac{\sum\limits_{A \cap B \cap D = E} m_1(A)\, m_2(B)\, m_3(D)}{1 - k_{12} - \sum\limits_{A \cap B \cap D = \phi, A \cap B \neq \phi} m_1(A)\, m_2(B)\, m_3(D)}\,.$$

Substituting k_{12} with the right-hand side of (31), we get

$$= \frac{\sum\limits_{A \cap B \cap D = E} m_1(A)\, m_2(B)\, m_3(D)}{1 - \left[\sum\limits_{A \cap B \cap D = \phi} m_1(A)\, m_2(B) + \sum\limits_{A \cap B \cap D = \phi, A \cap B \neq \phi} m_1(A)\, m_2(B)\, m_3(D)\right]}$$

Since $\Sigma_D m_3(D) = 1$, we can reformulate the normalization factor:

$$= \frac{\sum\limits_{A \cap B \cap D = E} m_1(A)\, m_2(B)\, m_3(D)}{1 - \left[\sum\limits_{A \cap B \cap D = \phi, A \cap B \phi} m_1(A)\, m_2(B)\, m_3(D) + \sum\limits_{A \cap B \cap D = \phi, A \cap B \neq \phi} m_1(A)\, m_2(B)\, m_3(D)\right]}$$

Finally, we get

$$\begin{aligned} &= \frac{\sum\limits_{A \cap B \cap D = E} m_1(A)\, m_2(B)\, m_3(D)}{1 - \sum\limits_{A \cap B \cap D \phi} m_1(A)\, m_2(B)\, m_3(D)} \\ &= N\left[(m_1 \oplus m_2) \oplus m_3\right]. \end{aligned}$$

Hence, we have shown that the normalization step in Dempster's rule can be delayed without changing the result of combination.

Our generalized rule of combination consists of two operations: a cross-product operation and a normalization process. Fuzzy BPA's are first combined by performing the following generalized cross product:

$$m'_{12}(C) = m_1 \oplus m_2(C) = \sum_{A \cap B = C} m_1(A)\, m_2(B)\,. \tag{32}$$

where $\cap$ denotes the fuzzy intersection operator and C is an unnormalized intersection of focal elements, which could be a subnormal fuzzy subset of the frame of discernment. The empty set is a special kind of subnormal focal elements. To compute the normalized combined BPA (e.g., for computing its belief function), we apply the following normalization process (discussed in Sect. 5.6-2) to the unnormalized combined BPA:

$$N[m'](D) = \frac{\sum\limits_{\bar{C} = D} \max\limits_{x_i} \mu_C(x_i)\, m'(C)}{1 - \sum\limits_{C \subset T} \left(1 - \max\limits_{x_1} \mu_C(x_i)\right) m'(C)}. \tag{33}$$

For example, if we need to combine three bpa's of the frame of discernment T, the result of combination is computed by first combining the three bpa's without normalization using (32), and then normalizing the final result:

$$m_1 \oplus m_2 \oplus m_3 = N\left[(m_1 \oplus m_2) \oplus m_3\right].$$

It is obvious that the generalized cross-product operation is commutative, e.g.,

$$N\left[(m_1 \oplus m_2) \oplus m_3\right] = N\left[m_1 \oplus (m_2 \oplus m_3)\right].$$

Thus, through delaying the normalization process, we are able to combine fuzzy BPA's in an order-independent fashion.

In the special case where there are only two fuzzy BPA's to be combined, the combined BPA using the generalized Dempster's rule of combination is

$$\begin{aligned} m_1 \oplus m_2\,(C) &= N\left[m_1 \oplus m_2\right](C) \\ &= \frac{\sum\limits_{(A\cap B)=C} \max\limits_{x_i} \mu_{A\cap B}\,(x_i)\, m_1\,(A)\, m_2\,(B)}{1 - \sum\limits_{A,B} \left(1 - \max\limits_{x_i} \mu_{A\cap B}\,(x_i)\right) m_1\,(A)\, m_2\,(B)}. \end{aligned} \tag{34}$$

The normalization process (i.e., (33)) generalizes the notion of *conflicting evidence* in the D–S theory to that of *partially conflicting evidence.* In Dempster's original rule, two pieces of evidence are either in conflict (i.e., the intersection of their focals is empty) or not in conflict at all (i.e., the intersection of their focals is not empty). In our generalized combining rule, two pieces of evidence are partially in conflict if the intersection of their focals is subnormal. The *degree of conflict* is measured by the difference between one and the peak (i.e., the maximum value) of the focal's membership function. The case of peak being zero corresponds to the case of total conflict in the D–S theory.

Our extension to Dempster's rule differs from Ishizuka's extension (discussed in Sect. 4) in its handling of subnormal intersections of focal elements. Ishizuka's degree of intersection $J(A, B)$ becomes $\max_{x_i} \mu_{A\cap B}(x_i)$ in (34) when both fuzzy set A and fuzzy set B are normal; therefore, it is analogous to the factor that scales down the basic probability in the normalization step of our approach. While we use the reciprocal of the factor to scale up the membership function of the focals' intersection, Ishizuka does not normalize the intersection. More importantly, Ishizuka's approach appeals to intuition without rigorous justification, whereas our approach is derived from the principle that the normalization step should preserve the relative probabilistic constraints imposed by focal elements, whether it is normal or not.[4]

[4] Obviously, the absolute probabilistic constraints of non-empty focal elements are not preserved by the normalization process because their basic probabilities are increased by the normalization factor (i.e., the denominator in (33)).

One of the most controversial issues regarding Dempster's rule of combination has been its normalization process. Zadeh, for instance, has questioned the validity of discarding the probability mass assigned to the empty set because the probability mass is an indication of the degree of conflict between the evidential sources that are combined [19]. However, to be consistent with axioms of probability theory, the probability of empty set has to be zero. In our approach, this dilemma is solved by delaying the normalization process. By computing the unnormalized BPA of the frame of discernment, our generalized rule of combination is able to use the basic probability of the empty set as a measure of the degree of conflict, which influences the credibility of the combined evidential sources. In the meantime, we can obtain the normalized BPA, which is needed for computing the belief function, by applying the normalization step to the unnormalized BPA. Hence, the generalized Dempster's rule not only allows the combination of vague evidential opinions, but also provides information regarding the credibility of the combined opinion.

6 Discussion

6.1 Consonant Focals and Fuzzy Focals

Several authors have discussed the similarity between possibility distribution and one specific instance of the D–S plausibility function called *consonant support function*—when the focal elements are nested, i.e., when they can be arranged in order so that each focal is contained in the following one [10]. Based on this observation, we have defined the probabilistic constraint of a fuzzy focal to be that of a set of consonant crisp focals in Sect. 5.4. Here, we will focus on the differences between the consonant focal elements and the fuzzy focal element.

A set of consonant focal elements differs from a fuzzy focal element in two important ways. First, consonant focal elements are more restrictive in the kinds of fuzzy evidential support they can represent. More specifically, they are limited to representing single vague evidential support. A fuzzy basic probability assignment (BPA), however, may consist of several fuzzy focal elements. Hence, it can express multiple fuzzy evidential supports. Second, each fuzzy focal element is induced by single evidential elements, while consonant focals are induced by several evidential elements that form an *inferential evidence* [9]. This difference between fuzzy focals and consonant focals explains their different comoination results. The combination of two consonant BPA's is a result of combining their evidential elements pairwise. Therefore, the combined focals are, in general, no longer consonant. However, the combination of two fuzzy focal elements, which involves the combination of underlying fuzzy compatibility relations, always yields another fuzzy focal element.

Due to these significant differences between fuzzy focals and consonant crisp focals, we should emphasize that we do not view fuzzy focal elements as identical to consonant crisp focals. In other words, the decomposition of a fuzzy focal element is not equivalent to the fuzzy focal itself. A fuzzy focal and its decomposition are only equivalent in the probabilistic constraints they imposed on the probability distribution of the frame of discernment.

7 Conclusion

We have described a generalization of the Dempster–Shafer theory to fuzzy sets. Rather than generalizing the formula for computing belief function, we generalize the basic constructs of the D–S theory: the compatibility relations, the objective functions of the optimization problem for calculating belief functions, and the probabilistic constraints imposed by focal elements. As a result, we can compute the lower probability (i.e., the belief function) directly from these generalized constructs. Moreover, by employing the noninteractive assumption in possibility theory, we have modified Dempster's rule to combine evidence that may be partially in conflict.

Our approach offers several advantages over previous work. First, the semantics of the D–S theory is maintained. Belief functions are treated as lower probabilities in our extension. Second, we avoid the problem of "choosing the right inclusion operators" faced by all previous approaches. Third, the generalized belief function is determined by the whole membership function of the focal element, not just by some critical points as used in some of the previous work. Any change of the membership function of a focal element is directly reflected in a change of the focal's probabilistic constraint, which in turn affects the belief function. Fourth, the generalized rule of combination provides information about the degree of conflict between the evidence combined by delaying the normalization step in original Dempster's rule. Finally, our generalization is well-justified using possibility theory and probability theory. Therefore, it serves as a bridge that brings together the Dempster–Shafer theory and fuzzy set theory into a hybrid approach to reasoning under various kinds of uncertainty in intelligent systems.

Acknowledgment

The author would like to thank Prof. Lotfi Zadeh for his continuous support and encouragement, which have made this work possible. Many thanks also go to Henri Prade, Piero Bonissone, and the reviewers for their valuable comments that help me in revising the paper.

References

[1] J. Pearl, “On evidential reasoning in a hierarchy of hypothesis,” *Artificial Intell.*, vol. 28, no. 1, Feb. 1986, pp. 9–16.

[2] S. A. Lesh, *An Evidential Theory Approach to Judgment-based Decision Making*, Ph.D. dissertation, Department of Forestry and Environmental Studies, Duke Univ., 1986.

[3] “Fuzzy control ensures a smooth ride,” Hitachi 1987, pp. 12–13.

[4] L. A. Zadeh, “Fuzzy sets and information granularity,” in *Advances in Fuzzy Set Theory and Applications*, 1979, pp. 3–18.

[5] M. Ishizuka, K. S. Fu, and J. T. P. Yao, “Inference procedures and uncertainty for the problem-reduction method,” *Inform. Sci.*, vol. 28, 1982, pp. 179–206.

[6] R. Yager, “Generalized probabilities of fuzzy events from fuzzy belief structures,” *Inform. Sci.*, vol. 28, 1982, pp. 45–62.

[7] H. Ogawa and K. S. Fu, “An inexact inference for damage assessment of existing structures,” *International Journal of Man-Machine Studies*, vol. 22, 1985, pp. 295–306.

[8] A. P. Dempster, “Upper and lower probabilities induced by a multivalued mapping,” *Ann. Math. Stat.*, vol. 38, pp. 325–339, 1967.

[9] G. Shafer, *Mathematical Theory of Evidence.* Princeton, N.J.: Princeton Univ. Press, 1976.

[10] ——, “Belief functions and possibility measures,” tech. report working paper no. 163, University of Kansas, School of Business, 1984.

[11] L. A. Zadeh, “Possibility theory and soft data analysis,” in *Mathematical Frontiers of the Social and Policy Sciences*, L. Cobb and R. M. Thrall, Eds. Boulder, CO: Westview Press, 1981, pp. 69–129.

[12] ——, “Fuzzy sets as a basis for a theory of possibility,” *Fuzzy Sets and Systems*, vol. 1, pp. 3–28, 1978.

[13] D. Dubois and H. Prade, *Possibility Theory.* New York: Plenum Press, 1988.

[14] P. Smets, “The degree of belief in a fuzzy event,” *Inform. Sci.*, vol. 25, pp. 1–19, 1981.

[15] L. A. Zadeh, “The concept of a linguistic variable and its application to approximate reasoning-I,” *Inform. Sci.*, vol. 8, pp. 199–249, 1975.

[16] ——, “Fuzzy languages and their relation to human and machine intelligence,” *Proc. Int. Conf. Man Comput.*, S. Karger, Ed.Bordeaux, France: Basel, 1972, pp. 130–165.

[17] ——, “Similarity relations and fuzzy orderings,” *Information Sci.*, vol. 3, pp. 177–200, 1971.

[18] D. Dubois and H. Prade, “On several representations of an uncertain body of evidence,” in *Fuzzy Information and Decision Processes*, M. M. Gupta and E. Sanchez, Ed. New York: North-Holland, 1982, pp. 167–181.

[19] L. A. Zadeh, “A simple view of the Dempster-Shafer theory of evidence and its implication for the Rule of Combination,” *AI Magazine*, vol. 7, no. 2, pp. 85–90, Summer 1986.

22

Bayesian Updating and Belief Functions

Jean-Yves Jaffray

Abstract. In a wide class of situations of uncertainty, the available information concerning the event space can be described as follows: There exists a true probability that is only known to belong to a certain set $\mathcal{P}$ of probabilities; moreover, the lower envelope f of $\mathcal{P}$ is a belief function, i.e., a nonadditive measure of a particular type, and characterizes $\mathcal{P}$, i.e., $\mathcal{P}$ is the set of all probabilities that dominate f. This is in particular the case when data result from large-scale sampling with incomplete observations. This study is concerned with the effect of conditioning on such situations. The natural conditioning rule is here the Bayesian rule: there exists a posterior probability after the observation of event E, and it is known to be located in $\mathcal{P}^E$, the set of conditionals of the members of $\mathcal{P}$. An explicit expression for the Möbius transform ϕ^E of f^E in terms of ϕ, the transform of f, is found and Fagin and Halpern's earlier finding that the lower envelope f^E of $\mathcal{P}^E$ is itself a belief function is derived from it. However, f^E no longer characterizes $\mathcal{P}^E$ (not all probabilities dominating f^E belong to it), unless f satisfy further stringent conditions that are both necessary and sufficient. The difficulties resulting from this fact are discussed and suggestions to cope with them are made.

1 Introduction

Consider a situation of uncertainty in which the existence of a probability measure, P, on the events can be hypothesized but, data being imprecise, P is only known to belong to some set of probability measures $\mathcal{P}$. This kind of situation is typically met in large-scale sampling with incomplete or vague observations.

Suppose that a certain event, E, is observed. It is natural to conclude that the true probability measure now belongs to the set, $\mathcal{P}^E$, formed by the conditionals of the members of $\mathcal{P}$ with respect to E. This change-over from $\mathcal{P}$ to $\mathcal{P}^E$, is called (*convex*) *Bayesian updating* by Kyburg [6].

In the case of sampling and, more generally, in all situations in which $\mathcal{P}$ is generated by a random set, $\mathcal{P}$ is characterizable by its lower envelope f, since $\mathcal{P}$ can be retrieved as the set of all probability measures which dominate f. Moreover, f is then necessarily a *belief function* [14].

The idea of representing such situations of uncertainty $\mathcal{P}$ by belief functions f is attractive, since these are more mathematically tractable. The introduction of belief functions with this interpretation is called the *lower probability approach* by Shafer [14].

The question then arises whether the representation property is preserved by Bayesian updating, or not. The only eligible candidate for representing $\mathcal{P}^E$ being its lower envelope, f^E, this question amounts to the following twofold one: is f^E a belief function and does it characterize $\mathcal{P}^E$?

The first part of the question has already been answered affirmatively by Fagin and Halpern [6]. The aim of this paper is to present a thorough study of the properties of f^E and $\mathcal{P}^E$, which provides an alternative proof of Fagin and Halpern's result, shows the answer to the second part of the question to be negative and, more precisely, defines the stringent conditions under which f^E does characterize $\mathcal{P}^E$.

The paper is organized as follows. In Sect. 2, we briefly recall some basic properties of belief functions. Section 3 defines Bayesian updating in this context and explores some ramifications. In Sect. 4, we provide illustrative examples that highlight some of the differences between Bayesian updating and the updating rule that is an integral part of Dempster–Shafer belief function theory, the so-called Dempster conditioning rule. Section 5 derives an explicit expression for the Möbius transform ϕ^E of f^E in terms of the Möbius transform ϕ of f, and it immediately follows as a corollary that f^E is a belief function. In Sect. 6, the relationship between the structure of $\mathcal{P}$ and $\mathcal{P}^E$ is studied, and necessary and sufficient conditions found for the representability of $\mathcal{P}^E$ by f^E. Difficulties resulting from the imperfect representation of $\mathcal{P}^E$ by f^E are examined and an alternative representation is considered.

2 Definition and Properties of Belief Functions and Related Objects

Let $\mathcal{X}$ be the finite set of *states of nature*, $\mathcal{A} = 2^{\mathcal{X}}$ the set of events, and $\mathcal{L}$ the set of all probability measures on $(\mathcal{X}, \mathcal{A})$.

A *belief function* [14] is a mapping $f : \mathcal{A} \to \mathbb{R}$ (actually: $\mathcal{A} \to [0,1]$) satisfying

$$f(\emptyset) = 0; \qquad f(\mathcal{X}) = 1 \tag{1}$$

$$A \subseteq B \Rightarrow f(A) \le f(B), \ \forall A, B \in \mathcal{A} \tag{2}$$

and

$$f\left(\bigcup_{i=1}^{k} A_i\right) \geq \sum_{\substack{I \subseteq \{1,2,\cdots,k\} \\ I \neq \emptyset}} (-1)^{|I|+1} f\left(\bigcap_{i \in I} A_i\right),$$

$$\forall A_i \in \mathcal{A},\ \forall k \geq 2. \tag{3}$$

A mapping $F:\ \mathcal{A} \rightarrow [0,1]$ associated with a belief function f by

$$F(A) = 1 - f(A^c),\ \text{for All } A \in A \tag{4}$$

is a *plausibility function.*

Any mapping $f: \mathcal{A} \rightarrow \mathbb{R}$ has a *Möbius transform* [13] $\phi: \mathcal{A} \rightarrow \mathbb{R}$ defined by

$$\phi(A) = \sum_{B \subseteq A} (-1)^{|A \backslash B|} f(B),\ \forall A \in \mathcal{A} \tag{5}$$

which characterizes f since

$$f(A) = \sum_{B \subseteq A} \phi(B), \quad \forall A \in \mathcal{A} \tag{6}$$

Note that

$$F(A) = \sum_{B \cap A \neq \emptyset} \phi(B), \quad \forall A \in \mathcal{A} \tag{7}$$

Shafer [14] defines a basic probability assignment (BPA) on $\mathcal{A}$ as a mapping $\phi: \mathcal{A} \rightarrow [0,1]$ satisfying

$$\phi(\emptyset) = 0; \sum_{A \in \mathcal{A}} \phi(A) = 1 \tag{8}$$

and proves the following proposition.

Proposition 1. *A mapping $f: \mathcal{A} \rightarrow \mathbb{R}$ is a belief function if and only if its Möbius transform ϕ is a BPA.*

The *focal set* of a belief function f is $\mathcal{B} = \{B \in \mathcal{A} : \phi(B) \neq 0\}$ and $\mathcal{C} = \cup_{B \in \mathcal{B}} B$ is its *core*. Real numbers $\lambda(B,x), B \in \mathcal{B}, x \in B$, satisfying

$$\lambda(B,x) \geq 0, \forall\ x \in B,\ \text{and} \sum_{x \in B} \lambda(B,x) = 1, \forall\ B \in \mathcal{B}, \tag{9}$$

define a probability measure $P_\lambda \in \mathcal{L}$ by

$$P_\lambda(\{x\}) = \sum\nolimits_{B \in \mathcal{B}, B \supseteq \{x\}} \lambda(B,x)\, \phi(B), \quad \forall\ x \in \mathcal{X} \tag{10}$$

where P_λ is called an *allocation* of BPA ϕ.

On the other hand, a probability measure $P \in \mathcal{L}$ is said to be *compatible* with f (or to *dominate* f) when

$$P(A) \geq f(A), \quad \forall\, A \in \mathcal{A}$$

(for short: $P \geq f$). As proven by Strassen [18], we get the following.

Proposition 2. *A probability measure $P \in \mathcal{L}$ is compatible with a belief function f if and only if P is an allocation of f's Möbius transform ϕ.*

By considering suitable allocations, it can then easily be shown that:

Proposition 3. *For every decreasing sequence of events $(A_i, l \leq i \leq m)$ there exists a probability measure P compatible with belief function f and such that*

$$P(A_i) = f(A_i), \text{ for } 1 \leq i \leq m.$$

Introducing the set of all probability measures dominating f:

$$\mathcal{P}_f = \{P \in \mathcal{L} : P \geq f\} \tag{11}$$

a straightforward corollary of Proposition 3 can be stated as follows.

Corollary 1. *Let f be a belief function; then*

$$1)\ f = \inf_{P \in \mathcal{P}_f} P \left(\textit{short for: } f(A) = \inf_{P \in \mathcal{P}_f} P(A),\ \textit{for all } A \in \mathcal{A} \right) \tag{12}$$

and, more precisely, for every $A \in \mathcal{A}$, there exists $P \in \mathcal{P}_f$ such that $P(A) = f(A)$;

2) For all $A, B \in \mathcal{A}$ such that $A \cap B = \emptyset$, there exists $P \in \mathcal{P}_f$ such that $P(A) = f(A)$ and $P(B) = F(B)$.

3 Belief Function Representation of Uncertainty and Bayesian Updating

3.1 Representation of Uncertainty Situations by Belief Functions

Let data concerning the events be completely summarized by the specification of $\mathcal{P}$, a nonempty subset of $\mathcal{L}$. In other words, suppose a true probability measure be known to exist and to be located in $\mathcal{P}$, but no member of $\mathcal{P}$ be more likely to be the true one than another member (*complete ignorance* on the location of P in $\mathcal{P}$).

In general, $\mathcal{P}$ is not characterizable by its *lower envelope*

$$f = \inf_{P \in \mathcal{P}} P, \tag{13}$$

since the set that can be retrieved from f, by (11), is the set of all compatible measures, $\mathcal{P}_f$, which may strictly contain $\mathcal{P}$. Furthermore, f is not necessarily a belief function (see, e.g., Examples 6 and 7 in Chateauneuf and Jaffray [2]).

However, in a broad class of situations, the set $\mathcal{P}$ of probability measures that are consistent with data is generated by a random set [10] and thus its lower envelope $f = \inf_{P \in \mathcal{P}} P$ automatically satisfies the following: 1) f is a belief function; and 2) $\mathcal{P} = \mathcal{P}_f$, i.e., $\mathcal{P} = \{P \in \mathcal{L} : \{P \geq f\}$.

This is in particular the case whenever data are collected by large-scale sampling in an imprecise or incomplete way [8].

Note that, by Corollary 1, 1) and 2) imply that $f = \inf_{P \in \mathcal{P}} P$.

Whenever 1) and 2) are valid, the situation of uncertainty described by $\mathcal{P}$ can be represented by its lower envelope, belief function f.

Remark 1. Belief functions were originally introduced by Dempster ([3], [4]) in relation with random sets (see Sect. 3.4). This framework was later abandoned by Shafer who favored a purely subjective evaluation of the belief in an event. Dempster–Shafer theory is thoroughly developed in [14]; it is also described in Pearl [11, ch. 9] (with a particular interpretation). Shafer's [15] presentation of its own point of view and discussion of various interpretations, and Pearl's [12] critical remarks (with comments by others in the field) might also be found instructive.

3.2 Bayesian Updating

Given an initial situation characterized by set $\mathcal{P}$, let an event E be observed. Provided

$$P\left(E\right) > 0, \quad \forall\ P \in \mathcal{P} \tag{14}$$

the conditional probability measure P given E, denoted P^E, is well defined, for every $P \in \mathcal{P}$, on the subalgebra of events:

$$\mathcal{A}^E = \{A \in \mathcal{A} : A \subseteq E\} \tag{15}$$

by Bayes rule:

$$P^E\left(A\right) = \frac{P\left(A\right)}{P\left(E\right)}, \forall\ A \in \mathcal{A}^E. \tag{16}$$

In accordance with the interpretation of the initial information, the new situation of uncertainty, after the observation of E, can be summarized by

$$\mathcal{P}^E = \left\{P^E \in \mathcal{L}^E : P \in \mathcal{P}\right\} \tag{17}$$

($\mathcal{L}^E$ denotes the set of all probability measures on $(E, \mathcal{A}^E)$). The transformation $\mathcal{P} \mapsto \mathcal{P}^E$ is called *Bayesian updating.*[1]

[1] Spies [17] presents a justification of the same conditioning rule based on the theory of conditional objects.

3.3 Belief Function Representation and Bayesian Updating

When the initial situation of uncertainty is *representable* by a belief function f, i.e., when assumptions 1) and 2) are satisfied, the question arises whether $\mathcal{P}^E$, which describes the new situation of uncertainty after event E has been observed, can itself be represented by a belief function.

Note that $\mathcal{P}^E$ is defined if and only if $f(E) > 0$, since, by Corollary 1, this condition is equivalent to (14).

As already remarked in Sect. 3.1, the only candidate for representing $\mathcal{P}^E$ is its lower envelope

$$f^E = \inf_{Q \in \mathcal{P}^E} Q = \inf_{P \in \mathcal{P}} P^E. \tag{18}$$

There exists a simple relation, first noticed by Dempster [3], between the lower envelopes f and f^E.

Proposition 4. *Given belief function f on $\mathcal{A}$ and event $E \in \mathcal{A}$ such that $f(E) > 0$, let $\mathcal{P} = \{P \in \mathcal{L} : P \geq f\}$ and $f^E = \inf_{P \in \mathcal{P}} P^E$. Then*

$$f^E(A) = \frac{f(A)}{f(A) + F(E \backslash A)} = \frac{f(A)}{f(A) + 1 - f(A \cup E^c)}, \ \forall\, A \in \mathcal{A}^E \tag{19}$$

moreover: 1) for every $A \in \mathcal{A}$, there exists $P \in \mathcal{P}$ such that $P^E(A) = f^E(A)$; and 2) if $f(A) > 0$ and $F(E \backslash A) > 0$, then

$$P^E(A) = f^E(A) \text{ if } P(A) = f(A) \text{ and } P(E \backslash A) = F(E \backslash A).$$

Proof. Let $A \in \mathcal{A}^E$. By definition

$$f^E(A) = \inf_{P \in \mathcal{P}} \frac{P(A)}{P(A) + P(E \backslash A)}.$$

If $f(A) = 0$, there exists, by Corollary 1, $P \in \mathcal{P}$ such that $P(A) = 0$; thus $f^E(A) = 0$, (19) holds, and 1) is true.

If $f(A) > 0$ and $F(E \backslash A) = 0$, then $P(E \backslash A) = 0$ for all $P \in \mathcal{P}$; thus $f^E(A) = 1$, (19) holds, and 1) is true.

In the remaining case, $f(A) > 0$ and $F(E \backslash A) > 0$. Since $P(A) > 0$ for all $P \in \mathcal{P}$, $f^E(A) = [1 + \sup_{P \in \mathcal{P}} (P(E \backslash A))/(P(A))]^{-1}$; moreover, $(P(E \backslash A))/(P(A)) \leq (F(E \backslash A))/(f(A))$ for all $P \in \mathcal{P}$ and the equality is achieved if and only if $P(A) = f(A)$ and $P(E \backslash A) = F(E \backslash A)$. To complete the proof, it is sufficient to remark that Corollary 1 asserts the existence of $P \in \mathcal{P}$ with these properties.

3.4 Relation with Dempster's Rule

Dempster–Shafer theory [3], [4], [14] uses a different updating rule, Dempster's conditioning rule.

There is a possible justification of Dempster's rule, based on the random set interpretation of belief functions. In this interpretation, information comes from sources emitting random messages. A unique source emitting message "$x \in B$" with probability $\phi(B)$ generates belief $f(A) = \Sigma_{B \subseteq A}\phi(B)$ that $x \in A$.

Suppose now that, beside this first source, there is a second one, emitting message "$x \in E$" with certainty, and that messages must be consistent, i.e., $B' = B \cap E$ cannot be empty. Then, the probability of receiving messages jointly asserting that "$x \in B'$" given that messages are always consistent (an event that has probability $\sum_{\substack{B \in \mathcal{B} \\ B \cap E \neq \emptyset}} \phi(B) = F(E)$) is

$$\psi^E(B') = \frac{1}{F(E)} \sum_{\substack{B \in \mathcal{B} \\ B \cap E = B'}} \phi(B). \quad \forall\, B' \in \mathcal{A}^E \setminus \{\emptyset\}. \tag{20}$$

This message distribution generates belief function g^E given by

$$g^E(A) = \frac{f(A \cup E^c) - f(E^c)}{1 - f(E^c)} = \frac{F(E) - F(E \setminus A)}{F(E)} \ \forall\, A \in \mathcal{A}^E \tag{21}$$

(note that g^E is defined as soon as $F(E) > 0$).

Clearly, g^E is different from f^E in general. More precisely, as observed by Dempster [3] and Kyburg [6], it follows from (19) and (21), and inequality $f(A \cup E^c) \geq f(A) + f(E^c)$, that $g^E \geq f^E$.

In fact, Dempster's rule $f \mapsto g^E$ records the effect of additional information, whereas the Bayesian rule corresponds to the selection of relevant

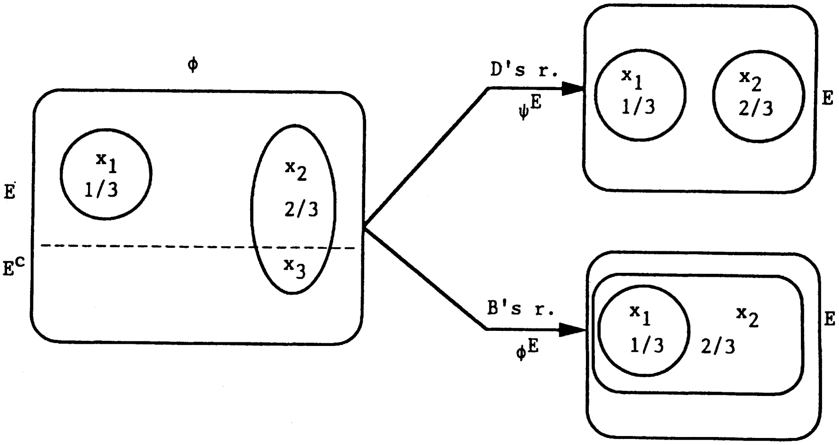


Fig. 1.

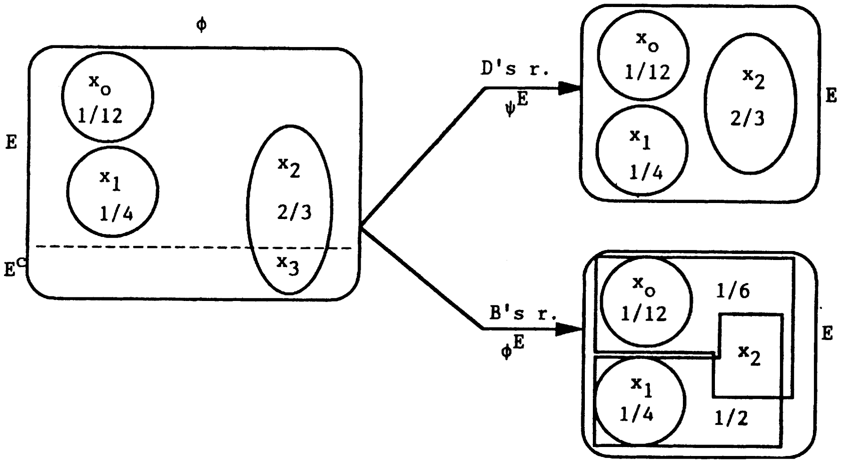


Fig. 2.

information. For example, g^E would describe updated knowledge after learning that a human population studied only consists of children and $\mathcal{P}^E$ would answer the question "what is known concerning children?"

Thus Dempster's rule and the Bayesian rule address different problems and should not be opposed. It is nonetheless clear, at least in the lower probability interpretation of belief functions, that the appropriate rule of conditioning proper is the Bayesian rule.

4 Examples

In Sect. 5, we shall prove that f^E is indeed a belief function by using Proposition 1; more precisely, we shall construct a certain BPA, ϕ^E, derived from the Möbius transform, ϕ, of f, and prove that ϕ^E is really the Möbius transform of f^E by merely checking the validity of (6).

The relation between ϕ^E and ϕ is more intricate than that between ψ^E and ϕ (relation (20)) and will be best understood by first studying a few examples. Their analysis will clarify the difference between the Bayesian rule and Dempster's rule.

Example 1 shows that Bayesian conditioning requires that, for $B \in \mathcal{B}$ that meets both E and E^C, weight $\phi(B)$ be transferred, not to $B \cap E$ as in Dempster's rule, but to larger subsets of $\mathcal{C} \cap E$.

Example 2 further indicates that, in Bayesian conditioning, where there is only one set $B'' \in \mathcal{B}$ that meets both E and E^C, its weight $\phi(B'')$ is allocated to sets $B' \cup (B'' \cap E)$, where $B' \in \mathcal{B}$ and $B' \subseteq E$, proportionally to weights $\phi(B')$, thus taking into account correctly the renormalization factor of Bayes' rule. However, when several members of $\mathcal{B}$ meet both E and E^C, the real location becomes involved as shown by a third example.

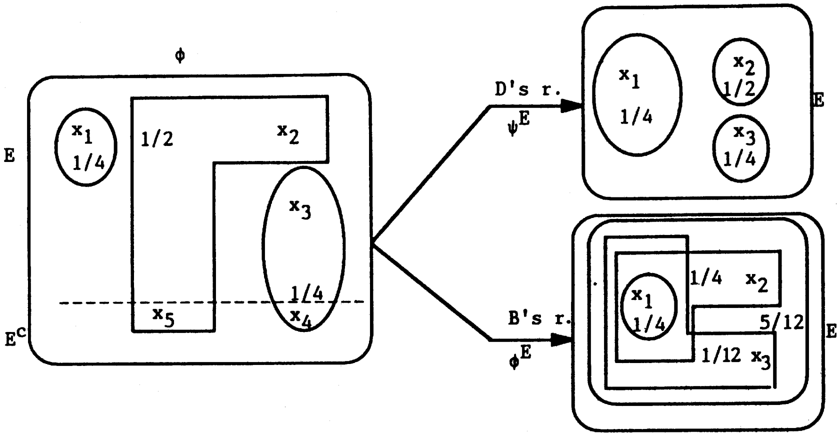


Fig. 3.

Moreover the second and third examples exhibit sets of conditional measures $\mathcal{P}^E$, which are proper subsets of $\mathcal{P}^E_{f^E} = \{Q \in \mathcal{L}^E : Q \geq f^E\}$, and, thereby prove that 2) is not satisfied in general.

Note: In the examples, we use abridged notations such as $x_1x_2x_3$ for $\{x_1, x_2x_3\}$ and thus $\mathcal{B} = \{x_1, x_2, x_3\}$ for $\mathcal{B} = \{\{x_1\}, \{x_2, x_3\}\}$, etc.

Example 1. $\chi = x_1x_2x_3$; $\mathcal{P} = \{P \in \mathcal{L} : P(x_1) = 1/3\}$; thus $\mathcal{P} = \mathcal{P}_f$, where f is given by $f(x_1) = 1/3$, $f(x_2) = f(x_3) = 0$, $f(x_1x_2) = f(x_1x_3) = 1/3$ and $f(x_2x_3) = 2/3$; hence f is a belief function since, by (5): $\mathcal{B} = \{x_1, x_2x_3\}$, $\phi(x_1) = 1/3$ and $\phi(x_2x_3) = 2/3$.

For $E = x_1x_2$, $f(E) = 1/3$, $F(E) = 1$, and $\mathcal{P}^E = \{Q \in \mathcal{L}^E : 1/3 \leq Q(x_1) \leq 1\}$; thus $f^E(x_1) = 1/3$ and $f^E(x_2) = 0$, and f^E is a belief function since, by (5), $\phi^E(x_1) = 1/3$, $\phi^E(x_2) = 0$ and $\phi^E(E) = 2/3$, hence $\phi^E \geq 0$. It is obvious here that $\mathcal{P}^E = \mathcal{P}^E_{f^E}$.

Note that Dempster's rule would lead to g^E, characterized by BPA ψ^E such that $\psi^E(x_1) = 1/3$ and $\psi^E(x_2) = 2/3$; thus g^E is the probability measure defined by $g^E(x_1) = 1/3$ and $g^E(x_2) = 2/3$, and is only compatible with itself.

Example 2. $\chi = x_0x_1x_2x_3$. $\mathcal{P} = \{P \in \mathcal{L} : P(x_0) = 1/12,\ P(x_1) = 1/4\}$; thus $\mathcal{P} = \mathcal{P}_f$, where f is given by $f(x_0) = 1/12$, $f(x_1) = 1/4$, $f(x_2) = f(x_3) = 0$, $f(x_0x_1) = 1/3$, $f(x_0x_2) = f(x_0x_3) = 1/12$, $f(x_1x_2) = f(x_1x_3) = 1/4$, $f(x_2x_3) = 2/3$, $f(x_1x_2x_3) = 11/12$, $f(x_0x_2x_3) = 3/4$ and $f(x_0x_1x_3) = f(x_0x_1x_2) = 1/3$; hence f is a belief function since, by (5), $\mathcal{B} = \{x_0, x_1, x_2x_3\}$, $\phi(x_0) = 1/12$, $\phi(x_1) = 1/4$, and $\phi(x_2x_3) = 2/3$.

For $E = x_0x_1x_2$, $f(E) = 1/3$, $F(E) = 1$, and $\mathcal{P}^E = \{Q \in \mathcal{L}^E : Q(x_0) = \alpha, Q(x_1) = 3\alpha,\ 1/12 \leq \alpha \leq 1/4\}$. Hence $f^E(x_0) = 1/12$, $f^E(x_1) = 1/4$, $f^E(x_2) = 0$, $f^E(x_0x_1) = 1/3$, $f^E(x_0x_2) = 1/4$, and $f^E(x_1x_2) = 3/4$, and f^E is a belief function since, by (5), $\mathcal{B}^E = \{x_0, x_1, x_0x_2, x_1x_2\}$ with $\phi^E(x_0) = 1/12,\ \phi^E(x_1) = 1/4,\ \phi^E(x_0x_2) = 1/6$, and $\phi^E(x_1x_2) = 1/2$.

$\mathcal{P}^E$ is only a proper subset of $\mathcal{P}^E_{f^E}$, since $\mathcal{P}^E_{f^E} = \{Q \in \mathcal{L}^E : Q(x_0) \geq 1/12,\ Q(x_1) \geq 1/4,\ Q(x_0x_2) \geq 1/4,\ Q(x_1x_2) \geq 3/4\}$, and $Q \in \mathcal{L}^E$ defined by $Q(x_0) = 1/12,\ Q(x_1) = 3/4$ and $Q(x_2) = 1/6$ belongs to the latter set but not to the former since $Q(x_1) \neq 3Q(x_0)$.

Note that Dempster's rule would lead to g^E, characterized by BPA ψ^E such that $\psi^E(x_0) = 1/12,\ \psi^E(x_1) = 1/4$, and $\psi^E(x_2) = 2/3$; thus g^E is the probability measure defined by $g^E(x_0) = 1/12,\ g^E(x_1) = 1/4$ and $g^E(x_2) = 2/3$ and is only compatible with itself.

Example 3. $x = x_1x_2x_3x_4x_5$. $\mathcal{P} = \{P \in \mathcal{L} : P(x_2x_5) = 1/2,\ P(x_3x_4) = 1/4\}$; thus $\mathcal{P} = \mathcal{P}_f$, where f is given by $f(x_1) = 1/4,\ f(x_j) = 0$ for $j \neq 1,\ f(x_1x_j) = 1/4$ for $j \neq 1,\ f(x_2x_3) = f(x_2x_4) = 0,\ f(x_2x_5) = 1/2,\ f(x_3x_4) = 1/4,\ f(x_3x_5) = f(x_4x_5) = 0,\ f(x_1x_2x_3) = f(x_1x_2x_4) = f(x_1x_3x_5) = f(x_1x_4x_5) = 1/4,\ f(x_1x_2x_5) = 3/4,\ f(x_1x_3x_4) = 1/2,\ f(x_2x_3x_4) = 1/4,\ f(x_2x_3x_5) = f(x_2x_4x_5) = 1/2,\ f(x_3x_4x_5) = 1/4,\ f(x_2x_3x_4x_5) = 3/4\ f(x_1x_3x_4x_5) = 1/2,\ f(x_1x_2x_4x_5) = f(x_1x_2x_3x_5) = 3/4$, and $f(x_1x_2\ x_3x_4) = 1/2$; hence f is a belief function since, by (5), $\mathcal{B} = \{x_1, x_2x_5, x_3x_4\}$, $\phi(x_1) = 1/4,\ \phi(x_2x_5) = 1/2$, and $\phi(x_3x_4) = 1/4$.

For $E = x_1x_2x_3,\ f(E) = 1/4,\ F(E) = 1$, and $\mathcal{P}^E = \{Q \in \mathcal{L}^E : Q(x_1) = k,\ Q(x_2) = \alpha k,\ Q(x_3) = \beta k,\ k = (1 + \alpha + \beta)^{-1},\ 0 \leq \alpha \leq 2,\ 0 \leq \beta \leq 1\}$. Hence, $f^E(x_1) = 1/4,\ f^E(x_2) = f^E(x_3) = 0,\ f^E(x_1x_2) = 1/2,\ f^E(x_1x_3) = 1/3$, and $f^E(x_2x_3) = 0$, and f^E is a belief function, since, by (5), $\mathcal{B}^E = \{x_1, x_1x_2, x_1x_3, E\}$, with $\phi^E(x_1) = \phi^E(x_1x_2) = 1/4,\ \phi^E(x_1x_3) = 1/12$, and $\phi^E(E) = 5/12$.

Here again, $\mathcal{P}^E \subset \mathcal{P}^E_{f^E}$, since for $Q(x_1) = 1/4,\ Q(x_2) = 2/3,\ Q(x_3) = 1/12,\ Q \in \mathcal{P}^E_{f^E}$ and $Q \notin \mathcal{P}^E$. Note also that Dempster's rule would lead to g^E, characterized by BPA ψ^E such that $\psi^E(x_1) = \psi^E(x_3) = 1/4$ and $\psi^E(x_2) = 1/2$; thus g^E is the probability measure defined by $g^E(x_1) = g^E(x_3) = 1/4$ and $g^E(x_2) = 1/2$, and is only compatible with itself.

Thus Bayesian conditioning seems to require that, when exactly two members B''_1 and B''_2 of $\mathcal{B}$ meet both E and E^c, the proportion of weight $\phi(B''_1)$, which would have been allocated to $(B''_1 \cap E) \cup B''_2$, had B''_2 be a subset of E, be in fact allocated to sets $B' \cup ((B''_1 \cap E) \cup (B'' \cap E) = B' \cup ((B''_1 \cup B''_2) \cap E)$, where $B' \in \mathcal{B}$ and $B' \subseteq E$; moreover, in the light of Example 2, the allocation should be proportional to $\phi(B')$, when there is more than one B' (which is not the case in Example 3).

Let us finally add that, like Dempster's rule, Bayes' rule clearly requires a normalization factor $1/(F(E))$ when $F(E) < 1$.

In the following theorem, it is proven that the general relation between ϕ and ϕ^E is indeed that suggested by the examples.

5 Lower Envelopes of Sets of Probability Measures Remain Belief Functions after Conditioning

Theorem 1. *Let f be a belief function on $\mathcal{A}$, and F the associated plausibility function; let ϕ be its Möbius transform, $\mathcal{B}$ its focal set, and let $\mathcal{P} = \{P \in \mathcal{L} : P \geq f\}$. Let event E of $\mathcal{A}$ satisfy $f(E) > 0$, and consider $\mathcal{P}^E = \{P^E \in \mathcal{L}^E : P \in \mathcal{P}\}$, where P^E denotes the conditional of P given E; then the lower envelope, $f^E = \inf_{Q \in \mathcal{P}^E} Q$, of $\mathcal{P}^E$ is a belief function.*[2]

Moreover: 1) The focal set, $\mathcal{B}^E$, of f^E is related to $\mathcal{B}$ and E as follows: let $\mathcal{B}' = \{B' \in \mathcal{B} : B' \subseteq E\}, \mathcal{B}'' = \{B'' \in \mathcal{B} : B'' \cap E \neq \emptyset$ and $B'' \cap E^c \neq \emptyset\}$, $K = |\mathcal{B}''|$, and let set of finite sequences of events $\mathcal{T}(B)$ be defined, for all $B \in \mathcal{A}^E$, by

$$\begin{aligned} \mathcal{T}(B) = \{T = (B', B_1'', B_2'', \cdots, B_k'') :& \\ B = B' \cup \left[\left(B_1'' \cup B_2'' \cup \ldots \cup B_k''\right) \cap E\right],& \\ B' \in \mathcal{B}', B_\ell'' \in \mathcal{B}'', \text{ all } B_\ell'' \text{ distinct},& \\ \ell = 1, 2, \cdots, k, 0 \leq k \leq K\};& \end{aligned} \tag{22}$$

then $\mathcal{B}^E = \{B \in \mathcal{A}^E : \mathcal{T}(B) \neq \emptyset\}$. 2) *The Möbius transform*, ϕ^E, of f^E is the BPA defined by

$$\phi^E(B) = \sum_{T \in T(B)} m(T), \quad \forall\, B \in \mathcal{B}^E \tag{23}$$

where, for $T = (B', B_1'', B_2'', \cdots, B_k'')$, (24) results.

$$m(T) = \frac{\phi(B') \times \phi(B_1'') \times \phi(B_2'') \times \cdots \times \phi(B_k'')}{F(E)\left[F(E) - \phi(B_1'')\right]\left[F(E) - \phi(B_1'') - \phi(B_2'')\right] \times \cdots \times \left[F(E) - \sum_{\ell=1}^k \phi(B_\ell'')\right]\Big]} \tag{24}$$

Proof. 1) According to (6), ϕ^E, defined by (23) and (24) on $\mathcal{B}^E$, and equal to zero elsewhere, is the Möbius transform of f^E if and only if

$$f^E(A) = \sum_{B \subseteq A} \phi^E(B) = \sum_{B \subseteq A} \sum_{T \in T(B)} m(T), \quad \forall\, A \in \mathcal{A}^E. \tag{25}$$

or, equivalently, by (19), if and only if

[2] Fagin and Halpern [6] have already given a proof of this result. However our proof (found independently), although based on the same idea as theirs, is quite different, since we find the explicit expression (23) and (24) of the Möbius transform ϕ^E of f^E, which drastically simplifies the rest of the proof. There is another proof, by Zhang [20], which does not involve Möbius inversion.

$$\frac{f\left(A\right)}{f\left(A\right)+F\left(E\backslash A\right)}=\sum_{B\subseteq A}\sum_{T\in T(B)}m\left(T\right),\quad\forall\ A\in\mathcal{A}^{E}.\tag{26}$$

Let us prove the validity of relation (26) for any $A\in\mathcal{A}^{E}$.

2) Given $A\in\mathcal{A}^{E}$, the double summation on the right-hand side of (26) can be reexpressed as

$$\sum_{B'\in\mathcal{B}'(A)}\ \sum_{\left(B_1'',B_2'',\cdots,B_k''\right)\in\mathcal{S}(\emptyset)}m\left(B',B_1'',B_2'',\cdots,B_k''\right),$$

where $\mathcal{B}'(A)=\{B'\in\mathcal{B}':B'\subseteq A\}$ and $\mathcal{S}(\emptyset)$ is the set of all finite sequences of distinct events of $\mathcal{B}''(A)=\{B''\in\mathcal{B}'':B''\cap E\subseteq A\}$, since $B=B'\cup[(B_1''\cup B_2''\cup\cdots\cup B_k'')\cap E]\subseteq A$ iff $B'\subseteq A$ and $B_\ell''\cap E\subseteq A$, for $1\leq\ell\leq k$.

Since $\mathcal{B}'(A)=\emptyset$ implies that $f(A)=f^{E}(A)=0$, hence that (26) is valid, it will be assumed henceforth that $\mathcal{B}'(A)\neq\emptyset$, but only that $|\mathcal{B}''(A)|=L\geq 0$.

3) In order to further decompose the summation, let us define subsets of $\mathcal{L}(\emptyset)$ with a common initial sequence of length $1\leq\ell\leq L$,

$$\begin{aligned}\mathcal{S}\left(B_1'',B_2'',\cdots,B_\ell''\right)=\Big\{\left(\overline{B_1}'',\overline{B_2}'',\cdots,\overline{B_k}''\right)&\in\mathcal{S}\left(\emptyset\right)\ :\\ k\geq\ell\text{ and }\overline{B_j}''&=B_j'',1\leq j\leq\ell\}.\end{aligned}$$

and denote the corresponding partial summation of weights, for a given $B'\in\mathcal{B}'(A)$:

$$\mathcal{S}\left(B',B_1'',B_2'',\cdots,B_\ell''\right)\qquad\sum_{=\frac{\left(\overline{B_1}'',\overline{B}_2'',\cdots,\overline{B}_k''\right)\in\mathcal{S}\left(B_1'',B_2'',\cdots,'',\cdots,B_\ell''\right)}{m\left(B',\overline{B}_1'',\overline{B}_2'',\cdots,\overline{B}_\ell''\right)}}\tag{27}$$

in particular, $S(B')$ corresponds to $\mathcal{S}(\emptyset)$, so that (26) becomes

$$\frac{f\left(A\right)}{f\left(A\right)+F\left(E\backslash A\right)}=\sum_{B'\in\mathcal{B}'(A)}\mathcal{S}(B')\tag{28}$$

We shall calculate $S(B')$ by induction:

4) Let $(B_1'',B_2'',\cdots,B_L'')$ be a given ordering of $\mathcal{B}''(A)$, and let B' be a given event of $\mathcal{B}'(A)$.

Every sequence of events in $\mathcal{S}(B_1'',B_2'',\cdots,B_\ell'')$, $0\leq\ell\leq L-1$, except $(B_1'',B_2'',\cdots,B_\ell'')$, has at least a $(\ell+1)$th term, B_j'', where $\ell+1\leq j\leq L$; therefore

$$\begin{aligned}S\left(B',B_1'',B_2'',\cdots,B_\ell''\right)&=m\left(B',B_1',B_2'',\cdots,B_\ell''\right)\\&\quad+\sum_{j=\ell+1}^{L}S\left(B',B_1'',B_2'',B_2'',\cdots,B_\ell'',\cdots,B_\ell'',B_j''\right).\\&\quad\text{for }0\leq\ell\leq L-1.\end{aligned}\tag{29}$$

This induction formula has a unique solution

$$\begin{aligned} S\left(B', B_1'', B_2'', \cdots, B_\ell''\right) = & m\left(B', B_1'', B_2'', \cdots, B_\ell''\right) \\ & \cdot \frac{F(E) - \sum_{j=1}^{\ell} \phi\left(B_j''\right)}{F(E) - \sum_{j=1}^{L} \phi\left(B_j''\right)}, \\ & \text{for } 0 \le \ell \le L \end{aligned} \tag{30}$$

since (α) $S(B', B_1'', B_2'', \cdots, B_L'') = m(B', B_1'', B_2'', \cdots, B_L'')$ (β) If (30) holds at order $(\ell + 1)$, then, by (24):

$$\begin{aligned} & m\left(B', B_1'', B_2'', \cdot, B_\ell''\right) + \sum_{j=\ell+1}^{L} \mathcal{S}\left(B', B_1'', B_2'', \cdot, B_\ell'', B_j''\right) \\ & = m\left(B', B_1'', B_2'', \cdot, B_\ell''\right) \left[1 + \sum_{j=\ell+1}^{L} \frac{\phi\left(B_j''\right)}{F(E) - \sum_{j=1}^{L} \phi\left(B_j''\right)}\right] \\ & = m\left(B', B_1'', B_2'', \cdots, B_\ell''\right) \frac{F(E) - \sum_{j=1}^{\ell} \phi\left(B_j''\right)}{F(E) - \sum_{j=1}^{L} \phi\left(B_j''\right)} \end{aligned}$$

and thus (30) also holds at order ℓ.
In particular, for $\ell = 0$, (30) states that

$$\begin{aligned} \mathcal{S}\left(B'\right) & = m\left(B'\right) \frac{F(E)}{F(E) - \sum_{j=1}^{L} \phi\left(B_j''\right)} \\ & = \frac{\phi\left(B'\right)}{F(E) - \sum_{j=1}^{L} \phi\left(B_j''\right)}. \end{aligned} \tag{31}$$

5) It remains only to compare

$$\sum_{B' \in \mathcal{B}'(A)} \mathcal{S}\left(B'\right) = \frac{\sum_{B' \in \mathcal{B}'(A)} \phi\left(B'\right)}{F(E) - \sum_{j=1}^{L} \phi\left(B_j''\right)}$$

and

$$\frac{f(A)}{f(A) + F(E \backslash A)}$$

It is straightforward that the numerator is equal to $f(A)$; as for the denominator:

$$\begin{aligned} F(E) - \sum_{j=1}^{L} \phi\left(B_j''\right) & = F(E) - \left[f\left(A \cup E^c\right) - f\left(E^c\right) - f(A)\right] \\ & = f(A) + F(E \backslash A). \end{aligned}$$

Thus, (28) or, equivalently (25), holds for all $A \in \mathcal{A}^E$, and ϕ^E is the Möbius transform of f; moreover, since ϕ^E is nonnegative, it is a BPA, and f^E is a belief function.

Examples 1, 2, and 3 (continued): Let us illustrate Theorem 1 with the preceding examples (abridged notations are still used).

Example 4. $\mathcal{B}' = \{x_1\};\ \mathcal{B}'' = \{x_2x_3\};\ \mathcal{B}^E = \{x_1, x_1x_2\}$.

$$m(x_1) = \phi(x_1) = 1/3; m(x_1, x_2x_3) = \frac{\phi(x_1)\,\phi(x_2x_3)}{1-\phi(x_2x_3)} = 2/3.$$

$$\phi^E(x_1) = m(x_1) = 1/3 \text{ and } \phi^E(x_1x_2) = m(x_1, x_2x_3) = 2/3.$$

Example 5. $\mathcal{B}' = \{x_0, x_1\};\ \mathcal{B}'' = \{x_2x_3\};\ \mathcal{B}^E = \{x_0, x_1, x_0x_2, x_1x_2\}$.

$$m(x_0) = \phi(x_0) = 1/12; m(x_0, x_2x_3) = \frac{\phi(x_0)\,\phi(x_2x_3)}{1-\phi(x_2x_3)} = 1/6;$$

$$m(x_1) = \phi(x_1) = 1/4; m(x_1, x_2x_3) = \frac{\phi(x_1)\,\phi(x_2x_3)}{1-\phi(x_2x_3)} = 1/2.$$

$$\phi^E(x_0) = m(x_0) = 1/12; \phi^E(x_0x_2) = m(x_0, x_2x_3) = 1/6;$$

$$\phi^E(x_1) = m(x_1) = 1/4; \phi^E(x_1x_2) = m(x_1, x_2x_3) = 1/2.$$

Example 6. $\mathcal{B}' = \{x_1\};\ \mathcal{B}'' = \{x_2x_5, x_3x_4\};\ \mathcal{B}^E = \{x_1, x_1x_2, x_1x_3; E\}$.

$$\begin{aligned}
m(x_1) &= \phi(x_1) = 1/4; m(x_1, x_2x_5)\\
&= \frac{\phi(x_1)\,\phi(x_2x_5)}{1-\phi(x_2x_5)}\\
&= 1/4;\\
m(x_1, x_3x_4) &= \frac{\phi(x_1)\,\phi(x_3x_4)}{1-\phi(x_3x_4)} = 1/12;\\
m(x_1, x_2x_5, x_3x_4) &= \frac{\phi(x_1)\,\phi(x_2x_5)\,\phi(x_3x_4)}{[1-\phi(x_2x_5)]\,[1-\phi(x_2x_5)-\phi(x_3x_4)]}\\
&= 1/4;\\
m(x_1, x_3x_4, x_2x_5) &= \frac{\phi(x_1)\,\phi(x_3x_4)\,\phi(x_2x_5)}{[1-\phi(x_3x_4)]\,[1-\phi(x_3x_4)-\phi(x_2x_5)]}\\
&= 1/6.\\
\phi^E(x_1) &= m(x_1) = 1/4; \phi^E(x_1x_2) = m(x_1, x_2x_5)\\
&= 1/4;\\
\phi^E(x_1x_3) &= m(x_1, x_3x_4) = 1/12;\\
\phi^E(E) &= m(x_1, x_2x_5, x_3x_4) + m(x_1, x_3x_4, x_2x_5)\\
&= 5/12
\end{aligned}$$

□

Remark 2. Since conditional probability P^E only exists for $P(E) > 0$, Bayes' rule has only been defined for $f(E) > 0$ and not, as in Dempster's rule, for $F(E) > 0$.

It is however possible to argue that, when $f(E) = 0$ and $F(E) > 0$, the set of conditional probabilities considered should be $\mathcal{P}_*^E = \{P^E : P \in \mathcal{P} \text{ and } P(E) > 0\}$, the observation of E excluding almost surely the possibility that the true probability satisfy $P(E) = 0$.

It is easily seen that the lower envelope, f_*^E of $\mathcal{P}_*^E$, is the elementary belief function with unique focal element $B_* = \cup_{B \in \mathcal{B}''}(B \cap E)$ and that $\mathcal{P}_*^E = \mathcal{P}_{f*}^E$.

Thus the natural extension of Bayes' rule to the $f(E) = 0,\ F(E) > 0$ case satisfies 1) and 2).

6 On the Representation of Sets of Conditional Probability Measures by Their Lower Envelopes

Example 2 demonstrates that Bayesian conditioning in general preserves neither 2) nor the representability of $\mathcal{P}^E$ by f^E. Nevertheless it does not preclude that the property 2) *might* be preserved for a class of belief functions containing, in addition to the additive ones (the probability measures) for which the property is trivial, other interesting families of belief functions. However the main result in this section (Theorem 2 below) shows this hope to be futile.

Theorem 2. *Let f be a belief function on $\mathcal{A}$ and $\mathcal{P}_f = \{P \in \mathcal{L} : P \geq f\}$. The following statements are equivalent.*

S_1) There exists $E \in \mathcal{A}$, such that $f(E) > 0$ and that

$$\mathcal{P}^E = \left\{P^E \in \mathcal{L}^E : P \in \mathcal{P}_f\right\}$$

and $\mathcal{P}_{f^E}^E = \{Q \in \mathcal{L}^E : Q \geq f_E = \inf_{Q \in \mathcal{P}^E} Q\}$ satisfy $\mathcal{P}^E \subset \mathcal{P}_{f^E}^E$.

S_2) There exist $A_1,\ A_2 \in \mathcal{A}$ such that:

1) $f(A_1 \cap A_2) > 0$;
2) $f(A_1 \cup A_2) < 1$; and
3) $f(A_1 \cup A_2) + f(A_1 \cap A_2) > f(A_1) + f(A_2)$.

The formulation of condition S_2) and the last part (part 3) of the proof of Theorem 2 given later (after some preliminary results), are adapted from ideas of Chateauneuf [1].

Lemma 1. *Let f be a belief function on $\mathcal{A} = 2^{\mathcal{X}}$. Then $\mathcal{P}_f = \{P \in \mathcal{L} : P \geq f\}$, identified to a subset of the simplex of $\mathbb{R}^n$, $n = |\mathcal{X}|$, is a bounded convex polyhedron.*

Moreover, its extreme points are probability measures $P_S,\ S = (x_{i_1}, x_{i_2}, \cdots, x_{i_n}) \in \sum$, set of all permutations of $\mathcal{X}$, defined by

$$P_S(S_\ell) = f(S_\ell) \text{ for } S_\ell = \{x_{i_1}, x_{i_2}, \cdots, x_{i_\ell}\}, 1 \leq \ell \leq n-1 \tag{32}$$

A proof of this lemma, which is basically a corollary of Proposition 3, can be found, e.g., in Chateauneuf and Jaffray [2].

Since $f^E = \inf_{P \geq f} P^E$ is a belief function when f is one, Lemma 1 in particular applies to $\mathcal{P}^E_{f^E} = \{Q \in \mathcal{L}^E : Q \geq f^E\}$. As for its subset $\mathcal{P}^E$, let us first note, with Kyburg [6], that it is convex, which results from the following:

Lemma 2. *Let $E \in \mathcal{A}$ and $P_j \in \mathcal{L}$ be such that $P_j(E) > 0,\ j = 1, 2, \cdots, m$. Then*

$$Q = \sum_{j=1}^{m} \beta_j P_j^E, \beta_j \geq 0 \text{ for } j = 1, 2, \cdots, m, \sum_{j=1}^{m} \beta_j = 1,$$

if and only if

$$Q = P^E, \text{ with } P = \sum_{j=1}^{m} \alpha_j P_j \text{ and}$$

$$\alpha_j = \frac{\beta_j / P_j\left(E\right)}{\sum\limits_{j=1}^{m} \beta_j / P_j\left(E\right)}, \text{ for } j = 1, 2, \cdots, m. \tag{33}$$

Proof. The conditional of $P = \sum_{j=1}^{m} \alpha_j P_j$ is given by

$$\begin{aligned} P^E\left(A\right) = \frac{P\left(A\right)}{P\left(E\right)} &= \frac{\sum_{j=1}^{m} \alpha_j P_j\left(A\right)}{\sum_{j=1}^{m} \alpha_j P_j\left(E\right)} \\ &= \frac{\sum_{j=1}^{m} \alpha_j P_j\left(E\right) P_j^E\left(A\right)}{\sum_{j=1}^{m} \alpha_j P_j\left(E\right)}. \end{aligned}$$

Thus $P^E = \sum_{j=1}^{m} \beta_j P_j^E$ with

$$\beta_j = \frac{\alpha_j P_j\left(E\right)}{\sum_{j=1}^{m} \alpha_j P_j\left(E\right)}, \text{ for } j = 1, 2, \cdots, m. \tag{34}$$

Moreover, it is easily seen that relations (33) and (34) are reciprocals. The lemma is now straightforward.

Further information on the structure of $\mathcal{P}^E$ is provided by Lemma 3.

Lemma 3. *Let f be a belief function on $\mathcal{A} = 2^{\mathcal{X}}$ and $\mathcal{P}_f = \{P \in \mathcal{L} : P \geq f\}$. Let $E \in \mathcal{A}$ be such that $f(E) > 0$. Then, $\mathcal{P}^E = \{P^E \in \mathcal{L}^E : P \in \mathcal{P}_f\}$, identified to a subset of the simplex of $\mathbb{R}^L$, $L = |E|$, is a bounded convex polyhedron.*

Moreover, its extreme points are conditionals of extreme points of $\mathcal{P}_f$.

Proof. By Lemma 1,

$$\mathcal{P}_f = \left\{ P = \sum_{S \in \Sigma} \alpha_S P_S : \alpha_S \geq 0, \text{ for } S \in \sum, \text{ and } \sum_{S \in \Sigma} \alpha_S = 1 \right\}$$

By Lemma 2,

$$\mathcal{P}^E = \left\{ Q = \sum_{S \in \Sigma} \beta_S P_S^E : \beta_S = \frac{\alpha_S P_S(E)}{\sum_{S \in \Sigma} \alpha_S P_S(E)}, \right. \\ \left. \alpha_S \geq 0, \text{ for } S \in \Sigma, \text{ and } \sum_{S \in \Sigma} \alpha_S = 1 \right\} \\ = \left\{ Q = \sum_{S \in \Sigma} \beta_S P_S^E : \beta_S \geq 0, \text{ for } S \in \Sigma, \right. \\ \left. \text{and} \sum_{S \in \Sigma} \beta_S = 1 \right\}.$$

which shows that 1) $\mathcal{P}^E$ is a bounded convex polyhedron; and 2) the set of its extreme points is a subset of $\{P_S^E, \ S \in \Sigma\}$. □

Proof of Theorem 2:

1) Let us first use the preceding lemmata to derive necessary and sufficient conditions on $\mathcal{P}$ for the validity of equality $\mathcal{P}^E = \mathcal{P}_{f^E}^E$ when $E \in \mathcal{A}$ is such that $f(E) > 0$.
Since $\mathcal{P}^E$ is a convex subset of $\mathcal{P}_{f^E}^E$, a bounded convex polyhedron, $\mathcal{P}^E$ is equal to $\mathcal{P}_{f^E}^E$ if and only if every extreme point of $\mathcal{P}_{f^E}^E$ belongs to $\mathcal{P}^E$. By Lemma 1, the extreme points of $\mathcal{P}_{f^E}^E$ are the probability measures $P_{S'} \in \mathcal{L}^E, S' = (x_{i_1}, x_{i_2}, \cdots, x_{i_L}) \in \Sigma'$, set of all permutations of E, determined by

$$P_{S'}(S'_\ell) = f^E(S'_\ell), \text{ for } S'_\ell = \{x_{i_1}, x_{i_2}, \cdots, x_\ell\}, 1 \leq \ell \leq L-1. \tag{35}$$

Thus $\mathcal{P}^E = \mathcal{P}_{f^E}^E$ if and only if, for every $S' \in \Sigma'$, there exists $P \in \mathcal{P}_f$ such that

$$P^E(S'_\ell) = f^E(S'_\ell), \text{ for } 1 \leq \ell \leq L-1. \tag{36}$$

According to Proposition 4, (36) is satisfied, if and only if

$$P(S'_\ell) = 0, \text{ when } f(S'_\ell) = 0, \text{ i.e., for } 1 \leq \ell \leq L_1 \tag{37a}$$

and

$$P\left(S'_\ell\right) = f\left(S'_\ell\right) \text{ and } P\left(E \backslash S'_\ell\right) = F\left(E \backslash S'_\ell\right), \tag{37b}$$

when $f(S'_\ell) > 0$ and $F(E\backslash S'_\ell) > 0$, i.e., for $L_1 < \ell \leq L_2$.
(Note that their properties uniquely define L_1 and L_2, that $0 \leq L_1 \leq L_2 \leq L-1$ and that there is no condition for $L_2 \leq \ell \leq L-1$).
Thus $\mathcal{P}^E = \mathcal{P}^E_{f^E}$ if and only if, for every $S' \in \Sigma'$ there is $P \in \mathcal{P}_f$ satisfying (37).

2) Let us now show that $(S_1) \Rightarrow (S_2)$.
It follows from Part (1) than when (S_1) is true, there exists $E \in \mathcal{A}$, with $f(E) > 0$, and $S' \in \Sigma'$ such that no $P \in \mathcal{P}_f$ satisfies (37).
However, according to Proposition 3, there exists $P \in \mathcal{P}_f$ such that

$$P\left(S'_\ell\right) = f\left(S'_\ell\right) = f\left(S'_\ell\right), \text{ for } 1 \leq \ell \leq L_2,$$

and

$$P\left(S'_{L_2} \cup E^c\right) = f\left(S'_{L_2} \cup E^c\right) \text{ or, equivalently,}$$
$$P\left(E \backslash S'_{L_2}\right) = F\left(E \backslash S'_{L_2}\right); \tag{38}$$

thus, since P cannot satisfy (37), it must be that $L_1 < L_2 - 1$ and

$$P\left(E \backslash S'_{\bar\ell}\right) < F\left(E \backslash S'_{\bar\ell}\right), \text{ for some } L_1 < \bar\ell < L_2. \tag{39}$$

From (38) and (39), it results then that

$$\begin{aligned} f\left(S'_{L_2}\right) - f\left(S'_{\bar\ell}\right) &= P\left(S'_{L_2}\right) - P\left(S'_{\bar\ell}\right) = P\left(E\backslash S'_{\bar\ell}\right) - P\left(E\backslash S'_{L_2}\right) \\ &< F\left(E\backslash S'_{\bar\ell}\right) - F\left(E\backslash S'_{L_2}\right) = f\left(S'_{L_2} \cup E^c\right) - f\left(S'_{\bar\ell} \cup E^c\right). \end{aligned}$$

It is straightforward that for $A_1 = S'_{L_2}$ and $A_2 = S'_{\bar\ell} \cup E^c$, hence for $A_1 \cap A_2 = S'_{\bar\ell}$ and $A_1 \cup A_2 = S'_{L_2} \cup E^c$, (S_2)'s requirements are satisfied.

3) Let us finally prove that, conversely $(S_2) \Rightarrow (S_1)$. Given A_1 and A_2 satisfying (S_2)'s requirements, let $B_1 = A_1 \cap A_2$, $B_2 = A_2$, and $E = (A_1 \cap A_2) \cup A_1^c$. Necessarily $B_1 \subset B_2$ and, therefore, there exists $S' \in \Sigma'$ and $\ell_1 < \ell_2$ such that $B_1 = S'_{\ell_1}$ and $B_2 = S'_{\ell_2}$. It follows moreover from (S_2) that

$$f\left(S'_{\ell_i}\right) > 0 \text{ and } F\left(E\backslash S'_{\ell_i}\right) > 0, \text{ for } i = 1, 2.$$

Thus any $P \in \mathcal{P}_f$ and satisfying (37) must, in particular, satisfy

$$\begin{aligned} P\left(B_1\right) = f\left(B_1\right), P\left(B_2\right) = f\left(B_2\right), P\left(E\backslash B_1\right) \\ = F\left(E\backslash B_1\right), \text{ and } P\left(E\backslash B_2\right) = F\left(E\backslash B_2\right) \end{aligned}$$

which are the same as

$$\begin{aligned} P\left(A_1 \cap A_2\right) = f\left(A_1 \cap A_2\right), P\left(A_2\right) \\ = f\left(A_2\right), P\left(A_1^c\right) = F\left(A_1^c\right) \left(\Leftrightarrow P\left(A_1\right) = f\left(A_1\right)\right) \end{aligned}$$

and

$$P\left(A_1^c \cap A_2^c\right) = F\left(A_1^c \cap A_2^c\right)\left(\Leftrightarrow P\left(A_1 \cup A_2\right) = f\left(A_1 \cup A_2\right)\right).$$

These relations however imply

$$\begin{aligned} F\left(A \cup A_2\right) + f\left(A_1 \cap A_2\right) &= P\left(A_1 \cup A_2\right) + P\left(A_1 \cap A_2\right) \\ &= P\left(A_1\right) + P\left(A_2\right) = f\left(A_1\right) + f\left(A_2\right) \end{aligned}$$

which contradicts (3). There is therefore no such P in $\mathcal{P}_f$, hence, by Part (1), $\mathcal{P}^E \subset \mathcal{P}^E_{f^E}$ and (S_1) is true.

Theorem 2 shows that only "almost" additive belief functions f have the property that $\mathcal{P}^E$ is representable by f^E for all E. In particular, it is worth noticing that necessity functions [5], [19], which are defined as mappings $\mathcal{A} \to [0,1]$ satisfying $f(A \cap B) = \inf(f(A), f(B))$, all $A,\ B \in \mathcal{A}$, or, equivalently, as belief functions with consonant, i.e. nested, focal elements, can only satisfy property (S_1) of Theorem 2 when $|\mathcal{B}| \leq 3$.

It can be concluded from Theorem 2 that, when the situation after conditioning is summarized by f^E (which represents $\mathcal{P}^E_{f^E}$ but not $\mathcal{P}^E$), there may exist an important loss of information. One of the unpleasant consequences of this loss concerns iterated conditioning is Corollary 2.

Corollary 2. *Let f be a belief function on $\mathcal{A}$ and $\mathcal{P}_f = \{P \in \mathcal{L} : P \geq f\}$. Statements (S_1) and (S_2) of Theorem 2 are both equivalent to statement (S_3). There exist $E_1,\ E_2 \in \mathcal{A}$, with $E_2 \subset E_1$ and $f(E_2) > 0$, such that $(f^{E_1})^{E_2} \neq f^{E_2}$.*

Proof. 1) For any B $\subseteq E_2$, there exists (by Proposition 4), $P \in \mathcal{P}_f$ such that $f^{E_2}(B) = P^{E_2}(B) = (P^{E_1})^{E_2}(B) \geq (f^{E_1})^{E_2}(B)$. Thus (S_3) states the existence of some $B \subseteq E_2$ such that $f^{E_2}(B) > (f^{E_1})^{E_2}(B)$.
(S_1) and (S_2) being equivalent, we shall prove that $(S_3) \Rightarrow (S_1)$ and $(S_2) \Rightarrow (S_3)$.

2) Suppose (S_3) is true. There exists (by Corollary 1) $Q \in \mathcal{P}^{E_1}$ such that $Q^{E_2}(B) = (f^{E_1})^{E_2}(B)$; however, $Q \notin \mathcal{P}^{\varepsilon_1}$ since the existence of $P \in \mathcal{P}_f$ such that $Q = P^{E_1}$ would imply $f^{E_2}(B) \leq (P^{E_1})^{E_2}(B) = Q^{E_2}(B)$. Thus (S_1) is true.

3) If (S_2) is true, define $E_1 = A_2 \cup A_1^c,\ E_2 = (A_1 \cap A_2) \cup (A_2^c \cap A_1^c)$, and $B = A_1 \cap A_2$ (thus $B\ U\ E_1^c = A_1, B\ U\ (E_1 \backslash E_2) = A_2$, and $B_\cup E_2^c = A_1 \cup A_2$). By relation (19):

$$f^{E_2}\left(B\right) = \frac{f\left(B\right)}{f\left(B\right) + 1 - f\left(B \cup E_2^c\right)} = \frac{f\left(A_1 \cap A_2\right)}{f\left(A_1 \cap A_2\right) + 1 - f\left(A_1 \cup A_2\right)}.$$

Similarly, from

$$\left(f^{E_1}\right)^{E_2}(B) = \frac{f^{E_1(B)}}{f^{E_1}(B) + 1 - f^{E_1}(B \cup (E_1 \backslash E_2))}$$

$$f^{E_1}(B) = \frac{f(B)}{f(B) + 1 - f(B \cup E_1^c)}$$

and

$$f^{E_1}(B_\cup (E_1 \backslash E_2)) = \frac{f(B_\cup (E_1 \backslash E_2))}{f(B_\cup (E_1 \backslash E_2)) + 1 - f(B_\cup E_2^c)}$$

it follows that (see equation at the bottom of the page).

Therefore, $f^{E_2}(B) > f^{(E_1)E_2}(B)$ since $[f(A_1 \cap A_2) + 1 - f(A_1)]\ [f(A_2) + 1 - f(A_1 \cup A_2)]^{-1} > 1$, if and only if inequality (3) of (S2) holds.

We shall use Example 2 again to illustrate the possible loss of information resulting from iterated conditioning, and the consequences of this loss in decision making.

Example 2 (continued): Since Property (S_2) is satisfied by $A_1 = x_o x_3$ and $A_2 = x_o x_2$, let us take $E_1 = x_o x_1 x_2 (= E)$, $E_2 = x_o x_1$ and $B = x_o$. It is easily seen that $\mathcal{P}^{E_2}$ is a singleton, f^{E_2} being the probability measure defined by $f^{E_2}(x_o) = 1/4$ and $f^{E_2}(x_1) = 3/4$. Thus $P^{E_2}(x_o)$ is exactly given by a direct calculation, but not by an iterated one since it follows from relation (19) applied to $f^{E_1} = f^E$ that $(f^{E_1})^{E_2}(x_o) = 1/10$ and $(f^{E_1})^{E_2}(x_1) = 1/2$, which only locates $p^{E_2}(x_o)$ in interval [1/10, 1/2].

$$\left(f^{E_1}\right)^{E_2}(B) = \frac{f(A_1 \cap A_2)}{f(A_1 \cap A_2) + [1 - f(A_1 \cup A_2)]\,[f(A_1 \cap A_2) + 1 - f(A_1)]\,[f(A_2) + 1 - f(A_1 \cup A_2)]^{-1}}$$

Consider then a decision maker (DM) who uses the MAXMIN-EU criterion (i.e., choses the decision that maximizes the smallest expected utility (EU) consistent with his information) and can, at any time, abandon the status quo (utility level zero) for decision d characterized by utility levels

$$u(d(x_o)) = 11, u(d(x_1)) = -2, u(d(x_2)) = u(d(x_3)) = -1.$$

In the initial state of information, the DM should prefer the status quo, since the EU offered by d is independent of P in $\mathcal{P}$ and equal to $(-1/4)$. Given E_1, updating rule $f \mapsto f^{E_1}$ (i.e., $\mathcal{P} \mapsto \mathcal{P}^{E_1}$) does not exclude probability Q defined by $Q(x_o) = 1/12$, $Q(x_1) = 3/4$, and $Q(x_2) = 1/6$, which associates an EU equal to $(-3/4)$ with d, and thus makes the status quo seem preferable. On the other hand, the correct updating rule, $\mathcal{P} \mapsto \mathcal{P}^{E_1}$ shows d to secure EU level 1/12 and thus to be the better decision. Finally, given E_2, updating rule $f^{E_1} \mapsto (f^{E_1})^{E_2}$ only shows d to secure EU level $(-7/10)$ (achieved when $P^{E_2}(x_o) = 1/10$), whereas the true EU level is in fact known and equal to 5/4, making d better than the status quo.

Conclusion

It follows from Theorem 2 that there exists no conditioning rule for belief functions that is also consistent with the lower envelope interpretation of belief functions.

The fact that the conditional belief function f^E does not correctly represent the situation of uncertainty given E, does not make it useless. First, f^E indicates the exact range of the conditional probability $p^E(A)$ of every event A. Second, f^E is useful for approximate calculations. For example, the "Choquet integral" $\int Y \, df^E$ provides a lower bound of $\underset{Q \in \mathcal{P}^E}{\inf} \int Y \, dQ$. Note that this calculation only requires the knowledge of ϕ^E which, due to Theorem 1, can be directly deduced from ϕ, and that ϕ is often the original data (rather than f). Still the systematic use of f^E for representing (approximately) $\mathcal{P}^E$ must be rejected on account of undesirable properties like (S_3) of Corollary 2. As a matter of fact, Smets [16] has proven that Dempster's rule $f \mapsto g^E$ (see 21) is the only rule possessing the iterated conditioning consistency property

$$\left(g^{E_1}\right)^{E_2} = q^{E_2} \quad \text{for} \quad E_2 \subseteq E_1$$

(and satisfying some other mild requirements; see also recent results by Gilboa and Schmeidler [7]); however g^E does not in general represent $\mathcal{P}^E$.

There are however other ways of representing $\mathcal{P}^E$. In particular, by using the fact that $\mathcal{P}$ is the set of all allocations of ϕ (see (8), (9), and Proposition 2), one can describe $\mathcal{P}^E$ as the set of all P_λ^E such that

$$P_\lambda^E(A) = \frac{P_\lambda(A)}{P_\lambda(E)} = \frac{\sum\limits_{x \in A} \overset{\lambda(B,x)\phi(B)}{\sum\limits_{B \in \mathcal{B}, B \supseteq \{x\}}}}{\sum\limits_{x \in E} \overset{\lambda(B,x)\phi(B)}{\sum\limits_{B \in \mathcal{B}, B \supseteq \{x\}}}}$$

$$\forall \; A \in \mathcal{A}^E$$

a ratio of two linear forms with respect to the variables, the unknown weights $\lambda(B, x)$. An alternative representation that uses f instead of ϕ is simply

$$\mathcal{P}^E = \left\{ Q \in \mathcal{L}^E : Q(A) = \frac{P(A)}{P(E)} \; \forall \; A : P \in \mathcal{L}, P \geq f \right\}.$$

These representations are promising since they make standard optimization techniques relevant for solving decision problems. For example, evaluating $\underset{Q \in \mathcal{P}^E}{\min} E_Q Y$, for a given r.v.Y., amounts in both cases to solving a fractional program (or the equivalent linear program). Results obtained by these representations will be presented in "Dynamic decision making with belief functions" (in prep.)

References

[1] A. Chateauneuf, Private communication, CERMSEM, Univ. Paris 1.

[2] A. Chateauneuf and J. Y. Jaffray, “Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion,” *Math. Soc. Sci.*, vol. 17, pp. 263–283, 1989.

[3] A. P. Dempster, “Upper and lower probabilities induced by a multivalued mapping,” *Ann. Math. Statist.*, vol. 38, pp. 325–339, 1967.

[4] A. P. Dempster, “A generalization of Bayesian inference,” *J. Royal Stat. Soc. B*, vol. 30, pp. 205–247, 1968.

[5] D. Dubois and H. Prade, *Théorie des Possibilités.* Paris: Masson, 1987.

[6] R. Fagin and J. Y. Halpern, “A new approach to updating beliefs,” I.B.M. Res. Rep. RJ 7222 (67989), (1989), in *Proc. 6th Conf. Uncertainty* in AI, 1990.

[7] I. Gilboa and D. Schmeidler, “Updating ambiguous beliefs,” working paper, Dept. Economics, Tel Aviv Univ., Tel Aviv, 1991.

[8] J. Y. Jaffray, “Belief functions, convex capacities and decision making,” in *Mathematical Psychology: Current Developments*, Doignon & Falmagne, Eds. New York: Springer, 1991.

[9] H. E. Kyburg, Jr., “Bayesian and non-Bayesian evidential updating,” *Artificial Intell.* vol. no. 31, 3, pp. 271–294, 1987.

[10] H. T. Nguyen, “On random sets and belief functions,” *J. Math. Anal. Appl.*, vol. 65, pp. 531–542, 1978.

[11] J. Pearl, *Probabilistic Reasoning in Intelligent Systems.* San Mateo, CA: Morgan Kaufmann, 1988.

[12] J. Pearl, “Reasoning with belief functions: An analysis of compatibility,” *Int. J. Approx. Reasoning*, vol. 4, pp. 363–389, 1990.

[13] G. C. Rota, “Theory of Möbius functions,” *Z. für Wahrscheinlichkeis-theorie und Verw. Gebiete*, vol. 2, pp. 340–368, 1964.

[14] G. Shafer, *A Mathematical Theory of Evidence.* Princeton, NJ: Princeton Univ. Press, 1976.

[15] G. Shafer, “Perspectives on the theory and practice of belief functions, *Int. J. Approx. Reasoning.*, vol. 4, pp. 323–362, 1990.

[16] P. Smets, “Belief functions and their combinations,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-12, pp. 447–458, 1990.

[17] M. Spies, “Combination of evidence with conditional objects and its application to cognitive modeling,” to appear in *Conditional Logic in Expert Systems*, Goodman *et al.*, Eds. Amsterdam: North Holland, 1990.

[18] V. Strassen, “Messfehler and information,” *Z. für Wahrscheinlichkeits-theorie und Verw. Gebiete*, vol. 2, pp. 273–305, 1964.

[19] L.A. Zadeh, “Fuzzy sets as a basis for a theory of possibility,” *Fuzzy Sets Syst.*, vol. 1 pp. 3–28, 1978.

[20] L. Zhang, “A new proof to Theorem 3.2 of Fagin and Halpern’s paper,” unpublished memorandum, Univ. Kansas, Business School, 1989.

23

Belief-Function Formulas for Audit Risk

Rajendra P. Srivastava and Glenn R. Shafer

Abstract. This article relates belief functions to the structure of audit risk and provides formulas for audit risk under certain simplifying assumptions. These formulas give plausibilities of error in the belief-function sense.

We believe that belief-function plausibility represents auditors' intuitive understanding of audit risk better than ordinary probability. The plausibility of a statement, within belief-function theory, measures the extent to which we lack evidence against the statement. High plausibility for error indicates only a lack of assurance, not positive evidence that there is error. Before collecting, analyzing, and aggregating the evidence, an auditor may lack any assurance that a financial statement is correct, and in this case will attribute very high plausibility to material misstatement. This high plausibility does not necessarily indicate any evidence that the statement is materially misstated, and hence, it is inappropriate to interpret it as a probability of material misstatement.

The SAS No. 47 formula for audit risk is based on a very simple structure for audit evidence. The formulas we derive in this article are based on a slightly more complex but still simplified structure, together with other simplifying assumptions. We assume a tree-type structure for the evidence, assume that all evidence is affirmative and that each variable in the tree is binary. All these assumptions can be relaxed. As they are relaxed, however, the formulas become more complex and less informative, and it then becomes more useful to think in terms of computer algorithms rather than in terms of formulas (Shafer and Shenoy 1988).

In general, the structure of audit evidence corresponds to a network of variables. We derive formulas only for the case in which each item of evidence bears either on all the audit objectives of an account or on all the accounts in the financial statement, as in Fig. 1, so that the network is a tree. Usually, however, there will be some evidence that bears on some but not all objectives for an account, on some but not all accounts, or on objectives at different levels; in this case, the network will not be a tree.

We assume that all evidence is affirmative because this is the situation treated by the SAS No. 47 formula and because belief-function formulas become significantly more complex when affirmative and negative evidence is combined. This complexity is due primarily to the renormalization involved in Dempster's rule for combining belief functions.

The variables in the network or tree represent various audit objectives, accounts, and the financial statement as a whole. We assume these variables are binary. For example, we assume that an account either is or is not materially misstated. This assumption is clearly too restrictive for most audit practice. Often, for example, an auditor must consider immaterial errors in individual accounts that could produce a material error in the financial statement when they are aggregated.

We derive formulas for plausibility of material misstatement at three levels: the financial statement level, the account level, and the audit objective level. The formula at the audit objective level resembles the SAS No. 47 formula,[1] but the formulas at the other two levels are significantly different. Because our model does distinguish evidence gathered at the three different levels, audits based on our formulas are sometimes significantly more efficient[2] than audits based on the SAS No. 47 model or on the simpler Bayesian models.

Key words: Audit risk, Belief functions, Planning model, Evaluation model

The remainder of this article is divided into seven sections and two appendices. In Sect. 1, we review the existing literature on the audit-risk model. In Sect. 2, we review the structure of various types of audit evidence. In Sect. 3, we review the belief-function treatment of audit evidence and discuss using

[1] With the evidential structure that we consider in this article, we obtain plausibility formulas that are closer in form to the SAS No. 47 formula than to Bayesian formulas (Leslie 1984). This is because the belief-function theory permits an auditor to have belief, say 0.6, based on the procedures performed, that a given objective is met, without having to assign the remaining 0.4 of his or her belief to whether the objective has been met or not. This unassigned belief, 0.4, in the case, represents the plausibility that the objective is not met.

Those accustomed to Bayesian thinking may prefer to express judgments about the effectiveness of procedures in terms of conditional probabilities for detection of error, given the presence or absence of such error. In our view, such conditional probabilities are only one way of expressing intuitive judgments that may alternatively be expressed in terms of belief functions. We would express our intuitive judgments as conditional probabilities only if we intended to carry out a completely Bayesian analysis. We would express our intuitive judgments directly in terms of $\boldsymbol{m}$-values if we intended to carry out a belief-function analysis. It is possible, nonetheless, to express intuitive judgments in terms of conditional probabilities and then to translate these conditional probabilities into belief functions. There is some arbitrariness involved in the translation, but it can be carried out in such a way that the Bayesian approach becomes a special case of the belief-function approach (Shafer 1982). If we combine the belief function representing the conditional probabilities with a Bayesian prior using Dempster's rule, we will obtain the appropriate Bayesian posterior.

[2] In general, the more accurately we model the structure of audit evidence, the more effective and efficient we can expect the audit to be. However, it is also important to recognize that if the necessary inputs to a more accurate model cannot be estimated accurately by the auditor, decreases in audit effort mandated by the model can decrease audit effectiveness.

belief-function plausibility to represent audit risk. In Sect. 4, we discuss the combination of beliefs (or m-values) at each level of the financial statement. This combination corresponds to aggregating the evidence that bears directly on the objectives at those levels. In Sect. 5, we present total plausibility formulas (i.e., belief-function formulas for the total audit risk) and numerical examples. In Sect. 6, we highlight the limitations of our formulas and suggest future directions for research. In Sect. 7, we summarize our results. We review the propagation of belief functions in a tree in appendix A, and we use these results to derive the belief-function formulas for the total audit risk in appendix B.

1 Review of the Literature

Here we will review the existing literature on audit risk formulas and the difficulties involved in interpreting the numbers in these formulas as probabilities.

Although the model of SAS No. 47 (AICPA 1983) has been used for planning purposes since June 1984, controversy about its applicability for planning and evaluation has persisted. Graham points out, for example, that "overall assessments of audit risk for the financial statements taken as a whole are usually impractical for audit planning and engagement control" (1985a, 14), since inherent and control risks can vary from one account to another and from one class of transactions to another. Graham further suggests that the audit-risk model should be decomposed into components that determine audit risk for management assertions related to each account balance (accounts receivable, inventory, etc.) or class of transactions (purchases, sales, etc.). Cushing and Loebbecke (1983) argue that the SAS No. 47 model provides no guidance on aggregating risks associated with various account balances, and Kinney (1989) has demonstrated recently that the model has properties that may significantly understate achieved audit risk.

Bayesian models have also been discussed in the literature for risk aggregation (see, e.g., Kinney 1984; Leslie 1984). The simplest Bayesian models do not consider different levels of the account, but Boritz and Jensen (1985) discuss the hierarchical structure of audit evidence and propose ways to combine such evidence by using probabilities to represent risks. Also, in discussing their assertion-based approach to auditing, Leslie et al. (1986) recognized the importance of the structure of audit evidence and emphasized that assurances from various items of evidence should be assessed at the management assertion level of the account and then combined. They considered the relationship between various accounts (e.g., accounts receivable depends on sales and cash receipts) in the aggregation process, but they assumed that the different items of evidence for different accounts were independent.[3]

[3] Leslie et al. (1986) assume Poisson error rates, and they base their analysis on assurance factors, which represent Poisson parameters. They do not explicitly

At a qualitative level, there has been increasing recognition of the importance of the structure of evidence and its relation to various audit objectives in assessing control risk. In SAS No. 55 entitled "The Auditor's Responsibility for Assessing Control Risk" (AICPA 1988a, par. 3), we find the following statement:

> After obtaining this understanding, the auditor assesses control risk for the assertions embodied in the account balance, transaction class, and disclosure components of the financial statements.

Although the AICPA has not yet required auditors to consider individual audit objectives when assessing other risks (e.g., inherent risk and detection risk), it appears that many are already doing so (Graham 1985a–1985e).

Boritz and Wensley (1990) have used the structure of audit evidence to develop their computer system for audit planning and evaluation, but their system uses heuristic rules rather than formal theory to combine uncertainties. Since such heuristics have been seen to fail in complex systems (Buchanan and Shortliffe 1984), we believe that more needs to be done to develop theoretical methods of combination that take account of the complexity of the structure of audit evidence.

We advance belief functions as a basis for this theoretical development because we believe that the usefulness of the Bayesian approach is limited by divergences between the intuitive and Bayesian interpretations of audit risk. For example, according to SAS No. 47, if an auditor decides not to consider inherent factors, then the inherent risk is set equal to 1. Since a probability of 1 means certainty, this seems to be saying that it is certain that the account is materially in error. But this is not what the auditor has in mind when deciding not to depend on inherent factors. The auditor's intention is represented better by a belief-function plausibility of 1 for material error, which says only that the auditor lacks evidence based on inherent factors.

In a less extreme situation, the auditor may believe, on the basis of inherent factors, that the account is fairly stated and yet be unwilling to rely on these factors past a certain point. In this case, the auditor may, as SAS No. 47 suggests, assign a value less than the maximum, say 70 percent, to inherent risk. If interpreted in probability terms, this number says that the inherent factors give a 30 percent chance that the account is not materially misstated and a 70 percent chance that it is materially misstated. This suggests that the evidence is negative, contrary to the auditor's intuition. The probability interpretation is even more confusing if the auditor sets the inherent risk at 50 percent. What does this mean? Does it mean that the auditor is completely ignorant about the state of the account, or does it mean there is more evidence that the account is not being materially misstated than when only 30 percent assurance was assumed?

assume the independence of different items of evidence, but this assumption appears to be the justification for their practice of summing individual assurance factors to obtain an overall level.

Belief functions, since they permit uncommitted belief, allow us to interpret the auditor's choices in a straightforward way. When the auditor sets the risk at 70 percent, a 30 percent degree of support from inherent factors is claimed, leaving 70 percent of the auditor's belief uncommitted. In this case, material misstatement has 70 percent plausibility, but absence of material misstatement has 100 percent plausibility. When the auditor sets the risk at 50 percent, stronger support from inherent factors is claimed. In this case, the plausibility of material misstatement is reduced to 50 percent, while the plausibility of no material misstatement remains at 100 percent.

We believe that belief functions provide a flexible and adaptable way to combine evidence from a variety of sources (Akresh et al. 1988). One aspect of this flexibility is that, when the belief functions representing individual items of evidence are probability measures, the belief-function analysis reduces to a Bayesian analysis (Shafer and Shenoy 1990).

2 The Structure of Audit Evidence

As we have seen, neglect of the structure of audit evidence has been a problem in some risk models. Before developing belief-function formulas we must, therefore, specify carefully the kind of structure we are considering.

We will adopt the structure currently assumed in auditing standards (AICPA 1988a; see also, e.g., Arens and Loebbecke 1988), with some simplifications. The standards generally divide audit evidence into four categories: (1) general knowledge about inherent risk, (2) evidence from analytical procedures, (3) knowledge of control factors and accounting systems, and (4) tests of details of balances. Within each general category, further structure arises because of the relevance of different items of evidence to different accounts and different objectives. In the following subsections, we review the structure within each of the four categories.

General Knowledge about Inherent Risk

In this category, we include general knowledge about risk factors that lie outside of the accounting system and also outside of the auditor's control. Examples include economic, political, business and regulatory environments, experience from the prior year's audit, management philosophy and style, organizational structure, and audit committee (see SAS No. 55 for more examples). Also included in this category are factors that make individual accounts more or less susceptible to error, such as the complexity of transactions, the volume of transactions processed, the susceptibility of assets to defalcation, and related party transactions. Such factors are important for the auditor's planning decisions, since the extent, nature, and timing of tests will depend on the auditor's assessment of the effect of these factors on the individual accounts and on the financial statement as a whole.

Some inherent factors affect entire financial statements, whereas others affect only certain accounts or classes of transactions. Some affect only a particular audit objective for an account or a class of transactions. Information about the competence and integrity of management, for example, will affect the entire financial statement. The auditor will have a higher level of assurance about the financial statement when management is of recognized competence and integrity than when management is known to have been involved previously in irregularities. In contrast, an auditor auditing a newspaper publisher realizes that libel suits against newspapers are common and will treat this knowledge as evidence affecting only accrued-contingent liability. The nature and complexity of an individual account (e.g., susceptibility to defalcation, volume of transactions, non-routine transactions, account balance based on management's judgment) also may affect only that account. As an example of evidence that affects only a certain audit objective, we might cite information about ralated party transactions from the minutes of board meetings. This information affects only the disclosure and classification objectives of the accounts involved in the transaction.

In summary, we see that evidence about inherent factors can bear on the financial statement at three different levels: (1) the financial statement level, (2) the individual account or class of transactions level, and (3) the audit objective level for individual accounts or classes of transactions. See Fig. 1 for details.

Analytical Procedures

Recently, the AICPA published SAS No. 56 (AICPA 1988b), requiring auditors to use analytical procedures on all audit engagements. According to SAS No. 56 (par. 2),

> Analytical procedures are an important part of the audit process and consist of evaluations of financial information made by a study of plausible relationships among both financial and non-financial data. Analytical procedures range from simple comparisons to the use of complex models involving many relationships and elements of data.

The statement proposes that analytical procedures be used for the following purposes (par. 4):

1. To assist the auditor in planning the nature, timing and extent of other auditing procedures.
2. As a substantive test to obtain evidential matter about particular assertions related to account balances or class of transactions.
3. As an overall review of the financial information in the final review stage of the audit.

According to SAS No. 56, the expected effectiveness and efficiency of an analytical procedure depends on (1) the nature of the audit objectives, (2)

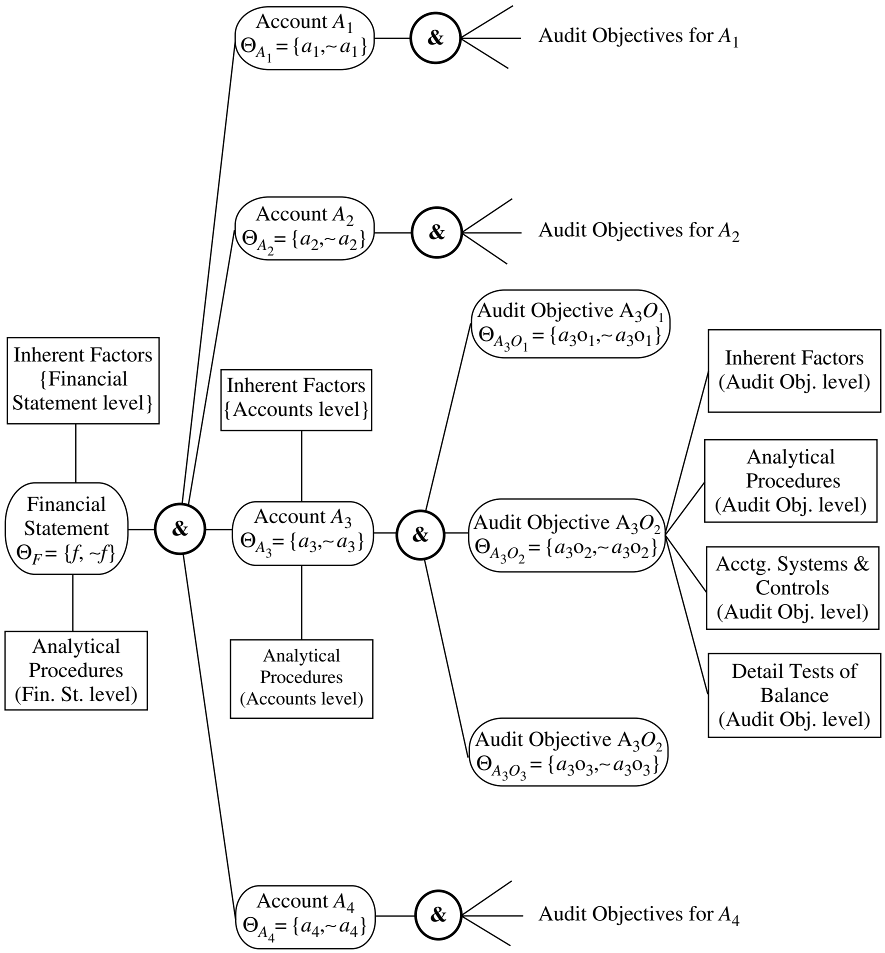

Account A_1
$\Theta_{A_1} = \{a_1, \sim a_1\}$
&
Audit Objectives for A_1
Account A_2
$\Theta_{A_2} = \{a_2, \sim a_2\}$
&
Audit Objectives for A_2
Audit Objective A_3O_1
$\Theta_{A_3O_1} = \{a_3o_1, \sim a_3o_1\}$
Inherent Factors (Audit Obj. level)
Inherent Factors {Financial Statement level}
Inherent Factors {Accounts level}
Analytical Procedures (Audit Obj. level)
Financial Statement
$\Theta_F = \{f, \sim f\}$
&
Account A_3
$\Theta_{A_3} = \{a_3, \sim a_3\}$
&
Audit Objective A_3O_2
$\Theta_{A_3O_2} = \{a_3o_2, \sim a_3o_2\}$
Acctg. Systems & Controls (Audit Obj. level)
Analytical Procedures (Fin. St. level)
Analytical Procedures (Accounts level)
Detail Tests of Balance (Audit Obj. level)
Audit Objective A_3O_2
$\Theta_{A_3O_3} = \{a_3o_3, \sim a_3o_3\}$
Account A_4
$\Theta_{A_4} = \{a_4, \sim a_4\}$
&
Audit Objectives for A_4

Fig. 1. An evidential network
Note: A rounded rectangle represents a variable (variables being the financial statements as a whole, various accounts, and the related audit objectives). A rectangle represents an item of audit evidence. The evidence is connected to a variable that it directly supports. A circle with "&" implies that the variable on the left is true if and only if the variables on the right of the circle are true

the plausibility and predictability of the relationship, (3) the reliability and availability of the data used to develop the expectation, and (4) the precision of the expectation.

Like general knowledge about inherent factors, analytical procedures can provide assurance at various levels. The more common analytical procedures seem to provide assurance at either the account level or the audit objective level. A comparison of the previous year's accounts payable with the current period's accounts payable provides assurance for the accounts payable balance as a whole. A comparison of the previous year's ratio of bad debt expense to accounts receivable balance with the current year's ratio would bear on collectibility of accounts receivable, a valuation objective.

In the present article, for the purpose of completeness, we will assume analytical procedures to be effective at all three levels: (1) the financial statement level, (2) the account or class of transactions level, and (3) the audit objective level for accounts or classes of transactions. When certain items of evidence are not to be considered in an audit, then those items are eliminated by setting the corresponding plausibilities to 1 (see the discussion in Sect. 5).

Control Factors and Accounting Systems

We include in this category all items of evidence related to accounting systems, control procedures, and tests of transactions. A test of controls typically bears on the audit objective level of an individual account, while a test of transactions typically bears on the audit objective level of a class of transactions. Controls built into a cost accounting system, for example, bear on the valuation objective of inventory, while the use of prenumbered bills of lading and sales invoices periodically accounted for bears on the completeness objective of sales.

Tests of Details of Balances

Tests of details of balances bear primarily on the audit objective level. Since it is costly to obtain this type of evidence, the auditor minimizes the need for it by maximizing the assurance to accounts and audit objectives from other sources.

Some tests of details balance bear on only one audit objective, while others may bear on more than one. A review of the minutes of board meetings to check whether receivables have been factored bears only on the ownership objective of accounts receivable, but confirmations of accounts receivable by customers provide assurance for both the existence and valuation objectives. In general, such assurances may vary in strength from objective to objective. For example, confirmation of accounts receivable may provide a higher level of support for the existence objective than for the valuation objective.

When an item of evidence bears equally on *all* the objectives of an account or *all* the accounts of the financial statement, we can represent it within a tree structure by linking it directly to the account or the financial statement, as the case may be. But when a test provides support to more than one audit objective, say, but not equally to all the objectives at once, we obtain a network of variables that is not a tree, and this makes the derivation of formulas cumbersome. The formulas given here are based on the assumption that the network is a tree, but they can be used as approximations in the non-tree case. One way to use them as approximations is to treat the evidence as if it consisted of independent items of evidence bearing on the different objectives.

The formula for the total audit risk (i.e., total plausibility of error) at the audit objective level will still be valid (i.e., it maintains its multiplicative form; see Sect. 5) when we do this, but the formulas at the account level and the financial statement level will provide only a conservative estimate of the total risk (i.e., plausibility of error). For example, suppose that confirmations of accounts receivable yield 0.9 level of assurance that both existence and valuation objectives are met. If we treat this as two items of evidence, one giving 0.9 degree of support for existence and one giving 0.9 for valuation, then our formulas give a total assurance, for the two objectives jointly, of $0.9 \times 0.9 = 0.81$, corresponding to a risk (plausibility of error) of 0.19. But the correct value for the assurance is 0.9, corresponding to a risk of 0.1.

3 Belief-Function Approach to Audit Evidence

In this section, we review the belief-function approach to representing uncertainties in audit evidence. The belief-function framework involves three related representations for beliefs concerning a topic: the belief function (***Bel***), the plausibility function (***PL***), and the basic probability assignment (***m***). As we will explain, the basic probability assignment is often convenient for expressing initial judgments, but the plausibility function is useful for expressing final judgments about audit risk.

The basic probability assignment is also called the ***m***-function, and its values are called ***m***-values (Shafer 1976). The basic difference between ***m***-values and probabilities is that probabilities are assigned to individual elements of a frame,[4] say Θ, whereas ***m***-values are assigned to a subset of elements of the frame. The sum of all the ***m***-values for all the subsets of the frame Θ is 1. Formally, the ***m***-function assigns a number $\boldsymbol{m}(B)$ to each subset B of Θ such that $\boldsymbol{m}(\varnothing) = 0(\varnothing$ being the empty set) and:

$$\sum_{B \subseteq \Theta} \boldsymbol{m}(B) = 1.$$

There are two ways to obtain ***m***-values on a frame: (1) they may be assigned directly by the decision maker on the basis of subjective judgment and (2) they may be derived from a compatibility relationship between a frame with known probabilities and the frame of interest.[5] We will use the first approach to discuss our example.

Suppose the auditor has performed a set of analytical procedures appropriate to account 'A' and finds no discrepancy or errors in the account. On

[4] We call an exhaustive and mutually exclusive set of possible answers to a question a *frame*. We will often use the symbol Θ to represent the frame in which we are interested. In the case of a yes-no question, the frame has only two elements; $\Theta = \{\text{yes}, \text{no}\}$, or $\Theta = \{a =$ account 'A' is not materially misstated, $\sim a =$ account 'A' is materially misstated$\}$, etc. But, in general, a frame may be a very large set, for its question may have many possible answers.

[5] In general, if we want to make probability judgments about the elements of a frame Θ for which we have no probability measures, we can do so by relating the

the basis of this observation, the auditor feels that the evidence is positive and provides a medium level of support, say 0.6, to 'a' that the account is not materially misstated.[6] However, at the same time, the auditor feels that there is nothing to indicate that the account is materially misstated ($\sim a$). This means that 0.6 degree of support is assigned to 'a', 0 to '$\sim a$', and the remaining 0.4 is the ignorance assigned to the entire frame $\Theta = \{a, \sim a\}$; that is,

$$\boldsymbol{m}_{PA}\,(a) = 0.6, \boldsymbol{m}_{PA}\,(\sim a) = 0, \text{ and } \boldsymbol{m}_{PA}\,(a, \sim a) = 0.4,$$

where the subscript PA stands for analytical procedures at the account level. The above set of $\boldsymbol{m}$-values represents affirmative evidence.[7]

elements of Θ to the elements of the frame S for which we have knowledge of its probability distribution. This relationship is called the compatibility relationship. The basic idea is that each probability $\boldsymbol{P}$(s), where s is an element of S, should contribute to a degree of belief in the subset Γ(s) of Θ consisting of elements with which s is compatible. If several s items have the same Γ(s)—in other words, Γ(s) is equal to B for several s items—then the probabilities of all these will contribute to our degree of belief that the answer to the question considered by Θ is somewhere in B. For each subset B of Θ, let $\boldsymbol{m}(B)$ be the total probability for all the s items whose Γ(s) is equal to B:

$$\boldsymbol{m}(B) = \sum_{\Gamma(\mathrm{s})=B} \boldsymbol{P}(s).$$

It follows from this formula that:

$$\sum_{B\subseteq\Theta} \boldsymbol{m}(B) = 1,$$

and

$$\boldsymbol{m}(\varnothing) = 0,$$

where $\varnothing$ is the empty set.

[6] As a general convention, we will use capital letters to denote names of accounts or audit objectives (nodes) and small letters in script to represent their values.

[7] Affirmative evidence implies that the evidence directly supports the assertion to a certain degree and provides no support for its negation. In our example, we have $\boldsymbol{m}_{PA}(a) = 0.6$, $\boldsymbol{m}_{PA}(\sim a) = 0$, and $\boldsymbol{m}_{PA}(\{a, \sim a\}) = 0.4$, which implies that we have direct evidence that 'a' is met with 0.6 degree of support and no evidence that 'a' is not met.

Negative evidence implies that the evidence directly supports the negation of the assertion and provides no support for the assertion, i.e.,

$$\boldsymbol{m}_{PA}(a) = 0, m_{PA}(\sim a) = 0.4, \text{ and } \boldsymbol{m}_{PA}(\{a, \sim a\}) = 0.6.$$

A mixed item of evidence can be defined as an item that provides some support for the assertion and some for its negation, i.e.,

$$\boldsymbol{m}_{PA}(a) = 0.5, m_{PA}(\sim a) = 0.3, \text{ and } \boldsymbol{m}_{PA}(\{a, \sim a\}) = 0.2.$$

As mentioned earlier, we will consider only affirmative evidence in our derivation of audit risk formulas. Although the approach of aggregating evidence discussed in this article is valid for any type of evidence, use of affirmative evidence avoids the renormalization procedure[8] in aggregating various items of evidence and thus yields simple analytical formulas.

Let us go back to our example of analytical procedures discussed above and express the auditor's judgment about the level of support obtained (or to be obtained when planning the audit) from the procedures for account 'A' in terms of algebraic expressions:

$$\boldsymbol{m}_{PA}\left(a\right) = 1 - APR_A, \tag{1}$$

$$\boldsymbol{m}_{PA}\left(\sim a\right) = 0, \tag{2}$$

and

$$\boldsymbol{m}_{PA}\left(\{a, \sim a\}\right) = APR_A, \tag{3}$$

where APR_A represents a number. Equation (1) implies that the analytical procedures performed by the auditor for account 'A' provide assurance that the account is not materially misstated with $(1 - APR_A)$ degree of support.

Belief Functions and Plausibility Functions

In general, the total belief in a subset B of the frame $\varTheta$ is given by:

$$\boldsymbol{Bel}(B) = \sum_{X \subseteq B} \boldsymbol{m}\left(X\right), \tag{4}$$

where X represents a set of elements of $\varTheta$, and the plausibility of B is given by:

$$\boldsymbol{PL}\left(B\right) = \sum_{B \cap X \neq \varnothing} \boldsymbol{m}\left(X\right) = 1 - \boldsymbol{Bel}(\sim B)\,. \tag{5}$$

[8] Consider two independent items of evidence with $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ representing the $\boldsymbol{m}$-values on a frame $\varTheta$. By Dempster's rule (Shafer 1976), the combined $\boldsymbol{m}$-value for a subset A of frame $\varTheta$ is:

$$\boldsymbol{m}(A) = K^{-1}\Sigma\{\boldsymbol{m}_1(B_1)\boldsymbol{m}_2(B_2)|B_1 \cap B_2 = A, A \neq \varnothing\},$$

where K is the renormalization constant;

$$K = 1 - \Sigma\{\boldsymbol{m}_1(B_1)\boldsymbol{m}_2(B_2)|B_1 \cap B_2 = \varnothing\}.$$

The second term in K represents the conflict between the two items of evidence. If the conflict term is 1, i.e., if the two items of evidence exactly contradict each other, then $K = 0$ and, in such a situation, the two items of evidence are not combinable. In other words, Dempster's rule cannot be used when $K = 0$.

Intuitively, the plausibility of B is the degree to which B is plausible in the light of the evidence—the degree to which we do not disbelieve B or assign belief to its negation $\sim B$. Complete ignorance or lack of opinion about B is represented by $\boldsymbol{Bel}(B) = 0$ and $\boldsymbol{PL}(B) = 1$.

Consider again the numerical example discussed above. We have $\boldsymbol{m}_{PA}(a) = 0.6$, $\boldsymbol{m}_{PA}(\sim a) = 0$, and $\boldsymbol{m}_{PA}(a, \sim a) = 0.4$. From (4) and (5), we obtain:

$$\boldsymbol{Bel}_{PA}(a) = \boldsymbol{m}_{PA}(a) = 0.6,$$
$$\boldsymbol{Bel}_{PA}(\sim a) = \boldsymbol{m}_{PA}(\sim a) = 0,$$
$$\boldsymbol{Bel}_{PA}(\{a, \sim a\}) = \boldsymbol{m}_{PA}(a) + \boldsymbol{m}_{PA}(\sim a) + \boldsymbol{m}_{PA}(a, \sim a) = 0.6 + 0 + 0.4 = 1.0,$$

and

$$\boldsymbol{PL}_{PA}(a) = 1 - \boldsymbol{Bel}_{PA}(\sim a) = 1,$$
$$\boldsymbol{PL}_{PA}(\sim a) = 1 - \boldsymbol{Bel}_{PA}(a) = 1 - .06 = 0.4.$$

The intuitive meaning of $\boldsymbol{Bel}_{PA}(a) = 0.6$ is that the auditor has direct evidence from analytical procedures relevant to account 'A' that 'a' is true with 0.6 degree of support (i.e., the account is not materially misstated with degree 0.6). $\boldsymbol{Bel}_{PA}(\sim a) = 0$ means that the auditor has no evidence from analytical procedures that the account is materially misstated (i.e., $\sim a$ is true).

Let us now consider $\boldsymbol{PL}_{PA}(a) = 1$. What does it mean? We know that analytical procedures provide no belief to $\sim a(\boldsymbol{Bel}_{PA}(\sim a) = 0)$. Since there is no support committed to just $\sim a$ all the probability mass could be assigned to a, which implies that $\boldsymbol{PL}_{PA}(a) = 1$. Similarly, since $\boldsymbol{Bel}_{PA}(a) = 0.6$ (i.e., 0.6 degree of belief is directly committed to a), the remaining amount 0.4 of uncommitted probability mass could be assigned to $\sim a$; that is, $\boldsymbol{PL}_{PA}(\sim a) = 0.4$.

Going back to the $\boldsymbol{m}$-values in (1)–(3), we obtain the following beliefs and plausibilities:

$$\boldsymbol{Bel}_{PA}(a) = 1 - APR_A, \boldsymbol{Bel}_{PA}(\sim a) = 0, \text{ and } \boldsymbol{Bel}_{PA}(\{a, \sim a\}) = 1, \quad (6)$$

and

$$\boldsymbol{PL}_{PA}(a) = 1, \text{ and } \boldsymbol{PL}_{PA}(\sim a) = APR_A. \quad (7)$$

The plausibility function, $\boldsymbol{PL}_{PA}(\sim a) = APR_A$, has an important interpretation. It provides a non-frequentist interpretation of the auditing concept of risk. This is a measure of how risky we feel it would be to stop with this evidence. According to the analytical procedures performed at the account level, we have $(1 - APR_A)$ degree of belief that a is true, leaving a plausibility of $\boldsymbol{PL}_{PA}(\sim a) = APR_A$ that the account is materially misstated. This is the audit risk associated with the analytical procedures performed at the account level (see Table 1 for definitions of other risks).

This plausibility interpretation of audit risk is conceptually in agreement with the thought process of the auditor when planning an audit. For example,

Table 1. List of symbols and their definitions

Propositions

a – Account 'A' is not materially misstated.

$\sim a$ – Account 'A' is materially misstated.

f – The financial statement, F, is not materially misstated.

$\sim f$ – The financial statement, F, is materially misstated.

ao – There is no material misstatement related to objective 'O' of account 'A'.

$\sim ao$ – There is material misstatement related to objective 'O' of account 'A'.

$\boldsymbol{m}$-Functions, Belief Functions, and Plausibility Functions

$\boldsymbol{m}_A(\bullet)$ – $\boldsymbol{m}$-values at the level of account A for the proposition(s) in the argument.

$\boldsymbol{m}_F(\bullet)$ – $\boldsymbol{m}$-values at the level of the financial statement for the proposition(s) in the argument.

$\boldsymbol{m}_{AO}(\bullet)$ – $\boldsymbol{m}$-values at the level of the audit objective O of account A for the proposition(s) in the argument.

$\boldsymbol{m}_{CO}(\bullet)$ – $\boldsymbol{m}$-values obtained from internal *controls* and accounting systems at the audit *objective* level for the proposition(s) in the argument.

$\boldsymbol{m}_{DO}(\bullet)$ – $\boldsymbol{m}$-values obtained from *detailed* test of balance at the audit *objective* level for the proposition(s) in the argument.

$\boldsymbol{m}_{IA}(\bullet)$ – $\boldsymbol{m}$-values obtained from *inherent* factors at the *account* level for the proposition(s) in the argument.

$\boldsymbol{m}_{IO}(\bullet)$ – $\boldsymbol{m}$-values obtained from *inherent* factors at the audit *objective* level for the proposition(s) in the argument.

$\boldsymbol{m}_{IF}(\bullet)$ – $\boldsymbol{m}$-values obtained from *inherent* factors at the *financial* statement level for the proposition(s) in the argument.

$\boldsymbol{m}_{PA}(\bullet)$ – $\boldsymbol{m}$-values obtained from analytical *procedures* performed at the *account* level for the proposition(s) in the argument.

$\boldsymbol{m}_{PF}(\bullet)$ – $\boldsymbol{m}$-values obtained from analytical *procedures* performed at the *financial statement* level for the proposition(s) in the argument.

$\boldsymbol{m}_{PO}(\bullet)$ – $\boldsymbol{m}$-values obtained from analytical *procedures* performed at the audit *objective* level for the proposition(s) in the argument.

$\boldsymbol{Bel}_x(\bullet)$ – A belief function, x represents various indices as described above in the case of $\boldsymbol{m}$.

$\boldsymbol{PL}_x(\bullet)$ – A plausibility function, x again represents various indices as described in the case of $\boldsymbol{m}$.

Plausibility Functions for Material Misstatements
(i.e., Audit Risks in Belief-Function Framework)*

IR_{AO} – Plausibility of material misstatement in objective 'O' of account 'A' obtained from inherent factors at 'O' of 'A', or the risk (in the belief-function framework) associated with inherent factors at 'O' of account 'A'.

CR_{AO} – Control risk (plausibility of material misstatement on the basis of internal controls and accounting systems) for account 'A' at audit objective 'O'.

APR_{AO} – The risk associated with analytical procedures performed at the audit objective level.

(Continued)

Table 1. (Continued)

DR_{AO} – The risk associated with detailed test of balance for account 'A' at objective 'O'.

$AR_{AO} = IR_{AO}APR_{AO}CR_{AO}DR_{AO}$, audit risk at the audit *objective* level when considering only the items of evidence that directly support the audit objective 'O' of account 'A'.

IR_A – The risk associated with inherent factors at the *account* level.

APR_A – The risk associated with analytical procedures performed at the *account* level.

IR_F – The risk associated with inherent factors at the *financial* statement level.

APR_F – The risk associated with analytical procedures performed at the *financial* statement level.

$AR^t_{AO} = IR_F APR_F IR_A APR_A AR_{AO}$, total audit risk at the *objective* level considering all the items of evidence.

$AR^t_A = IR_F APR_F IR_A APR_A \left[1 - \prod_O (1 - AR_{AO})\right]$, total audit risk at the *account* level considering all the items of evidence.

$AR^t_F = IR_F APR_F \left[1 - \prod_A (1 - AR_A)\right]$, total audit risk at the *financial statement* level.

* Note that we have used the term "risk" for plausibility of material misstatement in the table and also in the text.

if the auditor plans an audit of an account to obtain an overall assurance of 0.95 (i.e., $\boldsymbol{Bel}(a) = 0.95$) that the account is not materially misstated then, in plausibility terms, it means that the auditor is planning the audit at the 0.05 level of plausibility for material error in the account (i.e., $\boldsymbol{PL}(\sim a) = 0.05$). In other words, if the auditor had to stop after obtaining 0.95 level of assurance that 'a' was true, then the evidence gathered up to that point would suggest that '$\sim a$' is plausible with degree 0.05; that is, there is a *maximum* risk of 0.05 that the account is materially misstated.[9]

In general, $\boldsymbol{Bel}(B) \leq \boldsymbol{PL}(B)$ for every subset B of our frame Θ. If we believe B, then we think B is plausible, but the converse is not necessarily true. A zero plausibility for a proposition means that we are sure that it is false (like a zero probability in the Bayesian theory), but a zero degree of belief for a proposition means only that we see no reason to believe the proposition.

Similar explanations can be given to the $\boldsymbol{m}$-values, belief functions, and plausibility functions for the other seven items of evidence presented in

[9] Here is another example. Suppose the evidence gathered up to this point gives us the following $\boldsymbol{m}$-values: $\boldsymbol{m}(a) = 0.95$, $\boldsymbol{m}(\sim a) = 0.02$, $\boldsymbol{m}(\{a, \sim a\}) = 0.03$. From (4) and (5), we obtain $\boldsymbol{Bel}(a) = 0.95$, $\boldsymbol{PL}(a) = 0.98$, $\boldsymbol{Bel}(\sim a) = 0.02$, $\boldsymbol{PL}(\sim a) = 0.05$. This means that we have 0.95 degree of belief that 'a' is true and 0.02 degree of belief that '$\sim a$' is true. However, the plausibility of '$\sim a$' is 0.05, which means that if we had to stop at this point, we would be taking a maximum risk of 0.05 that '$\sim a$' is true although the belief in '$\sim a$' is only 0.02.

Table 2. The individual $\boldsymbol{m}$-values given in Table 2 combined with Dempster's rule (see fn. 8) will give us the overall belief that the financial statement is fairly presented. Since these $\boldsymbol{m}$-values are defined at different nodes in an evidential network (e.g., see Fig. 1), combining them becomes a problem of propagating $\boldsymbol{m}$-values (or belief functions) through the network. We have discussed this problem in a working paper[10] and have summarized the results in appendix A.

4 $\boldsymbol{m}$-Values Directly Defined at Each Node

To combine all the evidence in Fig. 1, we need to combine the $\boldsymbol{m}$-values given in Table 2. There are eight sets of $\boldsymbol{m}$-values (four at the audit objective level, two at the account level, and two at the financial statement level; see Table 2). Combining these $\boldsymbol{m}$-values, in general, is very complex. However, the process is simplified if we proceed in two steps. First, we combine the $\boldsymbol{m}$-values directly bearing on each node in Fig. 1. For example, we combine the $\boldsymbol{m}$-values obtained from the four items of evidence at the audit objective level. Similarly, the total $\boldsymbol{m}$-values bearing directly at the account level and the financial statement level will be the combination of two $\boldsymbol{m}$-values defined at each level. Second, we propagate the above $\boldsymbol{m}$-values obtained at each node using the results of appendix A. In the following paragraphs, we discuss the first step and provide analytical formulas for the resultant $\boldsymbol{m}$-values at each level.

$\boldsymbol{m}$-Values at the Financial Statement Level

We assume that the two items of evidence directly bearing at the financial statement level are: (1) inherent factors and (2) analytical procedures (see fig. 1). We further assume that these items of evidence are affirmative in nature. The corresponding $\boldsymbol{m}$-values are given in Table 2. Since there is no conflict among the evidence, we obtain the following values directly defined at the financial statement level using Dempster's rule (see fn. 8):

$$\begin{aligned}\boldsymbol{m}_F(f) &= \boldsymbol{m}_{IF}(f)\,\boldsymbol{m}_{PF}(f) + \boldsymbol{m}_{IF}(f)\,\boldsymbol{m}_{PF}(\{f, -f\}) + \boldsymbol{m}_{IF}(\{f, \sim f\})\,\boldsymbol{m}_{PF}(f)\\ &= (1 - IR_F)(1 - APR_F) + (1 - IR_F)\,APR_F + IR_F(1 - APR_F)\\ &= 1 - IR_F APR_F, \qquad (8)\end{aligned}$$

$$\begin{aligned}\boldsymbol{m}_F(\sim f) = \boldsymbol{m}_{IF}(\sim f)\,\boldsymbol{m}_{PF}(\sim f) + \boldsymbol{m}_{IF}(\sim f)\,\boldsymbol{m}_{PF}(\{f, \sim f\})\\ + \boldsymbol{m}_{IF}(\{f, \sim f\})\,\boldsymbol{m}_{PF}(\sim f) = 0, \qquad (9)\end{aligned}$$

and

$$\boldsymbol{m}_F(\{f, \sim f\}) = \boldsymbol{m}_{IF}(\{f, \sim f\})\,\boldsymbol{m}_{PF}(\{f, \sim f\}) = IR_F APR_F, \qquad (10)$$

[10] This paper is available on request from the authors.

Table 2. $\boldsymbol{m}$-Values, belief, and plausibility functions for different audit evidence given in Fig. 1

Evidence	$\boldsymbol{m}$*-values*	*Belief*	*Plausibility*
Audit Objective Level:			
Inherent factors	$\boldsymbol{m}_{IO}(ao) = 1 - IR_{AO}$	$\boldsymbol{Bel}_{IO}(ao) = 1 - IR_{AO}$	$\boldsymbol{PL}_{IO}(ao) = 1$
	$\boldsymbol{m}_{IO}(\sim ao) = 0$	$\boldsymbol{Bel}_{IO}(\sim ao) = 0$	$\boldsymbol{PL}_{IO}(\sim ao) = IR_{AO}$
	$\boldsymbol{m}_{IO}(\{ao, \sim ao\}) = IR_{AO}$	$\boldsymbol{Bel}_{IO}(\{ao, \sim ao\}) = 1$	
Analytical procedures	$\boldsymbol{m}_{PO}(ao) = 1 - APR_{AO}$	$\boldsymbol{Bel}_{PO}(ao) = 1 - APR_{AO}$	$\boldsymbol{PL}_{PO}(ao) = 1$
	$\boldsymbol{m}_{PO}(\sim ao) = 0$	$\boldsymbol{Bel}_{PO}(\sim ao) = 0$	$\boldsymbol{PL}_{PO}(\sim ao) = APR_{AO}$
	$\boldsymbol{m}_{PO}(\{ao, \sim ao\}) = APR_{AO}$	$\boldsymbol{Bel}_{PO}(\{ao, \sim ao\}) = 1$	
Control factors and accounting systems	$\boldsymbol{m}_{CO}(ao) = 1 - CR_{AO}$	$\boldsymbol{Bel}_{CO}(ao) = 1 - CR_{AO}$	$\boldsymbol{PL}_{CO}(ao) = 1$
	$\boldsymbol{m}_{CO}(\sim ao) = 0$	$\boldsymbol{Bel}_{CO}(\sim ao) = 0$	$\boldsymbol{PL}_{CO}(\sim ao) = CR_{AO}$
	$\boldsymbol{m}_{CO}(\{ao, \sim ao\}) = CR_{AO}$	$\boldsymbol{Bel}_{CO}(\{ao, \sim ao\}) = 1$	
Detailed tests of balance	$\boldsymbol{m}_{DO}(ao) = 1 - DR_{AO}$	$\boldsymbol{Bel}_{DO}(ao) = 1 - DR_{AO}$	$\boldsymbol{PL}_{DO}(ao) = 1$
	$\boldsymbol{m}_{DO}(\sim ao) = 0$	$\boldsymbol{Bel}_{DO}(\sim ao) = 0$	$\boldsymbol{PL}_{DO}(\sim ao) = DR_{AO}$
	$\boldsymbol{m}_{DO}(\{ao, \sim ao\}) = DR_{AO}$	$\boldsymbol{Bel}_{DO}(\{ao, \sim ao\}) = 1$	
Account Level:			
Inherent factors	$\boldsymbol{m}_{IA}(a) = 1 - IR_{A}$	$\boldsymbol{Bel}_{IA}(a) = 1 - IR_{A}$	$\boldsymbol{PL}_{IA}(a) = 1$
	$\boldsymbol{m}_{IA}(\sim a) = 0$	$\boldsymbol{Bel}_{IA}(\sim a) = 0$	$\boldsymbol{PL}_{IA}(\sim a) = IR_{A}$
	$\boldsymbol{m}_{IA}(\{a, \sim a\}) = IR_{A}$	$\boldsymbol{Bel}_{IA}(\{a, \sim a\}) = 1$	
Analytical procedures	$\boldsymbol{m}_{PA}(a) = 1 - APR_{A}$	$\boldsymbol{Bel}_{PA}(a) = 1 - APR_{A}$	$\boldsymbol{PL}_{PA}(a) = 1$
	$\boldsymbol{m}_{PA}(\sim a) = 0$	$\boldsymbol{Bel}_{PA}(\sim a) = 0$	$\boldsymbol{PL}_{PA}(\sim a) = APR_{A}$
	$\boldsymbol{m}_{PA}(\{a, \sim a\}) = APR_{A}$	$\boldsymbol{Bel}_{PA}(\{a, \sim a\}) = 1$	
Financial Statement Level:			
Inherent factors	$\boldsymbol{m}_{IF}(f) = 1 - IR_{F}$	$\boldsymbol{Bel}_{IF}(f) = 1 - IR_{F}$	$\boldsymbol{PL}_{IF}(f) = 1$
	$\boldsymbol{m}_{IF}(\sim f) = 0$	$\boldsymbol{Bel}_{IF}(\sim f) = 0$	$\boldsymbol{PL}_{IF}(\sim f) = IR_{F}$
	$\boldsymbol{m}_{IF}(\{f, \sim f\}) = IR_{F}$	$\boldsymbol{Bel}_{IF}(\{f, \sim f\}) = 1$	
Analytical procedures	$\boldsymbol{m}_{PF}(f) = 1 - APR_{F}$	$\boldsymbol{Bel}_{PF}(f) = 1 - APR_{F}$	$\boldsymbol{PL}_{PF}(f) = 1$
	$\boldsymbol{m}_{PF}(\sim f) = 0$	$\boldsymbol{Bel}_{PF}(\sim f) = 0$	$\boldsymbol{PL}_{PF}(\sim f) = APR_{F}$
	$\boldsymbol{m}_{PF}(\{f, \sim f\}) = APR_{F}$	$\boldsymbol{Bel}_{PF}(\{f, \sim f\}) = 1$	

Note: The symbols are defined in Table 1. Each item of audit evidence is assumed to be affirmative.

From the above results, one can generalize that if the frame consists of only two elements, such as f, and $\sim f$, and the $\boldsymbol{m}$-values for $\sim f$ are zero for all the evidence, then the combined $\boldsymbol{m}$-value for $\sim f$ will be zero, irrespective of the number of items of evidence, and the combined $\boldsymbol{m}$-value for the entire frame $\{f, \sim f\}$ will be the product of its $\boldsymbol{m}$-values from each item of evidence.

$\boldsymbol{m}$-Values at the Account Level

As shown in Fig. 1, we consider two items of evidence at the account level: (1) inherent factors and (2) analytical procedures. Again, it is assumed that these items of evidence are affirmative in nature, and the corresponding $\boldsymbol{m}$-values are given in Table 2. The combined $\boldsymbol{m}$-values are similar to those in (8) through (10):

$$\boldsymbol{m}_A\left(a\right) = 1 - IR_A APR_A, \tag{11}$$

$$\boldsymbol{m}_A\left(\sim a\right) = 0, \tag{12}$$

and

$$\boldsymbol{m}_A\left(\{a, \sim a\}\right) = IR_A APR_A, \tag{13}$$

$\boldsymbol{m}$-Values at the Audit Objective Level

In general, there are four items of evidence at the audit objective level (see Fig. 1). Again, assume that they are all affirmative in nature. Then the corresponding $\boldsymbol{m}$-values can be given as in Table 2. We want to combine all the $\boldsymbol{m}$-values obtained from these items of evidence. Since the $\boldsymbol{m}$-values for $\sim ao$ for all the evidence are assuumed to be zero, the combined $\boldsymbol{m}$-value for $\sim ao$ is zero (see [9]), that is,

$$\boldsymbol{m}_{AO}\left(\sim ao\right) = 0. \tag{14}$$

Also, as seen in (10), since there is no conflict (i.e., $K = 1$ in Dempster's rule: see fn. 8), the combined $\boldsymbol{m}$-value for the entire frame $\{ao, \sim ao\}$ is equal to the product of the $\boldsymbol{m}$-values for the frame from each item of evidence, that is,

$$\begin{aligned}\boldsymbol{m}_{AO}\left(\{ao, \sim ao\}\right) &= \boldsymbol{m}_{IO}\left(\{ao, \sim ao\}\right)\boldsymbol{m}_{PO}\left(\{ao, \sim ao\}\right)\\ &\quad\times \boldsymbol{m}_{co}\left(\{ao, \sim ao\}\right)\boldsymbol{m}_{DO}\left(\{ao, \sim ao\}\right)\\ &= IR_{AO}APR_{AO}CR_{AO}DR_{AO}. \end{aligned}\tag{15}$$

Let us define a new term, AR_{AO}, for convenience as:

$$AR_{AO} = IR_{AO}APR_{AO}CR_{AO}DR_{AO.} \tag{16}$$

From (14) through (16) and the definition of $\boldsymbol{m}$ function, we obtain:

$$\begin{aligned} \boldsymbol{m}_{AO}(ao) &= 1 - \boldsymbol{m}_{AO}(\sim ao) - \boldsymbol{m}_{AO}(\{ao, \sim ao\}) \\ &= 1 - AR_{AO}. \end{aligned} \tag{17}$$

The corresponding belief and plausibility functions are (from [4] and [5]):

$$\begin{gathered} \boldsymbol{Bel}_{AO}(ao) = 1 - AR_{AO}, \\ \boldsymbol{Bel}_{AO}(\sim ao) = 0, \text{ and } \boldsymbol{Bel}_{AO}(\{ao, \sim ao\}) = 1, \\ \boldsymbol{PL}_{AO}(ao) = 1, \end{gathered}$$

and

$$\boldsymbol{PL}_{AO}(\sim ao) = AR_{AO}. \tag{18}$$

Thus, AR_{AO} is the total plausibility for $\sim ao$ at the audit objective level.

It is interesting to note from (16) and (18) that the total plausibility at the audit objective level for $\sim ao$ is the product of the individual plausibilities for $\sim ao$ (see Table 2). As discussed earlier in Sect. 3, plausibility is one interpretation of audit risk. So (18) along with (16) represents the audit risk model at the audit objective level without considering any other evidence at the account level or the financial statement level. This formula is similar to that of SAS No. 47. However, it is incomplete as an overall model for the audit risk because it does not include evidence at the other levels.

In the remainder of this article, we develop analytical formulas for the overall plausibility of material misstatement in the financial statement and the account and compare and contrast them with the SAS No. 47 formula. Since we have now determined the $\boldsymbol{m}$-values directly defined at each node of Fig. 1, we can use the results of appendix A to derive our formulas (see appendix B for details).

5 Audit-Risk Formulas in the Belief-Function Framework

In this section, we give formulas for the overall plausibility of material misstatement at various levels of the financial statement. These formulas have been derived in appendix B.

Total Audit Risk (Plausibility of Material Misstatement) at the Financial Statement Level

We have the following expression for total plausibility of material misstatement at the financial statement level (B-7):

$$\boldsymbol{PL}_F^t(\sim f) = AR_F^t = IR_F APR_F \left[1 - \prod_A (1 - AR_A)\right], \tag{19}$$

where AR_A and AR_{AO} are defined in (B.4) and (16), respectively, as:

$$AR_A = IR_A APR_A \left[1 - \prod_o (1 - AR_{AO})\right], \tag{20}$$

and

$$AR_{AO} = IR_{AO} APR_{AO} CR_{AO} DR_{AO}. \tag{21}$$

Also, the total belief that the financial statement is fairly presented is given by (B.7):

$$\boldsymbol{Bel}^t_F(f) = 1 - IR_F APR_F \left[1 - \prod_A (1 - AR_A)\right]. \tag{22}$$

Equation (19) represents total plausibility of material misstatement in the financial statement or total audit risk, AR^t_F, at the financial statement level. The total belief that the financial statement is not materially misstated is given by (22). It should be noted that the algebraic form of (19) is very different from the formula discussed in SAS No. 47 or the Bayesian model. Unlike the audit risk model of SAS No. 47 or the Bayesian model, (19) takes into consideration all the evidence at all the levels of the financial statement. It also differs from SAS No. 47, of course, in the interpretation. Here, we interpret audit risk as a plausibility, not as a probability.

Total Audit Risk (Plausibility of Material Misstatement) at the Account Level

From (B.11), we have the following expression for total plausibility of material misstatement at the account level:

$$\boldsymbol{PL}^t_A(\sim a) = AR^t_A = IR_F APR_F IR_A APR_A \left[1 - \prod_o (1 - AR_{AO})\right], \tag{23}$$

and the total belief that the account 'A' is not materially misstated as:

$$\boldsymbol{Bel}^t_A(a) = 1 - AR^t_A = 1 - IR_F APR_F IR_A APR_A \left[1 - \prod_o (1 - AR_{AO})\right], \tag{24}$$

Here, (23) represents total plausibility of material misstatement or total audit risk, AR^t_A, at the account level. The total belief or assurance that the account is not materially misstated is given by (24). It is again the result of aggregating all the evidence at the account level, whether the evidence is coming from the audit objective level, the financial statement level, or directly bearing on the account. It again differs from the Bayesian or SAS No. 47 formula. In (23), we find that AR^t_A is the product of three types of plausibilities: (1) plausibility arising from inherent factors (i.e., the inherent risk, $IR_F IR_A$), (2) plausibility arising from analytical procedures (i.e., the analytical procedure risk, $APR_F APR_A$), and (3) plausibility arising from the evidence at the audit objective level for the account (i.e., the combined audit risk, $[1 - \Pi_o(1 - AR_{AO})]$). The third term represents 1 minus the level of support obtained from the procedures performed at the audit objective level. If no procedures are performed at that level, which means $AR_{AO} = 1$, then the support is zero and the third term equals 1. We will give a numerical example later.

Total Audit Risk (Plausibility of Material Misstatement) at the Audit Objective Level

The total plausibility of material misstatement and total belief at the audit objective level are given by (B.15) and (B.17):

$$\boldsymbol{PL}^t_{AO}\,(\sim ao) = AR^t_{AO} = IR_F APR_F IR_A APR_A AR_{AO}; \qquad (25)$$

$$\boldsymbol{Bel}^t_{AO}\,(ao) = 1 - IR_F APR_F IR_A APR_A AR_{AO}. \qquad (26)$$

Equation (25) represents the total plausibility that the audit objective '*AO*' will not be met when all the evidence at various levels has been aggregated. The total belief that the objective will be met is given by (26). As seen in (25), the total risk at the audit objective level is the product of three terms, $(IR_F APR_F)$, $(IR_A APR_A)$, and $(IR_{AO} APR_{AO} CR_{AO} DR_{AO})$, each defined at different levels. This formula resembles the multiplicative formula of SAS No. 47 if we separate the risks associated with inherent factors and analytical prodcedures:

$$AR^t_{AO} = (IR_F IR_A IR_{AO})\,(APR_F APR_A APR_{AO})\,(CR_{AO} DR_{AO})\,. \qquad (27)$$

The first factor in (27) determines the overall risk associated with inherent factors. Similarly, the second term represents the overall risk associated with analytical procedures performed at all levels. The third term is the product of control risk and detection risk. We must repeat that, although (27) is similar to the SAS No. 47 model, our interpretation of the risk is very different.

Numerical Example

Suppose we have only five accounts on the balance sheet and each account has five objectives. Suppose the auditor has gathered and evaluated all the relevant inherent factors at the level of the financial statement and the account and has assigned the following values for the respective plausibilities of material misstatement or risks: $IR_F = 0.7$ and $IR_A = 0.6$ for all the accounts. Also assume that the auditor has performed analytical procedures for various accounts and assigned a plausibility of material misstatement or risk of 0.4 to these procedures, but has not performed any analytical procedures at the financial statement level. Thus, $APR_A = 0.4$ for all the accounts and $APR_F = 1$. These values result in a total plausibility of error or risk at the financial statement level of 0.52, that is, $AR^t_F = 0.52$ (Table 3). This implies that, on the basis of

Table 3. Audit risk model in belief-function framework with IR_A, APR_A, and IR_F as inputs

Panel A. Input Risks:

(1) Risks at the Audit Objective Level

	IR_{AO}					APR_{AO}					CR_{AO}					DR_{AO}				
	Objective					Objective					Objective					Objective				
Account	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5
A_1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
A_2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
A_3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
A_4	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
A_5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

(2) Risks at the Account Level

Account	IR_A	APR_A
A_1	0.6	0.4
A_2	0.6	0.4
A_3	0.6	0.4
A_4	0.6	0.4
A_5	0.6	0.4

(3) Risks at the Financial Statement Level

$IR_F = 0.7$

$APR_F = 1.0$

Panel B. Output Risks:

	(1) *Total Risk at the Audit Objective Level (AR'_{AO}) (From eq. [25])*					(2) *Total Risk at the Account Level (AR'_A) (From eq. [23])*	(3) *Total Risk and Belief at the Financial Statement Level (From eq. [19])*
Account	O_1	O_2	O_3	O_4	O_5		
A_1	0.168	0.168	0.168	0.168	0.168	0.168	
A_2	0.168	0.168	0.168	0.168	0.168	0.168	$AR'_F = 0.52$
A_3	0.168	0.168	0.168	0.168	0.168	0.168	$Bel'_F(f) = 0.48$
A_4	0.168	0.168	0.168	0.168	0.168	0.168	
A_5	0.168	0.168	0.168	0.168	0.168	0.168	

Note: The term "risk" is used for plausibility of material misstatement. The symbols used in the table are defined in table 1.

evidence accumulated at the financial statement level and the account level, the auditor finds that the financial statement is not materially misstated with a belief or assurance of 0.48. This information would help the auditor plan a more efficient audit than is possible with the SAS No. 47 approach.

Let us now assume that the auditor plans to proceed to the detailed level and considers the following steps: (1) collect and evaluate inherent factors at the audit objective level, (2) perform analytical procedures at the audit objective level if appropriate, (3) study and evaluate client's accounting systems and control procedures and perform test of transactions, and (4) perform direct test of balance. The results of the auditor's judgment are shown in terms of plausibilities of material misstatement or risks as inputs in tables 4 through 7.

In Table 4, we see that there is almost no impact of inherent factors at the audit objective level on AR_F^t and AR_A^t. However, AR_{AO}^t reduces from 16.8 percent to 11.8 percent. Consideration of analytical procedures at the audit objective level reduces AR_F^t to 50 percent, AR_A^t to 15.7 percent, and AR_{AO}^t to 7.1 percent, as shown in Table 5. When accounting systems and control procedures are included in the model, the total plausibility of material misstatement (i.e., the total risk at various levels) is further reduced. $AR_F^t =$ 28 percent, AR_A^t varies between 6.0 and 10.1 percent, and AR_{AO}^t between 1.4 and 2.8 percent (see Table 6).

At this stage, the auditor decides about the extent, timing, and nature of the detailed test of balance so as to obtain the overall plausibility of material misstatement at 0.05 or an overall assurance of 0.95. As shown in Table 7, for the level of risk given in the table for each audit objective from the corresponding detailed test of balance, we obtain the desired 0.95 level of assurance or belief that the financial statement is not materially misstated (i.e., an overall plausibility or total audit risk of 0.05). The risks at the other levels vary as follows: AR_A^t varies between 1.0 and 1.4 percent, and AR_{AO}^t varies between 0.2 and 0.3 percent (see Table 7).

It is important to note that consideration of the structure of evidence in our plausibility models makes the audit process more efficient. In other words, the auditor will plan less extensive tests at the audit objective level when the evidence at the financial statement level and account level is positive. One should keep in mind that certain required procedures at the detailed level must be performed; the evidence at the financial and account levels would affect only the extent of testing. As seen from Table 8, when the evidence at the account level and the financial statement level is not included in the plausibility model (as done in SAS No. 47 for the audit risk model), the total plausibility of material misstatement or the total audit risk at the financial statement level is high ($AR_F^t = 29$ percent) or the total belief that the financial statement is not materially misstated is low (0.71). Thus, without an explicit treatment of the evidence at the financial statement level and the account level, the auditor will always under-estimate the overall assurance and will collect more evidence than necessary at the detail level.

Table 4. Audit risk model in belief-function framework with IR_{AO}, IR_A, APR_A, and IR_F as inputs

Panel A. Input Risks:

(1) Risks at the Audit Objective Level

	IR_{AO}					APR_{AO}					CR_{AO}					DR_{AO}				
	Objective					*Objective*					*Objective*					*Objective*				
Account	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5
A_1	0.7	0.7	0.7	0.7	0.7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
A_2	0.7	0.7	0.7	0.7	0.7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
A_3	0.7	0.7	0.7	0.7	0.7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
A_4	0.7	0.7	0.7	0.7	0.7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
A_5	0.7	0.7	0.7	0.7	0.7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

(2) Risks at the Account Level

Account	IR_A	APR_A
A_1	0.6	0.4
A_2	0.6	0.4
A_3	0.6	0.4
A_4	0.6	0.4
A_5	0.6	0.4

(3) Risks at the Financial Statement Level

$IR_F = 0.7$

$APR_F = 1.0$

Panel B. Output Risks:

	(1) *Total Risk at the Audit Objective Level* (AR'_{AO}) (*From eq.* [25])					(2) *Total Risk at the Account Level* (AR'_A) (*From eq.* [23])	(3) *Total Risk and Belief at the Financial Statement Level* (*From eq.* [19])
Account	O_1	O_2	O_3	O_4	O_5		
A_1	0.118	0.118	0.118	0.118	0.118	0.168	
A_2	0.118	0.118	0.118	0.118	0.118	0.168	$AR'_F = 0.52$
A_3	0.118	0.118	0.118	0.118	0.118	0.168	$\mathbf{Bel}'_F(f) = 0.48$
A_4	0.118	0.118	0.118	0.118	0.118	0.168	
A_5	0.118	0.118	0.118	0.118	0.118	0.168	

Note: The term "risk" is used for plausibility of material misstatement. The symbols used in the table are defined in table 1.

Table 5. Audit risk model in belief-function framework with IR_{AO}, APR_{AO}, IR_A, APR_A, and IR_F as inputs

Panel A. Input Risks:

(1) Risks at the Audit Objective Level

	IR_{AO}					APR_{AO}					CR_{AO}					DR_{AO}				
	Objective					Objective					Objective					Objective				
Account	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5
A_1	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
A_2	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
A_3	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
A_4	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
A_5	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

(2) Risks at the Account Level

Account	IR_A	APR_A
A_1	0.6	0.4
A_2	0.6	0.4
A_3	0.6	0.4
A_4	0.6	0.4
A_5	0.6	0.4

(3) Risks at the Financial Statement Level

$IR_F=0.7$

$APR_F=1.0$

Panel B. Output Risks:

	(1) *Total Risk at the Audit Objective Level* (AR'_{AO}) (From eq. [25])					(2) *Total Risk at the Account Level* (AR'_A) (From eq. [23])	(3) *Total Risk and Belief at the Financial Statement Level* (From eq. [19])
Account	O_1	O_2	O_3	O_4	O_5		
A_1	0.071	0.071	0.071	0.071	0.071	0.157	
A_2	0.071	0.071	0.071	0.071	0.071	0.157	$AR'_F=0.50$
A_3	0.071	0.071	0.071	0.071	0.071	0.157	$\mathbf{Bel}'_F(f)=0.50$
A_4	0.071	0.071	0.071	0.071	0.071	0.157	
A_5	0.071	0.071	0.071	0.071	0.071	0.157	

Note: The term "risk" is used for plausibility of material misstatement. The symbols used in the table are defined in table 1.

Table 6. Audit risk model in belief-function framework with IR_{AO}, APR_{AO}, CR_{AO}, IR_A, APR_A, and IR_F as inputs

Panel A. Input Risks:

(1) Risks at the Audit Objective Level

	IR_{AO}					*APR_{AO}*					*CR_{AO}*					*DR_{AO}*				
	Objective					*Objective*					*Objective*					*Objective*				
Account	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5
A_1	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.2	0.2	0.2	0.2	0.2	1.0	1.0	1.0	1.0	1.0
A_2	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.2	0.2	0.2	0.2	0.2	1.0	1.0	1.0	1.0	1.0
A_3	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.4	0.4	0.4	0.4	0.4	1.0	1.0	1.0	1.0	1.0
A_4	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.2	0.2	0.2	0.2	0.2	1.0	1.0	1.0	1.0	1.0
A_5	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.2	0.2	0.2	0.2	0.2	1.0	1.0	1.0	1.0	1.0

(2) Risks at the Account Level

Account	*IR_A*	*APR_A*
A_1	0.6	0.4
A_2	0.6	0.4
A_3	0.6	0.4
A_4	0.6	0.4
A_5	0.6	0.4

(3) Risks at the Financial Statement Level

$IR_F=0.7$

$APR_F=1.0$

Panel B. Output Risks:

	(1) *Total Risk at the Audit Objective Level (AR'_{AO}) (From eq. [25])*					(2) *Total Risk at the Account Level (AR'_A) (From eq. [23])*	(3) *Total Risk and Belief at the Financial Statement Level (From eq. [19])*
Account	O_1	O_2	O_3	O_4	O_5		
A_1	0.071	0.014	0.014	0.014	0.014	0.060	
A_2	0.071	0.014	0.014	0.014	0.014	0.060	$AR'_F=0.28$
A_3	0.071	0.028	0.028	0.028	0.028	0.101	**$Bel'_F(f)=0.72$**
A_4	0.071	0.014	0.014	0.014	0.014	0.060	
A_5	0.071	0.014	0.014	0.014	0.014	0.060	

Note: The term "risk" is used for plausibility of material misstatement. The symbols used in the table are defined in table 1.

Table 7. Audit risk model in belief-function framework with IR_{AO}, APR_{AO}, CR_{AO}, DR_{AO}, IR_A, APR_A, and IR_F as inputs

Panel A. Input Risks:

(1) Risks at the Audit Objective Level

	IR_{AO}					APR_{AO}					CR_{AO}					DR_{AO}				
	Objective					Objective					Objective					Objective				
Account	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5
A_1	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.2	0.2	0.2	0.2	0.2	0.15	0.15	0.15	0.15	0.15
A_2	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.2	0.2	0.2	0.2	0.2	0.15	0.15	0.15	0.15	0.15
A_3	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.4	0.4	0.4	0.4	0.4	0.10	0.10	0.10	0.10	0.10
A_4	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.2	0.2	0.2	0.2	0.2	0.15	0.15	0.15	0.15	0.15
A_5	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.2	0.2	0.2	0.2	0.2	0.15	0.15	0.15	0.15	0.15

(2) Risks at the Account Level

Account	IR_A	APR_A
A_1	0.6	0.4
A_2	0.6	0.4
A_3	0.6	0.4
A_4	0.6	0.4
A_5	0.6	0.4

(3) Risks at the Financial Statement Level

$IR_F = 0.7$

$APR_F = 1.0$

Panel B. Output Risks:

	(1) *Total Risk at the Audit Objective Level (AR'_{AO})* *(From eq. [25])*					(2) *Total Risk at the Account Level (AR'_A)* *(From eq. [23])*	(3) *Total Risk and Belief at the Financial Statement Level* *(From eq. [19])*
Account	O_1	O_2	O_3	O_4	O_5		
A_1	0.002	0.002	0.002	0.002	0.002	0.010	
A_2	0.002	0.002	0.002	0.002	0.002	0.010	$AR'_F = 0.05$
A_3	0.003	0.003	0.003	0.003	0.003	0.014	$Bel'_F(f) = 0.95$
A_4	0.002	0.002	0.002	0.002	0.002	0.010	
A_5	0.002	0.002	0.002	0.002	0.002	0.010	

Note: The term "risk" is used for plausibility of material misstatement. The symbols used in the table are defined in table 1.

Table 8. Audit risk model in belief-function framework with IR_{AO}, APR_{AO}, CR_{AO}, DR_{AO} as inputs

Panel A. Input Risks:

(1) Risks at the Audit Objective Level

	IR_{AO}					APR_{AO}					CR_{AO}					DR_{AO}				
	Objective					Objective					Objective					Objective				
Account	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5	O_1	O_2	O_3	O_4	O_5
A_1	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.2	0.2	0.2	0.2	0.2	0.15	0.15	0.15	0.15	0.15
A_2	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.2	0.2	0.2	0.2	0.2	0.15	0.15	0.15	0.15	0.15
A_3	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.4	0.4	0.4	0.4	0.4	0.10	0.10	0.10	0.10	0.10
A_4	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.2	0.2	0.2	0.2	0.2	0.15	0.15	0.15	0.15	0.15
A_5	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6	0.2	0.2	0.2	0.2	0.2	0.15	0.15	0.15	0.15	0.15

(2) Risks at the Account Level

Account	IR_A	APR_A
A_1	1.0	1.0
A_2	1.0	1.0
A_3	1.0	1.0
A_4	1.0	1.0
A_5	1.0	1.0

(3) Risks at the Financial Statement Level

$IR_F = 1.0$

$APR_F = 1.0$

Panel B. Output Risks:

	(1) *Total Risk at the Audit Objective Level (AR'_{AO}) (From eq. [25])*					(2) *Total Risk at the Account Level (AR'_A) (From eq. [23])*	(3) *Total Risk and Belief at the Financial Statement Level (From eq. [19])*
Account	O_1	O_2	O_3	O_4	O_5		
A_1	0.013	0.013	0.013	0.013	0.013	0.061	
A_2	0.013	0.013	0.013	0.013	0.013	0.061	$AR'_F = 0.29$
A_3	0.017	0.017	0.017	0.017	0.017	0.081	**$Bel'_F(f) = 0.71$**
A_4	0.013	0.013	0.013	0.013	0.013	0.061	
A_5	0.013	0.013	0.013	0.013	0.013	0.061	

Note: The term "risk" is used for plausibility of material misstatement. The symbols used in the table are defined in table 1.

6 Limitations of the Models

We must emphasize that there are strong limitations on the applicability of the formulas we have derived, limitations that are shared by the SAS No. 47 formula and existing Bayesian formulas.

First, since we have considered only binary variables, we have distinguished only whether an account is materially misstated or not. We have not distinguished between material misstatement due to overstatement or understatement. This limitation will make the audit process less efficient. For example, if there are two accounts, one materially overstated and the other materially understated by the same amount, and the auditor feels that the combination of the two accounts is fairly stated because of the off-setting errors, the present approach will suggest that the combination is materially misstated and hence lead to inefficiency.

Second, we have not considered immaterial errors in individual accounts that might add up to a material error in the whole financial statement. This will also make the audit less efficient.

Third, we have assumed that each account or audit objective is equally important. This assumption may make the audit process less efficient because the auditor may still have to obtain a high level of assurance for an unimportant audit objective.

We have already mentioned other limitations due to our simplifying assumptions. We considered only a tree-type evidential structure, and only affirmative evidence is considered. As we have already explained, it may be feasible to derive formulas with these assumptions relaxed to some extent, but an algorithmic approach to the more complex case is probably preferable.

When discussing the increased efficiency possible with more accurate representation of the structure of the audit evidence, we must always bear in mind that the decrease in audit effort that is implied by plans based on such structure can decrease audit effectiveness if the inputs to the model cannot be estimated accurately by auditors who employ it. We think this gives another advantage to the belief-function approach over the Bayesian approach, since the Bayesian approach demands, in general, probability judgments that may not be available or meaningful.

Conclusion

This article has used the structure of audit evidence currently assumed in auditing standards. We have represented this structure by a network of variables; including variables that represented the financial statement as a whole, various accounts, and various audit objectives of each account. For simplicity, we have considered only tree-type evidential structures and only variables with two possible values.

We have used the belief-function framework to relate audit risk to plausibility of material misstatement and have derived formulas[11] for the total plausibility at the three levels: financial statement level, account level, and audit objective level. The formula for the total plausibility of material misstatement at the audit objective level resembles the formula of SAS No. 47. The total plausibility of material misstatement associated with inherent factors is the product of risks at the financial statement level, the account level, and the audit objective level. Similarly, the plausibility of error associated with analytical procedures is the product of risks at the three levels.

The formulas at the other two levels are significantly different from the SAS No. 47 model. Unlike the SAS No. 47 model or the Bayesian model, all the formulas developed in this article aggregate all the evidence obtained from procedures performed at various levels of the account.

It must be emphasized that interpretation of audit risk in this article is significantly different from the probability interpretation used in the auditing literature. For example, in the belief-function framework, the detection risk, DR_{AO}, is the plausibility that objective 'O' of account 'A' is materially misstated, whereas the SAS No. 47 model interprets DR_{AO} as the probability that the auditor's procedures will fail to detect material misstatements related to objective 'O' of account 'A', given that the internal control procedures have failed to prevent or detect and correct such misstatements.

The formulas derived here provide an audit planning tool that can be used to determine the level of assurance to be obtained from various sources to achieve a desired level of overall assurance or an overall plausibility of material misstatement. Unlike the SAS No. 47 formula, these formulas take into consideraion all the items of evidence gathered by the auditor, whether that evidence bears on the objective level, the account level, or the financial statement level. Thus, our approach provides an improvement over the SAS No. 47 model for planning an efficient audit and evaluating the results.

There are many other issues that need to be addressed, including (1) the level of testing needed for a desired level of support or belief, (2) the level of support or belief obtained from a given statistical result of a test procedure for different variables, (3) how to integrate belief aggregation with the cost of evidence gathering to obtain the most effective and efficient audit strategy, and (4) whether a belief-function approach to planning and evaluation of an audit is cost effective. All of these issues require further research.

7 Appendix A Propagation of m-Values in an "And" Tree

Propagation of beliefs (or m-values) in a network is quite complex. Extensive work on this topic is reported in the artificial intelligence literature. To

[11] As Boritz and Wensley (1990) has shown, almost identical expressions for the overall audit risk can be obtained with probability theory.

mention a few studies, the work by Shafer (1976, 1986), Shafer and Shenoy (1988, 1990), Shenoy and Shafer (1988, 1990), Zarley et al. (1988), and Shenoy (1989) is important for this purpose.

We will summarize results of propagation of beliefs in an “and” tree. These results will be useful in deriving analytical formulas in appendix B. There are two important directions for propagation for the overall aggregation of beliefs in an “and” tree. One direction is from the subobjectives to the main objective. The second direction is to propagate to a given subobjective from the main objective and the other subobjectives. The details of this work are given in a working paper that is available from the authors on request.

Propagation of $\boldsymbol{m}$-Values from Audit Objectives to the Respective Accounts

Here, we summarize the results of propagation of $\boldsymbol{m}$-values in an “and” tree (Fig. 1) from subobjectives to the main objectives, that is, from audit objectives to corresponding accounts or from accounts to the financial statement. In the case of propagation from audit objectives (AOs) to an account ‘A’, one obtains the following values:

$$\boldsymbol{m}'_{A-\text{all }AO}\left(a\right) = \prod_{i=1}^{n} \boldsymbol{m}_{AO_i}\left(ao_i\right), \tag{A.1}$$

$$\begin{aligned}
\boldsymbol{m}'_{A-\text{all }AO}\left(\sim a\right) = & \sum_{i=1}^{n} \boldsymbol{m}_{AO_i}\left(\sim ao_i\right) \prod_{j\neq i}^{n} \boldsymbol{m}_{AO_j}\left(ao_j\right) \\
& + \sum_{i=1}^{n} \sum_{j\neq i}^{n} \boldsymbol{m}_{AO_i}\left(\sim ao_i\right) \boldsymbol{m}_{AQ_j}\left(\sim ao_j\right) \\
& \times \prod_{k\neq i,k\neq j}^{n} \boldsymbol{m}_{AO_k}\left(ao_k\right) \\
& + \ldots + \ldots + \prod_{i=1}^{n} \boldsymbol{m}_{AO_i}\left(\sim ao_i\right) \\
& + \sum_{i=1}^{n} \boldsymbol{m}_{AO_i}\left(\sim ao_i\right) \prod_{j\neq i}^{n} \boldsymbol{m}_{AO_i}\left(\{ao_j, \sim ao_j\}\right) \\
& + \sum_{i=1}^{n} \sum_{j\neq i}^{n} \boldsymbol{m}_{AO_i}\left(\sim ao_i\right) \boldsymbol{m}_{AO_j}\left(\sim ao_j\right) \\
& \times \prod_{k\neq j,k\neq i}^{n} \boldsymbol{m}_{AO_k}\left(|ao_k, \sim ao_k|\right) + \ldots + \\
& + \sum_{j=1}^{n} \prod_{i\neq i}^{n} \boldsymbol{m}_{AO_i}\left(\sim ao_i\right) \boldsymbol{m}_{AO_i}\left(\{ao_j, \sim ao_j\}\right),
\end{aligned} \tag{A.2}$$

and

$$\boldsymbol{m}'_{A-\text{all }AO}\left(\{a, \sim a\}\right) = 1 - \boldsymbol{m}'_{A-\text{all }AO}\left(a\right) - \boldsymbol{m}'_{A-\text{all }AO}\left(\sim a\right). \tag{A.3}$$

These values may look very complex but they are intuitive. For example, (A.1) represents the propagated $\boldsymbol{m}$-value for the state 'a' that the account 'A' is not materially misstated. This value is equal to the product of all the $\boldsymbol{m}$-values defined at each audit objective for 'ao' that the objective is met. This result is similar to the probability rule giving the probability that the account is not materially misstated as equal to the product of the individual probabilities that the audit objectives are met provided the "and" relationship is valid.

Equation (A.2) is also intuitive. It represents the resultant $\boldsymbol{m}$-value for '$\sim a$' received from all its audit objectives. The account is materially misstated when at least one or all of the objectives are not met. There are several situations that contribute to this condition: (1) at least one of the objectives is not met but the rest have been met, (2) all the objectives are not met, and (3) at least one of the objectives is not met but we have no knowledge whether the rest have been met or not. The first two terms in (A.2) represent the first situation. The third term represents the second situation. The last three terms represent the third situation.

For affirmative items of evidence (i.e., for $\boldsymbol{m}_A(\sim a) = 0$ and $\boldsymbol{m}_{AO_j}(\sim ao_j) = 0$ for all j), (A.2) reduces to the following simple form:

$$\boldsymbol{m}'_{A-\text{all }AO}\left(\sim a\right) = 0, \tag{A.4}$$

while (A.1) and (A.3) remain unchanged.

Propagation of $\boldsymbol{m}$-Values to a Given Subobjective from the Main Objective and the Other Subobjectives

In this section, we want to summarize the results of propagation to a given subobjective from the main objective and the other subobjectives. For example, the $\boldsymbol{m}$-values propagated to a given audit objective AO_i from account 'A' and the remaining audit objectives can be given by:

$$\boldsymbol{m}'_{AO_i-A\text{ \& all other }AO'\text{s}}\left(ao_i\right) = K_i^{-1}\boldsymbol{m}_A\left(a\right)\prod_{j\neq i}^{n}\left[1 - \boldsymbol{m}_{AO_j}\left(\sim ao_j\right)\right], \tag{A.5}$$

$$\boldsymbol{m}'_{AO_i-A\text{ \& all other }AO'\text{s}}\left(\sim ao_i\right) = K_i^{-1}\boldsymbol{m}_A\left(\sim a\right)\prod_{j\neq i}^{n}\boldsymbol{m}_{AO_j}\left(ao_j\right), \tag{A.6}$$

$$\begin{aligned}\boldsymbol{m}'_{AO_i-A\text{ \& all other }AO'\text{s}}\left(\{ao_i, \sim ao_i\}\right) &= 1 - \boldsymbol{m}'_{AO_i-A\text{ \& all other }AO'\text{s}}\left(ao_i\right)\\ &\quad - \boldsymbol{m}'_{AO_i-A\text{ \& all other }AO'\text{s}}\left(\sim ao_i\right),\end{aligned} \tag{A.7}$$

where K_i is the renormalization constant that can be written as $K_i = [1 - \boldsymbol{m}_A(a)C_i]$ and C_i is given by:

$$\begin{aligned}C_i =&\sum_{j\neq i}^{n} \boldsymbol{m}_{AO_j}\left(\sim ao_j\right)\prod_{k\neq j}^{n} \boldsymbol{m}_{AO_k}\left(ao_k\right)\\&+\sum_{j\neq i}^{n}\sum_{k\neq j}^{n} \boldsymbol{m}_{AO_j}\left(\sim ao_j\right)\boldsymbol{m}_{AO_k}\left(\sim ao_k\right)\prod_{l\neq j,l\neq k}^{n} \boldsymbol{m}_{AO_i}\left(ao_i\right)\\&+\ldots+\ldots+\prod_{j\neq i}^{n} \boldsymbol{m}_{AO_i}\left(\sim ao_j\right)\\&+\sum_{j\neq i}^{n} \boldsymbol{m}_{AO_j}\left(\sim ao_j\right)\prod_{k\neq j}^{n} \boldsymbol{m}_{AO_k}\left(\{ao_k,\sim ao_k\}\right)\\&+\sum_{j\neq i}^{n}\sum_{k\neq i}^{n} \boldsymbol{m}_{AO_i}\left(\sim ao_j\right)\boldsymbol{m}_{AO_k}(\sim ao_k)\\&\times\prod_{l\neq i,l\neq j}^{n} \boldsymbol{m}_{AO_i}\left(\{ao_l,\sim ao_l\}\right)+\cdots+\\&+\sum_{j\neq i}^{n}\prod_{k\neq j}^{n} \boldsymbol{m}_{AO_k}\left(\sim ao_k\right)\boldsymbol{m}_{AO_j}\left(\{ao_j,\sim ao_j\}\right).\end{aligned}\tag{A.8}$$

Again, the above equations may appear complex but the results are intuitive. For example, the $\boldsymbol{m}$-value expressed in (A.5) represents that the audit objective AO_i is met if account 'A' is not materially misstated ($\boldsymbol{m}_A(a) \neq 0$) and the other audit objectives are either met or their status is unknown ($\boldsymbol{m}_{AO_j}(\sim ao_j) < 1$). Equation (A.6) is easier to understand. It tells us that there is finite support for the audit objective's not having been met if the account is found to be materially misstated but the other audit objectives have been met. The renormalization constant, K_i, in (A.5) through (A.7) represents 1 minus the conflict term. In general, the conflict term is defined as the sum of all the products of $\boldsymbol{m}$-values of the sets of elements whose inter-section is a null set. In our case, conflict will occur when there is a finite support that the account is not materially misstated ($\boldsymbol{m}_A(a) \neq 0$) and at least one of the other audit objectives is not met ($C_i \neq 0$).

In the case of affirmative items of evidence (i.e., $\boldsymbol{m}_A(\sim a) = 0$ and $\boldsymbol{m}_{AO_j}(\sim ao_j) = 0$ for all j), which is the case assumed in the present work, the renormalization constant becomes 1 and (A.5) and (A.6) reduce to:

$$\boldsymbol{m}'_{AO,-A \text{ \& all other } AO\text{'s}}\left(ao_i\right) = \boldsymbol{m}_A\left(a\right),\tag{A.9}$$

$$\boldsymbol{m}'_{AO,-A \text{ \& all other } AO\text{'s}}\left(\sim ao_i\right) = 0,\tag{A.10}$$

while (A.7) remains unchanged. From (A.9), we find that, in the case of affirmative items of evidence, when $\boldsymbol{m}$-values propagate to a given subobjective from the main objective and the other subobjectives, the $\boldsymbol{m}$-values received by the subobjective are not a function of the $\boldsymbol{m}$-values of the other subobjectives.

8 Appendix B Derivation of Plausibility Formulas or Audit Risk Formulas

In this appendix, our main objective is to derive formulas for plausibility of material misstatement (i.e., audit-risk formulas; see fn. 2) within the belief-function framework for an evidential network (an "and" tree) given in Fig. 1. To achieve this objective, we need to aggregate all the items of evidence in Fig. 1. As discussed earlier, the process of combining various $\boldsymbol{m}$-values in a network is, in fact, a process of propagating $\boldsymbol{m}$-values through the network of variables. We will use the general results presented in appendix A to derive the audit risk formulas in the following sections. As mentioned earlier, the following assumptions are made in the derivation of the formulas.

First, each node (variable, such as an account or audit objective) in the tree is assumed to be a binary variable. In fact, the variables in an "and" tree are binary by definition. The general approach of aggregating evidence by using the belief-function framework is still valid for nonbinary variables, but the analytical formulas for such cases will be intractable. Second, each item of evidence is assumed to be affirmative in nature (i.e., the evidence provides positive support to the assertion and no support to its negation).

Financial Statement Level

The objective here is to derive a formula for the total plausibility of material misstatement or audit risk, at the financial statement level. This is achieved by combining all $\boldsymbol{m}$-values that are either defined directly at the financial statement node (F) or have been propagated to F from all other nodes in the network (see Fig. 2). We will proceed with the propagation process in two steps. First, we will consider propagation from audit objectives (AO's) to an account (A) and then combine these values from the audit objectives with the $\boldsymbol{m}$-values at the account. Second, we propagate the resultant $\boldsymbol{m}$-values from various accounts to the financial statement node and combine them with the $\boldsymbol{m}$-values at the financial statement level. This process will yield the desired audit risk (plausibility of material misstatement) formula.

For the first step above, as an illustration, let us consider propagation of $\boldsymbol{m}$-values defined at the three audit objectives of account 'A_3' in Fig. 2 to the account. These $\boldsymbol{m}$-values are the result of combining all four items of evidence at the audit objective level and are given, in general terms, in (14) through (4). When these $\boldsymbol{m}$-values are propagated to account A_3, we obtain the following values using (A.1), (A.3), and (A.4):

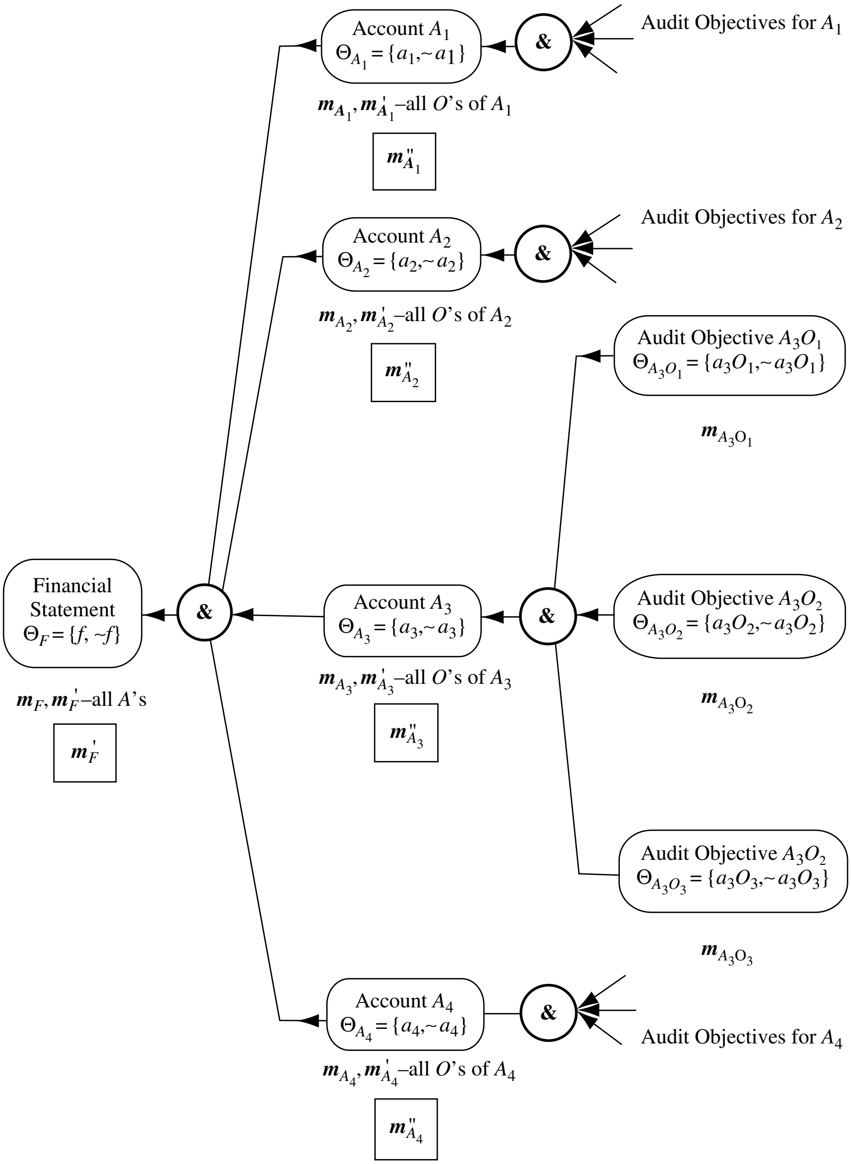

Account A_1
$\Theta_{A_1} = \{a_1, \sim a_1\}$
&
Audit Objectives for A_1
$\boldsymbol{m}_{A_1}, \boldsymbol{m}'_{A_1}$–all O's of A_1
$\boldsymbol{m}''_{A_1}$
Account A_2
$\Theta_{A_2} = \{a_2, \sim a_2\}$
&
Audit Objectives for A_2
$\boldsymbol{m}_{A_2}, \boldsymbol{m}'_{A_2}$–all O's of A_2
$\boldsymbol{m}''_{A_2}$
Audit Objective A_3O_1
$\Theta_{A_3O_1} = \{a_3O_1, \sim a_3O_1\}$
$\boldsymbol{m}_{A_3\mathrm{O}_1}$
Financial Statement
$\Theta_F = \{f, \sim f\}$
&
Account A_3
$\Theta_{A_3} = \{a_3, \sim a_3\}$
&
Audit Objective A_3O_2
$\Theta_{A_3O_2} = \{a_3O_2, \sim a_3O_2\}$
$\boldsymbol{m}_F, \boldsymbol{m}'_F$–all A's
$\boldsymbol{m}'_F$
$\boldsymbol{m}_{A_3}, \boldsymbol{m}'_{A_3}$–all O's of A_3
$\boldsymbol{m}''_{A_3}$
$\boldsymbol{m}_{A_3\mathrm{O}_2}$
Audit Objective A_3O_2
$\Theta_{A_3O_3} = \{a_3O_3, \sim a_3O_3\}$
$\boldsymbol{m}_{A_3\mathrm{O}_3}$
Account A_4
$\Theta_{A_4} = \{a_4, \sim a_4\}$
&
Audit Objectives for A_4
$\boldsymbol{m}_{A_4}, \boldsymbol{m}'_{A_4}$–all O's of A_4
$\boldsymbol{m}''_{A_4}$

Fig. 2. Propagation of $\boldsymbol{m}$-values from audit objectives and accounts to the financial statement

$$\begin{aligned}\boldsymbol{m}'_{A_3-\text{all O's of } A_3}(a_3) = \prod_{i=1}^{3} \boldsymbol{m}_{A_3O_i}(a_3O_i) &= \prod_{i=1}^{3}\left(1 - AR_{A_3O_i}\right),\\ \boldsymbol{m}'_{A_3-\text{all O's of } A_3}(\sim a_3) &= 0,\\ \boldsymbol{m}'_{A_3-\text{all O's of } A_3}(\{a_3o_i, \sim a_3O_i\}) = 1 - &\prod_{i=1}^{3}\left(1 - AR_{A_3O_i}\right), \qquad \text{(B.1)}\end{aligned}$$

where AR_{AO}'s are defined as in (16).

The next step is to combine the above $\boldsymbol{m}$-values with $\boldsymbol{m}_{A_3}$ at the account, as defined in general terms in (11) through (13), before propagating it to F. Since there is no conflict in the two sets of $\boldsymbol{m}$-values at the account level and there also is no support for $\sim a_3$, the combined values when using Dempster's rule becomes:

$$\boldsymbol{m}''_{A_3}(a_3) = 1 - \boldsymbol{m}''_{A_3}(\{a_3, \sim a_3\}) = 1 - IR_{A_3}APR_{A_3}\left[1 - \prod_{i=1}^{3}\left(1 - AR_{A_3O_1}\right)\right],$$

$$\boldsymbol{m}''_{A_3}(\sim a_3) = 0,$$

$$\begin{aligned}\boldsymbol{m}''_{A_3},(\{a_3, \sim a_3\}) &= \boldsymbol{m}_{A_3}(\{a_3, \sim a_3\})\,\boldsymbol{m}'_{A_3-\text{all O's of } A_3}(\{a_3, \sim a_3\})\\ &= IR_{A_3}APR_{A_3}\left[1 - \prod_{i=1}^{3}(1 - AR_{A_3O_1})\right]. \qquad \text{(B.2)}\end{aligned}$$

In general, one can write the above result for any account 'A' in an "and" tree similar to that in Fig. 1 as:

$$\boldsymbol{m}''_{A}(a) = 1 - \boldsymbol{m}''_{A}(\{a, \sim a\}) = 1 - IR_{A}APR_{A}\left[1 - \prod_{\text{O}}\left(1 - AR_{AO}\right)\right],$$

$$\boldsymbol{m}''_{A}(\sim a) = 0,$$

$$\boldsymbol{m}''_{A}(\{a, \sim a\}) = IR_{A}APR_{A}\left[1 - \prod_{\text{O}}\left(1 - AR_{AO}\right)\right]. \qquad \text{(B.3)}$$

Let us define a new term, AR_A, for simplicity as:

$$AR_A = IR_{A}APR_{A}\left[1 - \prod_{\text{O}}\left(1 - AR_{AO}\right)\right]. \qquad \text{(B.4)}$$

To combine all the items of evidence at F, we propagate the $\boldsymbol{m}$-values in (B.3) to F (see Fig. 2). Using (A.1), (A.3), and (A.4), we can write these values, in general form, as:

$$\begin{gathered}
\boldsymbol{m}'_{F-\text{all } A'\text{s}}(f) = \prod_A \boldsymbol{m}''_A(a) = \prod_A (1 - AR_A), \\
\boldsymbol{m}'_{F-\text{all } A'\text{s}}(\sim f) = 0, \\
\boldsymbol{m}'_{F-\text{all } A'\text{s}}(\{f, \sim f\}) = 1 - \prod_A (1 - AR'_A).
\end{gathered} \tag{B.5}$$

Now, to complete the process, the above $\boldsymbol{m}$-values are combined with $\boldsymbol{m}_F$ at F as defined in (8) through (10). Using Dempster's rule to combine $\boldsymbol{m}'_{F-\text{all } A'\text{s}}$ in (B.5) and $\boldsymbol{m}_F$, we obtain the total $\boldsymbol{m}$-values as:

$$\begin{gathered}
\boldsymbol{m}^t_F(f) = 1 - \boldsymbol{m}^t_F(\{f, \sim f\}) = 1 - IR_F APR_F \left[1 - \prod_A (1 - AR_A)\right]. \\
\boldsymbol{m}^t_F(\sim f) = 0, \\
\boldsymbol{m}^t_F(\{f, \sim f\}) = \boldsymbol{m}_F(\{f, \sim f\})\, \boldsymbol{m}'_{F-\text{all } A'\text{s}}(\{f, \sim f\}) \\
= IR_F APR_F \left[1 - \prod_A (1 - AR_A)\right].
\end{gathered} \tag{B.6}$$

This yields the following values for beliefs and plausibilities that are of interest (using [4], [5], and [B.6]):

$$\begin{gathered}
\boldsymbol{Bel}^t_F(f) = 1 - IR_F APR_F \left[1 - \prod_A (1 - AR_A)\right]. \\
\boldsymbol{Bel}^t_F(\sim f) = 0, \\
\boldsymbol{PL}^t_F(\sim f) = AR^t_F = IR_F APR_F \left[1 - \prod_A (1 - AR_A)\right].
\end{gathered} \tag{B.7}$$

Since we termed plausibility of error as audit risk, the desired formula for the audit risk is given by AR'_F in (B.7), where:

$$AR_A = IR_A APR_A \left[1 - \prod_O (1 - AR_{AO})\right]$$

as defined in (B.4) and $AR_{AO} = IR_{AO} APR_{AO} CR_{AO} DR_{AO}$ as defined in (16).

Accounts Level

In this section, we derive a formula for the total plausibility of material misstatement, or the total audit risk, at the account level. Similar to the procedure for deriving the formula for the financial statement level, this derivation combines the eight items of evidence: four at the audit objective level, two at the account level, and two at the financial statement level. This will be achieved by propagating $\boldsymbol{m}$-values from F, from the audit objectives, and from other accounts to the desired account. We will use the relationships developed in appendix A.

Let us consider, for the purpose of illustration, account A_3 in Fig. 3. The $\boldsymbol{m}$-values received by A_3 from the audit objectives were determined in the previous section ([B.1]). The $\boldsymbol{m}$-values received from the financial statement node is given below (use [8]-[10], [A.7], [A.9], and [A.10]):

$$\begin{gathered}
\boldsymbol{m}'_{A_3-\mathrm{F}\ \&\ \text{all other } A'\text{s}}\left(a_3\right) = \boldsymbol{m}_{\mathrm{F}}\left(f\right) = 1 - IR_F APR_F, \\
\boldsymbol{m}'_{A_3-\mathrm{F}\ \&\ \text{all other } A'\text{s}}\left(\sim a_3\right) = 0, \\
\boldsymbol{m}'_{A_3-\mathrm{F}\ \&\ \text{all other } A'\text{s}}\left(\{a_3, \sim a_3\}\right) = 1 - \boldsymbol{m}_{\mathrm{F}}\left(f\right) = IR_F APR_F, \qquad \text{(B.8)}
\end{gathered}$$

Similar $\boldsymbol{m}$-values will be received by the other accounts. At A_3, we now have three $\boldsymbol{m}$-values: $\boldsymbol{m}_A$, defined at A_3([11]-[13]), $\boldsymbol{m}'_{A_3-\mathrm{F}\ \&\ \text{all other } A'\text{s}}$ as given in (B.8), and $\boldsymbol{m}'_{A_3-\text{all } O'\text{s of } A_3}$ in (B.1). We combine these $\boldsymbol{m}$-values using Dempster's rule, which yields the following total $\boldsymbol{m}$-values:

$$\begin{gathered}
\boldsymbol{m}^t_{A_3}\left(\mathrm{a}_3\right) = 1 - IR_F APR_F IR_{A_3} APR_{A_3}\left[1 - \prod_{i=1}^{3}\left(1 - AR_{A_3 O_1}\right)\right], \\
\boldsymbol{m}^t_{A_3}\left(\sim a_3\right) = 0, \\
\boldsymbol{m}^t_{A_3}, \left(\{a_3 \sim a_3\}\right) = IR_F APR_F IR_{A_3} APR_{A_3}\left[1 - \prod_{i=1}^{3}\left(1 - AR_{A_3 O_1}\right)\right]. \\
\text{(B.9)}
\end{gathered}$$

In general, one can write (B.9) as:

$$\begin{gathered}
\boldsymbol{m}^t_A\left(a\right) = 1 - IR_F APR_F IR_A, APR_A\left[1 - \prod_O\left(1 - AR_{AO}\right)\right], \\
\boldsymbol{m}^t_A\left(\sim a\right) = 0, \\
\boldsymbol{m}^t_{\mathrm{A}}\left(\{a, \sim a\}\right) = IR_F APR_{\mathrm{F}} IR_{\mathrm{A}} APR_A\left[1 - \prod_O\left(1 - AR_{AO}\right)\right]. \qquad \text{(B.10)}
\end{gathered}$$

These values yield the following beliefs and plausibility values (using [4], [5], and [B.10]):

$$\begin{gathered}
\boldsymbol{Bel}^t_A\left(\mathrm{a}\right) = 1 - IR_F APR_F IR_A APR_A\left[1 - \prod_O\left(1 - AR_{AO}\right)\right], \\
\boldsymbol{Bel}^t_A\left(\sim a\right) = 0, \\
\boldsymbol{Pl}^t_A\left(\sim \mathrm{a}\right) = AR^t_A = IR_F APR_F IR_A APR_A\left[1 - \prod_O\left(1 - AR_{AO}\right)\right]. \qquad \text{(B.11)}
\end{gathered}$$

Since the total plausibility in $\sim a$ represents the total audit risk, AR^t_A in (B.11) is the desired formula at the account level.

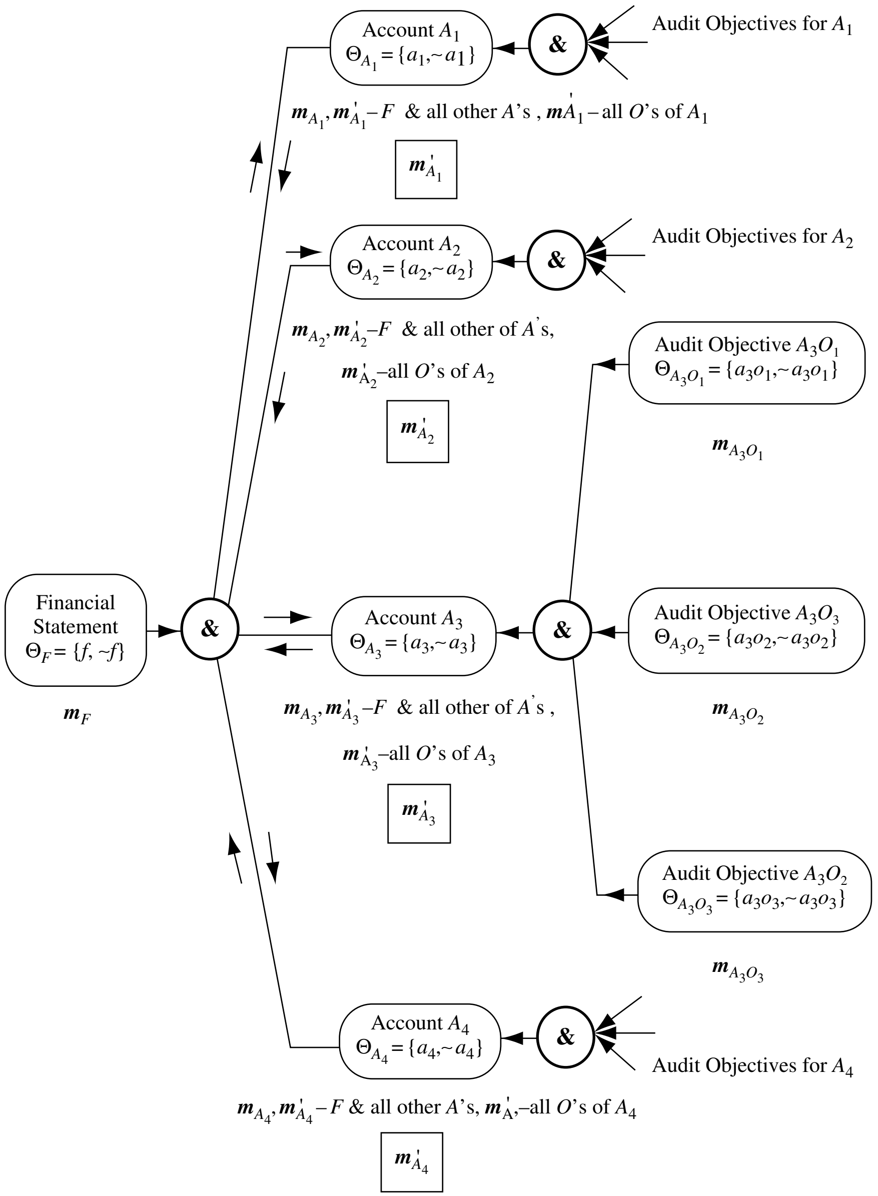

Account A_1
$\Theta_{A_1} = \{a_1, \sim a_1\}$
&
Audit Objectives for A_1
$\boldsymbol{m}_{A_1}, \boldsymbol{m}'_{A_1} - F$ & all other A's , $\boldsymbol{m}A'_1$ – all O's of A_1
$\boldsymbol{m}'_{A_1}$
Account A_2
$\Theta_{A_2} = \{a_2, \sim a_2\}$
&
Audit Objectives for A_2
$\boldsymbol{m}_{A_2}, \boldsymbol{m}'_{A_2} – F$ & all other of A's,
$\boldsymbol{m}'_{A_2}$–all O's of A_2
$\boldsymbol{m}'_{A_2}$
Audit Objective A_3O_1
$\Theta_{A_3O_1} = \{a_3o_1, \sim a_3o_1\}$
$\boldsymbol{m}_{A_3O_1}$
Financial Statement
$\Theta_F = \{f, \sim f\}$
$\boldsymbol{m}_F$
&
Account A_3
$\Theta_{A_3} = \{a_3, \sim a_3\}$
&
Audit Objective A_3O_3
$\Theta_{A_3O_2} = \{a_3o_2, \sim a_3o_2\}$
$\boldsymbol{m}_{A_3}, \boldsymbol{m}'_{A_3} – F$ & all other of A's ,
$\boldsymbol{m}'_{A_3}$–all O's of A_3
$\boldsymbol{m}_{A_3O_2}$
$\boldsymbol{m}'_{A_3}$
Audit Objective A_3O_2
$\Theta_{A_3O_3} = \{a_3o_3, \sim a_3o_3\}$
$\boldsymbol{m}_{A_3O_3}$
Account A_4
$\Theta_{A_4} = \{a_4, \sim a_4\}$
&
Audit Objectives for A_4
$\boldsymbol{m}_{A_4}, \boldsymbol{m}'_{A_4} - F$ & all other A's, $\boldsymbol{m}'_{\mathrm{A}}$,–all O's of A_4
$\boldsymbol{m}'_{A_4}$

Fig. 3. Propagation of m-values from the financial statement and audit objectives to the accounts

Audit Objective Level

In this section, we want to combine, again, all the items of evidence in a network in Fig. 1, but at the audit objective level. First, we propagate $\boldsymbol{m}$-values from the financial statement node to the accounts. As an illustration, we consider first propagation of $\boldsymbol{m}_F$ to A_3 (see Fig. 4). This yields $\boldsymbol{m}'_{A_3-F \text{ \& all other } A'\text{s}}$ given in (B.8). As a second step, we combine $\boldsymbol{m}'_{A_3-F \text{ \& all other } A'\text{s}}$ with $\boldsymbol{m}_3$ at A_3 ([11]–[13]). The result of combination yields $\boldsymbol{m}''_{A_3}$:

$$\begin{aligned}\boldsymbol{m}''_{A_1}(a) &= 1 - IR_F APR_F IR_{A_1} APR_{A_1},\\ \boldsymbol{m}''_{A_1}(\sim a) &= 0,\\ \boldsymbol{m}''_{A_1}(\{a, \sim a\}) &= IR_F APR_F IR_{A_1} APR_{A_1}.\end{aligned} \tag{B.12}$$

These $\boldsymbol{m}$-values are propagated to the various audit objectives. We use (A.7), (A.9), and (A.10) for this purpose. For audit objective A_3O_2, the result of propagation yields $\boldsymbol{m}'_{A_1O_2-A_1 \text{ \& all other } O'\text{s of } A_1}$:

$$\begin{aligned}\boldsymbol{m}'_{A_1O_2-A_1 \text{ \& all other } O'\text{s of } A_1}(a_3\mathrm{O}_2) = \boldsymbol{m}''_{A_1}(a_3) &= 1 - IR_F APR_F IR_{A_1} APR_{A_1},\\ \boldsymbol{m}'_{A_1O_2-A_1 \text{ \& all other } O'\text{s of } A_1}(\sim \mathrm{a}_3\mathrm{O}_2) &= 0,\\ \boldsymbol{m}'_{\mathrm{A}_1\mathrm{O}_2-\mathrm{A}_1 \text{ \& all other O's of } \mathrm{A}_1}(\{a_3O_2 \sim a_3O_2\}) &= IR_F APR_F IR_{A_1} APR_{A_1}.\end{aligned} \tag{B.13}$$

As shown in Fig. 4, we have two $\boldsymbol{m}$-values at each audit objective. For objective O_2 of account A_3, we have $\boldsymbol{m}'_{A_1O_2-A_1 \text{ \& all other } O'\text{s of} A_1}$ in (B.13) and $\boldsymbol{m}_{A_1O_2}$ in (14) through (4). We combine these values using Dempster's rule, which yields the total $\boldsymbol{m}$-values at the audit objective level. This represents an aggregation of all the items of evidence bearing on various levels.

$$\begin{aligned}\boldsymbol{m}^t_{A_1O_2}(a_3O_2) &= 1 - IR_F APR_F IR_{A_1} APR_{A_1} AR_{A_1O_2},\\ \boldsymbol{m}^t_{A_1O_2}(\sim a_3O_2) &= 0,\\ \boldsymbol{m}^t_{A_1O_2}(\{a_3O_2, \sim a_3O_2\}) &= IR_F APR_F IR_{A_1} APR_{A_1} AR_{A_1O_2}.\end{aligned} \tag{B.14}$$

From (4) and (5), one obtains the following values for beliefs and plausibilities for any audit objective AO:

$$\boldsymbol{Bel}^t_{AO}(aO) = 1 - IR_F APR_F IR_A APR_A AR_{AO}, \tag{B.15}$$

$$\boldsymbol{Bel}^t_{AO}(\sim aO) = 0, \tag{B.16}$$

$$\boldsymbol{PL}^t_{AO}(\sim aO) = AR^t_{AO} = IR_F APR_F IR_A APR_A AR_{AO}. \tag{B.17}$$

Equation (B.17) is the desired formula for this case.

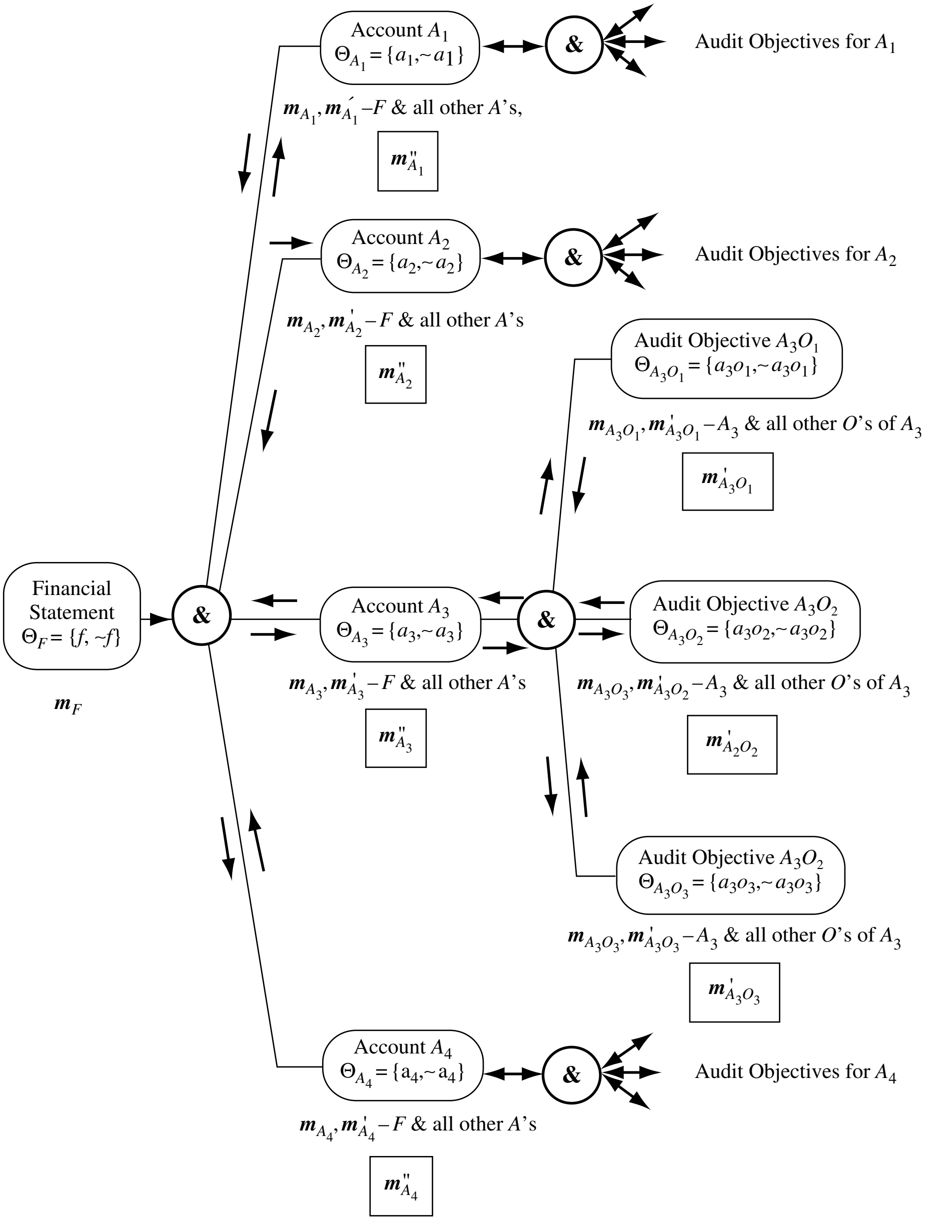

Account A_1
$\Theta_{A_1} = \{a_1, \sim a_1\}$
&
Audit Objectives for A_1
m_{A_1}, m'_{A_1} – F & all other A's,
m''_{A_1}
Account A_2
$\Theta_{A_2} = \{a_2, \sim a_2\}$
&
Audit Objectives for A_2
m_{A_2}, m'_{A_2} – F & all other A's
m''_{A_2}
Audit Objective A_3O_1
$\Theta_{A_3O_1} = \{a_3o_1, \sim a_3o_1\}$
$m_{A_3O_1}, m'_{A_3O_1}$ – A_3 & all other O's of A_3
$m'_{A_3O_1}$
Financial Statement
$\Theta_F = \{f, \sim f\}$
m_F
&
Account A_3
$\Theta_{A_3} = \{a_3, \sim a_3\}$
&
Audit Objective A_3O_2
$\Theta_{A_3O_2} = \{a_3o_2, \sim a_3o_2\}$
m_{A_3}, m'_{A_3} – F & all other A's
m''_{A_3}
$m_{A_3O_3}, m'_{A_3O_2}$ – A_3 & all other O's of A_3
$m'_{A_2O_2}$
Audit Objective A_3O_2
$\Theta_{A_3O_3} = \{a_3o_3, \sim a_3o_3\}$
$m_{A_3O_3}, m'_{A_3O_3}$ – A_3 & all other O's of A_3
$m'_{A_3O_3}$
Account A_4
$\Theta_{A_4} = \{a_4, \sim a_4\}$
&
Audit Objectives for A_4
m_{A_4}, m'_{A_4} – F & all other A's
m''_{A_4}

Fig. 4. Propagation of $\boldsymbol{m}$-values from the financial statement and accounts to the audit objectives

Acknowledgments

This research has been supported in part by grants from the Peat Marwick Foundation, the General Research Fund of the University of Kansas, and the National Science Foundation (grant IRI8902444). The authors would like to thank Andrew D. Bailey, Jr., Peter Gillett, Lynford E. Graham, William R. Kinney, Jr., Prakash P. Shenoy, W. R. Scott, Ira Solomon, and John Willingham for valuable comments on an earlier version of the article.

References

American Institute of Certified Public Accountants. 1983. *Statement on Auditing Standards No. 47: Audit Risk and Materiality in Conducting an Audit.* New York: AICPA.

——. 1988a. *Statement on Auditing Standards, No. 55: Consideration of the Internal Control Structure in a Financial Statement Audit.* New York: AICPA.

——. 1988b. *Statement on Auditing Standards, No. 56: Analytical Procedures.* New York: AICPA.

Akresh, A. D., J. K. Loebbecke, and W. R. Scott. 1988. Audit approaches and techniques. In *Research Opportunities in Auditing: The Second Decade*, edited by A. R. Abdel-khalik and Ira Solomon. Sarasota, FL: AAA, 13–55.

Arens, A. A., and J. K. Loebbecke. 1988. *Auditing: An Integrated Approach.* Englewood Cliffs, NJ: Prentice-Hall.

Boritz, J. E. 1990. Appropriate and inappropriate approaches to combining evidence in an assertion-based auditing framework. Working paper, School of Accountancy, University of Waterloo, Canada.

——, and R. E. Jensen. 1985. An hierarchical, assertion-oriented approach to planning audit evidence-gathering procedures. Presented at the *Symposium on Audit Judgment and Evidence Evaluation*, University of Southern California.

——, and A. K. P. Wensley. 1990. Structuring the assessment of audit evidence: An expert system approach. *Auditing: A Journal of Practice & Theory* 9 (Supplement): 49–109.

Buchanan, B. G., and E. H. Shortliffe. 1984. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Reading, MA: Addison-Wesley.

Canadian Institute of Chartered Accountants. 1980. *Extent of Audit Testing.* Toronto, Canada: CICA.

Cushing, B. E., and J. K. Loebbecke. 1983. Analytical approaches to audit risk: A survey and analysis. *Auditing: A Journal of Practice & Theory* 3 (Fall): 23–41.

Graham, L. E. 1985a. *Audit risk—Part I. The CPA Journal* 55 (August): 12–21.

——. 1985b. Audit risk—Part II. *The CPA Journal* 55 (September): 34–41.

——. 1985c. Audit risk—Part III. *The CPA Journal* 55 (October): 36–43.

——. 1985d. Audit risk—Part IV. *The CPA Journal* 55 (November): 38–47.

——. 1985e. Audit risk—Part V. *The CPA Journal* 55 (December): 26–35.

Kinney, Jr., W. R. 1984. Discussant's response to an analysis of the audit framework focusing on inherent risk and the role of statistical sampling in compliance testing. *Proceedings of the 1984 Touche Ross/University of Kansas Symposium on Auditing Problems.* Lawrence, KS: School of Business, University of Kansas,126–32.

——. 1989. Achieved audit risk and the audit outcome space. *Auditing: A Journal of Practice & Theory* 8 (Supplement): 67–97.

Leslie, D. A. 1984. An analysis of the audit framework focusing on inherent risk and the role of statistical sampling in compliance testing. *Proceedings of the 1984 Touche Ross/University of Kansas Symposium on Auditing Problems.* Lawrence, KS: School of Business, University of Kansas, 89–125.

——, S. J. Aldersley, D. J. Cockburn, and C. J. Reiter. 1986. An assertion-based approach to auditing (discussant's remarks). *Proceedings of the 1986 Touche Ross/University of Kansas Symposium on Auditing Problems.* Lawrence, KS: School of Business, University of Kansas, 31–67.

Shafer, G. 1976. *A Mathematical Theory of Evidence.* Princeton University Press.

——. 1982. Belief functions and parametric models. *Journal of the Royal Statistical Society* **44** (Series B): 322–52.

——. 1986. Probability judgment in artificial intelligence and expert systems. *Statistical Science* **2** (1): 3–16.

——, and P. P. Shenoy. 1988. Local computation in hypertrees. Working Paper No. 201, School of Business, University of Kansas, Lawrence, KS.

——, and ——. 1990. Probability propagation. *Annals of Mathematics and Artificial Intelligence* **2** (1–4): 327–52.

——, ——, and R. P. Srivastava. 1988. Auditor's assistant: A knowledge engineering tool for audit decisions. *Proceedings of the 1988 Touche Ross/University of Kansas Symposium on Auditing Problems.* Lawrence, KS: School of Business, University of Kansas, 61–84.

——, and R. P. Srivastava. 1990. The Bayesian and belief-function formalisms: A general perspective for auditing. *Auditing: A Journal of Practice & Theory* 9 (Supplement): 110–48.

Shenoy, P. P. 1989. A valuation-based language for expert systems. *International Journal of Approximate Reasoning* 3 (5): 383–411.

——, and G. Shafer. 1988. An axiomatic framework for Bayesian and belief-function propagation. *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence.* St. Paul, MN: AAAI Press, 307–14.

——, and ——. 1990. Axioms for probability and belief-function propagation. In *Uncertainty in Artificial Intelligence* 4, edited by R. D. Shachter et al. Amsterdam: North-Holland, 169–98.

Zarley, D., Yen-Teh Hsia, and G. Shafer. 1988. Evidential reasoning using DELIEF. *Proceedings of the Seventh National Conference on Artificial Intelligence*, Vol. 1. Cambridge, MA: AAAI Press, 205–09.

Decision Making Under Dempster–Shafer Uncertainties

Ronald R. Yager

Abstract. We are concerned here with the problem of selecting an optimal alternative in situations in which there exists some uncertainty in our knowledge of the state of the world. We show how the Dempster–Shafer belief structure provides a unifying framework for representing various types of uncertainties. We also show how the OWA aggregation operators provide a unifying framework for decision making under ignorance. In particular we see how these operators provide a formulation of a type epistemic probabilities associated with our degree of optimism.

Key words: Decision making, Dempster–shafer theory, Uncertainty

1 Introduction

The problem of decision making under uncertainty is a very important issue. In this problem we are concerned with the selection of an appropriate decision alternative, in the face of uncertainty with respect to the environment. The uncertainty manifests itself in that a different payoff is obtained for different states of nature. In this paper we provide a general formulation of this type of decision making. The Dempster–Shafer evidential structure[1, 2] plays a crucial role in providing a unifying framework for representing the uncertainty. The Ordered Weighted Averaging (OWA) operators[3] play a central role in providing a unifying framework for aggregation. The introduction of these OWA operators provides a more general formulation than that used by Yager[4, 5] in his previous work on decision making in the face of evidential knowledge.

We first discuss the classic problem of decision making under risk and ignorance.[6] In the environment of decision making under ignorance, we discuss the role of the decision maker's attitude in the selection of the procedure used to find the overall value associated with a particular alternative. In this environment we have a collection of possible outcomes, payoffs, but no probability associated with them. The value of this collection is determined by how optimistic or pessimistic the decision maker feels. We then show how the OWA operators provide a general framework for determining the value of a collection of outcomes.

We next show how the Dempster–Shafer belief structure provides a suitable framework for representing, in a unified manner, the information a decision maker may have in regards to the state of nature.

Finally we provide a methodology for selecting optimal alternatives in situations in which our knowledge about the uncertainty is contained in a Dempster–Shafer belief structure. The problem of making a decision in this environment is very important and has been considered by some authors. [4, 5, 7, 8, 9, 10]

2 Decision Making Under Uncertainty

Consider the following matrix provided to a decision maker:

$$\begin{array}{c|ccccc|}
 & S_1 & \dots & S_j & \dots & S_n \\
\hline
A_1 & C_{11} & \dots & C_{1j} & \dots & C_{1n} \\
\vdots & \vdots & & \vdots & & \vdots \\
A_i & C_{i1} & \dots & C_{ij} & \dots & C_{in} \\
\vdots & \vdots & & \vdots & & \vdots \\
A_q & C_{q1} & \dots & C_{qj} & \dots & C_{qn}
\end{array}$$

In the above each A_i corresponds to a possible action (alternative) available to the decision maker. Each S_j corresponds to a possible value of the variable called the *state of nature.* C_{ij} corresponds to the payoff to be received by the decision maker if he selects action A_i and the state of nature is S_j. The problem faced by the decision maker is to select the action which gives him the optimum payoff.

Since the payoff to the decision maker depends upon the state of nature his procedure for selecting the best alternative depends upon the type of knowledge he has about the state of nature. In the literature dealing with this problem,[6] three different decision making environments have been identified: decision making under certainty, decision making under risk and decision making under ignorance.

In decision making under certainty the decision maker knows exactly what is the state of nature. In this case the course of action is straightforward, he selects the alternative that has the maximum payoff for this state of nature.

In decision making under risk it is assumed that we have a probability distribution over the state of nature. In this case we know for each S_j, P_j the probability that S_j is the state of nature. The standard procedure in this case is to use expected values:

1. For each alternative A_i we calculate $C_i = \sum_j C_{ij} \times P_j$, its expected payoff.
2. Select as the optimal alternative, A^*, the one which has the largest expected payoff,

$$C^* = \text{Max}_i C_i.$$

It should be noted that decision making under certainty can be seen as a special case of decision making under risk. In particular, if we know that S_a is the state of nature, then we can consider $P_a = 1$.

In the third environment, decision making under ignorance, we assume no knowledge about the state of nature other than that it is an element in some set $S = \{S_1, S_2, \ldots, S_n\}$.

The methodology used in the selection of the optimal alternative in this environment requires the assumption of a particular *decision attitude* by the decision maker. Among the decision attitudes discussed in the literature are the following:[6]

1) *Pessimistic attitude*—Using this strategy the decision maker selects for each alternative the worst possible outcome and then selects the alternative that has the best worst. This strategy is sometimes called the *maximum* strategy.
2) *Optimistic attitude*—Under this strategy, the decision maker selects for each alternative the best possible outcome and then selects the alternative that has the best best. This strategy is called the *maximax* strategy.
3) *Hurwicz Approach*—In this approach the decision maker selects a value $\alpha \in [0, 1]$. Then for each alternative he takes a weighted average of the optimistic and pessimistic value

$$H = \alpha * \text{Opt.} + (1 - \alpha) * \text{Pess.}$$

He then chooses the alternative which has the highest H value.
4) *Normative Approach*—In this approach for each alternative the decision maker sums the payoffs across all possible outcomes and then selects the alternative with the highest total.

In the case of the decision making under ignorance the decision process is as follows:

1) For each A_i calculate $V_i = F(C_{i1}, C_{i2}, \ldots, C_{in})$. We note that F is some aggregation function whose form depends upon the decision makers assumed attitude.
2) Select the alternative A^* such that $V^* = \text{Max}_i[V_i]$.

The following table provides the F function for the four strategies discussed.

STRATEGY	*AGGREGATION FUNCTION*
Pessimistic	$F(C_{i1}, C_{i2}, \ldots C_{in}) = \text{Min}_j[C_{ij}]$
Optimistic	$F(C_{i1}, C_{i2}, \ldots C_{in}) = \text{Max}_j[C_{ij}]$
Hurwicz	$F(C_{i1}, C_{i2}, \ldots C_{in}) = \alpha * \text{Max}_j[C_{ij}] + (1 - \alpha)\text{Min}_j[C_{ij}]$
Normative	$F(C_{i1}, \ldots C_{in}) = \sum_j C_{ij}$

3 A General Approach to Alternative Selection Under Ignorance

In this section we shall suggest a general formulation to the optimal alternative selection problem under ignorance. This approach will be based upon the ordered weighted averaging (OWA) operators introduced by Yager.[3] We shall see that this general approach allows the four previously discussed methods as special cases.

In suggesting a general approach to alternative selection one should be concerned that it satisfies certain properties which one can consider as rational. A first criteria is that of Pareto optimality. This condition requires that given two alternatives A and B, where A has at least as high a payoff as B for each state of nature, then B should *not* be more preferred than A. A second condition is that it should treat the states of nature uniformily. Another desirable, though not necessary, requirement, is that the aggregation across the states of nature be an averaging like operator in the sense that if for a given alternative all the states of nature have the same payoff, a, then the overall value of that alternative should be a. Yager[3] introduced a new type of aggregation operator called OWA operators. He also suggested some extensions of these operators.[11] O'Hagan[12] has investigated their use in expert systems.

Definition 1. An ordered weighted averaging operator (OWA) of dimension n is a function

$$F : R^n \rightarrow R$$

that has associated with it a weighting vector W,

$$W = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_n \end{bmatrix}$$

such that

1) $w_i \in [0,1]$
2) $\Sigma_i w_i = 1$

and for any set of values $a_1, \ldots, a_n$

$$F(a_1, \ldots, a_n) = \sum_i (w_i * b_i)$$

where b_i is the ith largest element in the collection $a_1, a_2, \ldots, a_n$.

Example 1. If

$$W = \begin{bmatrix} 0.3 \\ 0.4 \\ 0.2 \\ 0.1 \end{bmatrix}$$

then $F(10, 0, 20, 30) = (0.3)*30 + (0.4)*20 + (0.2)*10 + 0.1*0 = 19$.

It should be noted that the weights in the OWA operator are associated with a position in the ordered arguments rather than a particular argument.

It is our suggestion that the OWA operators provide a family of operators, parameterized by W, which can be used to help in the selection of optimal alternatives in the face of ignorance. In particular we can use these operators to provide the aggregated value for each alternative. We can calculate $V_i = F(C_{i1}, C_{i2}, \ldots, C_{in})$ where F is an OWA aggregation operator. We then select the alternative that has the highest V value.

First we note that for any W the OWA aggregation operation satisfies the condition of pareto optimality. In particular if

$$a_j \geqq d_j \text{ for all } j = 1, \ldots, n$$

then

$$F\left(a_1, \ldots, a_n\right) \geqq F\left(d_1, \ldots, d_n\right).$$

Next we shall show that the four methods previously discussed are special cases of OWA operators.

1) *Pessimistic Attitude*: If we select W_* where

$$W_* = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

then $F_*(a_1, \ldots, a_n) = \text{Min}_j[a_j]$, which is the aggregation rule used in the pessimistic strategy.

2) *Optimistic Attitude*: If we select W^* where

$$W^* = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

then $F^*(a_1, \ldots, a_n) = \text{Max}_j[a_j]$ which is what is used in the optimistic strategy.

3) *Hurwicz Strategy*: If we select

$$W_H = \begin{bmatrix} \alpha \\ 0 \\ \vdots \\ 0 \\ 1-\alpha \end{bmatrix}$$

then $F_H(a_1, \ldots, a_n) = \alpha * \text{Max}[a_j] + (1-\alpha) * \text{Min}[a_j]$. This is exactly the formulation used in the Hurwicz strategy.

4) *Normative approach*: If we select

$$W_N = \begin{bmatrix} 1/n \\ 1/n \\ \vdots \\ 1/n \end{bmatrix}$$

then we get $F_N(a_1, \ldots, a_n) = 1/n\Sigma_i a_i$. This function is essentially the normative strategy.

We should note that the pessimistic and optimistic strategies provide limiting classes of OWA operators. It can be easily shown[3] that for any OWA operator F and any set of arguments $(a_1, \ldots, a_n)$ that $F_*(a_1, \ldots, a_n) \leqq F(a_1, \ldots, a_n) \leqq F^*(a_1, \ldots, a_n)$.

Yager[11] suggested a semantics that can be associated with the OWA aggregation procedure in this framework of decision making under ignorance. This semantics will provide a unifying interpretation of this operation. Yager suggests that we can view the OWA weights as a kind of probability distribution. In particular we can view w_i as the *probability that the ith best thing* will happen. We recall that the weights have the properties of a probability distribution in that each w_i lies in the unit interval and the sum of the w_i's is one. From this interpretation we see that the aggregation associated with each particular alternative can be seen as the expected value under this probability distribution. If $C_{i1}, C_{i2}, \ldots, C_{in}$ are payoffs corresponding to each of the states of nature under the selection of alternative A_i then $b_1, \ldots b_j$ are the ordered set of these payoffs. Then if $w_1, \ldots, w_n$ are the OWA weights interpreted as probabilities of the jth best thing happening under any selection of alternative we see that

$$V = \sum_j w_j * b_j$$

is the expected payoff in this case. Thus the OWA aggregation provides a kind of expected value similar to that used in decision making under risk. The difference between the two situations is that in later, decision making under risk, we have assigned a probability p_i to each particular state of nature S_i. In decision making under ignorance the probabilities, the weights, are assigned not to a particular state of nature, but to the preference ordered position of the payoff. Using this interpretation we can see that the pessimistic strategy is effectively a situation in which a probability of one is assigned to the worst thing happening given any selection of alternative. In the optimistic approach we are assuming a probability of one is assigned to the possibility of the best thing happening. In the normative case we are assuming equal probability for each of the preference positions. The Hurwicz strategy effectively assigns a probability α that the best thing will happen and $1 - \alpha$ probability that the worst thing will happen.

Yager[3] introduced a number of measures associated with the weights of an OWA operator, we briefly describe these. Assume W is a set of weights then the measure of *Optimism* associated with these weights is defined as

$$\mathrm{Opt}\,(W) = \sum_{j=1}^{n} w_j * h_n\,(j)$$

where $h_n(j) = (n - j)/(n - 1)$.

We note that $\mathrm{Opt}(W^*) = 1, \mathrm{Opt}(W_*) = 0, \mathrm{Opt}(W_n) = 0.5$ and Opt $(W_H) = \alpha$.

A second measure associated with these weights is a measure of entropy or dispersion

$$\mathrm{DISP}\,(W) = -\sum_{j} w_j \ln\,(w_j)\,.$$

We should note that the larger the Disp(W) the more the payoffs play a role in the determination of F. O'Hagan[12] has studied these measures in considerable detail.

A question that naturally arises, is, how does a decision maker obtain the weights he is going to use in solving a particular problem? At the fundamental level, the answer is that he subjectively decides, just as he does in deciding to be pessimistic or optimistic or normative. The most straightforward way of obtaining the weights is for the decision maker to directly select the values of the weights. In doing this, if he chooses to allocate, the allotted total of one, to weights near the top of the vector, he can be seen as being optimistic. If he allocates the weights to elements near the bottom he is being pessimistic.

An alternative method of selecting the weights has been suggested by O'Hagan.[8] With this approach the decision maker subjectively decides upon his coefficient of optimism β. He then inputs this value into a mathematical programming problem which is used to obtain the weights that have an appropriate degree of optimism while maximizing the dispersion.

The mathematical programming problem is
Maximize:

$$\sum_{j} w_j \ln\,(w_j) \quad \text{(entropy)}$$

Subject to:

$$\sum_j (h_n(j) * w_j) = \beta$$

$$\sum_j w_j = 1$$

$$w_j \geqq 0 \quad \text{for all } j = 1, \ldots, n.$$

This approach is closely related to the maximum entropy method used in probability theory.

One benefit of this approach is that we can consistently provide for weights corresponding to a given β for various different cardinalities of OWA operators.

4 A General Framework for Representing Uncertainty

Earlier we suggested that there were two distinct situations with respect to the knowledge about the state of nature, risk and ignorance. Actually we also discussed certainty but we suggested that this was a special case of risk, one in which the probability of some outcome is one. It actually can also be seen as a special case of ignorance where the set S consists of only one element.

In this section we introduce a more general framework for the representation of uncertainty. This scheme is called the Dempster-Shafer theory of evidence.[1, 2] We shall show that the two cases, risk and ignorance, are special cases of this more general formulation. In addition to being able to capture these classic formulations of our knowledge about uncertain environments the Dempster-Shafer structure allows us to represent various other forms of information a decision maker may have about the state of nature. The Dempster-Shafer framework has proved to be an important and useful tool in the development of expert systems.[13, 14, 15, 16]

A belief structure m on the set Y consists of a collection of non-empty subsets B_i of Y and an associated set of weights $m(B_i)$ such that:

1) $m(B_i) > 0$
2) $\Sigma_i m(B_i) = 1$.

The subsets B_i are called the focal elements of the belief structure.

The original formulation for these belief structures was suggested by Dempster[1] in the framework of multi-valued mappings. The semantics associated with this original formulation will provide useful for our future discussion. Let X be a set on which there exists a probability distribution such that $p_i = \text{prob}(\{x_i\})$. Let Y be another set, called the frame of discernment. We further assume that there exists a relationship R, called the compatibility relation, on $X \times Y$. This relationship connects elements in X with those in Y. The semantics of this relationship is that, if $(x, y) \in R$ then the value y

is an acceptable value for Y if x is the value of X. Our concern here is the determination of the value on the frame space Y. Given any $x_i \in X$ we can associate with this element a subset $B_i \subset Y$ such that $B_i = \{y|(x_i, y) \in R\}$. B_i is a focal element and is the set of outcomes of Y which are compatible with x_i. In this environment we can see that the occurrence of the outcome x_i induces an outcome on Y which is in the subset B_i. Furthermore we note that p_i is the probability associated with the outcome B_i, it is $m(B_i)$. The following example illustrates the Dempster-Shafer environment.

Example 2. Assume we have a lottery in which we sell ten tickets. If Joe buys three tickets, Bob buys 5 tickets and Ed buys two tickets then it is obvious that the probability of each winning is given by Prob(Joe)= 0.3, Prob(Bob)=0.5 and Prob(Ed)=0.2. However our concern is not with who is the winner but with the color of the hat that the winner is wearing. The following information is known about the hats owned by the various people:

	Red	*Blue*	*Black*	*Yellow*	*Green*
Joe	1	0	1	1	1
Bob	0	1	1	1	0
Ed	1	0	0	0	0

In the above table a one indicates that a particular color hat is owned by an individual. However, we don't know how many hats an individual has of a given color, nor do we know how an individual chooses which hat he shall wear. In this situation the focal elements are:

$$\begin{aligned} B_{\text{Joe}} &= \{\text{red, black, yellow, green}\} \\ B_{\text{Bob}} &= \{\text{blue, black, yellow}\} \\ B_{\text{Ed}} &= \{\text{red}\} \,. \end{aligned}$$

In this case we can say there is a 0.3 probability that the set B_{Joe} will be used as the set from which the hat is chosen, 0.5, that the set B_{Bob} is used and 0.2 probability that the set B_{Ed} is used.

A very appealing feature of this belief structure is that it can be used to represent in a unified manner various types of uncertainty we previously discussed. In the following we shall let Y be the set of possible states of nature.

If the belief structure consists of n focal elements such that $B_i = \{y_i\}$, each focal element is a singleton, then we essentially have the decision making under risk environment where $m(B_i) = P_i = \text{Prob}\{y_i\}$.

If our belief structure has only one focal element B, where $m(B) = 1$, then we essentially have the decision making under ignorance environment.

In addition to these two basic formulations of our knowledge the Dempster–Shafer formulation allows us to capture other more sophisticated forms of knowledge.

If our knowledge of state of nature is such that we know that there is a probability p that the state of nature lies in the set A and $1-p$ that it lies in not A then we can represent this by a belief structure with two focal elements as follows:

$$\begin{aligned} B_1 &= A \quad m(B_1) = p \\ B_2 &= \bar{A} \quad m(B_2) = 1-p. \end{aligned}$$

A closely related belief structure is one in which

$$\begin{aligned} B_1 &= A \quad m(B_1) = p \\ B_2 &= Y \quad m(B_2) = 1-p. \end{aligned}$$

With this belief structure we are essentially saying that the probability of A is at least p.

The essential point of this section is that the use of the Dempster-Shafer belief structure provides a unifying method for representing our knowledge about the state of nature in decision making problems.

5 Decision Making with Belief Structures

The Dempster-Shafer belief structures have proven to be a very useful representation scheme for expert and other knowledge based systems. In many cases the knowledge provided by these types of expert systems is in the form of a belief structure. A problem that is of considerable interest is that of selecting an appropriate course of action, alternative, in situations in which our knowledge about the state of nature is in the form of a belief structure. In this section we shall bring all the pieces together to provide a unified approach to decision making under uncertainty. This work provides a generalization of the ideas discussed by Yager.[4, 5]

Assume we have a decision problem in which we have a collection of q alternatives, we denote the set of alternatives as $A = \{A_1, \ldots, A_q\}$. In addition we assume the payoff depends upon the value of a variable which we call the state of nature. We assume the value of this variable is some element in the set S, where $S = \{S_1, \ldots, S_n\}$. We further assume that C_{ij} is the payoff to the decision maker if he selects alternative A_i and the state of nature is S_j. In addition we assume our knowledge of the state of nature is captured in terms of a belief structure m on S. The focal elements of m are $B_1, \ldots, B_r$ and associated with each of these is a probability mass value $m(B_i)$. The problem of concern is to select the alternative which maximizes the payoff to the decision maker.

The procedure for the determination of the best alternative is an extension of the previously described methods. It combines the schemes used for both decision making under risk and ignorance. We shall call this decision making under uncertainty. In a manner similar to decision making under risk we obtain a *generalized expected value*, C_i, for each alternative A_i. However, in obtaining this expected value we use the weights associated with the focal elements as the probabilities. The second step is to select the alternative which has the largest generalized expected value.

The generalized expected value, C_i, for a given alternative, A_i, is obtained using the evidential knowledge. The knowledge contained in the belief structure tells us that $m(B_k)$ is the probability that B_k will be the set that will determine the state of nature. In particular

$$C_i = \sum_{k=1}^{r} V(A_i, B_k) * m(B_k)\,.$$

In the above $V(A_i, B_k)$, which we shall denote as V_{ik}, is the payoff we get when we select alternative A_i and the state of nature lies in B_k. Thus we see that C_i is essentially the expected value of the payoffs under A_i.

The determination of the value V_{ik} can be seen as equivalent to the problem of decision making under ignorance. In particular for a given A_i and the knowledge that the state of nature lies in B_k we end up with a collection of possible payoffs. We shall let M_{ik} denote the collection (bag) of payoffs that can occur under B_k. In this case each element S_j in B_k contibutes one element to M_{ik}, its payoff under S_j, hence $M_{ik} = \langle C_{ij} | S_j \in B_k \rangle$. In order to determine the value of V_{ik} from M_{ik} we use the procedure developed for decision making under ignorance. First we obtain from the decision making his measure of optimism α. This measure of optimism is then used to solve the mathematical program problem described earlier to obtain the weights for the OWA vectors. Actually we must solve this problem for each different cardinality of M_{ik}.

Using these weights we can find $V_{ik} = F(M_{ik})$ where F is an OWA operator whose weights are determined above for a degree of optimism α and cardinality of M_{ik}.

The following summarizes the operations, assuming we have obtained the payoff matrix, the belief function m about the state of nature and the decision makers degree of optimism, α.

1) Solve for each different cardinality of focal elements the mathematical programming problem with the degree of optimism α. This gives us a collection of weights to be used in OWA aggregation.
2) For each alternative i do the following:
    a) For each focal element, B_k, find M_{ik}, the collection of payoffs corresponding to that focal element.
    b) For each M_{ik} calculate, using the appropriate OWA operator, $V_{ik} = F(M_{ik})$.
    c) Calculate $C_i = \sum_k V_{ik} * m(B_k)$.
3) Select the alternative which has its highest C_i as the optimal alternative.

The following example illustrates the procedure.

Example 3. Assume the payoff matrix is as follows

	S_1	S_2	S_3	S_4	S_5
A_1	7	5	12	13	6
A_2	12	10	5	11	2
A_3	9	13	3	10	9
A_4	6	9	11	15	4

Assume that our knowledge of the state of nature consists of the following belief structure, m:

Focal element	*Weights*
$B_1 = \{S_1, S_3, S_4\}$	0.6
$B_2 = \{S_2, S_5\}$	0.3
$B_3 = \{S_1, S_2, S_3, S_4, S_5\}$	0.1

We shall assume that the decision maker has a degree of optimism of 0.75. Solving the appropriate mathematical programming[12] problems we obtain the weights associated with the OWA operators for various numbers of arguments under the optimism value of 0.75:

No. of arguments	w_1	w_2	w_3	w_4	w_5
2	0.75	0.25			
3	0.62	0.27	0.11		
4	0.52	0.27	0.14	0.07	
5	0.46	0.26	0.15	0.08	0.05

We recall M_{ik} is the collection of payoffs that are possible if we select alternative A_i and the focal element B_k occurs. We next calculate the bags M_{ik}.

$$M_{11} = \langle 7, 12, 6\rangle,\ M_{12} = \langle 5, 13\rangle,\ M_{13} = \langle 7, 5, 12, 3, 6\rangle, M_{21} = \langle 12, 5, 2\rangle,$$
$$M_{22} = \langle 10, 11\rangle,\ M_{23} = \langle 12, 10, 5, 11, 2\rangle,\ M_{31} = \langle 9, 3, 9\rangle, M_{32} = \langle 13, 10\rangle,$$
$$M_{33} = \langle 9, 13, 3, 10, 9\rangle,\ M_{41} = \langle 6, 11, 4\rangle,\ M_{42} = \langle 9, 15\rangle, M_{43} = \langle 6, 9, 11, 15, 4\rangle.$$

Next we calculate V_{ik}, using the ordered weighting average operation: We recall that

$$V_{ik} = F(M_{ik})\,.$$

$$V_{11} = (0.62) * 12 + (0.27) * 7 + (0.11) * 6 = 9.99$$
$$V_{12} = (0.75) * 13 + (0.25) * 5 = 11$$
$$V_{13} = (0.46) * (13) + (0.26) * 12 + (0.15) * 7 + (0.08) * 6 + (0.04) * 5 = 10.88$$
$$V_{21} = 0.62 * 12 + 0.27 * 5 + 0.11 * 2 = 9.01$$
$$V_{22} = 0.75 * 111 + 0.25 * 10 = 10.75$$
$$V_{23} = 0.46 * 12 + 0.26 * 11 + 0.15 * 10 + 0.08 * 5 + 0.05 * 2 = 10.38$$

$V_{31} = 0.62 * 9 + (0.27) * 9 + (0.11) * 3 = 8.34$
$V_{32} = 0.75 * 13 + 0.25 * 10 = 12.25$
$V_{33} = 0.46 * 13 + 0.26 * 10 + 0.15 * 9 + 0.08 * 9 + 0.05 * 3 = 10.8$
$V_{41} = 0.62 * 11 + (0.27) * 6 + (0.11) * 4 = 8.88$
$V_{42} = 0.75 * 15 + 0.25 * 9 = 13.5$
$V_{43} = 0.46 * 15 + 0.26 * 11 + 0.15 * 9 + 0.08 * 6 + 0.05 * 4 = 11.79$

Finally we use these values to obtain the generalized expected value for each alternative:

$$C_i = V_{i1} m(B_1) + V_{i2} * m(B_2) + V_{i3} * m(B_3) = 0.6 * V_{i1} + 0.3 * V_{i2} + 0.1 * V_{i3}$$
$$C_1 = 10.382 \quad C_2 = 9.67 \quad C_3 = 9.759 \; C_4 = 10.557.$$

Given the above information the optimal choice is alternative A_4.

6 Conclusion

In this paper we have introduced a general approach to decision making with uncertain information. This methodology can be seen to provide a technique to augment the applications of the Dempster–Shafer theory in expert systems applications by providing a basis for making decisions. A number of possible connections exist between this approach and an approach suggested by Quiggin[17] and that of Jaffray[9] which merit future investigation. It can be shown that this approach satisfies most of the desirable properties of rational decision making, it is symmetric and Pareto optimal.

References

1. A. P. Dempster, "Upper and lower probabilities induced by a multi-valued mapping." *Ann. of Mathematical Statistics*, 1967, pp. 325–339.
2. G. A. Shafer, *Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J., 1976.
3. R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decision making." *IEEE Trans. on Systems, Man and Cybernetics*, **18**, 1988, pp. 83–190.
4. R. R. Yager, "A general approach to decision making with evidential knowledge." In: *Uncertainty in Artificial Intelligence*, edited by L. N. Kanal and J. L. Lemmer, North-Holland, Amsterdam, 1986, pp. 317–330.
5. R. R. Yager, "Optimal alternative selection in the face of evidential knowledge." In: *Optimization Models using Fuzzy Sets and Possibility Theory*, edited by J. Kacprzyk and S. A. Orlovski, D. Reidel, Dordrecht, 1987, pp. 123–140.
6. S. B. Richmond, *Operations Research for Management Decisions*, Ronald Press, New York, 1968.
7. M. J. Bolanos, M. T. Lamata and S. Moral, "Decision making problems in a general environment." *Fuzzy Sets and Systems*, **25**, 1988, pp. 135–144.

8. D. Dubois and H. Prade, “Evidence measures based on fuzzy information.” *Automatica*, **21**, 1985, pp. 547–562.
9. J. Y. Jaffray, “Application of linear utility theory to belief functions.” In: *Uncertainty and Intelligent Systems*, edited by B. Bouchon, L. Saitta and R. R. Yager, Springer-Verlag, Berlin, 1988, pp. 1–8.
10. T. M. Strat, “Making decisions with belief functions.” *Proceedings Fifth Workshop on Uncertainty in Artificial Intelligence*, Windsor, Ont., 1989, pp. 351–360.
11. R. R. Yager, “Applications and extensions of OWA aggregations.” *Int. J. of Man-Machine Studies* (to appear).
12. M. O’Hagan, ‘Aggregating template rule antecedents in real-time expert systems with fuzzy set logic.” *Int. J. of Man-Machine Studies*, (to appear).
13. J. Gordon and E. H. Shortliffe, “The Dempster–Shafer theory of evidence.” In: *Rule-Based Expert Systems: the MYCIN Experiments of the Stanford Heuristic Programming Project*, edited by B. G. Buchanan and E. H. Shortliffe, Addison Wesley, 1984, pp. 272–292.
14. J. Gordon and E. H. Shortliffe, “A method for managing evidential reasoning in a hierarchical hypothesis space.” *Artificial Intelligence*, **26**, 1985, pp. 323–357.
15. J. D. Lowrance and T. D. Garvey, “Evidential reasoning: A developing concept.” In: *IEEE Int. Conf. on Cybernetics and Society*, Seattle, WA, 1982, pp. 6–9.
16. J. D. Lowrance, T. D. Garvey and T. M. Strat, “A framework for evidential-reasoning systems.” *Proceedings of the 5th National Conference on Artificial Intelligence* (AAAI), Philadelphia, 1986, pp. 896–903.
17. V. Quiggin, “A theory of anticipated utility.” *J. of Economic Behavior and Organization*, **3**, 1982, pp. 323–343.

25

Belief Functions: The Disjunctive Rule of Combination and the Generalized Bayesian Theorem[*]

Philippe Smets

Abstract. We generalize the Bayes' theorem within the transferable belief model framework. The Generalized Bayesian Theorem (GBT) allows us to compute the belief over a space Θ given an observation $x \subseteq X$ when one knows only the beliefs over X for every $\theta_i \in \Theta$. We also discuss the Disjunctive Rule of Combination (DRC) for distinct pieces of evidence. This rule allows us to compute the belief over X from the beliefs induced by two distinct pieces of evidence when one knows only that one of the pieces of evidence holds. The properties of the DRC and GBT and their uses for belief propagation in directed belief networks are analysed. The use of the discounting factors is justfied. The application of these rules is illustrated by an example of medical diagnosis.

Key words: Belief functions, Bayes' theorem, Disjunctive rule of combination

1 Introduction

This paper presents the Disjunctive Rule of Combination (DRC) and the Generalized Bayesian Theorem (GBT) within the framework of the transferable belief model, a model for quantifying beliefs using belief functions. Their use is illustrated by a typical application in the domain of the medical diagnostic process.

Suppose $Bel_1 : 2^\Omega \rightarrow [0,1]$ is a belief function induced on the frame of discernment Ω by a piece of evidence E_1. Suppose $Bel_2 : 2^\Omega \rightarrow [0,1]$ is a belief function induced on the same frame of discernment Ω by another piece of evidence E_2. Suppose E_1 and E_2 are distinct pieces of evidence (Shafer 1976,

[*] This work has been partially funded by CEC-ESPRIT II Basic Research Project 3085 (DRUMS) And the Belgian National Incentive-Program for Fundamental Research in Artificial Intelligence.

Smets 1988, Smets 1992c). Shafer introduced Dempster's rule of combination to compute:

$$Bel_{12} = Bel_1 \oplus Bel_2$$

where Bel_{12} is the belief function induced on Ω by the conjunction 'E_1 and E_2'.

We present a combination rule, the DRC, that permits the derivation of the belief function induced on Ω by the disjunction of two pieces of evidence. It corresponds to a situation where you could assess your belief on Ω if E_1 were true, your belief on Ω if E_2 were true, but you only know that the disjunction 'E_1 or E_2' is true.

As an example of an application of the DRC, consider the medical diagnosis process. Let X be the domain of symptoms, each $x \in X$ being a particular symptom. Let Θ be the domain of diseases, each θ_i in Θ being a particular disease. The diseases θ_i are so defined that they are mutually exclusive and exhaustive. Suppose we have assessed our belief over the symptoms for every disease θ_i and we want to assess our belief over the symptoms knowing only that the patient has either disease θ_1 or disease θ_2. This is the case when it is known that all the diseases excepting θ_1 and θ_2 can be excluded. The DRC provides the solution when the *a priori* belief over θ_1 and θ_2 is vacuous. Its extension to the case where there is a non-vacuous *a priori* over θ_1 and θ_2 can also be obtained.

Simultaneously with the DRC, we derive the GBT. Bayes' theorem is central for probabilistic inference. In the medical diagnostic process considered, let $P(x|\theta_i)$ be the probability of the symptoms given each diagnostic $\theta_i \in \Theta$, and let our *a priori* belief over Θ be quantified by the probability distribution function P_0. After observing the symptom $x \subseteq X$, the probability distribution on Θ is updated into $P(\theta_i|x)$, the *a posteriori* probability distribution on Θ, by the application of Bayes' theorem:

$$P(\theta_i|x) = \frac{P(x|\theta_i)P_0(\theta_i)}{\sum_j P(x|\theta_j)P_0(\theta_j)} \qquad \forall \theta_i \in \Theta.$$

In other words, from the probability over X given each $\theta_i \in \Theta$ (and the *a prior* probability on Θ), Bayes' theorem allows us to derive the probability over Θ given any $x \subseteq X$.

The GBT is a generalization of Bayes' theorem where all conditional probabilities are replaced by belief functions and the *a priori* belief function on Θ is vacuous. *A* further generalization for non-vacuous *a priori* belief on Θ is also presented.

The use of the GBT for medical diagnosis resolves the problem of how to select uncommitted *a priori* probabilities on Θ that can represent the absence of any *a priori* commitment towards any disease. The *vacuous belief* that characterizes a state of total ignorance is used on the disease space Θ. Such a state of ignorance cannot be represented within probability theory; indeed total ignorance means that any strict subset of the disease set Θ should receive the *same* degree of belief. No probability function can describe such a belief state once $|\Theta| > 2$, as the *same* probability should be given to every θ_i, but also to every $\theta_i \cup \theta_j$....(any strict subset of Θ)

1.1 Belief Propagation in Directed Networks

Belief networks described by Shafer et al (1987) are undirected hyper-graphs. Hyper-nodes represent sets of variables (e.g., the symptoms and the diseases) and hyper-edges are weighted with belief functions defined on the product space of the variables represented by the nodes attached to the hyper-edges. In Pearl's approach (Pearl, 1988) - concerning only probability functions - the edges are directed and weighted by the conditional probabilities over (the variables represented by) the child node given (the variables represented by) the parent nodes.

In this paper, we provide the tools necessary to use belief functions (instead of probability functions) in directed graphs similar to those considered by Pearl. An edge between a parent node Θ and a child node X will be weighted by conditional belief functions over X for each value θ_i of Θ. Our approach is less general than Shafer's, but we feel that in practice the loss of generality is not important. Indeed we agree with Pearl (1988) who argues that it is more "natural" and "easier" to assess conditional probabilities (and conditional beliefs) over X given θ_i than the joint probabilities (and beliefs) over the space $X \times \Theta$, and that in most real life cases only conditional beliefs will be collected.

The DRC can be used for forward propagation in directed networks. Consider two parent nodes, Θ and Ψ, of node X and the conditional belief functions $Bel_X(.|\theta_i)$ and $Bel_X(.|\psi_j)$ on X given each $\theta_i \in \Theta$ and given each $\psi_j \in \Psi$. The conjunctive rule of combination provides the belief function on X given "θ_i and ψ_j". The disjunctive rule of combination provides the belief function on X given "θ_i or ψ_j".

The GBT can be used for backward propagation of beliefs in directed networks between a child node X and its parent node Θ. Given the conditional belief over X given each $\theta_i \in \Theta$, the GBT computes the belief induced on Θ for any $x \subseteq X$.

1.2 Content

In Sect. 2, we define the Principle of Minimal Commitment, the Generalized Likelihood Principle and the concept of Conditional Cognitive Independence. The first formalizes the idea that one should never give more belief to something than is justified. The second formalizes the idea that the belief induced by a disjunction of several pieces of evidence is a function of the beliefs induced by each piece of evidence. The third extends the idea of stochastic independence to belief functions.

In Sect. 3, we derive the DRC and the GBT. In Sect. 4, we show that they can also be derived through constructive approaches based on the Principle of Minimal Commitment. In Sect. 5, we present some properties of the GBT and some of its limitations. We show in particular that the GBT becomes the classical Bayes' theorem when all the belief functions happen to be probability functions. In Sect. 6, we present the use of the DRC and the GBT for the propagation of beliefs in directed belief networks. In Sect. 7, we present an example of the use of the DRC and GBT for a medical diagnosis problem. In Sect. 8, we summarize the major results and conclude.

1.3 Historical Notes

Smets (1978) derived initially both the DRC and the GBT by the technique presented in Sect. 4. Most theorems described here are proved in Smets (1978). The GBT was also presented in Smets (1981, 1986, 1988), discussed at full length in Shafer (1982). The DRC was presented in Moral (1985), Dubois and Prade (1986a, 1988), Smets (1988) and Cohen et al. (1987). The present paper not only details both rules and many of their properties, but it also provides normative requirements that justify them.

2 Belief Functions

We present some necessary material concerning belief functions and proceed to expound the following three principles: the Principle of Minimal Commitment, the Generalized Likelihood Principle and the Conditional Cognitive Independence. Belief functions are used to quantify someone's beliefs. They cover the same domain as subjective probabilities, but do not use the additivity axiom required for probability measures. The existence of 'basic belief masses' (bbm) allocated to subsets of a frame of discernment Ω is postulated. For $A \subseteq \Omega$, the bbm $m(A)$ quantifies the portion of belief that supports A without supporting any strict subset of A, and that could be transferred to subsets of A if further information justifies it. This model is at the core of the **transferable belief model**, our interpretation of Dempster-Shafer theory (Smets, 1988, 1990, Smets and Kennes 1994, Smets 1991). Our results can be easily transferred to other interpretations of Dempster-Shafer theory, like the hints theory (Kohlas, 1990) or the context model (Gebhardt and Kruse, 1993)

2.1 Background

Let Ω be a finite non empty set called the **frame of discernment**. The mapping $Bel : 2^\Omega \rightarrow [0, 1]$ is an (unnormalized) **belief function** iff there exists a **basic belief assignment** (bba) $m : 2^\Omega \rightarrow [0, 1]$ such that:

$$\sum_{A \subseteq \Omega} m(A) = 1$$

and

$$Bel(A) = \sum_{B \subseteq A; B \neq \emptyset} m(B).$$

Note that $Bel(\emptyset) = 0$. The values of $m(A)$ for A in Ω are called the **basic belief masses** (bbm). $m(\emptyset)$ may be positive; when $m(\emptyset) = 0$ (hence $Bel(\Omega) = 1$), Bel is called a *normalized belief function.* In Shafer's presentation, he asserts that $m(\emptyset) = 0$, or equivalently that $Bel(\Omega) = 1$, and consequently, belief combination and conditioning are normalized by dividing the results by appropriate scaling factors. The difference between Shafer's definition and ours was introduced when we considered the difference between the *open-world* and *closed-world* assumptions (Smets 1988). The nature of $m(\emptyset) > 0$ is fully discussed in Smets (1992b).

Our presentation is developed under the open-world assumption, as described in the transferable belief model. However the whole presentation is still valid under the more restrictive assumption of a closed-world.

Belief functions are in one-to-one correspondence with **plausibility functions** $Pl : 2^\Omega \to [0,1]$ and **commonality functions** $q : 2^\Omega \to [0,1]$ where for all $A \subseteq \Omega$, $A \neq \emptyset$,

$$\begin{aligned} Pl(A) &= Bel(\Omega) - Bel(\overline{A}) && \text{and } Pl(\emptyset) = 0 \\ q(A) &= \sum_{A \subseteq B} m(B) && \text{and } q(\emptyset) = 1 \end{aligned}$$

where $\overline{A}$ is the complement of A relative to Ω.

A **vacuous belief function** is a normalized belief function such that $Bel(A) = 0, \forall A \neq \Omega$. It quantifies our belief in a state of total ignorance as no strict subset of Ω receives any support.

Suppose Bel quantifies our belief about the frame of discernment Ω and we learn that $\overline{A} \subseteq \Omega$ is false. The resulting conditional belief function $Bel(. \mid A)$ is obtained through **the unnormalized rule of conditioning** (see remark 1 for the use of $\mid$ for the unnormalized conditioning. $Bel(B \mid A)$ can be read as the (degree of) belief of B given A or the belief of B in a context where A holds):

$$\begin{aligned} m(B \mid A) &= \sum_{X \subseteq \overline{A}} m(B \cup X) && \text{if } B \subseteq A \subseteq \Omega && 2.1 \\ &= 0 && \text{otherwise} \\ Bel(B \mid A) &= Bel(B \cup \overline{A}) - Bel(\overline{A}) && \forall B \subseteq \Omega \\ Pl(B \mid A) &= Pl(A \cap B) && \forall B \subseteq \Omega \end{aligned}$$

The origin of this relation is to be found in the nature of the transferable belief model itself. A mass $m(B)$ given to B is transferred by conditioning on A to $A \cap B$. Other justifications can also be advanced. $Bel(. \mid A)$ is the minimal commitment specialization of Bel, such that $Pl(\overline{A} \mid A) = 0$

(Klawonn and Smets 1992). It can also be derived as the minimal commitment solution where $Bel(\text{“}B \mid A\text{”})$ is considered to be the belief in the conditional object “$B \mid A$” (Nguyen and Smets, 1993). Note that these derivations are obtained without ever considering the concept of ‘combination of distinct pieces of evidence’, hence without requiring any definition of the notions of distinctness, combination and probability).

Consider two belief functions Bel_1 and Bel_2 induced by two distinct pieces of evidence on Ω. The belief function Bel_{12} that quantifies the combined impact of the two pieces of evidence is obtained through **the conjunctive rule of combination**: $Bel_{12} = Bel_1 \textcircled{\cap} Bel_2$ where $\textcircled{\cap}$ represents the conjunctive combination operator. Its computation is based on the basic belief assignment m_{12}:

$$m_{12}(A) = \sum_{B \cap C = A} m_1(B) m_2(C) \qquad \forall A \subseteq \Omega. \tag{2.2}$$

Expressed with the commonality functions, it becomes:

$$q_{12}(A) = q_1(A) q_2(A).$$

It can also be represented as: (Dubois and Prade (1986b) proved the relation for m_{12}.)

$$\begin{aligned} m_{12}(A) &= \sum_{B \subseteq \Omega} m_1(A \mid B) m_2(B) \\ Bel_{12}(A) &= \sum_{B \subseteq \Omega} Bel_1(A \mid B) m_2(B) \\ Pl_{12}(A) &= \sum_{B \subseteq \Omega} Pl_1(A \mid B) m_2(B) \\ q_{12}(A) &= \sum_{B \subseteq \Omega} q_1(A \mid B) m_2(B) \end{aligned} \tag{2.3}$$

Note that no **normalization factor** appears in these rules.

Remark 1. **Definitions and symbols.** Almost all authors working with belief functions consider only normalized belief functions, whereas we consider mainly unnormalized belief functions. In order to avoid confusion, we propose to keep the names of Dempster’s rule of conditioning and Dempster’s rule of combination for the normalized forms of conditioning and conjunctive combination, as was introduced by Shafer (1976). For the unnormalized rules, we propose to use the names of unnormalized rule of conditioning for 2.1, conjunctive rule of combination for 2.2 and disjunctive rule of combination for the rule introduced in Sect. 3.

We also propose to use the following symbols to represent these operations.

Dempster's rule of conditioning:	$\mid$	$Bel(A\mid B)$
unnormalized rule of conditioning:	$¦$	$Bel(A\,¦\,B)$
Dempster's rule of combination:	$\oplus$	$Bel_{12} = Bel_1 \oplus Bel_2$
conjunctive rule of combination:	$\textcircled{$\cap$}$	$Bel_{12} = Bel_1 \textcircled{$\cap$} Bel_2$
disjunctive rule of combination:	$\textcircled{$\cup$}$	$Bel_{12} = Bel_1 \textcircled{$\cup$} Bel_2$

The difference betwen the elements of the two pairs ($\mid$, $¦$) and ($\oplus$, $\textcircled{$\cap$}$) results only from the normalization factors applied in $\mid$ and $\oplus$. $\textcircled{$\cup$}$ does not have a specific counterpart in Shafer's presentation (indeed once Bel_1 and Bel_2 are normalized, $Bel_1 \textcircled{$\cup$} Bel_2$ is also normalized). Note that $Bel(.\,¦\,B)$ could be a normalized belief function. In fact $¦$ is a generalization of $\mid$.

Remark 2. **Notation.** Given two spaces Θ and X, we write $Bel_X(.\,¦\,\theta)$ and $Pl_X(.\,¦\,\theta)$ to represent the belief and plausibility functions induced on space X in a context where $\theta \subseteq \Theta$ is the case, and $Bel_{X\times\Theta}$, $Pl_{X\times\Theta}$ to represent belief and plausibility functions on the space $X \times \Theta$. We write $x \cap \theta$ as a shorthand for the intersection of the cylindrical extensions of $x \subseteq X$ and $\theta \subseteq \Theta$ over the product space $X \times \Theta$ (i.e., $x \cap \theta$ means $cyl(x) \cap cyl(\theta)$). Similarly $x \cup \theta$ means $cyl(x) \cup cyl(\theta)$... Subscripts of Bel and Pl represent their domain and are omitted when there is no ambiguity as in $Bel(x\,¦\,\theta)$, $Bel(\theta)$,...

Remark 3. Our notation will not distinguish between elements like θ_i where $\theta_i \in \Theta$ and their corresponding singleton $\{\theta_i\} \subseteq \Theta$. The context should always make it clear which is intended, and the notation is seriously lightened.

The following lemmas will be useful:

Lemma 1. *If $Pl : 2^{\Omega} \rightarrow [0, 1]$ is a plausibility function, then the corresponding commonality function q is $q(A) = \sum_{B\subseteq A}(-1)^{|B|+1}Pl(B)$.*

Proof. immediate by replacing $Bel(\overline{B})$ by $Pl(\Omega) - Pl(B)$ in the relation between q and Bel given in Shafer 1976, p. 41. QED

Lemma 2. *$\forall x \subseteq X$, $\forall\theta \subseteq \Theta$, $\forall\theta_i \in \theta$: $Pl(x\,¦\,\theta) \geq Pl(x\,¦\,\theta_i)$.*

Proof. Let $cyl(x)$ and $cyl(\theta)$ be the cylindrical extensions of x and θ on the space $X \times \Theta$. Then $Pl_X(x\,¦\,\theta) = Pl_{X\times\Theta}(cyl(x)\,¦\,cyl(\theta)) = Pl_{X\times\Theta}(cyl(x) \cap cyl(\theta)) \geq Pl_{X\times\Theta}(cyl(x) \cap cyl(\theta_k)) = Pl_X(x\,¦\,\theta_k)$ where $\theta_k \in \theta$. QED

2.2 The Principle of Minimal Commitment

We introduce the *Principle of Minimal Commitment.* Given a belief function derived on Ω, this principle induces the construction of new belief functions 1) on **refined spaces** Ω' where every element of Ω is split into several elements of Ω' and 2) on **extended spaces** Ω", where Ω" contains all the elements of Ω and some new elements. These two processes are called the **vacuous extension** and the **ballooning extension**, respectively. In this paper, the vacuous extension transforms a belief function over Θ into a belief function over $X \times \Theta$ and the ballooning extension transforms a conditional belief function $Bel_X(.\,|\,\theta_i)$ defined on X for $\theta_i \in \Theta$ into a new belief function over $X \times \Theta$.

In order to understand the Principle of Minimal Commitment, we must consider the meaning of $Bel(A)$ and $Pl(A)$. Within the transferable belief model, the degree of belief $Bel(A)$ given to a subset A quantifies the amount of *justified specific support* to be given to A, and the degree of plausibility $Pl(A)$ given to a subset A quantifies the maximum amount of *potential specific support* that could be given to A.

$$Bel(A) = \sum_{\emptyset \neq X \subseteq \Omega} m(X) \qquad Pl(A) = \sum_{A \cap X \neq \emptyset} m(X) = Bel(\Omega) - Bel(\overline{A}).$$

We say *specific* because $m(\emptyset)$ is neither included in $Bel(A)$ nor in $Pl(A)$. The bbms $m(X)$ included in $Bel(A)$ are only those given to the subsets of A that are not subsets of $\overline{A}$. $m(\emptyset)$ is not included because $\emptyset$ is a subset of both A and $\overline{A}$.

We say *justified* because we include in $Bel(A)$ *only* the bbms given to subsets of A. For instance, consider two distinct elements x and y of Ω. The bbm $m(\{x, y\})$ given to $\{x, y\}$ could support x if further information indicates this. However given the available information the bbm can only be given to $\{x, y\}$.

We say *potential* because the bbm included in $Pl(A)$ could be transferred to non empty subsets of A if some new information could justify such a transfer. It would be the case if we learn that $\overline{A}$ is impossible. After conditioning on A, note that $Bel(A\,|\,A) = Pl(A)$. Large plausibilities given to all subsets reflect the lack of commitment of our belief; we are ready to give a large belief to *any* subset.

Consider now the case where there is ambiguity about the amount of plausibility that should be given to the subsets of Ω. The ambiguity could be resolved by giving the largest possible plausibility to every subsets.

The Principle of Minimal Commitment formalizes this idea: *one should never give more support than justified* to any subset of Ω. It satisfies a form of scepticism, noncommitment, or conservatism in the allocation of belief. In spirit, it is not far from what probabilists attempt to achieve with the maximum entropy principle. The concept of commitment was already introduced to create an ordering on the set of belief functions defined on a frame of discernment Ω (see Moral 1986, Yager 1986, Dubois and Prade 1986a, 1987, Delgado and Moral 1987, Kruse and Schwecke 1990, Hsia 1991).

To define the principle, let Pl_1 and Pl_2 be two plausibility functions on Ω such that:

$$Pl_1(A) \leq Pl_2(A) \qquad \forall A \subseteq \Omega. \tag{2.4}$$

We say that Pl_2 is **no more committed** than Pl_1 (and less committed if there is at least one strict inequality). The same qualification is extended to the related bbas and belief functions. The least committed belief function is the vacuous belief function ($m(\Omega) = 1$). The most committed belief function is the contradictory belief function ($m(\emptyset) = 1$).

The ***principle of minimal commitment*** *indicates that, given two equally supported beliefs, only one of which can apply, the most appropriate is the least committed.*

For unnormalized belief functions, the principle is based on the plausibility function. The inequalities 2.4 expressed in terms of belief functions become:

$$Bel_1(A) + m_1(\emptyset) \geq Bel_2(A) + m_2(\emptyset) \qquad \forall A \subseteq \Omega. \tag{2.5}$$

To define the principle by requiring that:

$$Bel_1(A) \geq Bel_2(A) \qquad \forall A \subseteq \Omega \tag{2.6}$$

is inappropriate as seen in the following example. Let: $Bel_1(A) = 0, \forall A \neq \Omega$, and $Bel_1(\Omega) = .7$. If Bel_2 is a vacuous belief function, it is less committed than Bel_1. It is not the case that $Bel_2(A) \leq Bel_1(A), \forall A \subseteq \Omega$. However, one has $Pl_1(A) = .7 \leq Pl_2(A) = 1, \forall A \subseteq \Omega$ as required.

Under the closed-world assumption, the principle can be similarly defined with plausibility inequalities 2.4 or belief function inequalities 2.6. The last definition is historically the oldest. This explains why we maintain the "Minimal Commitment" name even though it could be argued that the principle would be better named the principle of "maximal plausibility" or "maximal scepticism".

The Principle of Minimal Commitment is not used to derive the DRC and the GBT in Sect. 3. However during the constructive derivations of the GBT in Sect. 4, we will encounter plausibility functions Pl whose values are known only for a set $\mathcal{F}$ of subsets of Ω. In most cases, one can build a plausibility function Pl^* such that $Pl^*(A) = Pl(A), \forall A \subseteq \mathcal{F}$ and Pl^* is nevertheless known everywhere on Ω. This is achieved by committing the largest possible plausibility to every subset of Ω that is not an element of $\mathcal{F}$. This application of the principle of minimal commitment is translated into the following property.

The Principle of Minimal Commitment for partially defined plausibility functions. Let $\mathcal{F}$ be a set of subsets of a frame of discernment Ω, and let Pl be a plausibility function whose value is known only for those subsets of Ω in F. Let $\mathcal{P}$ be the set of all the plausibility functions Pl' on Ω such that $Pl'(A) = Pl(A)$ for all A in $\mathcal{F}$. The maximal element Pl^* of $\mathcal{P}$, when it exists, is the plausibility function Pl^* such that $\forall Pl'$ in $\mathcal{P}$: $Pl^*(B) \geq Pl'(B)$, $\forall B \subseteq \Omega$.

Two special cases of the principle will be used here: the vacuous extension and the "ballooning" extension.

1) Let Ω be a frame of discernment and let Pl be defined for every subset of Ω. Let Ω' be a refinement[1] R of Ω. The plausibility function Pl' on Ω'

[1] The mapping R from Ω to Ω' is a *refinement* if every element of Ω is mapped by R into one or more elements of Ω' and the images $R(\omega)$ of the elements ω of Ω under the refinement R partition Ω'.

induced by Pl that satisfies the Principle of Minimal Commitment is the **vacuous extension** of Pl on Ω via R. Its bbms are defined as follows (Shafer 1976, p. 146 et seq.). Let m and m' be the bbas underlying Pl and Pl'. Then $m'(R(A)) = m(A), \forall A \subseteq \Omega$, and $m'(B) = 0$ otherwise.

2) Let Θ and X be two finite spaces, $Bel_X(.|\theta)$ be a conditional belief function on X given some $\theta \in \Theta$ and $\mathcal{B}el^*$ be the set of belief functions $Bel_{X\times\Theta}$ over space $X \times \Theta$ such that their conditioning given θ is equal to $Bel_X(.|\theta)$. The element of $\mathcal{B}el^*$ that satisfies the Principle of Minimal Commitment is the belief function $Bel^*_{X\times\Theta}$ such that:

$$Bel^*_{X\times\Theta}((cyl(x) \cap cyl(\theta)) \cup cyl(\overline{\theta})) - Bel^*_{X\times\Theta}(cyl(\overline{\theta})) = Bel_X(x|\theta)$$

where $cyl(x)$ and $cyl(\theta)$ are the cylindrical extensions of x and θ on the space $X \times \Theta$, and $Bel^*_{X\times\Theta}(cyl(\overline{\theta})) = m_X(\emptyset|\theta)$. It can be informally rewritten as:

$$Bel^*_{X\times\Theta}(x \cup \overline{\theta}) = Bel_X(x|\theta) + m_X(\emptyset|\theta).$$

We call this transformation between Bel and Bel^* the **deconditionalization process** (Smets 1978). Bel^* is called the **"ballooning extension"** of $Bel(x|\theta)$ on $X \times \Theta$ as each mass $m(x|\theta)$ is given after deconditionalization to the largest subset of $X \times \Theta$ so that its intersection with $cyl(\theta)$ is the set $cyl(x) \cap cyl(\theta)$ (see Fig. 1). Shafer (1982) called Bel^* the 'conditional embedding' of $Bel(x|\theta)$. (Note the similarity between this ballooning extension and the passage from a conjunction $cyl(x) \cap cyl(\theta)$ to a material implication $cyl(x) \rightarrow cyl(\theta)$.)

2.3 Conditional Cognitive Independence

In our derivation of the GBT and the DRC, we need to determine the belief induced by two 'independent' observations given the belief induced by each observation. The concept of 'independence' is defined as follows. Let X and Y be two spaces from which we collect observations (pieces of evidence).

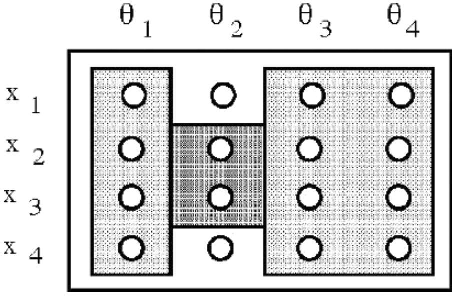


Fig. 1. Ballooning of the bbm $m(x_2 \cup x_3 | \theta_2)$ (*dark area*) onto $X \times \Theta$ (*shaded area*). The white dots correspond to the 16 elements of $X \times \Theta$

The two variables X and Y are said to be 'independent' if *the knowledge of the particular value taken by one of them does not change our belief about the value that the second could take*, i.e. $Bel_X(A|y) = Bel_X(A|y'), \forall A \subseteq X, \forall y, y' \in Y, y \neq y'$ and $Bel_Y(B|x) = Bel_Y(B|x')$, $\forall B \subseteq Y, \forall x, x' \in X, x \neq x'$.

We use this concept of independent observations in order to derive the DRC and the GBT as we claim that two independent observations induce two belief functions that can be combined by the conjunctive rule of combination. More specifically, suppose a set $\Theta = \{\theta_i : i = 1...n\}$ of contexts θ_i. Suppose we collect two observations that are independent whatever the context θ_i. Such two observations are said to be conditionally independent. Each observation induces a belief on Θ and constitutes thus a piece of evidence relative to Θ. We claim that two observations that are conditionally independent constitute two pieces of evidence relative to Θ that are distinct. The satisfaction of that claim was often asked for, it motivated the development of the GBT in Smets (1978), and authors complain of its non satisfaction by other attempts to define an equivalent of the GBT (e.g. see Halpern and Fagin, 1990).

Once that claim is admitted, the properties underlying the concept of Cognitive Independence, detailled here below, are deduced as a spin-off of the DRC. But in fact the concept of independent observations is already sufficient to deduce the properties underlying the concept of Cognitive Independence within the TBM, therefore without regard to the DRC and the GBT.

In the transferable belief model framework, the concept of two independent variables X and Y translates as follows: the ratio of the plausibilities on X should not depend on $y \subseteq$y :

$$\frac{Pl_X(x_1|y)}{Pl_X(x_2|y)} = \frac{Pl_X(x_1)}{Pl_X(x_2)} \qquad \forall x_1, x_2 \subseteq X, \forall y \subseteq Y. \tag{2.7}$$

As $Pl_X(x|y) = Pl_{X \times Y}(x \cap y)$, the independence requirement becomes:

$$\frac{Pl_{X \times Y}(x_1 \cap y)}{Pl_{X \times Y}(x_2 \cap y)} = \frac{Pl_X(x_1)}{Pl_X(x_2)} \qquad \forall x_1, x_2 \subseteq X, \forall y \subseteq Y.$$

These ratio constraints imply that (the proof is given under lemma 4 in the appendix):

$$Pl_{X \times Y}(x \cap y) = Pl_X(x) Pl_Y(y) \qquad \forall x \subseteq X, \forall y \subseteq Y. \tag{2.8}$$

Two variables (X and Y) that satisfy this requirement are said to satisfy the **Cognitive Independence** property. This definition was introduced in Shafer (1976, p. 150). It extends the classical stochastic independence.

The Cognitive Independence concept can be extended in a straighforward manner when the plausibility functions are conditonal plausibility functions. If the two variables X and Y are independent in each context θ_i, for all

$\theta_i \in \Theta$, then they satisfy the **Conditional Cognitive Independence** (CCI) property if:

$$Pl_{X\times Y}(x \cap y \text{¦} \theta_i) = Pl_X(x \text{¦} \theta_i) Pl_Y(y \text{¦} \theta_i) \qquad \forall x \subseteq X, \forall y \subseteq Y, \forall \theta_i \in \Theta \quad 2.9$$

The previous independence definitions are based on plausibility functions. They could have been based as well on belief functions. Two variables X and Y are CCI iff the ratio of their belief functions satisfy the dual of (2.7)

$$\frac{Bel_X(x_1 \text{¦} y)}{Bel_X(x_2 \text{¦} y)} = \frac{Bel_X(x_1)}{Bel_X(x_2)} \qquad \forall x_1, x_2 \subseteq X, \forall y \subseteq Y. \qquad 2.10$$

In fact, both definitions are equivalent as (2.7) is equivalent to (2.10). A proof is given in the appendix (see lemma 5).

2.4 The Generalized Likelihood Principle

In order to derive the DRC and the GBT, we need to generalize the likelihood principle within the transferable belief model. It simply postulates that the belief function induced by the disjunction of two pieces of evidence is only a function of the belief functions induced by each piece of evidence. We will build $Pl_X(. \text{¦} \theta)$ on X for any subset θ of Θ, even though we only know the conditional plausibility functions $Pl_X(. \text{¦} \theta_i)$ over X, $\forall \theta_i \in \Theta$.

To help in understanding the principle, we present the likelihood principle as described in probability theory. The likelihood $l(\theta_i|x)$ (sometimes called the relative plausibility) of the "single" hypothesis θ_i, $\forall \theta_i \in \Theta$, given the data $x \subseteq X$ is defined as being equal to the conditional probability $p(x|\theta_i)$ of the data x given the single hypothesis θ_i (Edwards, 1972)

$$l(\theta_i|x) = p(x|\theta_i).$$

The likelihood of the disjunction $\theta \subseteq \Theta$ of several single hypotheses $\theta_i, i = 1, 2...k$ where $\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$ is defined as a function of the likelihoods of the single hypotheses $\theta_i \in \theta$:

$$l(\theta|x) = f(\{l(\theta_i|x) : \theta_i \in \theta\})$$

where f is the maximum operator ($f(a, b, ..) = \max(a, b, ...)$). The link between the likelihood functions extended to disjunction of hypotheses and possibility functions (Zadeh, 1978, Dubois and Prade, 1985) was shown in Smets (1982).

A form of this principle was already proposed in Shafer (1976, p. 239) when he studied statistical inference in the context of belief functions. He proposed to define $Pl(\theta|x) = \max_{\theta_i \in \theta} Pl(\theta_i|x)$. This solution is not satisfactory for statistical inference, as it does not satisfy Requirement R1 in Sect. 3, a requirement which satisfaction is often asked for (Smets 1978, Halpern and Fagin, 1990).

The likelihood principle is defined for probability functions. We broaden it into the **Generalized Likelihood Principle** applicable to plausibility function within the transferable belief model:

$$\forall \theta \subseteq \Theta, \forall x \subseteq X, Pl(x|\theta) \text{ depends only on } \{Pl(x|\theta_i), Pl(\overline{x}|\theta_i) : \theta_i \in \theta\}.$$

The maximum operator is not assumed. The need of both $Pl(x|\theta_i)$ and $Pl(\overline{x}|\theta_i)$ reflects the non additivity of the plausibility functions.

The origin of the Principle can be justified by requiring that:

1. $Pl(x|\theta)$ is the same after the frame X has been transformed by coarsening into the frame with only two elements: x and $\overline{x}$. This explains why only those values of $Pl(.|\theta_i)$ for x and $\overline{x}$ are used.
2. the values of $Pl(x|\theta_j)$ for $\theta_j \notin \theta$ are irrelevant to the values of $Pl(x|\theta)$. Hence only the $\theta_i \in \theta$ are used.

3 The Disjunctive Rule of Combination and the Generalized Bayesian Theorem

We proceed with the derivation of the DRC and the GBT. Let X and Θ be two finite non empty sets. Suppose all we know about X is represented initially by the set $\{Bel_X(.|\theta_i) : \theta_i \in \Theta\}$ of belief functions $Bel_X(.|\theta_i)$ on X. We only know the beliefs on X when we know which element of Θ holds. We do not know these beliefs on X when we only know that the prevailing element of Θ belongs to a given subset θ of Θ. The DRC permits to build the belief function $Bel_X(.|\theta)$ on X for any $\theta \subseteq \Theta$.

Simultaneously we derive the GBT that permits to build $Bel_\Theta(.|x)$ for any $x \subseteq X$ from the conditional belief functions $Bel_X(.|\theta_i)$, as the DRC and the GBT are linked through the relation:

$$Pl_X(x|\theta) = Pl_\Theta(\theta|x), \quad \forall \theta \subseteq \Theta, \forall x \subseteq X.$$

The derivation of the DRC and the GBT is based on the following ideas. Let X and Y be two frames of discernment. For each $\theta_i \in \Theta$, let $Bel_X(.|\theta_i)$ quantify our belief on X given θ_i, and $Bel_Y(.|\theta_i)$ quantify our belief on Y given θ_i. θ_i can be interpreted as a context. We assume there is no other knowledge about X and Y except these conditional belief functions on X and Y known for each $\theta_i \in \Theta$. It implies among others that we do not have any *a priori* belief on Θ, i.e. we have the vacuous *a priori* belief function Bel_0 on Θ (this condition will be relaxed in Sect. 5).

Suppose we learn then that $x_0 \subseteq X$ holds. What is the belief function $Bel_\Theta(.|x_0)$ on Θ induced by the knowledge of the conditional belief functions $Bel_X(.|\theta_i)$, $\forall \theta_i \in \Theta$, and of the fact that x_0 holds? As we assume that every state of knowledge induces a unique belief on any variable, the belief function

$Bel_\Theta(.|x_0)$ on Θ exists and is unique. Hence $Bel_\Theta(.|x_0)$ is a function F of x_0 and the $Bel_X(.|\theta_i)$ for $\theta_i \in \Theta$:

$$Bel_\Theta(.|x_0) = F(x_0, \{Bel_X(.|\theta_i) : \theta_i \in \Theta\}).$$

Similarly if we learn that $y_0 \subseteq Y$ holds, the belief function $Bel_\Theta(.|y_0)$ on Θ is a function F of y_0 and the $Bel_Y(.|\theta_i)$ for $\theta_i \in \Theta$:

$$Bel_\Theta(.|y_0) = F(y_0, \{Bel_Y(.|\theta_i) : \theta_i \in \Theta\}).$$

Finaly, if we learn that the joint observation $(x_0, y_0) \subseteq X \times Y$, $x_0 \subseteq X, y_0 \subseteq Y$, is the case, we could build the belief function $Bel_\Theta(.|x_0, y_0)$ on Θ based on (x_0, y_0) if we knew the conditional belief functions $Bel_{X\times Y}(.|\theta_i)$ for $\theta_i \in \Theta$:

$$Bel_\Theta(.|x_0, y_0) = F((x_0, y_0), \{Bel_{X\times Y}(.|\theta_i) : \theta_i \in \Theta\}).$$

Suppose the observations $x_0 \subseteq X$ and $y_0 \subseteq Y$ are conditionaly independent whatever context $\theta_i \in \Theta$ holds. The conditional independence of X and Y implies that the observations x_0 and y_0 are two distinct pieces of evidence relative to Θ. Each piece of evidence induces a belief on Θ: $Bel_\Theta(.|x_0)$ and $Bel_\Theta(.|y_0)$. The belief $Bel_\Theta(.|x_0, y_0)$ that x_0 and y_0 jointly induce on Θ can be obtained by the conjunctive rule of combination: $Bel_\Theta(.|x_0, y_0) = Bel_\Theta(.|x_0) \text{\textcircled{$\cap$}} Bel_\Theta(.|y_0)$.

In Requirement R, we ask that the belief function $Bel_\Theta(.|x_0, y_0)$ induced on Θ by two pieces of evidence x_0 and y_0 that corespond to two independent observations $x_0 \subseteq X$ and $y_0 \subseteq Y$ is the same as the belief function $Bel_\Theta(.|x_0) \text{\textcircled{$\cap$}} Bel_\Theta(.|y_0)$ on Θ computed by the conjunctive combination of the individual belief functions $Bel_\Theta(.|x_0)$ and $Bel_\Theta(.|y_0)$. We also ask that $Pl_X(.|\theta)$, $Pl_Y(.|\theta)$ and $Pl_{X\times Y}(.|\theta)$, $\theta \subseteq \Theta$, satisfy the Generalized Likelihood Principle.

Requirement R. Given

- three frames of discernment X, Y and Θ.
- our knowledge on X, Y and Θ is represented by $Bel_X(.|\theta_i)$ and $Bel_Y(.|\theta_i)$, $\forall\theta_i \in \Theta$.
- X and Y are conditionally independent given θ_i, $\forall\theta_i \in \Theta$
- $\forall x \subseteq X$ and $\forall y \subseteq Y$, there is a function F such that

$$\begin{aligned} Bel_\Theta(.|x) &= F(x, \{Bel_X(.|\theta_i) : \theta_i \in \Theta\}) \\ Bel_\Theta(.|y) &= F(y, \{Bel_Y(.|\theta_i) : \theta_i \in \Theta\}) \\ Bel_\Theta(.|x, y) &= F((x, y), \{Bel_{X\times Y}(.|\theta_i) : \theta_i \in \Theta\}) \end{aligned}$$

Then:

Requirement R1:

$$Bel_\Theta(.|x, y) = Bel_\Theta(.|x) \text{\textcircled{$\cap$}} Bel_\Theta(.|y).$$

Requirement R2:

$$Pl_X(x ¦ \theta) = g(\{Pl_X(x ¦ \theta_i), Pl_X(\overline{x} ¦ \theta_i) : \theta_i \in \theta\}) \qquad \forall x \subseteq X, \forall \theta \subseteq \Theta$$

$$Pl_Y(y ¦ \theta) = g(\{Pl_Y(y ¦ \theta_i), Pl_Y(\overline{y} ¦ \theta_i) : \theta_i \in \theta\}) \qquad \forall y \subseteq Y, \forall \theta \subseteq \Theta$$

$$Pl_{X\times Y}(w ¦ \theta) = g(\{Pl_{X\times Y}(w ¦ \theta_i), Pl_{X\times Y}(\overline{w} ¦ \theta_i) : \theta_i \in \theta\}) \forall w \subseteq X \times Y, \forall \theta \subseteq \Theta.$$

The functions F and g will be deduced from Requirement R in Theorems 1 to 4. This allows us to build:

1. $Bel_X(. ¦ \theta)$ and $Bel_Y(. ¦ \theta), \theta \subseteq \Theta$, (the DRC)
2. $Bel_\Theta(. ¦ x)$ and $Bel_\Theta(. ¦ y)$, (the GBT) and
3. $Bel_{X\times Y}(. ¦ \theta), \theta \subseteq \Theta$, (the CCI),

from the set of conditional belief functions $Bel_X(. ¦ \theta_i)$, and $Bel_Y(. ¦ \theta_i), \theta_i \in \Theta$.

The derivation of the DRC and the GBT are presented successively, first when the belief functions $Bel_X(.|\theta_i)$, and $Bel_Y(.|\theta_i)$, $\theta_i \in \Theta$, are normalized (i.e. $Bel_X(X|\theta_i) = 1$ and $Bel_Y(Y|\theta_i) = 1$), then when they are not. The CCI is a by-product of the DRC derivation. All proofs are given in the appendix. We present only the formulas for $Bel_X(. ¦ \theta), \theta \subseteq \Theta$, and $Bel_\Theta(. ¦ x), x \subseteq X$, (and their related Pl, m and q functions). The same formulas can be written for $Bel_Y(. ¦ \theta), \theta \subseteq \Theta$ and $Bel_\Theta(. ¦ y), y \subseteq Y$.

Theorem 1. *The Disjunctive Rule of Combination, normalized beliefs. Given the Requirement R and its antecedents.*

Given $Bel_X(X|\theta_i) = 1$ and $Bel_Y(Y|\theta_i) = 1$, $\forall \theta_i \in \Theta$. Then $\forall \theta \subseteq \Theta$, $\forall x \subseteq X$,

$$Bel_X(x|\theta) = Bel_X(x ¦ \theta) = \prod_{\theta_i \in \theta} Bel_X(x|\theta_i) \tag{3.1}$$

$$Pl_X(x|\theta) = Pl_X(x ¦ \theta) = 1 - \prod_{\theta_i \in \theta} (1 - Pl_X(x|\theta_i) \tag{3.2}$$

$$m_X(x|\theta) = m_X(x ¦ \theta) = \sum_{(\cup_{i:\theta_i \in \theta} x_i) = x} \; \prod_{i:\theta_i \in \theta} m_X(x_i|\theta_i) \tag{3.3}$$

The relation 3.3 shows the dual nature of the conjunctive and disjunctive rules of combination (Dubois and Prade, 1986a). Suppose two belief functions with their basic belief assignments m_1 and m_2 on Ω. When combined, the product $m_1(A)m_2(B)$, $A \subseteq \Omega$, $B \subseteq \Omega$, is allocated to $A \cap B$ in the conjunctive rule of combination, and to $A \cup B$ in the disjunctive rule of combination. One has $\forall C \subseteq \Omega$:

1) conjunctive rule of combination (CRD)

$$m_1 \bigcirc\!\!\!\!\!\cap\, m_2(C) = \sum_{A \cap B = C} m_1(A) m_2(B)$$

$$q_1 \bigcirc\!\!\!\!\!\cap\, q_2(C) = q_1(C) q_2(C)$$

2) disjunctive rule of combination (DRC)

$$m_1 \circledcup m_2(C) = \sum_{A \cup B = C} m_1(A) m_2(B)$$
$$Bel_1 \circledcup Bel_2(C) = Bel_1(C) Bel_2(C)$$

The $\cap$ and $\cup$ operators encountered in the relations for the basic belief assignments explain the origin of the symbols $\circledcap$ and $\circledcup$. These relations show also the dual role of Bel and q. Indeed $Bel(C)$ is the sum of the basic belief masses given to the subsets of C and $q(C)$ as the sum of the basic belief masses given to the supersets of C (beware of the comments after theorem 3).

Once the DRC is known, the GBT is derived thanks to the relation:

$$Pl_\Theta(\theta \shortmid x) = Pl_X(x \shortmid \theta) \qquad \forall \theta \subseteq \Theta, \forall x \subseteq X$$

as confirmed by the equality between 3.2 and 3.5.

Theorem 2. *The Generalized Bayesian Theorem, normalized beliefs. Given the Requirement R and its antecedents. Given* $Bel_X(X|\theta_i) = 1$ *and* $Bel_Y(Y|\theta_i) = 1$, $\forall \theta_i \in \Theta$*. Then* $\forall \theta \subseteq \Theta$, $\forall x \subseteq X$,

$$Bel_\Theta(\theta \shortmid x) = \prod_{\theta_i \in \overline{\theta}} Bel_X(\overline{x}|\theta_i) - \prod_{\theta_i \in \Theta} Bel_X(\overline{x}|\theta_i) \tag{3.4}$$
$$Bel_\Theta(\theta | x) = K.Bel_\Theta(\theta \shortmid x)$$
$$Pl_\Theta(\theta \shortmid x) = 1 - \prod_{\theta_i \in \theta} (1 - Pl_X(x|\theta_i)) \tag{3.5}$$
$$Pl_\Theta(\theta | x) = K.Pl_\Theta(\theta \shortmid x)$$
$$q_\Theta(\theta \shortmid x) = \prod_{\theta_i \in \theta} Pl_X(x|\theta_i) \tag{3.6}$$
$$q_\Theta(\theta | x) = K.q_\Theta(\theta \shortmid x)$$

where

$$K^{-1} = 1 - \prod_{\theta_i \in \Theta} Bel_X(\overline{x}|\theta_i) = 1 - \prod_{\theta_i \in \theta} (1 - Pl_X(x|\theta_i)).$$

As announced the CCI is derived as a by-product of the DRC. Note that 3.2 and 3.5 are identical, reflecting the equality between $Pl_X(x \shortmid \theta)$ and $Pl_\Theta(\theta \shortmid x)$.

Lemma 3. *the Conditional Cognitive Independence. Under theorem 1 conditions:*

$$Pl_{X \times Y}(x \cap y \shortmid \theta_i) = Pl_X(x \shortmid \theta_i) Pl_Y(y \shortmid \theta_i) \qquad \forall x \subseteq X, \forall y \subseteq Y, \theta_i \in \Theta.$$

We proceed with the derivation of the DRC and the GBT when the initial conditional belief functions are not normalized. Given a belief function $Bel : 2^{\Omega} \rightarrow [0,1]$, we define a function $b : 2^{\Omega} \rightarrow [0,1]$ such that $b(A) = Bel(A) + m(\emptyset)$. This b function is the real dual of the commonality function q. The real difference between theorems 1-2 and 3-4 concerns the computation of $Bel_X(x | \theta)$ and $Bel_{\Theta}(\theta | x)$.

Theorem 3. *The Disjunctive Rule of Combination, general case.*

Given the Requirement R and its antecedents. Then $\forall \theta \subseteq \Theta$, $\forall x \subseteq X$,

$$b_X(x | \theta) = \prod_{\theta_i \in \theta} b_X(x | \theta_i) \tag{3.7}$$

$$Bel_X(x | \theta) = b_X(x | \theta) - b_X(\emptyset | \theta) \tag{3.8}$$

$$Pl_X(x | \theta) = 1 - \prod_{\theta_i \in \theta} (1 - Pl_X(x | \theta_i)) \tag{3.9}$$

$$m_X(x | \theta) = \sum_{(\cup_{i:\theta_i \in \theta} x_i) = x} \; \prod_{i:\theta_i \in \theta} m_X(x_i | \theta_i) \tag{3.10}$$

The real dual of q is b, not Bel: indeed in the disjunctive rule of combination one multiplies the b functions, not the Bel functions. $b(C)$ is the sum of the basic belief masses given to the subsets of C, including $\emptyset$. Another way to see the dual nature of the DRC and CRC consists in building the 'complementary' basic belief assignment $\overline{m} : 2^{\Omega} \rightarrow [0,1]$ of a basic belief assignment $m : 2^{\Omega} \rightarrow [0,1]$ with $\overline{m}(A) = m(\overline{A})$ for every $A \subseteq \Omega$. Then $\overline{b}(A) = q(\overline{A})$ (Dubois and Prade, 1986a).

Theorem 4. *The Generalized Bayesian Theorem, general case.*

Given the Requirement R and its antecedents. Then $\forall \theta \subseteq \Theta$, $\forall x \subseteq X$,

$$b_{\Theta}(\theta | x) = \prod_{\theta_i \in \overline{\theta}} b_X(\overline{x} | \theta_i) \tag{3.11}$$

$$Bel_{\Theta}(\theta | x) = b_{\Theta}(\theta | x) - b_{\Theta}(\emptyset | x) \tag{3.12}$$

$$Pl_{\Theta}(\theta | x) = 1 - \prod_{\theta_i \in \theta} (1 - Pl_X(x | \theta_i)) \tag{3.13}$$

$$q_{\Theta}(\theta | x) = \prod_{\theta_i \in \theta} Pl_X(x | \theta_i) \tag{3.14}$$

4 Constructive Derivations of Theorems 3 and 4 Results

In theorems 3 and 4 we derive the DRC and the GBT from general principles (see (Smets 1978)). These relations can also be obtained in a constructive way by the application of the Principle of Minimal Commitment. We present three different ways to derive both the DRC and the GBT. These constructions help in understanding the nature of the solutions.

4.1. For each $\theta_i \in \Theta$, build the ballooning extension $Bel^{(i)}_{X\times\Theta}$ of $Bel_X(.\mid\theta_i)$ on $X\times\Theta$. Combine these belief functions $Bel^{(i)}_{X\times\Theta}$ by the conjunctive rule of combination. Let $Bel_{X\times\Theta} = Bel^{(1)}_{X\times\Theta} \bigcirc\!\!\!\!\!\cap\, Bel^{(2)}_{X\times\Theta} \bigcirc\!\!\!\!\!\cap \ldots \bigcirc\!\!\!\!\!\cap\, Bel^{(n)}_{X\times\Theta}$ be the resulting belief function on $X\times\Theta$. Let $\omega \subseteq X\times\Theta$ and let x_i be the projection of $\omega \cap cyl(\theta_i)$ on X. Then

$$\begin{aligned}
Bel_{X\times\Theta}(\omega) &= \prod_{\theta_i\in\Theta} b_X(x_i\mid\theta_i) - \prod_{\theta_i\in\Theta} b_X(\emptyset\mid\theta_i)\\
m_{X\times\Theta}(\omega) &= \prod_{\theta_i\in\Theta} m_X(x_i\mid\theta_i)\\
q_{X\times\Theta}(\omega) &= \prod_{\theta_i\in\Theta} q_X(x_i\mid\theta_i)
\end{aligned}$$

(all proofs are given in Smets 1978, p. 163 et seq.)

The relations of Theorems 3 and 4 are obtained by conditioning $Bel_{X\times\Theta}$ on $cyl(x)$ or $cyl(\theta)$ and marginalizing the results on X or Θ.

Suppose the conditional belief functions $Bel_X(.|\theta_i)$ are normalized for all $\theta_i \in \Theta$, then any subset of $X\times\Theta$ whose projection on Θ is not Θ itself receives a zero belief, i.e. the only knowledge of the normalized conditional belief functions $Bel_X(.\mid\theta_i)$ induces a vacuous belief on Θ.

4.2. Theorems 3 and 4 results can also be derived by individually considering the ballooning extension Bel_i of each conditional belief function $Bel_X(.\mid\theta_i), i = 1, 2...n$, $(n = |\Theta|)$, on space $X\times\Theta$. Then the Bel_i are conditioned on $x \subseteq X$. The marginalization on Θ of the resulting conditional belief function is the (normalized) simple support function with basic belief masses

$$\begin{aligned}
m(\overline{\theta}_i\mid x) &= Bel_X(\overline{x}\mid\theta_i) + m_X(\emptyset\mid\theta_i)\\
m(\Theta\mid x) &= Bel_X(X\mid\theta_i) - Bel_X(\overline{x}\mid\theta_i)
\end{aligned}$$

The conjunctive combination of these simple support functions on Θ obtained for each $\theta_i \in \Theta$ are the relations 3.11 to 3.13.

4.3. Finally one can also consider that each θ_i, $(i = 1, 2...n)$, is the value of a variable Θ_i that can take only two values: θ_i and $\overline{\theta}_i$. Given $Bel_X(.\mid\theta_i)$, apply the Principle of Minimal Commitment to build the belief function on the space $X\times\Theta_i$ (i.e. build the ballooning extension). Then vacuously extend these belief functions obtained on each $X\times\Theta_i$ onto the space $X\times\Theta_1\times\Theta_2\times\ldots\times\Theta_n$ by again applying the Principle of Minimal Commitment (i.e. build their vacuous extensions on $X\times\Theta_1\times\Theta_2\times\ldots\times\Theta_n$). Combine all these belief functions on $X\times\Theta_1\times\Theta_2\times\ldots\times\Theta_n$ by the conjunctive rule of combination and call the resulting belief function Bel_{Xn}. Let Θ be the space whose elements τ_i

are the intersections (of the cylindrical extensions) of the complements of all $\theta_\nu : \nu \neq i$ and θ_i: so $\tau_i = \overline{\theta}_1 \cap \overline{\theta}_2 ... \cap \theta_i ... \cap \overline{\theta}_n$. Condition Bel_{Xn} on the space $X \times \Theta$. The belief function induced on that space $X \times \Theta$ is the same as the one deduced in Sect. 4.1.

Note that the belief function Bel_X on X induced by the conditioning of Bel_{Xn} on $\theta_1 \cap \theta_2 .... \cap \theta_n$ is the belief function one would have derived by applying the conjunctive rule of combination to the individual conditional belief functions: $Bel_X = Bel_X(.|\theta_1) \textcircled{\cap} Bel_X(.|\theta_2) \textcircled{\cap} \ldots \textcircled{\cap} Bel_X(.|\theta_n)$.

5 Properties of the GBT

5.1. Assume there exists some ***a priori*** **belief $\mathbf{Bel_0}$ over** Θ distinct from the belief induced by the set of conditional belief functions $Bel_X(.|\theta_i)$, $\theta_i \in \Theta$. Combining Bel_0 with the belief function induced on the space $X \times \Theta$ leads to a generalization of the DRC. By (2.3)

$$Bel_X(x) = \sum_{\theta \subseteq \Theta} m_0(\theta) Bel_X(x|\theta) \tag*{5.1}$$

$$= \sum_{\theta \subseteq \Theta} m_0(\theta) \Big(\prod_{\theta_i \in \theta} b_X(x|\theta_i) - \prod_{\theta_i \in \theta} b_X(\emptyset|\theta_i) \Big) \tag*{5.2}$$

$$Pl_X(x) = \sum_{\theta \subseteq \Theta} m_0(\theta) Pl_X(x|\theta) \tag*{5.3}$$

$$= \sum_{\theta \subseteq \Theta} m_0(\theta) \Big(1 - \prod_{\theta_i \in \theta} (1 - Pl_X(x|\theta_i))\Big) \tag*{5.4}$$

Proof. The solution is obtained by $\textcircled{\cap}$-combining the vacuous extension of Bel_0 on $X \times \Theta$ with $Bel_{X \times \Theta}$ and marginalizing them on X, using then $Bel_X(x|\theta)$ as given by 3.8. The full proof is given in Smets 1978, p. 178. QED

Equations 5.1 and 5.3 are particular cases of 2.3. They can be used to speed up computation of beliefs in beliefs networks.

To obtain the belief function induced on Θ given some $x \subseteq X$, we $\textcircled{\cap}$-combine Bel_0 with the belief function deduced on Θ by the GBT. The results are the same as those obtained if we combine the vacuous extension of Bel_0 with the belief function $Bel_{X \times \Theta}$ induced on $X \times \Theta$ by the set of conditional belief functions $Bel_X(.|\theta_i)$, $\theta_i \in \Theta$ (see section 4.1) and then condition the result on x. (Proofs in Smets 1978, p. 177)

5.2. Assume we have some **belief $\mathbf{Bel_{X0}}$ on** X. The GBT becomes

$$Bel_\Theta(\theta) = \sum_{x \subseteq X} m_{X0}(x) Bel_\Theta(\theta|x) \tag*{5.5}$$

where $Bel_\Theta(\theta|x)$ is given by 3.11.

Proof. build the vacuous extension of Bel_{X0} on $X \times \Theta$, $\textcircled{\cap}$-combine it with $Bel_{X\times\Theta}$ as derived in Sect. 4.1., and marginalize the result on Θ. QED

Note that (5.5) enables the backward propagation of belief based on doubtful observations.

5.3. If each $Bel_X(.\text{¦}\theta_i)$ happens to be a **probability function** $P(.|\theta_i)$ **on X**, then the GBT for $|\theta| = 1$ becomes:

$$Pl_\Theta(\theta\text{¦}x) = P(x|\theta) \qquad \forall x \subseteq X.$$

That is, on the singletons θ of Θ, $Pl_\Theta(.\text{¦}x)$ reduces to the likelihood of θ given x. The analogy stops there as the solutions for the likelihood of subsets of Θ are different (see Sect. 2.4).

If, furthermore, the *a priori* belief on θ is also a probability function $P_0(\theta)$, then the normalized GBT becomes:

$$Bel_\Theta(\theta|x) = \frac{\sum_{\theta_i\in\theta} P(x|\theta_i)P_0(\theta_i)}{\sum_{\theta_i\in\Theta} P(x|\theta_i)P_0(\theta_i)} = P(\theta|x)$$

i.e. the (normalized) GBT reduces itself into the classic Bayesian theorem, which explains the origin of its name.

5.4. Assume $Bel_X(.\text{¦}\theta)$ is known not on each singleton of Θ, but on the elements of a partition of Θ. Then redefine Θ by creating the coarsening Θ' of Θ such that the elements of Θ' are the elements of the partition of Θ and proceed as before on the space Θ'.

5.5. Assume $Bel_X(.\text{¦}\theta)$ is known on subsets of Θ which are not mutually exclusive. For instance assume one knows $Bel_X(.\text{¦}\theta_1)$, $Bel_X(.\text{¦}\theta_2)$ and $Bel_X(.\text{¦}\theta_1 \cup \theta_2)$. We must determine whether $Bel_X(.\text{¦}\theta_1 \cup \theta_2)$ is compatible with the Generalized Likelihood Principle (accepting some *a priori* belief on Θ) i.e., does there exist some *a priori* belief function Bel_0 on Θ such that for all $x \subseteq X$:

$$Bel_X(x\text{¦}\theta_1 \cup \theta_2) = m_0(\theta_1)Bel_X(x\text{¦}\theta_1) + m_0(\theta_2)Bel_X(x\text{¦}\theta_2)$$

$$+m_0(\theta_1 \cup \theta_2)(b_X(x\text{¦}\theta_1)b(x\text{¦}\theta_2) - b(\emptyset\text{¦}\theta_1)b(\emptyset\text{¦}\theta_2))$$

(see Sect. 5.1.). A m_0 must be found that satisfies these constraints. This search will not always be successful in which case the DRC and the GBT do not apply. Failure reflects the fact that $Bel_X(\text{¦}\theta_1 \cup \theta_2)$ is based on more information than the one represented by $Bel_X(.\text{¦}\theta_1)$, $Bel_X(.\text{¦}\theta_2)$ and some Bel_0. Difficulties can also appear when there are several solutions m_0 that satisfy the constraints. We will not discuss them further here as, fortunately, in typical cases, $Bel_X(.\text{¦}\theta)$ is known for the singletons θ of Θ (or for subsets θ of Θ that constitutes a partition of Θ). Then both the DRC and the GBT apply.

5.6. When one has an *a priori* belief function Bel_{X0} on X, one could compute

$$Bel^*_{Xi} = Bel_X(.\vdots\theta_i)\textcircled{$\cap$}Bel_{X0}$$

for each θ_i, i.e. our belief over X that combines both pieces of evidence, the one related to the θ_i and the one related to the prior on X. But it is erroneous to use the Bel^*_{Xi} in the GBT directly. Indeed, Bel^*_{Xi} and Bel^*_{Xj}, $i \neq j$, do not result from distinct pieces of evidence as they share the same *a priori* Bel_{X0}. The correct computation consists in isolating each $Bel_X(.\vdots\theta_i)$, ballooning them on $X \times \Theta$, $\textcircled{$\cap$}$-combining them and marginalizing them on X and then $\textcircled{$\cap$}$-combining the result with Bel_{X0}. Through this technique, each piece of evidence is taken into consideration once and only once.

5.7. Discounting a Belief Function Consider an evidence that induces a normalized belief function Bel_Ω on Ω. When the evidence as a whole is itself affected by some uncertainty (unreliability), Shafer (1976, p. 251 et seq.) suggested 'discounting' Bel_Ω in order to take this new uncertainty into account. Let $1-\alpha$ be the degree of trust (reliability) in the evidence as a whole, where $0 \leq \alpha \leq 1$. The discounted belief function Bel^α_Ω on Ω is defined by Shafer (1976 p. 251) such that :

$$\begin{array}{ll} \forall A \subseteq \Omega, A \neq \Omega, & Bel^\alpha_\Omega(A) = (1-\alpha)Bel_\Omega(A) \\ \text{and} & Bel^\alpha_\Omega(\Omega) = Bel_\Omega(\Omega) = 1 \end{array}$$

Shafer considers this concept of discounting as simple and useful but did not explain the origin of within his theory. It can be explained using the same ideas as those that lead to the GBT.

Let $\mathcal{E}$ be a frame with two elements E and $\overline{E}$, where E means 'I know the evidence', and $\overline{E}$ means 'I do not know the evidence'. Assume that these are the only pieces of evidence available. By definition, the belief function $Bel_\Omega(.|E)$ induced on Ω by E is Bel_Ω. The belief function $Bel_\Omega(.|\overline{E})$ induced by on Ω is vacuous: not knowing an evidence leaves us in a state of total ignorance. Thus for each element in $\mathcal{E}$, one has a belief over Ω : $Bel_\Omega(.|E) = Bel_\Omega(.)$ and $Bel_\Omega(.|\overline{E})$ is the vacuous belief function. Lemma 2 shows that $Bel_\Omega(.|E \text{ or } \overline{E})$ is vacuous as $Bel_\Omega(.|\overline{E})$ is vacuous (and this irrespective of the DRC).

Let $1-\alpha$ be my degree of belief over $\mathcal{E}$ that E holds (i.e. my degree of belief that the source of the evidence E is reliable). So one has the bba over $\mathcal{E}$ with $m_\mathcal{E}(E) = 1-\alpha$ and $m_\mathcal{E}(\mathcal{E}) = \alpha$.

Let Bel^*_Ω be the belief induced on Ω by the conditional belief functions $Bel_\Omega(.|E)$, $Bel_\Omega(.|\overline{E})$ and $Bel_\Omega(.|E \text{ or } \overline{E})$, and the prior bba $m_\mathcal{E}$ on $\mathcal{E}$. The application of (5.1) leads to

$$\begin{array}{rll} Bel^*_\Omega(A) & = m_\mathcal{E}(E)Bel_\Omega(.|E) + m_\mathcal{E}(\overline{E})Bel_\Omega(.|\overline{E}) + m_\mathcal{E}(\mathcal{E})Bel_\Omega(.|E \text{ or } \overline{E}) & \\ & = (1-\alpha)Bel_\Omega(A) & \forall A \subseteq \Omega, A \neq \Omega \\ & = 1 & A = \Omega \end{array}$$

Hence $Bel_{\Omega}^{*} = Bel_{\Omega}^{\alpha}$. The relation is always true as it is derived from (5.1) which always holds and not from (5.2) which is derived from the GBT. The discounted belief function Bel_{Ω}^{α} can thus be justified within the TBM.

Informally, the discounted belief function Bel_{Ω}^{α} results from the idea that I have a degree of belief $(1 - \alpha)$ that E is a legitimate (reliable) piece of evidence, in which case my belief on Ω is quantified by Bel_{Ω}. The remaining bbm α is given to the fact thet E might be but is not necessarily a legitimate piece of evidence, in which case my belief on Ω can be quantified by any belief function, including Bel_{Ω}. In such a state of ignorance, the Principle of Minimal Commitment justifies the use of the vacuous belief function to quantify my belief on Ω. Bel_{Ω}^{α} results from the combination of the initial belief function Bel_{Ω} on Ω and the belief built on $\mathcal{E}$.

Discounting can also be seen as the result of the impact of a meta-belief over the set $\mathcal{B}$ of belief functions on Ω. It fits with a very special but important case of a general theory of meta-beliefs. α is the meta-bbm (the basic belief mass related to the meta-belief function) given to the particular element Bel_{Ω} of the set $\mathcal{B}$ of belief functions on Ω. 1-α is the meta-bbm given to $\mathcal{B}$ itself. The discounting operation corresponds to the collapse of the meta-beliefs over the set of belief functions on Ω into a belief function on Ω.

6 Belief Networks

We now introduce some possible applications of the GBT and the DRC. All belief functions considered here are induced by distinct pieces of evidence.

Consider the simplest directed belief network with two nodes Θ and X representing binary variables. The weights on the edge are the conditional plausibility functions on X given θ and $\overline{\theta}$.

$$\begin{tikzcd}
\Theta \arrow[r, "\begin{vmatrix} Pl(x \mid \theta) & Pl(\overline{x} \mid \theta) & Pl(x \cup \overline{x} \mid \theta) \\ Pl(x \mid \overline{\theta}) & Pl(\overline{x} \mid \overline{\theta}) & Pl(x \cup \overline{x} \mid \overline{\theta}) \end{vmatrix}"'] & X
\end{tikzcd}$$

Forward propagation: Assume there is some basic belief masses on Θ: $m(\theta)$, $m(\overline{\theta})$ and $m(\theta \cup \overline{\theta})$. Then we can compute the plausibility induced on X by 5.4:

$$\begin{aligned}
Pl(x) &= m(\theta)Pl(x \mid \theta) + m(\overline{\theta})Pl(x \mid \overline{\theta}) \\
&\qquad + m(\theta \cup \overline{\theta})(1 - (1 - Pl(x \mid \theta))(1 - Pl(x \mid \overline{\theta}))) \\
Pl(\overline{x}) &= m(\theta)Pl(\overline{x} \mid \theta) + m(\overline{\theta})Pl(\overline{x} \mid \overline{\theta}) \\
&\qquad + m(\theta \cup \overline{\theta})(1 - (1 - Pl(\overline{x} \mid \theta))(1 - Pl(\overline{x} \mid \overline{\theta}))) \\
Pl(x \cup \overline{x}) &= m(\theta)Pl(x \cup \overline{x} \mid \theta) + m(\overline{\theta})Pl(x \cup \overline{x} \mid \overline{\theta}) \\
&\qquad + m(\theta \cup \overline{\theta})(1 - (1 - Pl(x \cup \overline{x} \mid \theta))(1 - Pl(x \cup \overline{x} \mid \overline{\theta})))
\end{aligned}$$

Backward propagation: Should we receive a plausibility on X instead, we could compute the belief on Θ by (3.3)

$$\begin{aligned}
Pl(\theta) &= m(x)Pl(x \,|\, \theta) + m(\overline{x})Pl(\overline{x} \,|\, \theta) + m(x \cup \overline{x})Pl(x \cup \overline{x} \,|\, \theta) \\
Pl(\overline{\theta}) &= m(x)Pl(x \,|\, \overline{\theta}) + m(\overline{x})Pl(\overline{x} \,|\, \overline{\theta}) + m(x \cup \overline{x})Pl(x \cup \overline{x} \,|\, \overline{\theta}) \\
Pl(\theta \cup \overline{\theta}) &= m(x)(1 - (1 - Pl(x \,|\, \theta))(1 - Pl(x \,|\, \overline{\theta}))) \\
&\qquad + m(\overline{x})(1 - (1 - Pl(\overline{x} \,|\, \theta))(1 - Pl(\overline{x} \,|\, \overline{\theta}))) \\
&\qquad + m(x \cup \overline{x})(1 - (1 - Pl(x \cup \overline{x} \,|\, \theta))(1 - Pl(x \cup \overline{x} \,|\, \overline{\theta})))
\end{aligned}$$

Propagation in both directions: Should one receive both a belief Bel_{Θ} on Θ and a belief Bel_X on X, then

- *for the X node:* apply forward propagation using Bel_{Θ} and the conditional plausibilities and $\textcircled{\cap}$-combine the result with Bel_X.
- *for the Θ node:* apply backward propagation using Bel_X and the conditional plausibilities and $\textcircled{\cap}$-combine the result with Bel_{Θ}.

Notice the strong symmetry between the above two sets of formula; it reflects the fact that unnormalized conditional plausibilities are symmetrical in their two arguments. Computing the corresponding belief function is immediate. Computing the corresponding basic belief masses or the commonality function should be done with the Fast Moebius Transform (Kennes and Smets 1990) to optimize computation time.

For more complicated acyclic belief networks, the computation is similar. Each node stores the beliefs induced by its immediate neighbours. Once a node X indicates that its belief has changed, it propagates its new belief to all its neighbours. Each neighbour updates the belief induced by X by $\textcircled{\cap}$-combining with its stored beliefs, using commonality functions for efficiency reasons. They then propagate the updated belief to Y's neighbours that have not yet been updated. This propagation is in fact identical to the one encountered in Shafer, Shenoy and Mellouli's algorithm (Shafer et al. 1987). The advantage of our method is that storage on the edge is smaller (at most $|\Theta|2^{|X|}$ values) and propagation between nodes is accelerated. The only weakness of our method is that it does not cover *all* possible belief functions between two variables, it is restricted to those belief functions that can be represented through the set of conditional belief functions, thus a subset of the set of all belief functions. We believe that this loss of generality is not serious, as far as most natural cases correspond to those where only the conditional belief functions are received. Finally, our computation is faster and requires less memory than the Shafer-Shenoy-Mellouli algorithm.

7 Example

In order to illustrate the use of the GBT and the DRC, we consider an example of a medical diagnosis process. Let $\Theta = \{\theta_1, \theta_2, \theta_\omega\}$ be a set of diseases with

Table 1. Conditional beliefs (Bel) and bbm (m) on the symptoms $x \subseteq X$ within each of the mutualy exclusive and exhaustive diagnosis θ_1, θ_2 and $\theta_\omega \in \Theta$. The right part of the table presents the beliefs (and bbm) on X given the disease is either θ_1 or θ_2

	$\{\theta_1\}$		$\{\theta_2\}$		$\{\theta_\omega\}$		$\{\theta_1, \theta_2\}$	
X	m	Bel	m	Bel	m	Bel	m	Bel
$\{x_1\}$	.0	.0	.0	.0	.0	.0	.00	.00
$\{x_2\}$	.0	.0	.0	.0	.0	.0	.00	.00
$\{x_3\}$	.5	.5	.2	.2	.0	.0	.10	.10
$\{x_1, x_2\}$	.2	.2	.6	.6	.0	.0	.12	.12
$\{x_1, x_3\}$	.0	.5	.1	.3	.0	.0	.05	.15
$\{x_2, x_3\}$	.0	.5	.1	.3	.0	.0	.05	.15
$\{x_1, x_2, x_3\}$	.3	1.0	.0	1.0	1.0	1.0	.68	1.00

three mutually exclusive and exhaustive diseases. θ_1 and θ_2 are two 'well known' diseases, i.e. we have some beliefs on what symptoms could hold when θ_1 holds or when θ_2 holds. θ_ω corresponds to the complement of $\{\theta_1, \theta_2\}$ relative to all possible diseases. θ_ω represents not only all the 'other' diseases but also those not yet known. In such a context, our belief on the symptoms can only be vacuous. What do we know about the symptoms caused by a still unknown disease? Nothing of course, hence the vacuous belief function.

We consider two sets X and Y of symptoms with $X = \{x_1, x_2, x_3\}$ and $Y = \{y_1, y_2\}$. Tables 1 and 2 present the beliefs over X and Y when each of the individual diseases holds. They also show the beliefs over the symptoms when we only know that either θ_1 or θ_2 holds. They are derived from theorem 3. The beliefs translate essentially the facts that θ_1 'causes' (supports) x_3 and y_2, and θ_2 'causes' x_2 or x_3 (without preference) and y_1. When we only know that θ_1 or θ_2 holds, then we have a balanced support over X, and some support in favor of y_1.

Table 3 presents the beliefs induced on Θ by the individual observation of symptom x_3 or of symptom y_2, respectively. We assume that the symptoms are independent within each disease, hence the GBT can be applied. The indepen-

Table 2. Conditional beliefs (Bel) and bbm (m) on the symptoms $y \subseteq Y$ within each of the mutualy exclusive and exhaustive diagnosis θ_1, θ_2 and $\theta_\omega \in \Theta$. The right part of the table presents the beliefs (and bbm) on Y given the disease is either θ_1 or θ_2

	$\{\theta_1\}$		$\{\theta_2\}$		$\{\theta_\omega\}$		$\{\theta_1, \theta_2\}$	
Y	m	Bel	m	Bel	m	Bel	m	Bel
$\{y_1\}$	.1	.1	.6	.6	.0	.0	.12	.12
$\{y_2\}$	.7	.7	.0	.0	.0	.0	.00	.00
$\{y_1, x_2\}$	.1	.9	.4	1.0	1.0	1.0	.88	1.00

Table 3. Left part: the basic belief masses (m) and the related commonality functions (q) induced on $\varTheta$ by the observation of symptom x_3 or of symptom y_2. **Right part**, the basic belief masses (m) and the related belief function (Bel), plausibility function (Pl) and commonality function (q) induced on $\varTheta$ by the joint observation of x_3 and y_2

	$¦ x_3$		$¦ y_2$		$¦ x_3, y_2$			
$\varTheta$	m	q	m	q	m	q	Bel	Pl
$\{\theta_1\}$	.00	.80	.00	.80	.00	.64	.00	.64
$\{\theta_2\}$	.00	.40	.00	.60	.00	.24	.00	.24
$\{\theta_\omega\}$	.12	1.00	.08	1.00	.27	1.00	.27	1.00
$\{\theta_1, \theta_2\}$	.00	.32	.00	.48	.00	.15	.00	.73
$\{\theta_1, \theta_\omega\}$	.48	.80	.32	.80	.49	.64	.76	1.00
$\{\theta_2, \theta_\omega\}$	.08	.40	.12	.60	.09	.24	.36	1.00
$\{\theta_1, \theta_2, \theta_\omega\}$	.32	.32	.48	.48	.15	.15	1.00	1.00

dence assumption means that if we knew which disease holds the observation of one of the symptoms would not change our belief about the status of the other symptom. The right half of table 3 presents the beliefs induced on $\varTheta$ by the joint observation of symptom x_3 and of symptom y_2. The beliefs are computed by the application of theorem 4. The symptoms individually and jointly support essentially $\{\theta_1, \theta_\omega\}$. The meaning of $Bel(\theta_\omega ¦ x_3, y_2) = 0.27$ merits some consideration. It quantifies our belief that the joint symptoms x_3 and y_2 are neither 'caused' by θ_1 nor by θ_2. It supports the fact that the joint observation is 'caused' by another disease or by some still unknow disease. A large value for $Bel(\theta_\omega ¦ x_3, y_2)$ somehow supports the fact that we might be facing a new disease. In any case it should induce us in looking for other potential causes to explain the observations.

Table 4 presents the beliefs induced on $\{\theta_1, \theta_2\}$ when we condition our beliefs on $\varTheta$ on $\{\theta_1, \theta_2\}$, or when we have some *a priori* belief on $\varTheta$. The results are obtained by the application of the conjunctive rule of combination applied

Table 4. The basic belief masses (m) and the related (normalized) belief function (Bel_n) induced on $\varTheta$ by the joint observation of x_3 and y_2, and based on three different *a priori* beliefs on $\varTheta$: an *a priori* that reject θ_ω, a probabilistic *a priori* on $\{\theta_1, \theta_2\}$ and a simple support function on $\{\theta_1, \theta_2\}$

$¦ x_3, y_2$	$m(\theta_1, \theta_2) = 1$		$m(\theta_1) = .3$		$m(\theta_1) = .3$	
			$m(\theta_2) = .7$		$m(\theta_1, \theta_2) = .7$	
$\varTheta$	m	Bel_n	m	Bel_n	m	Bel_n
$\{\}$	.30	.00	.70	.00	.32	.00
$\{\theta_1\}$	.54	.77	.19	.63	.57	.84
$\{\theta_2\}$	.06	.09	.11	.37	.04	.06
$\{\theta_1, \theta_2\}$	.10	1.00	.00	1.00	.07	1.00

to the *a priori* belief on Θ and the belief induced by the joint observations. The belief functions presented are normalized.

8 Conclusions

We have presented the GBT and the DRC built on the knowledge of a set of conditional belief functions $Bel_X(.|\theta)$ on X for each θ in Θ where the θ's constitute a partition of Θ. Distinct beliefs on X and/or Θ can be included. Beside the direct relevance of these theorems for inference and the combination of distinct disjunctive pieces of evidence, they are also useful when building belief networks: the assessment of conditional beliefs on X given each θ is more natural and easier than the direct assessment of the joint belief on the space $X \times \Theta$. The loss of generality does not appear to be of any practical importance. In any case, even for the general one, one can always speed up computation and reduce memory requirements thanks to 5.1 and 5.3 that are always valid. Instead of storing the general belief function $Bel_{X\times\Theta}$, store the set of conditional belief functions $Bel_X(.|\theta)$ $\forall\theta \subseteq \Theta$. The total amount of stored data is at most $2^{|X|+|\Theta|}$ instead of $2^{|X|.|\Theta|}$, a serious gain once $|X|$ and $|\Theta|$ become large.

The appropriate use of the GBT and the DRC resolves many of the problems that were raised in Pearl (1990) as supposedly counter-examples against the Dempster-Shafer theory (see Smets (1992a) for an in-depth re-analysis of these examples).

One should take care not to apply the GBT and the DRC blindly. The Generalized Likelihood Principle is not always satisfied. Its applicability must be verified. As a **counterexample**, consider a set of urns with ten balls among which some (n) are white, the others black. Suppose an urn with six white balls ($n = 6$). Let $Bel(W|n = 6)$ be your belief that the next ball extracted from that urn is white knowing there are 6 white balls. You are free to give any value to $Bel(W|n = 6)$. Hacking's frequency principle (Hacking 1965) supports that $Bel(W|n = 6)$ should be 6/10. It provides a reference scale to quantify beliefs, but any monotonous transformation could be as good. Nevertheless $Bel(W|n = 6)$ and $Bel(W|n = 7)$ are related: once $Bel(W|n = 6)$ is given, $Bel(W|n = 7)$ may not be smaller, (if you have the least amount of coherence). Only in the world of "Absurdia" could one accept that the knowledge of $Bel(W|n = 6)$ does not induce any constraint on the value of $Bel(W|n = 7)$. We accept - hopefully - that we are not living in Absurdia. Hence $Bel(W|n = 6)$ and $Bel(W|n = 7)$ are related by extra constraints and these constraints must be incorporated into the model. Applying blindly the GBT is such a context without due regard to the constraints that exist between the conditional belief functions would lead to erroneous answers.

Appendix:

Lemma 4. *If*

$$\frac{Pl_X(x \mid z)}{Pl_Y(y \mid z)} = \frac{Pl_X(x)}{Pl_Y(y)} \qquad \forall x \subseteq X, \forall y \subseteq Y, \forall z \subseteq Z,$$

then $Pl_X(x \mid z) = Pl_X(x)Pl_Z(z)$*.*

Proof. By hypothesis,

$$\frac{Pl_X(x \mid z)}{Pl_X(x)} = \frac{Pl_Y(y \mid z)}{Pl_Y(y)}.$$

So these ratios do not depend on x (nor on y). Let the ratio be equal to $f(z)$. Hence $Pl_X(x \mid z) = Pl_X(x)f(z)$. As $Pl_X(x \mid z) = Pl_{X\times Z}(x \cap z) = Pl_Z(z \mid x)$, then $f(z) = Pl_Z(z)$. QED

Lemma 5. *Let* X *and* Y *be two frames of discernment. Let* Pl_X *and* Pl_Y *be plausibility functions over the frames of discernment* X *and* Y*, respectively. Let* $Pl_{X\times Y}$ *be the plausibility function on* $X \times Y$ *such that:* $Pl_{X\times Y}(x \cap y) = Pl_X(x)Pl_Y(y)$*. Then*

$$Bel_X(x \mid y) = Bel_X(x)Pl_Y(y) \qquad \forall x \subseteq X, \forall y \subseteq Y$$

and

$$\frac{Bel_X(x_1 \mid y)}{Bel_X(x_2 \mid y)} = \frac{Bel_X(x_1)}{Bel_X(x_2)} \qquad \forall x_1, x_2 \subseteq X, \forall y \subseteq Y.$$

Proof. One has:

$$\begin{aligned} Pl_Y(y) &= \frac{Pl_Y(y)(Pl_X(X) - Pl_X(\overline{x}))}{Pl_X(X) - Pl_X(\overline{x})} = \frac{Pl_{X\times Y}(X \cap y) - Pl_{X\times Y}(\overline{x} \cap y)}{Bel_X(x)} \\ &= \frac{Pl_X(X \mid y) - Pl_X(\overline{x} \mid y)}{Bel_X(x)} = \frac{Bel_X(x \mid y)}{Bel_X(x)} \end{aligned}$$

what proves the first equality. The second is then immediate. QED

Proof of Theorem 1:

Let X and Y be two finite spaces. Let $\{Bel_X(. \mid \theta_i), \theta_i \in \Theta\}$ and $\{Bel_Y(. \mid \theta_i), \theta_i \in \Theta\}$, be two sets of normalized belief functions on X and Y, respectively. Let $Pl(\theta \mid x)$, $q(\theta \mid x)$ and $Pl(\theta \mid y)$, $q(\theta \mid y)$ be the plausibility and commonality functions induced on Θ by the two distinct pieces of evidence $x \subseteq X$ and $y \subseteq Y$. Requirement R1 is:

$$q(\theta \mid x, y) = q(\theta \mid x).q(\theta \mid y) \qquad \forall \theta \subseteq \Theta \qquad \text{A.1}$$

It becomes by lemma 1:

$$\sum_{\theta'\subseteq\theta}(-1)^{|\theta'|+1}Pl(\theta'\mid x,y) = \Big(\sum_{\theta'\subseteq\theta}(-1)^{|\theta'|+1}Pl(\theta'\mid x)\Big)\Big(\sum_{\theta'\subseteq\theta}(-1)^{|\theta'|+1}Pl(\theta'\mid y)\Big) \qquad \text{A.2}$$

We analyse successively the cases $|\theta| = 1, 2$ and n.
1) When $|\theta| = 1$, A.2 becomes:

$$Pl(\theta\mid x,y) = Pl(\theta\mid x)Pl(\theta\mid y)$$

or equivalently

$$Pl_{X\times Y}(x\cap y\mid\theta) = Pl_X(x\mid\theta)Pl_Y(y\mid\theta) \qquad \text{A.3}$$

So x and y are CCI (see Sect. 2.3).
2) Assume $\theta = \theta_1\cup\theta_2$ with θ_1,$\theta_2 \in \Theta$, $\theta_1 \neq \theta_2$. For$i$ = 1, 2, let

$$\alpha_i = Pl_X(x\mid\theta_i), \gamma_i = Pl_Y(y\mid\theta_i), \overline{\alpha}_i = Pl_X(\overline{x}\mid\theta_i), \overline{\gamma}_i = Pl_Y(\overline{y}\mid\theta_i),$$
$$f_i = Pl_{X\times Y}(\overline{x}\cup\overline{y}\mid\theta_i).$$

By A. 3, $Pl_{X\times Y}(x\cap y\mid\theta_1) = \alpha_1\gamma_1$ and $Pl_{X\times Y}(x\cap y\mid\theta_2) = \alpha_2\gamma_2$. By the Generalized Likelihood Principle, there exists a g function such that

$$Pl_X(x\mid\theta) = g(\alpha_1, \overline{\alpha}_1, \alpha_2, \overline{\alpha}_2)$$

and

$$Pl_X(y\mid\theta) = g(\gamma_1, \overline{\gamma}_1, \gamma_2, \overline{\gamma}_2)$$

Equation A. 2 becomes:

$$\alpha_1\gamma_1 + \alpha_2\gamma_2 - g(\alpha_1\gamma_1, f_1, \alpha_2\gamma_2, f_2) =$$
$$(\alpha_1 + \alpha_2 - g(\alpha_1, \overline{\alpha}_1, \alpha_2, \overline{\alpha}_2)).(\gamma_1 + \gamma_2 - g(\gamma_1, \overline{\gamma}_1, \gamma_2, \overline{\gamma}_2)) \qquad \text{A.4}$$

Let $Pl_X(.\mid\theta_1)$ be vacuous. Hence $\alpha_1 = \overline{\alpha}_1 = 1$ and $f_1 = 1$ as $f_1 = Pl_{X\times Y}(\overline{x}\cup\overline{y}\mid\theta_1) \geq Pl_X(\overline{x}\mid\theta_1) = 1$. One has also $g(1, 1, \alpha_2, \overline{\alpha}_2) = 1$ as $Pl_X(x\mid\theta) \geq Pl_X(x\mid\theta_1) = 1$.

Equation A.4 becomes:

$$\gamma_1 + \alpha_2\gamma_2 - g(\gamma_1, 1, \alpha_2\gamma_2, f_2) = \alpha_2(\gamma_1 + \gamma_2 - g(\gamma_1, \overline{\gamma}_1, \gamma_2, \overline{\gamma}_2)$$

So g does not depend on its second parameter. Identically g does not depend on its fourth parameter. Let: $k(\alpha, \gamma) = g(\alpha, ., \gamma, .)$.

One has $Pl(\theta_1\cup\theta_2\mid x) = k(Pl(\theta_1\mid x), Pl(\theta_2\mid x))$, or identically, $Pl_X(x\mid\theta_1\cup\theta_2) = k(Pl_X(x\mid\theta_1), Pl_X(x\mid\theta_2))$. Let $Pl_X(x\mid\theta_1) = 1$. As $Pl_X(x\mid\theta_1\cup\theta_2) \geq Pl_X(x\mid\theta_1)$ by lemma 2, then $k(1, \gamma) = 1 = k(\gamma, 1)$ as k is symmetrical in its arguments.

Let $\alpha_1 = \gamma_2 = 1$. Then A.4 becomes:

$$\gamma_1 + \alpha_2 - k(\gamma_1, \alpha_2) = (1 + \alpha_2 - 1).(\gamma_1 + 1 - 1))$$

hence,

$$k(\gamma_1, \alpha_2) = \gamma_1 + \alpha_2 - \alpha_2\gamma_1 = 1 - (1 - \gamma_1)(1 - \alpha_2)$$

and

$$Pl_X(x|\theta){=}k(\alpha_1, \alpha_2){=}1{-}(1{-}\alpha_1)(1{-}\alpha_2){=}1{-}(1{-}Pl_X(x|\theta_1)).(1{-}Pl_X(x|\theta_2)).$$

3) By iteration one gets $Pl_X(x|\theta)$. Assume $\theta = \cup_{i=1}^{n}\theta_i$ where $\theta_i \cap \theta_j = \emptyset$, $\forall i \neq j$. Assume

$$Pl_X(x|\theta) = 1 - \prod_{\theta_i \in \theta}(1 - Pl_X(x|\theta_i)) = 1 - \prod_{i=1}^{n}(1 - \alpha_i).$$

Consider part 2 of the proof, but replace θ_1 by θ and θ_2 by θ_{n+1}. The proof proceeds as in 2). One gets:

$$\begin{aligned} Pl_X(x|\theta \cup \theta_{n+1}) &= Pl_X(x|\theta) + Pl_X(x|\theta_{n+1}) - Pl_X(x|\theta)Pl_X(x|\theta_{n+1}) \\ &= 1 - \prod_{\theta_i \in \theta \cup \theta_{n+1}} (1 - Pl_X(x|\theta_i)) \qquad \text{A.5} \end{aligned}$$

The relation for $Bel_X(x|\theta)$ and $m_X(x|\theta)$ are deduced from A.5. The results are normalized. QED

Proof of theorem 2.

Derive directly from $Pl(\theta|x) = Pl_X(x|\theta)$ and $Bel(\theta|x) = Pl(\Theta|x) - Pl(\overline{\theta}|x)$ and normalize by dividing by $Bel(\Theta|x)$ QED.

Proof of theorem 3.

$m_X(\emptyset|\theta_i)$ (and/or $m_Y(\emptyset|\theta_i)$) might be non null. To see the impact of such non null basic belief masses, enlarge the X space into X' where $X' = X \cup \omega$ where $X \cap \omega = \emptyset$. Apply the same proof as for theorem 1 with normalized belief functions on X' and condition all results on X. As such conditioning is idempotent, one can apply it at the level of $Pl_X(.|\theta)$ or at the level of each $Pl_X(.|\theta_i)$. For all $x \subseteq X$, the plausibilities before and after conditioning are the same. So the Generalized Likelihood Principle still applies for all $x \subseteq X$. But after the conditioning has been applied, the functions $Pl_X(.|\theta_i)$ are unnormalized plausibility functions. QED

Proof of theorem 4.

Derive directly from $Pl(\theta|x) = Pl_X(x|\theta)$ and $Bel(\theta|x) = Pl(\Theta|x) - Pl(\overline{\theta}|x)$. QED.

Acknowledgements

The author is indebted to Didier Dubois, Yen-Teh Hsia, Frank Klawonn, Rudolf Kruse, Victor Poznanski and an anonymous referee for many suggestions for improving the presentation, and Glenn Shafer who, in 1978, indicated to the author that the ballooning extension satisfies the Principle of Minimal Commitment.

Bibliography

COHEN, M.S., LASKEY, K.B. and ULVILA, J.W. (1987) The management of uncertainty in intelligence data: a self-reconciling evidential database. Falls Church, VA: Decision Science Consortium, Inc.

DELGADO M. and MORAL S. (1987) On the concept of possibility-probabilty consistency. Fuzzy Sets and Systems 21: 311–3018.

DUBOIS D. and PRADE H. (1985) *Théorie des possibilités.* Masson, Paris.

DUBOIS D. and PRADE H. (1986a) *A* set theoretical view of belief functions. Int. J. Gen. Systems, 12: 193–226.

DUBOIS D. and PRADE H. (1986b) On the unicity of Dempster rule of combination. Int. J. Intelligent Systems, 1: 133–142.

DUBOIS D. and PRADE H. (1987) The principle of minimum specificity as a basis for evidential reasoning. in: Uncertainty in knowledge-based systems, Bouchon B. and Yager R. eds, Springer Verlag, Berlin, p. 75–84.

DUBOIS D. and PRADE H. (1988) Representation and combination of uncertainty with belief functions and possibility measures. Computational Intelligence, 4: 244–264.

EDWARDS A.W.F. (1972) Likelihood. Cambridge University Press, Cambridge, UK.

GEBHARDT F. and KRUSE R. (1993) The context model: an integrating view of vagueness and uncertainty. Int. J. Approx. Reas. 9(3), 283–314.

HACKING I. (1965) Logic of statistical inference. Cambridge University Press, Cambridge, U.K.

HALPERN J.Y. and FAGIN R. (1990) Two views of belief: Belief as Generlaized Probability and Belief as Evidence. Proc. Eighth National Conf. on AI, 112–119.

HSIA Y.-T. (1991) Characterizing Belief with Minimum Commitment. IJCAI-91: 1184–1189.

KENNES R. and SMETS Ph. (1990) Computational Aspects of the Mšbius Transform. Procs of the 6th Conf. on Uncertainty in AI, Cambridge, USA.

KLAWONN F. and SCHWECKE E. (1992) On the axiomatic justification of Dempster's rule of combination. Int. J. Intel. Systems 7: 469–478.

KLAWONN F. and SMETS Ph. (1992) The dynamic of belief in the transferable belief model and specialization-generalization matrices. in Dubois D., Wellman M.P., d'Ambrosio B. and Smets P. Uncertainty in AI 92. Morgan Kaufmann, San Mateo, Ca, USA, 1992, p. 130–137.

KRUSE R. and SCHWECKE E. (1990) Specialization: a new concept for uncertainty handling with belief functions. Int. J. Gen. Systems 18: 49–60.

KOHLAS J. and MONNEY P. *A.* (1990) Modeling and reasoning with hints. Technical Report. Inst. Automation and OR. Univ. Fribourg.

MORAL S. (1985) Información difusa. Relationes entre probabilidad y possibilidad. Tesis Doctoral, Universidad de Granada.

NGUYEN T. H. and SMETS Ph. (1993) On Dynamics of Cautious Belief and Conditional Objects. Int. J. Approx. Reas. 8(2), 89–104.

PEARL J. (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Pub. San Mateo, Ca, USA.

PEARL J. (1990) Reasoning with Belief Functions: an Analysis of Compatibility. Intern. J. Approx. Reasoning, 4: 363–390.

SHAFER G. (1976) *A* mathematical theory of evidence. Princeton Univ. Press. Princeton, NJ.

SHAFER G. (1982) Belief functions and parametric models. J. Roy. Statist. Soc. B44 322-352.

SHAFER G., SHENOY P.P. and MELLOULI K. (1987) Propagating belief functions in qualitative Markov trees. Int. J. Approx. Reasoning, 1: 349–400.

SMETS Ph. (1978) Un modèle mathématico-statistique simulant le processus du diagnostic médical. Doctoral dissertation, Université Libre de Bruxelles, Bruxelles, (Available through University Microfilm International, 30–32 Mortimer Street, London W1N 7RA, thesis 80-70,003)

SMETS Ph. (1981) Medical Diagnosis : Fuzzy Sets and Degrees of Belief. Fuzzy Sets and systems, 5 : 259–266.

SMETS P. (1982) Possibilistic Inference from Statistical Data. In : Second World Conference on Mathematics at the Service of Man. A. Ballester, D. Cardus and E. Trillas eds. Universidad Politecnica de Las Palmas, pp. 611–613.

SMETS Ph. (1986) Bayes' theorem generalized for belief functions. Proc. ECAI-86, vol. II. 169–171, 1986.

SMETS Ph. (1988) Belief functions. in Smets Ph, Mamdani *A.* , Dubois D. and Prade H. ed. Non standard logics for automated reasoning. Academic Press, London p. 253–286.

SMETS Ph. (1990) The combination of evidence in the transferable belief model. IEEE Trans. Pattern analysis and Machine Intelligence, 12: 447–458.

SMETS Ph. (1991) The Transferable Belief Model and Other Interpretations of Dempster-Shafer's Model. in Bonissone P.P., Henrion M., Kanal L.N. and Lemmer J.F. eds. Uncertainty in Artificial Intelligence 6, North Holland, Amsteram, 375–384.

SMETS Ph. (1992a) Resolving misunderstandings about belief functions: A response to the many criticisms raised by J. Pearl. Int. J. Approximate Reasoning.6: 321–344.

SMETS Ph. (1992b) The nature of the unnormalized beliefs encountered in the transferable belief model. in Dubois D., Wellman M.P., d'Ambrosio B. and Smets P. Uncertainty in AI 92. Morgan Kaufmann, San Mateo, Ca, USA, 1992, p. 292–297.

SMETS Ph. (1992c) The concept of distinct evidence. IPMU 92 Proceedings, p. 789–794.

SMETS P. and KENNES R. (1994) The transferable belief model. Artificial Intelligence, 66(2), 191–234.

YAGER R. (1986) The entailment principle for Dempster-Shafer granules. Int. J. Intell. Systems 1: 247–262

ZADEH L.A. (1978) Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems. 1: 3–28.

26

Representation of Evidence by Hints

Jürg Kohlas and Paul-André Monney

Abstract. This paper introduces a mathematical model of a hint as a body of imprecise and uncertain information. Hints are used to judge hypotheses: the degree to which a hint supports a hypothesis and the degree to which a hypothesis appears as plausible in the light of a hint are defined. This leads in turn to support- and plausibility functions. Those functions are characterized as set functions which are normalized and monotone or alternating of order ∞. This relates the present work to G. Shafer's mathematical theory of evidence. However, whereas Shafer starts out with an axiomatic definition of belief functions, the notion of a hint is considered here as the basic element of the theory. It is shown that a hint contains more information than is conveyed by its support function alone. Also hints allow for a straightforward and logical derivation of Dempster's rule for combining independent and dependent bodies of information. This paper presents the mathematical theory of evidence for general, infinite frames of discernment from the point of view of a theory of hints.

Key words: Hints, Evidence, Support functions, Plausibility functions, Dempster's rule

1 Hints—An Intuitive Introduction

Intuitively, a hint is a body of information relative to some question which is in general **imprecise** in that it does not point to a precise answer but rather to a range of possible answers. It is also often **uncertain** in the sense that the information allows for several possible interpretations and it is not entirely sure which is the correct one. There may be internal conflict within a hint because different interpretations may lead to contradictory answers. Also there can be external contradictions between distinct and different hints relative to the same question. The goal of this paper is to develop a mathematical model of this intuitive notion of a hint and to study some of its basic properties. It takes as its starting point A. Dempster's (1967) multivalued mapping and develops into similar lines as G. Shafer's (1976) mathematical theory of evidence. The theory will however be developed for the most general case and not be limited to the case of finite frames as in Shafer's book.

For an introduction and as a motivation the simpler case of finite hints will first be discussed. Let Θ be an arbitrary finite set whose elements θ represent the possible answers to a given question which has to be considered. One of the elements of Θ represents the true, but unknown answer. Θ is called the **frame of discernment**. The subsets of Θ represent possible propositions about the answer to the question considered. Let Ω denote the finite **set of possible interpretations** of the information contained in the hint to be represented. One of the elements $\omega \in \Omega$ must be the **correct** interpretation, but it is unknown which one. However, not all possible interpretations are equally likely. Thus, a probability $p(\omega)$ for the interpretations $\omega \in \Omega$ is introduced.

Each possible interpretation ω restricts the possible answers within Θ somehow. If ω is the correct interpretation, then the correct answer θ is known to be within some nonempty subset $\Gamma(\omega)$ of Θ, the **focal set** of the interpretation. Alternatively, for any possible interpretation ω, the family $\mathcal{S}$ of the propositions (subsets of Θ) **implied** by the interpretation ω can be considered. $\mathcal{S}$ is simply the family of supersets of the focal set $\Gamma(\omega)$. It has thus trivially the following properties:

(1) $H \in \mathcal{S}$ and $H \subseteq H'$ imply $H' \in \mathcal{S}$
(2) $H_1 \in \mathcal{S},\ H_2 \in \mathcal{S}$ imply $H_1 \cap H_2 \in \mathcal{S}$.
(3) Θ belongs to $\mathcal{S}$, $\emptyset$ does not belong to $\mathcal{S}$.

In addition, the intersection of all implied sets of an interpretation equals $\Gamma(\omega)$. Furthermore, for any possible interpretation, one can also look at the family $\mathcal{P}$ of propositions which are **possible** under the interpretation. A subset $H \subseteq \Theta$ is possible, when H intersects the focal set $\mathit{\Gamma}(\omega)$ of the interpretation. Equivalently, H is possible, iff its complement is not implied, $H^c \notin \mathcal{S}$. $\mathcal{P}$ has the following properties:

(1′) $H \in \mathcal{P}$ and $H \subseteq H'$ imply $H' \in \mathcal{P}$
(2′) $H_1 \in \mathcal{P},\ H_2 \in \mathcal{P}$ imply $H_1 \cup H_2 \in \mathcal{P}$.
(3′) Θ belongs to $\mathcal{P}, \emptyset$ does not belong to $\mathcal{P}$.

Furthermore, if $H \in \mathcal{S}$, then $H^c \notin \mathcal{S}$ and thus $\mathcal{S} \subseteq \mathcal{P}$.

A **hint** is thus defined by a frame of discernment Θ to which it refers, a set of possible interpretations Ω together with a probability $p(\omega)$ and finally a multivalued mapping Γ from the set of interpretations into the frame Θ. If the interpretation ω happens to be the correct one, then the answer to the question considered is restricted to the set $\Gamma(\omega)$. So far, any hint $\mathcal{H}$ is a quadruple $(\Omega, p, \Gamma, \Theta)$.

If a proposition $H \subseteq \Theta$ is fixed as a hypothesis about the correct answer, then it will be interesting to judge this hypothesis in the light of a hint $\mathcal{H}$. Let $\mathcal{S}(\omega)$ and $\mathcal{P}(\omega)$ denote the families of implied and possible propositions of an interpretation ω. Then one can look at the subsets of interpretations under which H is **implied**, $u(H)$, or **possible**, $v(H)$

$$\begin{aligned} u(H) &= \{\omega \in \Omega : H \in \mathcal{S}(\omega)\} \\ \upsilon(H) &= \{\omega \in \Omega : H \in \mathcal{P}(\omega)\} \end{aligned} \tag{1}$$

A hypothesis H, which is implied or supported by many possible interpretations, or more important, by very probable interpretations, is very **credible** in the light of the hint. Also, if the hypothesis is possible under many interpretations, or under very probable interpretations, then the hypothesis is very **plausible** in the light of the hint. Thus, in order to measure the **degree of credibility** or **support** $sp(H)$ and the **degree of plausibility** $Pl(H)$, the probabilities of $u(H)$ and $\upsilon(H)$ can be considered:

$$\begin{aligned} sp(H) &= P(u(H)) \\ pl(H) &= P(\upsilon(H)) . \end{aligned} \tag{2}$$

The values $sp(H)$ and $Pl(H)$ are defined for all subsets of Θ. sp is called a **support** (or belief) **function** and pl a **plausibility function** (or upper probability). These concepts were introduced by A. Dempster (1967) and extensively studied by Shafer (1976) for finite frames of discernment.

The goal of this contribution is to study hints with respect to arbitrary, especially infinite frames. To the best of our knowledge, only very few papers study evidence theory in this general case (Goodman, Nguyen, 1985; Nguyen, 1978; Shafer, 1979; Strat, 1984) The case of belief functions on infinite frames of discernment was in particular studied by Shafer (1979). In this paper belief functions are axiomatically defined as Choquet capacities, monotone of order ∞. Using an integral representation theorem of Choquet (1953, 1969) an **allocation of probability** for belief functions is derived. This concept provides for an interpretation of the meaning of belief. However, with this interpretation, the definition of Dempster's rule for the combination of belief functions is less straightforward. In an unpublished paper G. Shafer (1978) defines first product belief functions on a product space $\Theta \times \Theta$ and then Dempster's rule as a conditioning of the product belief function to the diagonal of $\Theta \times \Theta$. This seems somehow to be a detour. Hints on the other hand allow for a straightforward and logical derivation of Dempster's rule for combining independent and also dependent bodies of information.

Furthermore and more importantly, it will be seen that in the general case a hint contains more information than is conveyed by its support function alone. Therefore, hints cannot be combined on the base of their support functions alone as proposed in Shafer's paper (1978)! This would result in a loss of information. This will be one of the main results of this paper. Another main result is that support- and plausibility functions as defined by (2) can be characterized as Choquet capacities, monotone of order ∞. The proof of this result rejoins Shafer's (1979) development and will only be sketched here. Finally, a new inclusion relation between hints will be introduced in this paper which generalizes a similar relation between support functions introduced by Yager (1985, see also Dubois, Prade, 1986).

In Sect. 2 the general mathematical concept of a hint will be defined. In Sect. 3 support- and plausibility functions will be introduced. A process of refining hints is presented in Sect. 4. It leads to a relation of inclusion between hints. Section 5 studies inclusion relations between hints which are equivalent in the sense that they define partially the same support- and plausibility functions. Finally, in Sect. 6, the combination of hints will be discussed and Dempster's rule derived. In particular, it will be shown that inclusion of hints is maintained under Dempster's rule. The results of this section show that Dempster's rule cannot be defined in terms of support functions only.

2 The Mathematical Model of Hints

The frame of discernment Θ is now an arbitrary set and in particular it can be infinite. The set of possible interpretations Ω can then also be arbitrary. However, Ω will be a probability space $(\Omega, \mathcal{A}, P)$ with a σ-algebra $\mathcal{A}$ and a probability measure P on it. As before (Sect. 1) any possible interpretation $\omega \in \Omega$ restricts the possible answers in Θ somehow. It will be assumed here that to any $\omega \in \Omega$ a family $\mathcal{S}(\omega)$ of implied propositions $H \subseteq \Theta$, satisfying conditions (1) to (3) of Sect. 1, is assigned. A family of subsets satisfying conditions (1) to (3) of Sect. 1 is called a **filter**. The family $\mathcal{P}(\omega) = \{H \subseteq \Theta : H^c \notin \mathcal{S}(\omega)\}$ of possible propositions satisfies conditions (1′) to (3′) of Sect. 1 above. A pair of such dual families $\mathcal{R} = (\mathcal{S}, \mathcal{P})$ will be called a **restriction**.

A restriction $\mathcal{R}$ is called **vacuous**, if $\mathcal{S}$ contains only Θ (and $\mathcal{P}$ all subsets of Θ except the empty set). A vacuous restriction does not restrict at all the possible answers. It is used to represent the situation that, under some interpretations, a hint contains possibly no information at all concerning the question considered.

The set $R = \cap\{H : H \in \mathcal{S}\}$ is called the **base** of the restriction $\mathcal{R}$. One might wonder, whether $\mathcal{S}$ should not be closed under arbitrary intersections and thus $R \in \mathcal{S}$. This will not be assumed here — for reasons which become clear later. However, a restriction $\mathcal{R}$ with $R \in \mathcal{S}$ will be called **set-based**, because in this case $\mathcal{S} = \{H \subseteq \Theta : R \subseteq H\}$ and $\mathcal{P} = \{H \subseteq \Theta : R \cap H \neq \emptyset\}$. For a set-based restriction we write $\mathcal{R} = R$. Similarly, if (2) and (2′) of Sect. 1 hold for countable families, the restriction will be called a **σ-restriction**.

To go back to the model of a hint, it will thus be assumed, that every possible interpretation $\omega \in \Omega$ has assigned a **nonempty** restriction $\Gamma(\omega) = (\mathcal{S}(\omega), \mathcal{P}(\omega))$ describing its implied and possible propositions. Γ is a mapping from Ω into the set $\mathcal{R}(\Theta)$ of restrictions on Θ. This is a generalization of the multivalued mappings considered by A. Dempster (1967). A **hint** $\mathcal{H}$ is thus finally a quintuple $\mathcal{H} = (\Omega, \mathcal{A}, P, \Gamma, \Theta)$ of elements as described above.

A hint $\mathcal{H} = (\Omega, \mathcal{A}, P, \Gamma, \Theta)$ is called **set-focussed**, iff its restrictions $\Gamma(\omega)$ are set-based for all $\omega \in \Omega$. The bases of $\Gamma(\omega)$ are then called **focal sets**. If Θ is a finite set, then all restrictions and thus all hints are set-based. But even in the general case many important classes of hints are set-focussed. For all $\omega \in \Omega$, if $\Gamma(\omega)$ is either a fixed set-based restriction R or the vacuous restriction, then the hint is called **simple**. If $\Gamma(\omega)$ equals the vacuous restriction for all $\omega \in \Omega$, then the hint is called **vacuous**; it represents full ignorance about the question at hand. If $\mathcal{H}$ is a set-focussed hint whose focal sets $\Gamma(\omega)$ all contain only one single point $\theta(\omega)$ of Θ, then the hint is called **precise**. A precise hint corresponds essentially to a random variable (under reserve of the appropriate measurability condition).

Restrictions are fundamental to the theory. In many respects they behave like ordinary subsets of Θ. Especially the operation of **intersection** or **conjunction** can be defined: If $\mathcal{R}_1$ and $\mathcal{R}_2$ are two restrictions known to hold on Θ, then their conjunction forms a new restriction $\mathcal{R} = \mathcal{R}_1 \cap \mathcal{R}_2$ defined by $\mathcal{S} = \{H_1 \cap H_2 : H_1 \in \mathcal{S}_1, H_2 \in \mathcal{S}_2\}$. It is easily verified, that $\mathcal{S}$ is a filter if $\emptyset$ does not belong to $\mathcal{S}$. If $\emptyset \in \mathcal{S}$, then $\mathcal{R}_1$ and $\mathcal{R}_2$ are called **contradictory**. If $\mathcal{R}_1 = R_1$ and $\mathcal{R}_2 = R_2$, then $\mathcal{R}_1 \cap \mathcal{R}_2 = R_1 \cap R_2$. In the same way, the intersection is defined for arbitrary families of restrictions, not only for finite ones.

In order to judge hypotheses $H \subseteq \Theta$ in the light of a hint $\mathcal{H}$, the subset $u(H)$ of interpretations which imply H and the subset $v(H)$ of interpretations under which H is possible are defined as in (1). This defines mappings u and v from the power set $\mathcal{P}(\Theta)$ to the power set $\mathcal{P}(\Omega)$. The following theorem lists some of their elementary properties:

Theorem 1. *(1)* $u(\emptyset) = v(\emptyset) = \emptyset$.
(2) $u(\Theta) = v(\Theta) = \Omega$.
(3) $u(H) = v(H^c)^c$.
(4) $v(H) = u(H^c)^c$.
(5) $u(\cap\{H_i : i \in \mathcal{C}\}) = \cap\{u(H_i) : i \in \mathcal{C}\}$, *where* $\mathcal{C}$ *is* **finite** *in general,* **countable** *for hints with* σ*-restrictions* $\Gamma(\omega), \omega \in \Omega$, *and* **arbitrary** *for set-focussed hints.*
(6) $u(\cup\{H_i : i \in \mathcal{C}\}) \supseteq \cup\{u(H_i) : i \in \mathcal{C}\}$ *for an* **arbitrary** $\mathcal{C}$.
(7) $v(\cup\{H_i : i \in \mathcal{C}\}) = \cup\{v(H_i) : i \in \mathcal{C}\}$, *where* $\mathcal{C}$ *is* **finite** *in general,* **countable** *for hints with* σ*-restrictions* $\Gamma(\omega), \omega \in \Omega$, *and* **arbitrary** *for set-focussed hints.*
(8) $v(\cap\{H_i : i \in \mathcal{C}\}) \subseteq \cap\{v(H_i) : i \in \mathcal{C}\}$ *for an* **arbitrary** $\mathcal{C}$.
(9) $u(H') \subseteq u(H'')$ *if* $H' \subseteq H''$.
(10) $v(H') \subseteq v(H'')$ *if* $H' \subseteq H''$.

Proof. (1) and (2) are trivial. By definition, $v(H)^c = \{\omega \in \Omega : H^c \in \mathcal{S}(\omega)\} = u(H^c)$ and (4) is proved. (3) follows by applying (4) to H^c. (5): If $\omega \in u(H_i)$ for all $i \in \mathcal{C}$, then $H_i \in \mathcal{S}(\omega)$, thus $\cap\{H_i : i \in \mathcal{C}\} \in \mathcal{S}(\omega)$ and therefore $\omega \in u(\cap\{H_i : i \in \mathcal{C}\})$. Inversely, $\omega \in u(\cap\{H_i : i \in \mathcal{C}\})$ implies $\cap\{H_i : i \in \mathcal{C}\} \in \mathcal{S}(\omega)$, hence $H_i \in \mathcal{S}(\omega)$ and $\omega \in u(H_i)$ for all $i \in \mathcal{C}$. (6): If $\omega \in u(H_i)$ for some $i \in \mathcal{C}$, then $H_i \in \mathcal{S}(\omega)$, thus $\cup\{H_i : i \in \mathcal{C}\} \in \mathcal{S}(\omega)$ and $\omega \in u(\cup\{H_i : i \in \mathcal{C}\})$. (7) and (8) are proved using (3),(4),(5) and (6) together with de Morgan laws. (9) and (10) follow immediately from the definitions of u and v. Q.E.D.

In view of (5) u is called a $\cap$ - homomorphism and in view of (7) v is called a $\cup$ - homomorphism.

3 Support and Plausibility Functions

For a hint $\mathcal{H} = (\Omega, \mathcal{A}, P, \Gamma, \Theta)$ the degree of support $sp(H)$ and the degree of plausibility $Pl(H)$ are defined by (2) for any subset H of Θ for which $u(H) \in \mathcal{A}$ and $v(H) \in \mathcal{A}$ respectively. Let $\mathcal{E}_s$ be the class of all subsets H of Θ for which $u(H) \in \mathcal{A}$, i.e. for which the degree of support is defined. The sets of $\mathcal{E}_s$ are called **s-measurable** and $\mathcal{E}_s$ is the domain of the set-function *sp*. Similarly let $\mathcal{E}_p$ be the class of all subsets $H \subseteq \Theta$ for which $v(H) \in \mathcal{A}$, i.e. for which the degree of plausibility is defined. The sets of $\mathcal{E}_p$ are called **p-measurable** and $\mathcal{E}_p$ is the domain of the set-function *pl*.

Note that there is a strong link between the support- and the plausibility function. In fact, according to theorem 1 (4) and (3)

$$\begin{aligned} pl\,(H) &= P\,(v\,(H)) = P\,(u\,(H^c)^c) = 1 - sp\,(H^c) \\ sp\,(H) &= P\,(u\,(H)) = P\,(v\,(H^c)^c) = 1 - pl\,(H^c) \end{aligned} \tag{3}$$

whenever the corresponding probabilities are defined.

Theorem 2. *(1) $\mathcal{E}_s$ is a multiplicative class (i.e. closed under finite intersections) or a σ-multiplicative class (closed under countable intersections) depending on whether $\Gamma(\omega)$, $\omega \in \Omega$ are general restrictions or σ-restrictions.*

(2) $\mathcal{E}_p$ is an additive class (i.e. closed under finite unions) or a σ-additive class (closed under countable unions) depending on whether $\Gamma(\omega)$, $\omega \in \Omega$ are general restrictions or σ-restrictions.

(3) $\mathcal{E}_p = \{H \subseteq \Theta : H^c \in \mathcal{E}_s\}$, $\mathcal{E}_s = \{H \subseteq \Theta : H^c \in \mathcal{E}_p\}$ and $\emptyset, \Theta$ belong to both $\mathcal{E}_s$ and $\mathcal{E}_p$.

Proof. (1) and (2) are direct consequences of theorem 1 (5) and (7) and the fact that $\mathcal{A}$ is a σ-algebra. (3): $H \in \mathcal{E}_s$ is equivalent to $u(H) \in \mathcal{A}$, which is equivalent to $v(H^c) \in \mathcal{A}$ (theorem 1 (3) and (4)) which finally is equivalent to $H^c \in \mathcal{E}_p$. $\emptyset, \Theta$ belong to $\mathcal{E}_s$ and $\mathcal{E}_p$ because of theorem 1 (1) and (2).

Q.E.D.

$\mathcal{E}_s$ and $\mathcal{E}_p$ are called **dual** classes of s- and p-measurable sets. If Ω is a finite set, then **all** subsets of Θ are s- and p-measurable. However, in general $\mathcal{E}_s$ and $\mathcal{E}_p$ are strict subclasses of the power set of Θ. Let's illustrate theorem 2 by a simple, albeit somewhat pathological example: If $(\Omega, \mathcal{A}, P)$ is a probability space and $B \subseteq \Omega$ a subset which does not belong to $\mathcal{A}, \Gamma(\omega) = F \subseteq \Theta$ for all $\omega \in B, \Gamma(\omega) = \Theta$ otherwise, then $\mathcal{E}_s$ contains all subsets of Θ which do not contain F plus the set Θ. We have $u(H) = \emptyset$ for all $H \in \mathcal{E}_s,\ H \neq \Theta$ and thus $sp(H) = 0$, unless $H = \Theta$. $\mathcal{E}_p$ contains all subsets of Θ which are not contained in F^c plus $\emptyset$.

Theorem 3. *The* **support**- *and* **plausibility functionns** *of a hint* $\mathcal{H} = (\Omega, \mathcal{A}, P, \Gamma, \Theta)$*, sp:* $\mathcal{E}_s \rightarrow [0,1]$ *and pl:* $\mathcal{E}_p \rightarrow [0,1]$ *respectively, satisfy the following conditions:*

(1) $sp(\emptyset) = pl(\emptyset) = 0$ *and* $sp(\Theta) = pl(\Theta) = 1$*.*
(2) sp is **monotone of order** ∞*, i.e.*

$$sp\,(E) \geq \sum \left\{ (-1)^{|I|+1}\, sp\,(\cap_{i \in I} E_i) : \emptyset \neq I \subseteq \{1, \ldots, n\} \right\} \tag{4}$$

for all $n \geq 1$ *and sets* $E, E_i \in \mathcal{E}_s$*, such that* $E \supseteq E_i$*; and pl is* **alternating of order** ∞*, i.e.*

$$pl\,(E) \leq \sum \left\{ (-1)^{|I|+1}\, pl\,(\cup_{i \in I} E_i) : \emptyset \neq I \subseteq \{1, \ldots, n\} \right\} \tag{5}$$

for all $n \geq 1$ *and sets* $E, E_i \in \mathcal{E}_p$*, such that* $E \subseteq E_i$*.*

Furthermore, if all $\Gamma(\omega)$*,* $\omega \in \Omega$ *are* σ*-restrictions, then the following conditions hold:*

(3) sp and pl are **continuous***, i.e. if* $E_1 \supseteq E_2 \supseteq \ldots$ *is a monotone decreasing sequence of sets of* $\mathcal{E}_s$*, then*

$$sp\,(\cap_{i=1}^{\infty} E_i) = \lim_{i \rightarrow \infty} sp\,(E_i) \tag{6}$$

and if $E_1 \subseteq E_2 \subseteq \ldots$ *is a monotone increasing sequence of sets of* $\mathcal{E}_p$*, then*

$$pl\,(\cap_{i=1}^{\infty} E_i) = \lim_{i \rightarrow \infty} pl\,(E_i)\,. \tag{7}$$

Proof. (1) follows from theorem 1 (1) and (2). In order to prove (2) for the support function, the well-known inclusion-exclusion formula of probability theory, together with theorem 1 (5), (6) and (9) is used:

$$\begin{aligned} sp\,(E) &= P\,(u\,(E)) \geq P\,(u\,(\cup \{E_i : i = 1, 2, \ldots, n\})) \\ &\geq P\,(\cup\,[u\,(E_i) : i = 1, 2, \ldots, n]) \\ &= \sum \left\{ (-1)^{|I|+1}\, P\,(\cap_{i \in I} u\,(E_i)) : \emptyset \neq I \subseteq \{1, \ldots, n\} \right\} \\ &= \sum \left\{ (-1)^{|I|+1}\, P\,(u\,(\cap_{i \in I} E_i)) : \emptyset \neq I \subseteq \{1, \ldots, n\} \right\} \\ &= \sum \left\{ (-1)^{|I|+1}\, sp\,(\cap_{i \in I} E_i) : \emptyset \neq I \subseteq \{1, \ldots, n\} \right\}. \end{aligned}$$

Condition (2) for the plausibility function is proved in the same way or by using (4) together with (3).

$E_1 \supseteq E_2 \supseteq \ldots$ implies $u(E_1) \supseteq u(E_2) \supseteq \ldots$ (theorem 1 (9)) and $\cap_{i=1}^{\infty} E_i \in \mathcal{E}_s$ (theorem 2 (1)). By the continuity of probabilities and theorem 1 (5)

$$\begin{aligned} sp\left(\cap_{i=1}^{\infty} E_i\right) &= P\left(u(\cap_{i=1}^{\infty} E_i)\right) \\ &= P\left(\cap_{i=1}^{\infty} u(E_i)\right) \\ &= \lim_{i\to\infty} P\left(u\left(E_i\right)\right) = \lim_{i\to\infty} sp\left(E_i\right) \end{aligned}$$

and condition (3) is proved.

Q.E.D.

Note that in particular set-focussed hints have continuous support-and plausibility functions.

Does it make sense to define the degree of support for a hypothesis $H \subseteq \Theta$ outside the class $\mathcal{E}_s$ of s-measurable subsets? If $u(H) \subseteq \Omega$ is not measurable, the model of the hint $\mathcal{H}$ does not contain the necessary information to determine the probability of the set of interpretations supporting H. But any measurable set of interpetations $A \subseteq \Omega$ which is contained in $u(H)$ is a support for H. Hence one may say that the unknown support for H must be at least $P(A)$, for any $A \subseteq u(H)$ and $A \in \mathcal{A}$. Thus, in the absence of further information the support of H could be defined as

$$sp_e\left(H\right) = sup\left\{P\left(A\right) : A \subseteq u\left(H\right), A \in \mathcal{A}\right\} = P_*\left(u\left(H\right)\right) \tag{8}$$

where P_* is the inner probability to P. This is an extension of the support function *sp* onto the whole power set $\mathcal{P}(\Theta)$ because the restriction of sp_e to $\mathcal{E}_s$ equals *sp*. We call sp_e the **vacuous extension** of *sp* to underline that no information not contained in the hint $(\Omega, \mathcal{A}, P, \Gamma, \Theta)$ has been added.

By duality, we may also extend the plausibility functions *pl* from $\mathcal{E}_p$ to $\mathcal{P}(\Theta)$:

$$pl_e\left(H\right) = 1 - sp_e\left(H^c\right). \tag{9}$$

This is similarly called the vacuous extension of *pl*. This name is justified by the following proposition:

Theorem 4. *The equality*

$$pl_e\left(H\right) = inf\ \left\{P\left(A\right) : A \supseteq v\left(H\right), A \in \mathcal{A}\right\} = P^*\left(v\left(H\right)\right) \tag{10}$$

holds. P^* *is the outer probability to* P.

Proof. From the definitions (8) and (9) and theorem 1 (4) it follows that

$$\begin{aligned} pl_e(H) &= 1 - sp_e\left(H^c\right) \\ &= 1 - sup\left\{P\left(A\right) : A \in \mathcal{A}, A \subseteq u\left(H^c\right)\right\} \\ &= 1 - sup\left\{P\left(A\right) : A \in \mathcal{A}, u\left(H^c\right)^c \subseteq A^c\right\} \\ &= inf\left\{P\left(A^c\right) : A \in \mathcal{A}, v\left(H\right) \subseteq A^c\right\} \\ &= inf\left\{P\left(A\right) : A \in \mathcal{A}, v\left(H\right) \subseteq A\right\}. \end{aligned}$$

Q.E.D.

Furthermore, it turns out that sp_e and Pl_e satisfy also the conditions of theorem 3.

Theorem 5. *Let sp_e and Pl_e be the extended support- and plausibility functions of a hint $\mathcal{H} = (\Omega, \mathcal{A}, P, \Gamma, \Theta)$. Then*

(1) sp_e and Pl_e are **monotone** *and* **alternating of order** ∞ *respectively on $\mathcal{P}(\Theta)$.*
(2) If $\Gamma(\omega)$ is a σ-restriction for all ω, then sp_e and Pl_e are also **continuous***.*

The proof of this theorem will not be given here. It seems to be surprisingly difficult and relies on the notion of an allocation of probability (Shafer, 1979). See Kohlas (1990) for a proof of this theorem. The connection between inner probability measures and support or belief functions have also been noted by Ruspini (1987) and Fagin and Halpern (1989), see also Shafer (1990).

4 Refining Hints

A hint $\mathcal{H} = (\Omega, \mathcal{A}, P, \Gamma, \Theta)$ can be refined in several respects by adding supplementary information to it:

(1) The restrictions $\Gamma(\omega)$ associated with the interpretations ω may become more precise: A restrictions $(\mathcal{S}', \mathcal{P}')$ is said to be more precise than (or included in) a restriction $(\mathcal{S}, \mathcal{P})$ iff $\mathcal{S}' \supseteq \mathcal{S}$ (or equivalently $\mathcal{P}' \subseteq \mathcal{P}$), i.e. if it implies more propositions and if less propositions are possible. We write then $(\mathcal{S}', \mathcal{P}') \subseteq (\mathcal{S}, \mathcal{P})$.
(2) Some interpretations which originally are considered as possible may become known as impossible: The new set of possible interpretations Ω' becomes a subset of Ω. This implies also that the original probability P must be conditionned on Ω'. This leads to a new probability space $(\Omega', \mathcal{A}', P')$ of possible interpretations, where $\mathcal{A}' = \mathcal{A} \cap \Omega'$ and $P'(A) = P^*(A \cap \Omega')/P^*(\Omega')$, provided that $P^*(\Omega') > 0$. Note that Ω' is not necessarily measurable; P' is still a probability measure on $\mathcal{A}'$ (Neveu, 1964).
(3) The probability measure P' on the set of possible interpretations Ω' may be extended from the σ-algebra $\mathcal{A}'$ to a probability measure P'' on a larger σ-algebra $\mathcal{A}''$ containing $\mathcal{A}'$. Let's note that in this case

$$P'_*(A) \le P''(A) \le P'^*(A) \tag{11}$$

for all $A \in \mathcal{A}''$.

Thus, combining all three refining steps in the above sequence, a new, refined hint $\mathcal{H}'' = (\Omega'', \mathcal{A}'', P'', \Gamma'', \Theta)$ may be obtained, such that $\Omega'' \subseteq \Omega, \mathcal{A}'' \supseteq \mathcal{A} \cap \Omega''$, P'' is an extension to $\mathcal{A}''$ of the probability measure $P'(A) = P^*(A \cap \Omega'')/P^*(\Omega'')$ on $\mathcal{A} \cap \Omega''$ and $\Gamma''(\omega) \subseteq \Gamma(\omega)$ for all $\omega \in \Omega''$. In this case we write $\mathcal{H}'' \subseteq \mathcal{H}$ and say that $\mathcal{H}''$ is **included** in or is **finer** than

$\mathcal{H}$ (and $\mathcal{H}$ is **coarser** then $\mathcal{H}''$). Of course, many times not all three refining steps are present; in particular often only step (1) or steps (1) and (3) are considered. These particular cases correspond to Yager's (1985) definition of inclusion.

This notion of inclusion of hints leads to the following comparison of the corresponding support- and plausibility functions:

Theorem 6. *Let $\mathcal{H}'' = (\Omega'', \mathcal{A}'', P'', \Gamma'', \Theta)$ and $\mathcal{H} = (\Omega, \mathcal{A}, P, \Gamma, \Theta)$ be two hints such that $\mathcal{H}'' \subseteq \mathcal{H}$ and with sp''_e, pl''_e and sp_e, pl_e as their respective extended support- and plausibility functions. If $k = P^*(\Omega'')$, then*

(1) $sp_e(H) \leq k \cdot sp''_e(H) + (1-k)$ for all $H \subseteq \Theta$
(2) $Pl_e(H) \geq k \cdot pl''_e(H) for\ all\ H \subseteq \Theta$.

Proof. Let $v''(H)$ and $v(H)$ be the subsets of interpretations of Ω'' and Ω respectively under which H is possible. Then clearly $v''(H) \subseteq v(H) \cap \Omega''$ by the refining step (1).

Now, for any $H \subseteq \Theta$,

$$pl_e(H) = P^*(v(H)) \geq P^*(v(H) \cap \Omega'') \geq P^*(v''(H)) = P^*(v''(H) \cap \Omega'').$$

Let $P'(A) = P^*(A \cap \Omega'')/P^*(\Omega'')$ for $A \in \mathcal{A} \cap \Omega''$ and $P'^*(A)$ denote the outer probability measure with respect to P'. Then it follows easily that $P'^*(v''(H) \cap \Omega'') = P^*(v''(H) \cap \Omega'')/P^*(\Omega'')$ and hence

$$pl_e(H) \geq P'^*(v''(H) \cap \Omega'')\, P^*(\Omega'').$$

If $P''^*(A)$ is the outer measure with respect to the probability measure P'' on $\mathcal{A}''$, then clearly $P'^*(A) \geq P''^*(A)$ for any $A \subseteq \Omega''$. Thus

$$\begin{aligned} pl_e(H) &\geq P''^*(v''(H) \cap \Omega'')\, P^*(\Omega'') = P''^*(v''(H))\, P^*(\Omega'') \\ &= pl''_e(H)\, P^*(\Omega'') = k \cdot pl''_e(H). \end{aligned}$$

This proves (2).

By (9) we have

$$\begin{aligned} sp_e(H) &= 1 - pl_e(H^c) \leq 1 - k \cdot pl''_e(H^c) \\ &= 1 - k \cdot (1 - sp''_e(H)) = k \cdot sp''_e(H) + (1-k). \end{aligned}$$

This proves (1).

Q.E.D.

If only refining steps (1) and possibly (3) are present, then $k = 1$ and $[sp''_e(H), pl''_e(H)] \subseteq [sp_e(H), pl_e(H)]$.

To any hint $\mathcal{H} = (\Omega, \mathcal{A}, P, \Gamma, \Theta)$ a vacuous hint $\mathcal{V} = (\Omega, \mathcal{A}, P, \Gamma_{vac}, \Theta)$ can be associated, where $\Gamma_{vac}(\omega)$ is the vacuous restriction for all ω. Clearly $\Gamma(\omega) \subseteq \Gamma_{vac}(\omega)$ for all ω and therefore we have always $\mathcal{H} \subseteq \mathcal{V}$.

5 Families of Hints Related to a Support Function

A hint generates a support function *sp* on some multiplicative class $\mathcal{E}_s$. This function has the properties (1) and (2), possibly (3) as stated in theorem 3. If now *sp* is a function on a multiplicative class $\mathcal{E}_s$, satisfying conditions (1) and (2) of theorem 3, is there always a hint which generates this support function? The answer is affirmative. This is a consequence of an integral theorem of Choquet (1953) as was noted by Shafer (1979). But it can easily be seen that different hints may generate the **same** support function *sp* on $\mathcal{E}_s$, but with **different** extensions sp_e to $\mathcal{P}(\Theta)$. In fact, let $(\Omega, \mathcal{A}, P)$ be a probability space and let B_1, B_2 be two different non-measurable subsets of Ω which have different inner probabilities. Furthermore, let Θ be a frame of discernment and F a strict subset of Θ. This allows to define two distinct hints $\mathcal{H}_i = (\Omega, \mathcal{A}, P, \Gamma_i, \Theta), i = 1, 2$, where

$$\Gamma_i(\omega) = \begin{cases} F & \text{if } \omega \in B_i \\ \Theta & \text{otherwise.} \end{cases}$$

For both hints, the class $\mathcal{E}_s$ equals all subsets of Θ which do not contain F plus the set Θ and the support functions of $\mathcal{H}_1$ and $\mathcal{H}_2$ coincide. But if sp_{1e} and sp_{2e} denote their respective extended support functions, then

$$sp_{1e}(F) = P_*(B_1) \neq P_*(B_2) = sp_{2e}(F).$$

Thus there exists a whole family of hints related to a support function *sp* on $\mathcal{E}_s$. The goal of this section is to study this family of hints. In a similar vain, Shafer (1979) studied various extensions of support (or belief) functions. This section puts some of his results into the perspective of hints.

In the context of the theory of hints Choquet's theorem can be stated as follows:

Theorem 7. *Let $\mathcal{E}_s$ be a multiplicative class and sp: $\mathcal{E}_s \to [0,1]$ a function satisfying conditions (1) and (2) of theorem 3. Then there exists a hint whose support function is sp. If furthermore $\mathcal{E}_s$ is a σ-multiplicative class and sp satisfies condition (3) of theorem 3 (continuity), then there exists a hint whose restrictions are all σ-restrictions and whose support function is sp.*

For a formal proof we refer to Choquet (1953) (see also Shafer, 1978 and Kohlas, 1990). Let's only describe the hint constructed in this proof: As set of possible interpretations the set $\mathcal{R}(\mathcal{E}_s)$ of all filters on the multiplicative class $\mathcal{E}_s$ is selected. Note that to any restriction $\mathcal{R} = (\mathcal{S}, \mathcal{P})$ in $\mathcal{R}(\Theta)$ can be associated a filter $\varphi(\mathcal{R}) = \mathcal{S} \cap \mathcal{E}_s$ on $\mathcal{E}_s$. The maping φ from $\mathcal{R}(\Theta)$ to $\mathcal{R}(\mathcal{E}_s)$ is onto because for any filter $\mathcal{F} \in \mathcal{R}(\mathcal{E}_s)$ the restriction $\mathcal{R}_c(\mathcal{F}) \in \mathcal{R}(\Theta)$ defined by its class of implied propositions $\mathcal{S} = \{H \subseteq \Theta$: *there is an* $E \in \mathcal{F}$ such that $E \subseteq H\}$ is in $\varphi^{-1}(\mathcal{F})$. This shows that $\{\varphi^{-1}(\mathcal{F}) : \mathcal{F} \in \mathcal{R}(\mathcal{E}_s)\}$ is a partition of $\mathcal{R}(\Theta)$. Moreover, $\mathcal{R}_c(\mathcal{F})$ is the coarsest restriction in $\varphi^{-1}(\mathcal{F})$: if $\mathcal{R}' \in \varphi^{-1}(\mathcal{F})$, then $\mathcal{R}' \subseteq \mathcal{R}_c(\mathcal{F})$. Define $\Gamma''(\mathcal{F}) = \mathcal{R}_c(\mathcal{F})$ for any $\mathcal{F} \in \mathcal{R}(\mathcal{E}_s)$. Then there is according to Choquet (1953) a σ-algebra $\mathcal{A}''$ in $\mathcal{R}(\mathcal{E}_s)$ and a probability measure P'' defined on it such that the hint $(\mathcal{R}(\mathcal{E}_s), \mathcal{A}'', P'', \Gamma'', \Theta)$ has *sp* as support function.

Note that using φ the probability space $(\mathcal{R}(\mathcal{E}_s), \mathcal{A}'', P'')$ induces a probability space $(\mathcal{R}(\Theta), \mathcal{A}', P')$. If we define $\Gamma_c(\mathcal{R}) = \mathcal{R}_c(\varphi(\mathcal{R}))$, then the hint $(\mathcal{R}(\Theta), \mathcal{A}', P', \Gamma_c, \Theta)$ generates clearly also the support function *sp* on $\mathcal{E}_s$. Let $u_c(H), v_c(H)$ be the functions (1) defined with respect to Γ_c and let $u_c(\mathcal{E}_s), v_c(\mathcal{E}_p)$ (where $\mathcal{E}_p$ is the dual class to $\mathcal{E}_s$) be the images of $\mathcal{E}_s$ and $\mathcal{E}_p$ with respect to u_c and v_c respectively. By theorem 1 (5) and (7), $u_c(\mathcal{E}_s)$ is a multiplicative class and $v_c(\mathcal{E}_p)$ an additive class. Both $u_c(\mathcal{E}_s)$ and $v_c(\mathcal{E}_p)$ are contained in $\mathcal{A}'$. Now, let $\mathcal{A}_c$ be the smallest σ-algebra containing $u_c(\mathcal{E}_s)$ and $v_c(\mathcal{E}_p)$; $\mathcal{A}_c$ is a subalgebra of $\mathcal{A}'$. Let finally P_c be the restriction of P' to $\mathcal{A}_c$. Then the hint $\mathcal{H}_c = (\mathcal{R}(\Theta), \mathcal{A}_c, P_c, \Gamma_c, \Theta)$ still has *sp* on $\mathcal{E}_s$ as support function. This hint is called the **canonical hint** of the support function *sp* on $\mathcal{E}_s$. We shall see that $\mathcal{H}_c$ is in some sense the coarsest hint which generates *sp* on $\mathcal{E}_s$: among all hints generating *sp*, it contains the least information. This will be formulated more precisely using the inclusion relation between hints introduced in the previous section.

Thus, let $\mathcal{H} = (\Omega, \mathcal{A}, P, \Gamma, \Theta)$ be any hint, which defines the support function *sp* on $\mathcal{E}_s$. More precisely, suppose that the class of s-measurable sets of $\mathcal{H}$ contains $\mathcal{E}_s$ and that on $\mathcal{E}_s$ its support function equals *sp*. Hints which define in this sense identical support functions on $\mathcal{E}_s$ are called **equivalent**. In order to compare equivalent hints among themselves and in particular with the canonical hint, they must be represented with respect to an identical set of possible interpretations. By the mapping Γ, the σ-algebra $\mathcal{A}$ and the probability measure P can be transported to the set $\mathcal{R}(\Theta)$ in the usual way: Consider the σ-algebra $\mathcal{A}'$ of all subsets $B \subseteq \mathcal{R}(\Theta)$ for which $\Gamma^{-1}(B) \in \mathcal{A}$ and define a probability P' on $\mathcal{A}'$ by $P'(B) = P(\Gamma^{-1}(B))$. This leads to an equivalent hint $(\mathcal{R}(\Theta), \mathcal{A}', P', id, \Theta)$ where *id* stands for the identical mapping $id(\mathcal{R}) = \mathcal{R}$. This is called the **canonical representation** of $\mathcal{H}$. In particular, note that this new hint defines the same extended support function sp_e' as $\mathcal{H}$. In this sense $\mathcal{H}$ and its canonical representation $\mathcal{H}_{cr}$ contain exactly the same information.

The following theorem states now that the canonical hint is the coarsest hint among all equivalent hints with respect to a support function *sp* on $\mathcal{E}_s$.

Theorem 8. *Let $\mathcal{H}_c$ be the canonical hint with respect to a support function sp on a multiplicative class $\mathcal{E}_s$. If $\mathcal{H}$ is any equivalent hint with respect to this support function and $\mathcal{H}_{cr}$ its canonical representation, then $\mathcal{H}_{cr} \subseteq \mathcal{H}_c$.*

Proof. Both $\mathcal{H}_{cr}$ and $\mathcal{H}_c$ have the same set of possible interpretations $\mathcal{R}(\Theta)$. Moreover, clearly $id(\mathcal{R}) \subseteq \mathcal{R}_c(\varphi(\mathcal{R})), \mathcal{A}' \supseteq \mathcal{A}_c$ and the restriction of P' to $\mathcal{A}_c$ equals P_c.

Q.E.D.

As a consequence of this theorem, it follows that $[sp_e(H), pl_e(H)] \subseteq [sp_{ce}(H), pl_{ce}(H)]$ for all $H \subseteq \Theta$, if sp_{ce}, pl_{ce} denote the extended support and plausibility functions of the canonical hint and sp_e, pl_e the extended support and plausibility functions of the hint $\mathcal{H}$. Shafer (1979) studied extensions of support functions and identified among others the minimal extension of a support function sp on $\mathcal{E}_s$. It turns out that this minimal extension is in fact as one expects the extension of the canonical hint with respect to sp on $\mathcal{E}_s$.

Theorem 9. *If sp_{ce}, pl_{ce} are the extended support and plausibility functions of the canonical hint $\mathcal{H}_c$ with respect to a support and plausibility function sp and pl on a multiplicative class $\mathcal{E}_s$ and its dual additive class $\mathcal{E}_p$, then*

$$\begin{aligned} sp_{ce}\left(H\right) = sup \Big\{ \sum \Big\{ \left(-1\right)^{|I|+1} sp\left(\cap_{i \in I} E_i\right) : \emptyset \neq I \subseteq \{1,\ldots,n\} \Big\} : \\ E_i \subseteq H, E_i \in \mathcal{E}_s, i = 1,\ldots,n; n = 1,2,\ldots \Big\}, \qquad (12) \end{aligned}$$

$$\begin{aligned} pl_{ce}(H) = inf \Big\{ \sum \Big\{ (-1)^{|I|+1} pl(\cup_{i \in I} E_i) : \emptyset \neq I \subseteq \{1,\ldots,n\} \Big\} : \\ E_i \supseteq H, E_i \in \mathcal{E}_p, i = 1,\ldots,n; n = 1,2,\ldots \Big\}. \qquad (13) \end{aligned}$$

Proof. Note that by theorem 1 (6) $\cup_{i=1}^n u_c(E_i) \subseteq u_c(\cup_{i=1}^n E_i)$. Furthermore

$$\begin{aligned} & sp_{ce}\left(H\right) = P_{c*}\left(u_c\left(H\right)\right) \\ & \geq sup\left\{P_c\left(\cup_{i=1}^n u_c\left(E_i\right)\right) : E_i \subseteq H, E_i \in \mathcal{E}_s, i = 1,\ldots,n; n = 1,2,\ldots\right\} \\ & = sup \Big\{ \sum \Big\{ \left(-1\right)^{|I|+1} P_c\left(\cap_{i \in I} u_c\left(E_i\right)\right) : \emptyset \neq I \subseteq \{1,\ldots,n\} \Big\} : \\ & \quad E_i \subseteq H, E_i \in \mathcal{E}_s, i = 1,\ldots,n; n = 1,2\ldots \Big\} \\ & = sup \Big\{ \sum \Big\{ \left(-1\right)^{|I|+1} sp\left(\cap_{i \in I} u_c E_i\right) : \emptyset \neq I \subseteq \{1,\ldots,n\} \Big\} : \\ & \quad E_i \subseteq H, E_i \in \mathcal{E}_s, i = 1,\ldots,n; n = 1,2\ldots \Big\} \end{aligned}$$

On the other hand, Shafer (1979) proves that the right hand side of (12) defines indeed a support function sp_m on the power set $\mathcal{P}(\Theta)$ satisfying the conditions of theorem 7. There exists therefore a hint $\mathcal{H}'$ which generates this support function and let $\mathcal{H}'_{cr}$ its canonical representation. But theorem 8 implies that $\mathcal{H}'_{cr} \subseteq \mathcal{H}_c$ and by theorem 6 $sp_{ce}(H) \leq sp_{cre}(H) = sp_m(H)$ since $k = 1$. Thus we obtain finally $sp_{ce}(H) = sp_m(H)$ which proves (12).

(13) is deduced from (12) using (3) and theorem 1 (3) and (4) together with the de Morgan laws.

Q.E.D.

Theorem 9 together with theorems 6 and 8 show that sp_m is the smallest support function which extends sp from $\mathcal{E}_s$ to all of $\mathcal{P}(\Theta)$.

If the support function *sp* on a σ-multiplicative class $\mathcal{E}_s$ is **continuous** (satisfies condition (3) of theorem 3), then a canonical hint associated to this support function can be constructed in a similar way with respect to the set of σ-restrictions $\mathcal{R}_\sigma(\Theta)$ on Θ. For any hint for which all restrictions are σ-restrictions, a canonical representation with respect to $\mathcal{R}_\sigma(\Theta)$ can be defined along similar lines as above. Then two further results corresponding to theorems 8 and 9 can be proved:

Theorem 10. *Let $\mathcal{H}_c$ be the canonical hint with respect to a* **continuous** *support function sp on a σ-multiplicative class $\mathcal{E}_s$. If $\mathcal{H}$ is any equivalent hint with respect to this support function and $\mathcal{H}_{cr}$ its canonical representation, then $\mathcal{H}_{cr} \subseteq \mathcal{H}_c$.*

Theorem 11. *If sp_{ce}, pl_{ce} are the extended support and plausibility functions of the canonical hint $\mathcal{H}_c$ with respect to continuous support and plausibility functions sp and pl on a σ-algebra $\mathcal{E}_s = \mathcal{E}_p$, then*

$$sp_{ce}(H) = \sup\left\{\lim_{i\to\infty} sp(E_i) : E_1 \supseteq E_2 \supseteq \ldots, E_i \in \mathcal{E}_s, \cap E_i \subseteq H\right\} \qquad (14)$$

$$pl_{ce}(H) = \inf\left\{\lim_{i\to\infty} pl(E_i) : E_1 \subseteq E_2 \subseteq \ldots, E_i \in \mathcal{E}_p, \cup E_i \supseteq H\right\}. \qquad (15)$$

These theorems will not be proved here. The proofs develop along similar lines as those of theorems 8 and 9. Note that for theorem 11 Shafer (1979) showed that the right hand side of (14) is indeed a continuous support function. This theorem shows that it is the smallest continuous support function which extends the continuous support function *sp* from $\mathcal{E}_s$ to $\mathcal{P}(\Theta)$.

6 Combining Hints

Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two hints relative to the same frame Θ and defined by $(\Omega_1, \mathcal{A}_1, P_1, \Gamma_1, \Theta)$ and $(\Omega_2, \mathcal{A}_2, P_2, \Gamma_2, \Theta)$. The basic idea for the combination of these hints into a combined body of information is that in each hint there must be exactly one correct interpretation $\omega_i, i = 1, 2$ such that — looking at both hints together — ω_1 and ω_2 must be simultaneously correct interpretations. Hence (ω_1, ω_2) must be the correct combined correct interpretation. Therefore, in order to combine the two hints $\mathcal{H}_1$ and $\mathcal{H}_2$ into one new combined hint, we form first the product space of the combined interpretations from the two hints $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2, P')$ where P' is any probability measure on $\mathcal{A}_1 \otimes \mathcal{A}_2$ reflecting the common likelihood of combined interpretations. The two hints are called **independent**, if the interpretations of the two hints are stochastically independent. Then P' is the product measure of P_1 and P_2. This is the case which will be pursued here although other cases would be equally possible.

If the combined interpretation (ω_1, ω_2) is the correct one, then the restriction

$$\Gamma\left(\omega_1,\omega_2\right)=\Gamma_1\left(\omega_1\right)\cap\Gamma_2\left(\omega_2\right) \tag{16}$$

must necessarily hold. Note that it is possible that $\Gamma_1(\omega_1)$ and $\Gamma_2(\omega_2)$ are contradictory. Then ω_1 and ω_2 are called contradictory interpretations.

Define now

$$\begin{aligned} u'\left(H\right)&=\left\{\left(\omega_1,\omega_2\right)\in\Omega_1\times\Omega_2: \ H \textit{ is implied by } \Gamma\left(\omega_1,\omega_2\right)\right\}\\ v'\left(H\right)&=\left\{\left(\omega_1,\omega_2\right)\in\Omega_1\times\Omega_2: \ H \textit{ is possible under } \Gamma\left(\omega_1,\omega_2\right)\right\}. \end{aligned} \tag{17}$$

Theorem 1 — except (1) and (2) — clearly applies to u' and v'; (1) is replaced by $v'(\emptyset)=\emptyset$ and (2) by $u'(\Theta)=\Omega_1\times\Omega_2$. $u'(\emptyset)$ represents the set of contradictory interpretation pairs. Such a pair can never be the correct one because contradictions are not possible. Therefore contradictory interpretations must be eliminated and the probability must be conditioned on the event that there is no contradiction. Provided that $u'(\emptyset)$ is measurable, i.e. $u'(\emptyset)\in\mathcal{A}_1\otimes\mathcal{A}_2$ and $P'(u'(\emptyset))<1$, the new combined hint $\mathcal{H}_1\oplus\mathcal{H}_2=(\Omega,\mathcal{A},P,\Gamma,\Theta)$ can be formed, where

$$\begin{aligned} \Omega&=u'\left(\emptyset\right)^c=v'\left(\Theta\right),\\ \mathcal{A}&=u'\left(\emptyset\right)^c\cap\mathcal{A}_1\otimes\mathcal{A}_2,\\ P\left(A\right)&=P'\left(A\right)/P'\left(u'\left(\emptyset\right)^c\right) \end{aligned}$$

and Γ is defined by (16) (and restricted to Ω). This way to combine hints is called **Dempster's rule** (A. Dempster (1967)).

Let u and v be defined by (1) relative to the hint $\mathcal{H}_1\oplus\mathcal{H}_2$. Then $u(H)=u'(H)\cap\Omega=u'(H)-u'(\emptyset)$ and $v(H)=v'(H)$.

Dempster's rule may be extended even to the case where $u'(\emptyset)$ is not measurable. In this case the conditional probability space $(\Omega,\mathcal{A},P)$ can be considered, where $(\Omega,\mathcal{A})$ is defined as above and $P(A)=P'^{*}(A\cap u'(\emptyset)^c/P'^{*}(u'(\emptyset)^c)$, provided that $P'^{*}(u'(\emptyset)^c)>0$. This leads to the combined hint $\mathcal{H}_1\oplus\mathcal{H}_2=(\Omega,\mathcal{A},P,\Gamma,\Theta)$.

As before, we have $u(H)=u'(H)\cap\Omega=u'(H)-u'(\emptyset)$ and $v(H)=v'(H)$. Let $\mathcal{E}_s$ and $\mathcal{E}_p$ be the classes of s- and p-measurable sets relative to the hint $\mathcal{H}_1\oplus\mathcal{H}_2$. Denote by $\mathcal{E}'_s$ and $\mathcal{E}'_p$ the classes of sets H such that $u'(H)$ and $v'(H)$ are measurable with respect to $\mathcal{A}_1\otimes\mathcal{A}_2$. From $u'(H)\in\mathcal{A}_1\otimes\mathcal{A}_2$ it follows that $u(H)\in\Omega\cap\mathcal{A}_1\otimes\mathcal{A}_2$ and thus $\mathcal{E}'_s\subseteq\mathcal{E}_s$. Similarly, because $v'(H)\subseteq\Omega, v'(H)\in\mathcal{A}_1\otimes\mathcal{A}_2$ implies $v(H)\in\Omega\cap\mathcal{A}_1\otimes\mathcal{A}_2$ or $\mathcal{E}'_p\subseteq\mathcal{E}_p$. If $u'(\emptyset)$ is measurable, then $\mathcal{E}'_s=\mathcal{E}_s$ and $\mathcal{E}'_p=\mathcal{E}_p$.

The next theorem states that inclusion of hints is maintained under Dempster's rule:

Theorem 12. *Let $\mathcal{H}_1,\mathcal{H}_2,\mathcal{H}''_1,\mathcal{H}''_2$ be four hints such that $\mathcal{H}''_1\subseteq\mathcal{H}_1$ and $\mathcal{H}''_2\subseteq\mathcal{H}_2$. Then $\mathcal{H}''_1\oplus\mathcal{H}''_2\subseteq\mathcal{H}_1\oplus\mathcal{H}_2$.*

Proof. $\Gamma''_1(\omega_1)\subseteq\Gamma_1(\omega_1)$ and $\Gamma''_2(\omega_2)\subseteq\Gamma_2(\omega_2)$ imply $\Gamma''_1(\omega_1)\cap\Gamma''_2(\omega_2)\subseteq\Gamma_1(\omega_1)\cap\Gamma_2(\omega_2)$. This, together with $\Omega''_1\subseteq\Omega_1$ and $\Omega''_2\subseteq\Omega_2$ implies $\Omega''\subseteq\Omega$. Also $\mathcal{A}''_1\supseteq\mathcal{A}_1\cap\Omega''_1$ and $\mathcal{A}''_2\supseteq\mathcal{A}_2\cap\Omega''_2$ imply that

$$\begin{aligned}\mathcal{A}'' &= \mathcal{A}_1'' \otimes \mathcal{A}_2'' \cap \Omega'' \supseteq (\mathcal{A}_1 \cap \Omega_1'') \otimes (\mathcal{A}_2 \cap \Omega_2'') \cap \Omega'' \\ &= (\mathcal{A}_1 \otimes \mathcal{A}_2) \cap (\Omega_1'' \times \Omega_2'') \cap \Omega'' \\ &= \mathcal{A}_1 \otimes \mathcal{A}_2 \cap \Omega'' = (\mathcal{A}_1 \otimes \mathcal{A}_2 \cap \Omega) \cap \Omega'' = \mathcal{A} \cap \Omega''.\end{aligned}$$

It remains to show that

$$P''(A) = P^*(A)/P^*(\Omega'')$$

for any $A \in \mathcal{A} \cap \Omega''$. Let Q'', Q denote the product measures of P_1'' and P_2'' and P_1 and P_2 on the product spaces $(\Omega_1'' \times \Omega_2'',\ \mathcal{A}_1'' \otimes \mathcal{A}_2'')$ and $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ respectively. Then by definition $P''(A) = Q''^*(A)/Q''^*(\Omega'')$ for any $A \in \mathcal{A} \cap \Omega''$. It is thus sufficient to show that $Q''^*(A) = k \cdot P^*(A)$ for some constant k independent of A.

To begin with, let's suppose that the sets $\Omega_1'', \Omega_2'', \Omega''$ and Ω are measurable with respect to $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}$ and $\mathcal{A}_1 \otimes \mathcal{A}_2$ respectively. Then $P''(A) = Q''(A)/Q''(\Omega'')$ and $P^*(A)/P^*(\Omega'') = P(A)/P(\Omega'')$ for $A \in \mathcal{A} \cap \Omega''$ and we must prove that $Q^*(A) = k \cdot P(A)$. Let X_A denote the indicator function of A. Then

$$Q''(A) = \int P_1''(d\omega_1)\, P_2''(d\omega_2)\, X_A.$$

Because X_A is a measurable function with respect to $\mathcal{A}$, it is sufficient to take the restrictions of the probability measures P_1'' and P_2'' to $\mathcal{A}_1$ and $\mathcal{A}_2$. But there these probabilities are conditional probabilities such that

$$\begin{aligned}Q''(A) &= \int P_1(d\omega_1)\, P_2(d\omega_2)\, X_A / P_1(\Omega_1'')\, P_2(\Omega_2'') \\ &= Q(A)/P_1(\Omega_1'')\, P_2(\Omega_2'') = P(A)\,(Q(\Omega)/P_1(\Omega_1'')\, P_2(\Omega_2'')).\end{aligned}$$

This proves the theorem in the case of measurable sets $\Omega_1'', \Omega_2'', \Omega''$ and Ω. If Ω is not measurable, then there exists a measurable set $\bar{\Omega}$, containing Ω, such that $Q^*(\Omega) = Q(\bar{\Omega})$. If $A \in \mathcal{A} \cap \Omega$, then $\bar{A} = A \cap \bar{\Omega}$ is measurable, contains A, and $Q^*(A) = Q(\bar{A})$.

Thus $P(\bar{A}) = P(A)$ for all $A \in \mathcal{A} \cap \Omega$ and Ω may be replaced by $\bar{\Omega}$ and $\mathcal{A} \cap \Omega$ by $\mathcal{A} \cap \bar{\Omega}$ without changing the relevant probability values. In this way the case where some or all sets $\Omega_1'', \Omega_2'', \Omega''$ and Ω are not measurable can be reduced to the former case. This proves the theorem.

Q.E.D.

In the case of theorem 12, the constant k appearing in theorem 6 equals $P^*(\Omega'')$, where Ω'' contains all combined interpretations (ω_1, ω_2) which are not contradictory under $\mathcal{H}_1'' \oplus \mathcal{H}_2''$. Some combined interpretations, which are not contradictory under $\mathcal{H}_1 \oplus \mathcal{H}_2$ may however be contradictory under $\mathcal{H}_1'' \oplus \mathcal{H}_2''$. This accounts for the possible difference between Ω and Ω''. If the situation is such that $\Omega'' = \Omega$, then $k = 1$ and $[sp_e''(H),\ pl_e''(H)] \subseteq [sp_e(H),\ pl_e(H)]$.

Let $\mathcal{V}$ be the vacuous hint associated with $\mathcal{H}_2$. Then theorem 12 implies that $\mathcal{H}_1 \oplus \mathcal{H}_2 \subseteq \mathcal{H}_1 \oplus \mathcal{V}$. Similarily $\mathcal{H}_1 \oplus \mathcal{H}_2 \subseteq \mathcal{V} \oplus \mathcal{H}_2$. As the combination of a hint with a vacuous hint does not add new information to the hint, this result shows that a combined hint $\mathcal{H}_1 \oplus \mathcal{H}_2$ is always finer than each of the two hints $\mathcal{H}_1$ and $\mathcal{H}_2$ alone. And in particular, if *sp* is the support function of $\mathcal{H}_1 \oplus \mathcal{H}_2$, then we have $[sp_e(H), pl_e(H)] \subseteq [sp_{1e}(H), pl_{1e}(H)]$ and $[sp_e(H), pl_e(H)] \subseteq [sp_{2e}(H), pl_{2e}(H)]$, if $\mathcal{H}_1$ and $\mathcal{H}_2$ have no contradictory interpretations.

Bibliography

1. Choquet G. (1953): Theory of capacities. *Ann. Inst. Fourier (Grenoble)* **5**, 131–295.
2. Choquet G. (1969): *Lectures on analysis.* Benjamin, New York.
3. Dempster A.P. (1967): Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, **38**, 325–339.
4. Dubois D., and Prade H. (1986): A Set-Theoretic View of Belief Functions. Logical Operations and Approximations by Fuzzy Sets. *Int. J. General Systems*, **12**, 193–226
5. Fagin R., and Halpern J.Y. (1989): Uncertainty, Belief, and Probability. *IJCAI-89*, 1161–1167.
6. Goodman I.R., and Nguyen H.T. (1985): *Uncertainty Models for Knowledge-Based Systems.* North Holland, New York.
7. Kohlas J. (1990): A Mathematical Theory of Hints. *Working Paper, Institute for Automation and Operations Research*, University of Fribourg (Switzerland), No. 173.
8. Neveu J. (1964): *Bases mathématiques du calcul des probabilités.* Masson, Paris.
9. Nguyen H.T. (1978): On Random Sets and Belief Functions. *J. Math. Anal. Appl.*, **65**, 531–542.
10. Ruspini E.H. (1987): Epistemic Logics, Probability, and the Calculus of Evidence. *IJCAI-87*, 924–931.
11. Shafer G. (1976): *A mathematical theory of evidence.* Princeton University Press.
12. Shafer G. (1978): Dempster's rule of combination. *Unpublished Manuscript.* The University of Kansas, School of Business, 202 Summerfield, Lawrence, Kansas 66045.
13. Shafer G. (1979): Allocations of probability. *The Annals of Probability*, **7**, 827–839.
14. Shafer G. (1990): Perspectives on the Theory and Practice of Belief Functions. *Int. J. Approx. Reas.*, **4**, 323–362.
15. Strat T. (1984): Continuous Belief Functions for Evidential Reasoning. *AAAI-84*, 308–313.
16. Yager R.R. (1985): The entailment principle for Dempster-Shafer granule. *Tech. Report MII-512*, Iona college, New Rochelle, N.Y.

27

Combining the Results of Several Neural Network Classifiers

Galina Rogova

Abstract. Neural networks and traditional classifiers work well for optical character recognition; however, it is advantageous to combine the results of several algorithms to improve classification accuracies. This paper presents a combination method based on the Dempster–Shafer theory of evidence, which uses statistical information about the relative classification strengths of several classifiers. Numerous experiments show the effectiveness of this approach. The method allows 15–30% reduction of misclassification error compared to the best individual classifier.

Key words: Classifier, Neural network, Character recognition, The Dempster–Shafer theory of evidence, Evidence

1 Introduction

Pattern recognition problems, such as classification of machine or handprinted characters, are currently solved with acceptable accuracy by using traditional classifiers or neural networks of different architectures and based on different sets of features. We may suppose that many of them tend to make recognition errors of different types; that is, they may be regarded as error independent. It is easier in many cases to apply several error-independent classifiers to the same recognition task and use their "error independence" to improve recognition performance of a combined system instead of inventing a new architecture or a feature extractor to achieve the same accuracy.

Recently, various combination techniques were proposed by different authors, where majority voting scheme, neural net, Bayesian, and the Dempster–Shafer theories were employed (Mandler and Schurmann, 1988; Xu, Krzyzak, & Suen, 1991, 1992). It appears that using the Dempster–Shafer theory of evidence is very productive, but the result depends considerably on a function that is used as a basic probability assignment. Xu, Krzyzak, and Suen (1992) applied the Dempster–Shafer theory of evidence to combine so-called "syntactic classifiers" that produce only a class label as output. They used the recognition rate and the substitution rate of each individual classifier to calculate basic probability assignments. On the other hand, neural networks as well as a number of traditional classifiers generate an output vector that can supply additional information on a "measurement level." For this type of classifier, posterior class-conditional probabilities can be calculated, providing a natural basic probability assignment. The calculation of posterior probabilities, however, demands numerous approximations that pose very difficult problems, especially in the situation when the number of classes is large.

In this paper, we present two new sets of support functions for the calculation of evidences. They permit us to obtain a considerable improvement of classification accuracy without complex computations.

2 The Dempster–Shafer Theory of Evidence

The Dempster–Shafer theory of evidence is a tool for representing and combining measures of evidences. This theory is a generalization of Bayesian reasoning and it is more flexible than Bayesian when our knowledge is incomplete, and we have to deal with uncertainty and ignorance. We introduce its basic concepts in this section, following Barnett (1981) and Shafer (1976).

Let $\varTheta$ be a set of mutually exhaustive and exclusive atomic hypotheses, $\varTheta = \{\theta_1, \ldots, \theta_K\}$. $\varTheta$ is called the *frame of discernment*. Let $2^{\varTheta}$ denote the set of all subsets of $\varTheta$. A function $\mathbf{m}$ is called *basic probability assignment* if:

$$\mathbf{m} : 2^{\varTheta} \to [0,1] \quad \mathbf{m}(\varnothing) = 0, \quad \text{and} \quad \sum_{A \subseteq \varTheta} \mathbf{m}(A) = 1. \tag{1}$$

Whereas the probability theory assigns a measure of *probability* to atomic hypotheses $\theta_i, \mathbf{m}(A)$ represents *belief* in a not necessarily atomic hypothesis A. For $A \neq \theta_i, \mathbf{m}(A)$ reflects our ignorance because it is a belief we cannot further subdivide among the subsets of A. $\mathbf{m}(A)$ is a measure of support we are willing to assign to a composite hypothesis A at the expense of support $\mathbf{m}(\theta_i)$ of atomic hypotheses θ_i. If for the frame of discernment $\varTheta$ we set $\mathbf{m}(\theta_i) \neq 0$ for all θ_i and $\mathbf{m}(A) = 0$ for all $A \neq \theta_i$, we find ourselves in the situation of probability theory with $\Sigma_i\ \mathbf{m}(\theta_i) = 1$ and $m(\theta_i)$ that may be regarded as a probability of θ_i.

Because $\mathbf{m}(A) + \mathbf{m}(\neg A) \leq 1$, the amount of belief committed, neither to A nor to compliment of A is the degree of ignorance. Therefore, the Dempster–Shafer theory of evidence allows us to represent only our actual knowledge "without being forced to overcommit when we are ignorant."

If $\mathbf{m}$ is a basic probability assignment, then a function $\mathbf{Bel} : 2^{\varTheta} \to [0,1]$ satisfying:

$$\mathrm{Bel}(B) = \sum_{A \subseteq B} \mathbf{m}(A) \tag{2}$$

is called *a belief function.* We can consider a basic probability assignment as a generalization of a probability density function whereas a belief function is a generalization of a probability function.

There is one-to-one correspondence between the belief function and the basic probability assignment. If A is an atomic hypothesis, $\mathbf{Bel}(A) = \mathbf{m}(A)$.

If $\mathbf{m}_1$ and $\mathbf{m}_2$ are basic probability assignments on Θ, their combination or *orthogonal sum*, $\mathbf{m} = \mathbf{m}_1 \oplus \mathbf{m}_2$, is defined as:

$$\mathbf{m}\left(A\right) = C^{-1} \sum_{D \cap B = A} \mathbf{m_1}\left(B\right) \cdot \mathbf{m}_2\left(D\right), \tag{3}$$

where

$$C = \sum_{D \cap B \neq \varnothing} \mathbf{m}_1\left(B\right) \cdot \mathbf{m}_2\left(D\right), \mathbf{m}\left(\varnothing\right) = 0, \text{and } A \neq \varnothing. \tag{4}$$

Obviously, the combination rule may be generalized to combine multiple evidence.

Because there is one-to-one correspondence between **Bel** and **m**, the orthogonal sum of belief functions $\mathbf{Bel} = \mathbf{Bel}_1 \oplus \mathbf{Bel}_2$ is defined in the obvious way.

Special kinds of **Bel** functions are very good at representing evidence. These functions are called *simple* and *separable support* functions. **Bel** is a *simple support* function if there exists an $F \subseteq \Theta$ called *focus* of **Bel**, such that $\mathbf{Bel}(\Theta) = 1$ and

$$\mathbf{Bel}\left(A\right) = \begin{cases} \mathrm{s} \neq 0 & \text{if} \quad F \subseteq A \text{ and } A \neq \Theta \\ 0 & \text{if} \quad \mathrm{F} \nsubseteq \mathrm{A} \end{cases}, \tag{5}$$

where $\mathbf{s}$ is called **Bel**'s *degree of support.*

A *separable support function* is either a simple support function or an orthogonal sum of simple support functions. Separable support functions are very useful when we want to combine evidences from several sources. If **Bel** is a simple support function with focus $F \neq \Theta$, then $\mathbf{m}(F) = \mathbf{s}, \mathbf{m}(\Theta) = 1 - \mathbf{s}$, and $\mathbf{m}$ is 0 elsewhere.

Let F be a focus for two simple support functions with degrees of support $\mathbf{s}_1$ and $\mathbf{s}_2$, respectively. If $\mathbf{Bel} = \mathbf{Bel}_1 \oplus \mathbf{Bel}_2$ then $\mathbf{m}(F) = 1 - (1 - \mathbf{s}_1)(1 - \mathbf{s}_2), \mathbf{m}(\Theta) = (1 - \mathbf{s}_1)(1 - \mathbf{s}_2)$, and $\mathbf{m}$ is 0 elsewhere.

3 Classification

Assume that we have an unlabeled input vector $\bar{\mathbf{x}}$. Let N be the number of different classifiers $f^n, n = 1, \ldots, N$. Also assume that each classifier produces an output vector $\bar{\mathbf{y}}^n \in \mathbf{R}^K, \bar{\mathbf{y}}^n = f^n(\bar{\mathbf{x}})$. Here K is the number of classes (in case of character recognition, $K = 10$ for digit classifiers and $K = 36$ for alphanumeric classifiers). For an individual classifier we assign class j to the input vector $\bar{\mathbf{x}}$ if $y_j = \max_{1 \le k \le K} y_k$. This decision rule does not give us a chance to say what the measure of confidence of our classification result is. For example, the output vectors $\bar{\mathbf{y}}_1 = (0, 0, 0, 1)$ and $\bar{\mathbf{y}}_2 = (0.2, 0.2, 0.2, 0.202)$ both yield the same class assignment, class 4, but the quality of this decision may be considered quite poor for $\bar{\mathbf{y}}_2$. So, if we want to combine N classifiers, this decision rule permits us to use the majority voting scheme only, which cannot take into account the quality of each vote. Suppose that for each classifier f^n and each candidate class k, we calculated the value $e_k(\bar{\mathbf{y}}^n) = e_k(f^n(\bar{\mathbf{x}}))$, which represents some measure of evidence for the proposition "$\bar{\mathbf{y}}^n$ is of class k." If we introduced these values in terms of the Dempster–Shafer theory we could combine these evidences according to this theory and choose the class with the highest evidence.

4 Existing Methods for Computation of Evidences

To calculate evidence for a neural network output, we might consider a posterior probability of each class, given an output vector, as a basic probability assignment. To estimate class-conditional probability distribution for all K classes, we have to produce multidimensional distribution for output vector given each class. For this purpose, K histograms for K-dimensional output vectors of the training set have to be built. In these histograms, the bin size should be small enough to yield sufficient precision. If the histogram has m bins in each coordinate, the total number of bins, m^K, is too large even for a training set of substantial size. Such a histogram in practice cannot be regarded as a realization of a continuous probability density function without rather arbitrary simplifications and approximations.

Mandler and Schurmann (1988) used a combination method based on the Dempster–Shafer theory for nearest neighbor classifiers with different distance measures. Statistical analysis of distances between learning data and a number of reference points in the input space was carried out to estimate distributions of intra- and interclass distances. These distributions were used to calculate class-conditional probabilities that were transformed into evidences and combined.

Attempts to apply a similar approach to neural network outputs brought forward questions about a choice of reference vectors and a distance measure. In addition, we would prefer to avoid approximations associated with estimation of parameters of statistical models for intra- and interclass distances. We shall present our method in the next section.

5 Proposed Method

Now we introduce a different method for calculation of evidences. We would like these values to reflect the classification abilities of each classifier. It appears that sets of outputs $\{f^n(\bar{\mathbf{x}})\}, n = 1, \ldots N$ computed for the training data can provide relevant information and we shall use it in our computations.

Let $\{\bar{\mathbf{x}}_k\}$ be a subset of the training data corresponding to a class k. Let $\bar{\mathbf{E}}_k^n$ be the mean vector for a set $\{f^n(\bar{\mathbf{x}}_k)\}$ for each classifier f^n and each class k. $\bar{\mathbf{E}}_k^n$ is a reference vector for each class k and $d_k^n = \phi(\bar{\mathbf{E}}_k^n, \bar{\mathbf{y}}^n)$ is a proximity measure for $\bar{\mathbf{E}}_k^n$ and $\bar{\mathbf{y}}^n$. We want the values of this function to vary between 1 and 0 with the maximum when output vector coincides with a reference vector. We shall discuss a specific form for the function ϕ later. Now we need to transform these proximity measures into evidences $e_k(\bar{\mathbf{y}}^n)$.

Consider a frame of discernment $\Theta = \{\theta_1, \ldots, \theta_K\}$, where θ_k is the hypothesis that "$\bar{\mathbf{y}}^n$ is of class k." For any classifier f^n and each class k, a proximity measure d_k^n can represent evidence *pro*-hypothesis θ_k, and all d_i^n, with $i \neq k$, can represent evidences *pro* $\neg\theta_k$ or *contra* θ_k. We can use d_k^n as a degree of support for a simple support function with focus θ_k. This yields the basic probability assignment

$$\mathbf{m}_k\left(\theta_k\right) = d_k^n \text{ and } \mathbf{m_k}\left(\Theta\right) = 1 - d_k^n. \tag{6}$$

In a similar manner, d_i^n are degrees of support for simple support functions with a common focus $\neg\theta_k$, if $i \neq k$. The combination of these simple support function with focus $\neg\theta_k$ is a separable support function with the degree of support $1 - \Pi_{i\neq k}(1 - d_k^n)$. The corresponding basic probability assignment is

$$\begin{aligned}\mathbf{m}_{\neg k}\left(\neg\theta_k\right) =& 1 - \prod_{i\neq k}\left(1 - d_k^n\right) \text{ and} \\ &\mathbf{m}_{\neg k}\left(\Theta\right) = 1 - \mathbf{m}_{\neg k}\left(\neg\theta_k\right) = \prod_{i\neq k}\left(1 - d_k^n\right).\end{aligned} \tag{7}$$

Combining our knowledge about θ_k we obtain the evidence $e_k(\bar{\mathbf{y}}^n) = \mathbf{m}_k \oplus \mathbf{m}_{\neg k}$ *pro* θ_k for class k and classifier n:

$$e_k\left(\bar{y}^n\right) = \frac{d_k^n \prod_{i\neq k}\left(1 - d_i^n\right)}{1 - d_k^n\left[1 - \prod_{i\neq k}\left(1 - d_i^n\right)\right]}. \tag{8}$$

Finally, evidences for all classifiers may be combined according to the Dempster–Shafer rule to obtain a measure of confidence for each class k for the input vector $\bar{\mathbf{x}}$: $e_k(\bar{\mathbf{x}}) = e_k(\bar{\mathbf{y}}^1) \oplus \ldots \oplus e_k(\bar{\mathbf{y}}^N)$. $e_k(\bar{\mathbf{y}}^n)$, after an appropriate normalization, can be considered as Bayesian evidence function with nonzero basic probability assignments only on atomic hypotheses. Hence $e_k(\bar{\mathbf{x}}) = C\Pi_n e_k(\bar{\mathbf{y}}^n)$, where C is the normalizing constant. Now we assign class j to the input vector $\bar{\mathbf{x}}$ if $e_j = \max_{1\leq k\leq K} e_k(\bar{\mathbf{x}})$.

The major problem now is to find the most effective form of the function ϕ. Several candidate functions for a proximity measure d_k^n for $\bar{\mathbf{E}}_k^n$ and $\bar{\mathbf{y}}^n$ were considered:

$$
\begin{aligned}
d_k^n &= 1 - ||\bar{E}_k^n - \bar{y}^n||;\\
d_k^n &= 1 - ||\bar{E}_k^n - \bar{y}^n||;\\
d_k^n &= \exp\left(-||\bar{E}_k^n - \bar{y}^n||^m\right);\\
d_k^n &= \frac{\left(1 + ||\bar{\mathbf{E}}_k^n - \bar{y}^n||^m\right)^{-1}}{\sum_{1\le i\le K}\left(1 + ||\bar{\mathbf{E}}_i^n - \bar{y}^n||^m\right)^{-1}};\\
d_k^n &= \frac{1}{1 + ||\bar{\mathbf{E}}_k^n - \bar{y}^n||^m};\, d_k^n = \cos^m\left(\alpha_k^n\right),
\end{aligned}
$$

where α_k^n is the angle between $\bar{\mathbf{E}}_k^n$ and $\bar{\mathbf{y}}^n$.

Of the functions tried, two were found to have the best performance on validation sets. One of them is $\cos^2(\alpha_k^n)$:

$$\phi_1\left(\bar{E}_k^n, \bar{y}^n\right) = \frac{\left(\sum_{1\le i\le k} E_{ik}^n y_i^n\right)^2}{||\bar{E}_k^n||^2||\bar{y}_k^n||^2} \tag{9}$$

The second one is a function based on the Euclidian distance between $\bar{\mathbf{E}}_k^n$ and $\bar{\mathbf{y}}^n$:

$$\phi_2\left(\bar{\mathbf{E}}_k^n, \bar{y}^n\right) = \frac{\left(1 + ||\bar{\mathbf{E}}_k^n - \bar{y}^n||^2\right)^{-1}}{\sum_{1\le i\le k}\left(1 + ||\bar{\mathbf{E}}_i^n - \bar{y}^n||^2\right)^{-1}}. \tag{10}$$

Our approach has a useful property of punishing overconfident, overtrained classifiers: their averages of output activations over the training set will be close to zero or 1, and this automatically means that both our proximity measures will be smaller for "fuzzier" activation vectors corresponding to the test data.

6 Experiments and Results

Our first experiment was conducted with handprinted digits from a private data base. The training set contained 25,000 characters and the test set contained 4000. Output activations of three classifiers were used. The first was a two-hidden-layer neural network trained by back propagation with local receptive fields (LRF) using direct bitmap input of 20 × 30 pixels (Pawlicki, 1991). The two other neural networks used as input a set of units corresponding to features extracted by projecting the original pixel input onto a basis of Gabor wavelets (Shustorovich, 1994). One of them was a two-hidden-layer neural network with 144 input units trained by back propagation with LRF, the other was a one-hidden-layer neural network with 113 input units trained by back propagation with global receptive fields (GRF). We refer to these classifiers as Bitmap-LRF, Gabor-LRF, and Gabor-GRF, respectively. The individual and combination results for digits are given in Tables 1 and 2. The proximity measures ϕ_1 and ϕ_2 are defined as in (9) and (10),

Table 1. Performance of Individual Classifiers for Digits

Classifier	Reject 0%	Reject 5%
Gabor-LRF	95.7%	97.9%
Bitmap-LRF	94.7%	98.0%
Gabor-GRF	93.4%	96.0%

respectively. The combination of these three classifiers was not any better than the combination of the best pair (Bitmap-LRF and Gabor-LRF), which means that the third one could not add anything new to the combination of the two.

Alphanumeric classifiers were trained using a data base contained 27,720 characters. The test set contained 12,960 characters. We used output activations of three classifiers: the above-mentioned Bitmap-LRF and two polynomial classifiers, namely, a polynomial classifier with simple quadratic features (SQF) and a polynomial classifier with "fuzzy" features (FF) (Anderson & Gaborski, 1993). In both cases, the polynomial classifier is a combination of the classical least square method and a neural network-type supervised training algorithm. Characters are converted nonlinearly to feature vectors using different quadratic polynomials of the pixels. We refer to these classifiers as Poly-SQF and Poly-FF, respectively. The individual and combination results for alphanumerics are shown in Tables 3 and 4. The proximity measures ϕ_1 and ϕ_2 are the same as those in Table 2. As we can see in the tables, the best combination of classifiers allowed 30% reduction in error rates for digits and 25% for alphanumerics compared to the best individual classifier. These results can be favorably compared with those of the majority voting scheme. When applied to the outputs of all three digits, it decreased misclassification error by 10% for corresponding testing set. The same result was obtained when the scheme was used for all three alphanumeric classifiers.

There is a very important question related to a problem of combination of several classifiers. Suppose we have a set of classifiers of different architectures and based on different sets of features. All these classifiers have different

Table 2. Performance of Combinations of Classifiers for Digits

	Proximity Measure ϕ_1		Proximity Measure ϕ_2	
Classifiers	Reject 0%	Reject 5%	Reject 0%	Reject 5%
Bitmap-LRF	96.4%	98.7%	97.0%	99.1%
Bitmap-LRF and Gabor-GRF	95.7%	98.2%	96.4%	98.5%
Gabor-LRF and Gabor-GRF	95.8%	98.2%	95.8%	98.5%

Table 3. Performance of Individual Classifiers for Alphanumeric Characters

Classifier	Reject 0%	Reject 5%
Poly-SQF	83.7%	85.9%
Bitmap-LRF	86.1%	88.3%
Poly-FF	86.3%	88.5%

recognition power. The question is, which subset of these classifiers is the most advantageous for the combination. Experiments with all our classifiers showed that a better result is not necessarily achieved on the combination of classifiers with better individual performance. In some cases it turns out that it is more important to combine more "independent" classifiers than those with better performance. For example, Table 2 shows that the combination of the results of Gabor-GRF and Bitmap-LRF is the same as the combination of the results of Gabor-GRF and Gabor-LRF in spite of the fact that the individual performance of the latter is better. Apparently, different feature extractors used during the preprocessing stage provide more independent results than different architectures of neural networks.

More experiments were conducted for the U.S. Census Bureau/NIST First OCR Systems Competition (1992). There were three categories of isolated hand-printed characters: digits, and lowercase and uppercase letters. Three-quarters of the NIST data base were used for training individual classifiers, and the last quarter was divided between a validation set and an internal test set. We entered the competition with combined algorithms in all three categories. For digits and lowercase letters, we integrated the results of Bitmap-LRF, Gabor-LRF, and Poly-FF classifiers. Gabor-LRF and Poly-FF were used for uppercase letters. The combination of the algorithms decreases misclassification error by 23% for digits, by 15% for uppercase, and by 25% for lowercase letters (on our designated test) compared to the best individual algorithm used in the combinations. The performance of the algorithms allowed Eastman Kodak Company to finish the competition among the tight group of leaders.

Table 4. Performance of Combinations of Classifiers for Alphanumeric Characters

	Proximity Measure ϕ_1		Proximity Measure ϕ_2	
Classifiers	Reject 0%	Reject 5%	Reject 0%	Reject 5%
Poly-SQF and Poly-FF	87.4%	89.7%	87.4%	89.5%
Poly-SQF and Bitmap-LRF	88.9%	91.3%	88.8%	90.7%
Poly-FF and Bitmap-LRF	89.7%	92.1%	89.6%	90.4%
Poly-SQF and Poly-FF and Bitmap-LRF	90.1%	92.4%	89.5%	91.4%

Acknowledgments

The author expresses her gratitude to all members of the Recognition and Information Processing Group of Eastman Kodak, especially P. Anderson, L. Barski, T. Pawlicki, and A. Shustorovich for providing her with data produced by individual classifiers and for valuable discussions. This paper was prepared while the author was with Eastman Kodak Company.

Bibliography

1. Anderson, P.G., & Gaborski, R.S. (1993). The polynomial method augmented by supervised training for handprinted character recognition. *Proceedings of the International Conference on Neural Networks and Genetic Algorithms* (pp. 417–422). Innsbruck, Austria: Springer-Verlag.
2. Barnett, J.A. (1981). Computational methods for mathematical theory of evidence. *Proceedings of Seventh International Joint Conference on Artificial Intelligence* (pp. 868–875). Vancouver, BC.
3. Mandler, E.J., & Schurmann, J. (1988). Combining the classification results of independent classifiers based on the Dempster/Shafer theory of evidence. *Pattern Recognition and Artificial Intelligence*, **X**, 381–393.
4. Pawlicki, T.F. (1991). Personal communication.
5. Shafer, G. (1976). *A mathematical theory of evidence.* Princeton: MIT Press.
6. Shustorovich, A. (in press). A subspace projection approach to feature extraction: the two-dimensional Gabor transform for character recognition. *Neural Networks.*
7. U.S. Census Bureau/NIST First OCR Systems Conference, May 27–28, 1992, Gaithersburg, MD.
8. Xu, L., Krzyzak, A., & Suen, C.Y. (1991). Associative switch for combining multiple classifiers. *Proceedings of the International Joint Conference on Neural Networks* (pp. I-43–48). Seattle, WA: IEEE Press.
9. Xu, L., Krzyzak, A., & Suen, C.Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, **22**(3), 418–435.

28

The Transferable Belief Model*

Philippe Smets and Robert Kennes

Abstract. Smets, P. and R. Kennes, The transferable belief model, Artificial Intelligence 66 (1994) 191–234.

We describe the transferable belief model, a model for representing quantified beliefs based on *belief functions*. Beliefs can be held at two levels: (1) a credal level where beliefs are entertained and quantified by belief functions, (2) a pignistic level where beliefs can be used to make decisions and are quantified by probability functions. The relation between the belief function and the probability function when decisions must be made is derived and justified. Four paradigms are analyzed in order to compare Bayesian, upper and lower probability, and the transferable belief approaches.

Key words: Belief function; Dempster–Shafer theory; quantified beliefs

1 Introduction

The aim of this paper is to present the transferable belief model (TBM) i.e. our interpretation of the Dempster–Shafer model. The TBM is a model for representing the quantified beliefs held by an agent at a given time on a given frame of discernment. It concerns the same concepts as considered by the Bayesian model, except it does not rely on probabilistic quantification, but on a more general system based on belief functions.

Since Shafer introduced his model based on belief functions in his book [33], many interpretations of it have been proposed. Three main interpretations have been developed: the random set, the generalized Bayesian, and the upper and lower probability interpretations. However, great confusion and

* This work has been partially funded by the CEC-ESPRIT II Basic Research Project 3085 (DRUMS), by the CEC-ESPRIT II Project 2256 (ARCHON) and the Belgian National Incentive-Program for Fundamental Research in Artificial Intelligence.

even blatant errors pervade the literature about the meaning and applicability of these models [29, 47]. We personally develop a model for point-wise quantified beliefs—the transferable belief model—and show how belief functions can be used for such a quantification. Bayesian probability is the most classical model for quantified beliefs. So our presentation focuses on comparing the TBM with its real contender: the Bayesian model. In particular we will discuss the problem of decision making within the TBM because it is necessary to explain how the model is used in real situations where decisions must be made, and because it is central to any Bayesian presentation. We even argue that Dutch Books—a betting strategy that would lead to a sure loss—cannot be raised against TBM users. In fact when decisions must be made, we require that beliefs be quantified by probability functions in order to avoid Dutch Books.

Several paradigms are analyzed in order to provide some insight into the nature of the TBM. These paradigms are used to contrast the TBM solution with the Bayesian, upper and lower probabilities, likelihood and fiducial solutions. The TBM is compared with random sets in [46], with possibility functions in [41], and with upper and lower probabilities in [15, 38]. The major differences between these models can be found in the way updating/conditioning must be performed. Axiomatic justifications are not given here but are developed in [49, 53].

We also argue in this paper that the TBM should not be considered as just a generalized probability model: indeed there are no *necessary* links between the TBM and any underlying probability model. Hence we dissociate ourselves from Dempster's model where some underlying probability is essential. Any decisions as to the nature of Shafer's model are left to Shafer himself (see [34]), but in our opinion, the TBM is very close to what Shafer described in his book [33]. In later work, Shafer creates confusion by speaking about random sets and upper and lower probabilities interpretations. Recently Shafer [34] clarified his position, rejected these interpretations and defended essentially the Dempster interpretation based on the random codes (a one-to-many mapping with an underlying probability distribution). We depart from this interpretation in that we do not *require* any underlying probability distribution, even though they *may* exist.

Not all interpretations of Dempster-Shafer theory are analysed here (see [43]). We do not discuss the interpretation of a belief as being the probability of a modal proposition [31] or the probability of provability [28].

Transferable Belief Model

The transferable belief model is based on:

- a two-level model: there is a *credal level* where beliefs are entertained and a *pignistic level* where beliefs are used to make decisions (from *pignus* = a bet in Latin [50]);

- at the credal level beliefs are quantified by belief functions;
- the credal level precedes the pignistic level in that, at any time, beliefs are entertained (and updated) at the credal level; the pignistic level appears only when a decision needs to be made;
- when a decision must be made, beliefs at the credal level induce a probability measure at the pignistic level, i.e. there is a *pignistic transformation* from belief functions to probability functions.

Bayesians do not consider an autonomous credal level. The introduction of a two-level model would be useless if decisions were the same as those derived within the Bayesian model. We will show in the "Mr. Jones" paradigm (Sect. 4) that this is not the case. The introduction of a credal level therefore is not merely an academic subtlety.

The TBM essentially fits with the model developed in Shafer's book [33] except for some differences and explanations such as:

- the complete dissociation from any *necessary* underlying probability model that precedes the construction of the belief functions at the credal level, as encountered in Dempster's approach (we do not mean the pignistic probabilities used at the pignistic level and that are derived from the belief functions);
- the fundamental concept of transferable "parts of belief";
- the two-level model and the pignistic transformation;
- the "open-world" and "closed-world" assumptions and the introduction of the unnormalized belief functions [39];
- the precedence of the conditioning process over the combination process;
- the justification of Dempster's rule of combination as the unique compositional rule to combine two belief functions [19, 20, 27, 40].

The TBM is Unrelated to a Probability Model

The TBM is intended to model subjective, personal beliefs, i.e. what the Bayesians claim to be their domain of application. The major point of the TBM is its complete dissociation from any model based on probability functions. This contrasts with what has been done in some of Shafer's more recent publications that favor the random set interpretation [26, 35], and most publications on Dempster–Shafer's model [2, 22]. The TBM is neither a random sets model [46] nor a generalization of the Bayesian model nor of some upper and lower probability (ULP) models [15]. It is another model whose aim is to quantify someone's degree of belief. The model is normative, supposedly simulating the behavior of a reasonable and consistent agent, the "stat rat" of Barnard (see discussion in [50, p. 26]).

To support our case that the TBM is different from the Bayesian model, we present an example, the "Mr. Jones" case, that leads to different results according to which model is used to analyse it. Such an example might provide a tool for discriminating between the two models: according to which result fits your requirements, you can select the model.

Other examples have already been provided to show the difference between the Bayesian model and the TBM. But their power of persuasion as a discriminating tool is weak as the TBM answer can usually also be derived from a Bayesian analysis. The interest of the "Mr. Jones" example is that the TBM solution can only be obtained by a Bayesian analysis by introducing some unpalatable assumptions.

Summary of the Content

The TBM is presented in Sect. 2. We then present a theory for decision making (Sect. 3). That decisions are based on probability functions (and expected utilities) is not disputed. Whenever a decision has to be made by an agent, he/she constructs a probability distribution derived from the belief function that describes his/her credal state. Bear in mind that the existence of such a probability distribution when decisions are made does not imply that this probability function quantifies our belief at the credal level (i.e. outside of any decision context).

We show (1) the impact of a *betting frame* on bet, (2) how someone's betting behavior could be used to assess a belief function, (3) how conditioning acts on the betting behavior, and (4) how Dutch Books are avoided.

We then proceed by analyzing several paradigms in detail. We know from experience that these paradigms are very useful in appreciating the particularity of the TBM, especially when compared with other approaches. Each paradigm enables the difference to be shown between the TBM and some of its contenders.

In Sect. 4 we present the "Mr. Jones" example, a very pointed example that shows the difference between the TBM approach and the Bayesian approach.

In Sect. 5, we present the "guards and posts" paradigm. It clarifies the nature of the conditioning process in the TBM.

In Sects. 6 and 7, we present two other paradigms to illustrate situations where the TBM leads to results different from those of its contenders: the Bayesian model, the ULP model, the likelihood model, and fiducial model. Some of these comparisons have also been attempted by Hunter [17] and Laskey [23]. Other comparisons are presented in [43].

In Sect. 8, we discuss the origin of the basic belief assignment used in our paradigms.

In Sect. 9, we show the difference between the credal level where someone's beliefs are quantified by belief functions and the pignistic level where "pignistic" probabilities must be constructed. Revision of beliefs is performed at the credal level by Dempster's rule of conditioning, not at the pignistic level by probability conditioning.

In Sect. 10, we conclude by answering some potential criticisms of the TBM.

2 The Transferable Belief Model

2.1 The Model

The necessary background information on belief functions is summarized hereafter. A full description can be found in Shafer's book [33]. A somehow revised version appears in [39]. Further results on Bayes' theorem and the disjunctive rule of combination appear in [37, 48].

Let L be a finite *propositional language*, and $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ be the set of *worlds* that correspond to the interpretations of L. *Propositions* identify subsets of Ω. Let $\top$ be the tautology and $\bot$ be the contradiction. For any proposition X, let $[[X]] \subseteq \Omega$ be the set of worlds identified by X. Let A be a subset of Ω, then f_A is any proposition that identifies A. So A is $[[f_A]], \emptyset = [[\bot]]$, and $\Omega = [[\top]]$. By definition there is an *actual world* ϖ and it is an element of Ω. In L, two propositions A and B are *logically equivalent*, denoted $A \equiv B$, iff $[[A]] = [[B]]$.

Let Π be a partition of Ω. Given the elements of the partition Π, we build $\mathcal{R}$, the *Boolean algebra* of the subsets of Ω generated by Π. We call Ω the *frame of discernment* (the frame for short). The elements of the partition Π are called the *atoms* of $\mathcal{R}$. Given $\mathcal{R}$, the number of atoms in a set $A \in \mathcal{R}$ is the number of atoms of $\mathcal{R}$ that are included in A. We call the pair $(\Omega, \mathcal{R})$ a *propositional space*.

By abuse of language but for the sake of simplicity, we do not distinguish between the subsets of Ω and the propositions that denote them. We use the same notation for both of them. So the same symbol (like A, B, $C, \ldots$) is used for a subset of Ω and for any proposition that denotes that subset. The standard Boolean notation is used. Let $A, B \in \mathcal{R}$. $\bar{A}$ stands for the complement of A relative to Ω. $A \cup B$ and $A \cap B$ denote the set-theoretic union and intersection respectively of the (subsets denoted by the) propositions A and B. $A \subseteq B$ means that all the elements of A (the subset denoted by A) are elements of B (the subset denoted by B) (or equivalently, that proposition A implies the proposition B). Any algebra $\mathcal{R}$ defined on Ω contains two special propositions: $\top$ and $\bot$ denoted by their corresponding sets Ω and $\emptyset$.

All beliefs entertained by You[1] at time t about which world is the actual world ϖ are defined relative to a given *evidential corpus* ($\mathrm{EC}_t^{\mathrm{Y}}$) i.e., the set of pieces of evidence in Your mind at time t. Our approach is normative: You is an ideal rational agent and Your evidential corpus is deductively closed. The *credal state* on a propositional space $(\Omega, \mathcal{R})$ describes Your subjective, personal judgment that $\varpi \in A$ for every proposition $A \in \mathcal{R}$. By a classical abuse of language, the actual world ϖ is called the "true" world, and we say that "A is true" or "truth is in A" to mean that $\varpi \in A$. Your credal state results from $\mathrm{EC}_t^{\mathrm{Y}}$ that induces in You some partial beliefs on the propositions of $\mathcal{R}$ (note that we did not say Ω). These partial beliefs quantify the strength

[1] "You" is the agent that entertains the beliefs considered in this presentation.

of Your belief that $\varpi \in A, \forall A \in \mathcal{R}$. It is an epistemic construct as it is relative to Your knowledge included in Your evidential corpus $\mathrm{EC}_t^{\mathrm{Y}}$.

Basic Assumption. The TBM postulates that the impact of a piece of evidence on an agent is translated by an allocation of parts of an initial unitary amount of belief among the propositions of $\mathcal{R}$. For $A \in \mathcal{R}$, $m(A)$ is a part of the agent's belief that supports A, i.e. that the actual world ϖ is in A, and that, due to lack of information, does not support any strict *sub*proposition of A.

The $m(A)$-values, $A \in \mathcal{R}$, are called the *basic belief masses* and the m-function is called the *basic belief assignment.*[2]

Let $m : \mathcal{R} \rightarrow [0, 1]$ with

$$\sum_{A \in \mathcal{R}} m(A) = 1, \qquad m(\emptyset) = 0.$$

Every $A \in \mathcal{R}$ such that $m(A) > 0$ is called a focal proposition. The difference with probability models is that masses can be given to any proposition of $\mathcal{R}$ instead of only to the atoms of $\mathcal{R}$.

As an example, let us consider a somehow reliable witness in a murder case who testifies to You that the killer is a male. Let $\alpha = 0.7$ be the reliability You give to the testimony. Suppose that *a priori* You have an equal belief that the killer is a male or a female. A classical probability analysis would compute the probability $P(M)$ of M where $M =$"the killer is a male". $P(M) = 0.85 = 0.7 + 0.5 \times 0.3$ (the probability that the witness is reliable (0.7) plus the probability of M given the witness is not reliable (0.5) weighted by the probability that the witness is not reliable (0.3)). The TBM analysis will give a belief 0.7 to M. The 0.7 can be viewed as the *justified* component of the probability given to M (called the belief or the support) whereas the 0.15 can be viewed as the *aleatory* component of that probability. The TBM deals only with the justified components. (Note: the Evidentiary Value Model [9, 11] describes the same belief component, but within a strict probability framework, and differs thus from the TBM once conditioning is introduced.)

If some further evidence becomes available to You and implies that B is true, then the mass $m(A)$ initially allocated to A is transferred to $A \cap B$. Hence the name TBM.

Continuing with the murder case, suppose there are only two potential male suspects: Phil and Tom. Then You learn that Phil is not the killer. The testimony now supports that the killer is Tom. The reliability 0.7 You gave to the testimony initially supported "the killer is Phil or Tom". The new information about Phil implies that 0.7 now supports "the killer is Tom".

The transfer of belief described in the TBM satisfies the so-called *Dempster rule of conditioning.* Let m be a basic belief assignment on the propositional space $(\Omega, \mathcal{R})$ and suppose the conditioning evidence tells You that the truth

[2] Shafer speaks about basic probability masses and assignment. To avoid confusion, we have banned the word "probability" whenever possible.

is in $B \in \mathcal{R}$, then the basic belief assignment m is transformed into $m_B : \mathcal{R} \to [0,1]$, where

$$m_B(A) = \begin{cases} c \sum_{X \subseteq \bar{B}} m(A \cup X) & \text{for } A \subseteq B, \\ 0 & \text{for } A \not\subseteq B, \\ 0 & \text{for } A = \emptyset, \end{cases} \tag{1}$$

where

$$c = \frac{1}{1 - \sum_{X \subseteq \bar{B}} m(X)}.$$

In this presentation we have assumed that one and only one element of Ω is true (closed-world assumption). In [39] we generalized the model and accepted that none of the elements could be true (open-world assumption). In that last case, positive basic belief masses could be given to $\emptyset$ and the normalization coefficient c in (1) is 1. The closed-world assumption is accepted hereafter. The meaning of the basic belief mass given to $\emptyset$ is analyzed in [45].

Degrees of Belief and Plausibility

Given $(\Omega, \mathcal{R})$, the degree of belief of *A, bel(A)*, quantifies the total amount of *justified specific support* given to A. It is obtained by summing all the basic belief masses given to proposition $X \in \mathcal{R}$ with $X \subseteq A$ (and $X \neq \emptyset$)

$$Bel(A) = \sum_{\emptyset \neq X \subseteq A} m(X).$$

We say *justified* because we include in *Bel(A) only* the basic belief masses given to subsets of A. For instance, consider two distinct atoms x and y of $\mathcal{R}$. The basic belief mass $m(\{x, y\})$ given to {x, y} could support x if further information indicates this. However given the available information the basic belief mass can only be given to {x, y}. (Note that under open-world assumption, $m(\emptyset)$ might be positive. The basic belief mass $m(\emptyset)$ should not be included in *Bel(A)* nor in *pl(A)*, as it is given to the subset $\emptyset$ that supports not only A but also $\bar{A}$. This is the origin of the *specific* support.)

The function $Bel : \mathcal{R} \to [0,1]$ is called a belief function. The triple $(\Omega, \mathcal{R}, Bel)$ is called a credibility space. Belief functions satisfy the following inequalities [33]:

$$\begin{aligned} &\forall_n \geq 1, A_1, A_2, \ldots, A_n \in \mathcal{R}, \\ &\quad Bel(A_1 \cup A_2 \cup \ldots \cup A_n) \\ &\qquad \geq \sum_i Bel(A_i) - \sum_{i>j} Bel(A_i \cap A_j) \\ &\qquad\quad - \cdots - (-1)^n Bel(A_1 \cap A_2 \cap \cdots \cap A_n). \end{aligned} \tag{2}$$

The *degree of plausibility* of A, *pl(A)*, quantifies the maximum amount of *potential specific support* that could be given to $A \in \mathcal{R}$. It is obtained by adding all those basic belief masses given to propositions X compatible with A, i.e. such that $X \cap A \neq \emptyset$:

$$pl\left(A\right) = \sum_{X \cap A \neq \emptyset} m\left(X\right) = Bel\left(\Omega\right) - Bel\left(\bar{A}\right).$$

We say *potential* because the basic belief masses included in *pl(A)* could be transferred to non-empty subsets of A if new information could justify such a transfer. It would be the case if we learn that $\bar{A}$ is impossible.

The function *pl* is called a plausibility function. It is in one-to-one correspondence with belief functions. It is just another way of presenting the same information and could be forgotten, except inasmuch as it provides a convenient alternate representation of our beliefs.

Dempster's rule of conditioning expressed with *Bel* and *Pl* is:

$$Bel\left(A|B\right) = \frac{Bel\left(A \cup \bar{B}\right) - Bel\left(\bar{B}\right)}{1 - Bel\left(\bar{B}\right)}, \qquad pl\left(A|B\right) = \frac{pl\left(A \cap B\right)}{pl\left(B\right)}.$$

Doxastical Equivalence

Besides the logical equivalence already mentioned, there is another concept of equivalence related to Your evidential corpus $\mathrm{EC}_t^{\mathrm{Y}}$. Let $[[\mathrm{EC}_t^{\mathrm{Y}}]]$ represent the set of worlds in Ω where all propositions deduced on Ω from $\mathrm{EC}_t^{\mathrm{Y}}$ are true. All the worlds in Ω that are not in $[[\mathrm{EC}_t^{\mathrm{Y}}]]$ are accepted as "impossible" by You at time t. Two propositions A and B are said to be *doxastically equivalent* for You at time t, denoted $A \cong B$, iff $[[\mathrm{EC}_t^{\mathrm{Y}}]] \cap [[A]] = [[\mathrm{EC}_t^{\mathrm{Y}}]] \cap [[B]]$. Logical equivalence implies doxastic equivalence. This is important as it implies that A and B should get the same support, the same degree of belief. Hence the Consistency Axiom that expresses the equi-credibility of doxastically equivalent propositions.

Consistency Axiom. *Let us consider two credibility spaces* $(\Omega, \mathcal{R}_i, Bel_i)$, $i = 1, 2$, *that represent Your beliefs on two algebras* $\mathcal{R}_1$ *and* $\mathcal{R}_2$ *as induced by Your* $\mathrm{EC}_t^{\mathrm{Y}}$. *Let* $A_1 \in \mathcal{R}_1$ *and* $A_2 \in \mathcal{R}_2$. *If* $A_1 \cong A_2$, *then* $Bel_1(A_1) = Bel_2(A_2)$.

This consistency axiom means that doxastically equivalent propositions share the same degree of belief, which is required since they share the same truth value. It also means that the belief given to a proposition does not depend on the structure of the algebra to which the proposition belongs. This consistency axiom is usually postulated for probability distributions, when they quantify degrees of belief [21]. Here it is postulated only for those functions that quantify beliefs at the credal level.

Total Ignorance

Total ignorance is represented by a vacuous belief function, i.e. a belief function such that $m(\Omega) = 1$, hence $Bel(A) = 0 \; \forall \; A \in \mathcal{R}, \; A \neq \Omega$, and $Bel(\Omega) =$

1. The origin of this particular quantification for representing a state of total ignorance can be justified. Suppose that there are three propositions labeled A, B, and C, and You are in a state of total ignorance about which is true. You only know that one and only one of them is true but even their content is unknown to You. You only know their number and their label. Then You have no reason to believe any one more than any other; hence, Your beliefs about their truth are equal: $Bel(A) = Bel(B) = Bel(C) = \alpha$ for some $\alpha \in [0, 1]$. Furthermore, You have no reason to put more or less belief in $A \cup B$ than in C : $Bel(A \cup B) = Bel(C) = \alpha$ (and similarly $Bel(A \cup C) = Bel(B \cup C) = \alpha$). The vacuous belief function is the only belief function that satisfies equalities like $Bel(A \cup B) = Bel(A) = Bel(B) = \alpha$. Indeed the inequalities (2) are such that $Bel(A \cup B) \geq bel(A) + Bel(B) - Bel(A \cap B)$. As $A \cap B = \emptyset$, $Bel(A \cap B) = 0$. The inequality becomes $\alpha \geq 2\alpha$ where $\alpha \in [0, 1]$ hence $\alpha = 0$.

Remark 1. The TBM also includes a description of *Dempster's rule of combination* - a rule for the conjunctive combination of two belief functions that somehow generalizes Dempster's rule of conditioning. This rule is not used in this presentation. An axiomatic justification of Dempster's rule of combination within the TBM is given in [14, 19, 20, 40]. There also exists a disjunctive rule of combination of two belief functions, which is described in [48].

Remark 2. It is important to note that the TBM includes *two components*: one *static*, the basic belief assignment, and one *dynamic*, the transfer process. Many authors on the Dempster–Shafer model consider only the basic belief assignment and discover that the basic belief masses are probabilities on the power set of Ω. But usually they do not study the dynamic component, i.e. how beliefs are updated. Their comparisons are therefore incomplete, if not misleading.

2.2 Refinements and Consistent Beliefs

Let us consider two propositional languages L_1 and L_2. It is always possible to build a common underlying propositional language L such that each proposition of L_1 (and of L_2) is a proposition of L. Let Ω_1, Ω_2, and Ω be the sets of worlds that correspond to the interpretations of L_1, L_2, and L, respectively. Each world of Ω_1 (and of Ω_2) corresponds to a set of worlds of Ω, and the images of the worlds of Ω_1 (and of Ω_2) constitute a partition of Ω. Hence, whenever we describe two propositional spaces, we can always use a common underlying Ω without loss of generality. In fact the concept in a propositional space $(\Omega, \mathcal{R})$ that is important for this presentation is the algebra $\mathcal{R}$, not the set of worlds Ω. All beliefs are built on the algebras $\mathcal{R}$, not on Ω. The granularity of Ω is irrelevant once Ω is fine enough to allow for a definition of the atoms of $\mathcal{R}$ (i.e., each atom of $\mathcal{R}$ contains at least one element of Ω). Therefore, the definition of two propositional spaces $(\Omega_i, \mathcal{R}_i)$, $i = 1, 2$, with different sets Ω_i is equivalent to a definition of two propositional spaces

$(\Omega,\ \mathcal{R}_i),\ i=1,2$, sharing the same Ω. From now on, we will not worry about the Ω, they will be adapted such that each $\mathcal{R}$ is non-ambiguously defined.

Consider two propositional spaces $(\Omega,\ \mathcal{R}_1)$ and $(\Omega,\ \mathcal{R})$ (see the left half of Fig. 1, where the elements of Ω are four individuals characterized by name and age, the atoms of $\mathcal{R}_1$ are male and female and $\mathcal{R}$ is the power set of Ω). Let Λ_1 be a one-to-many mapping from $\mathcal{R}_1$ to $\mathcal{R}$ such that each atom of $\mathcal{R}_1$ is mapped on a proposition of $\mathcal{R}$, the images of the atoms of $\mathcal{R}_1$ constitute a partition of Ω, and this mapping is additive. Λ_1 is called a *refining* from $\mathcal{R}_1$ to $\mathcal{R}$. $\mathcal{R}$ is called a *refinement* of $\mathcal{R}_1$. $\mathcal{R}_1$ is called a *coarsening* of $\mathcal{R}$ (see [33, p. 115]). For $B \in \mathcal{R}$, let

$$\bar{\Lambda}_1^{-1}(B) = \cup\{A : A \in \mathcal{R}_1, \Lambda_1(A) \cap B \neq \emptyset\}.$$

Let us consider two propositional spaces $(\Omega,\ \mathcal{R}_i),\ i=1,2$, and two refinings Λ_i to a common refinement $\mathcal{R}$. By construction, if $B \in \mathcal{R}$ is true (the actual world $\varpi \in B$), then $\bar{\Lambda}_1^{-1}(B)$ and $\bar{\Lambda}_2^{-1}(B)$ are true. The two credibility spaces $(\Omega,\ \mathcal{R}_i,\ Bel_i),\ i=1,2$, are said to be *consistent* if there exists a belief function Bel on $\mathcal{R}$ such that $Bel_i(\bar{\Lambda}_i^{-1}(B)) = Bel(B)$ for all $B \in \mathcal{R},\ i=1,2$.

2.3 Least Committed Belief on $\mathcal{R}_2$ Induced by a Belief on $\mathcal{R}_1$

Let the credibility space $(\Omega,\ \mathcal{R}_1,\ Bel_1)$ be induced by Your $\mathrm{EC}_t^{\mathrm{Y}}$. Let Λ be a relation between $\mathcal{R}_1$ and a new algebra $\mathcal{R}_2$ defined on Ω such that $\Lambda(\omega) \neq \emptyset$ for every atom ω of $\mathcal{R}_1$ and $\Lambda(A \cup B) = \Lambda(A) \cup \Lambda(B)\ \forall\ A, B \in \mathcal{R}_1$. The question is to construct a belief function Bel_2 on $\mathcal{R}_2$, consistent with Bel_1, that conveys on $\mathcal{R}_2$ the same "information" as Bel_1 does on $\mathcal{R}_1$. Let $\mathcal{B}$ be the family of belief functions Bel on $\mathcal{R}_2$ consistent with Bel_1. By the Consistency

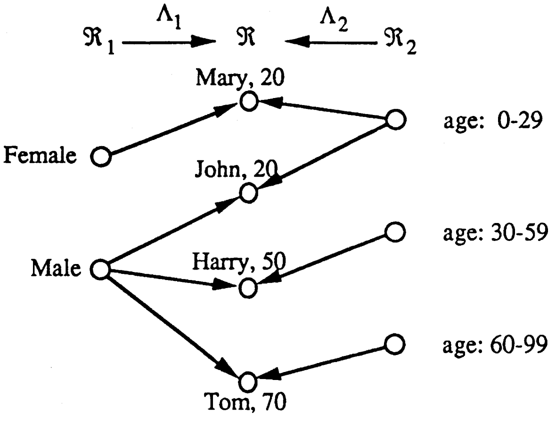


Fig. 1. Example of two propositional spaces $(\Omega,\ \mathcal{R}_i),\ i=1,2$. Λ_i are the refinings from $\mathcal{R}_i$ to $\mathcal{R}$. Each circle is an atom. The atoms of $\mathcal{R}$ are those of a refinement common to $\mathcal{R}_1$ and $\mathcal{R}_2$

Axiom, one has $Bel_1(A) = Bel(B)$, for every pair $(A,\ B)$ with $A \in \mathcal{R}_1$ and $B \in \mathcal{R}_2$ such that $A \cong B$. But this requirement does not provide the value of Bel for those $B \in \mathcal{R}_2$ that are not doxastically equivalent to some $A \in \mathcal{R}_1$. The *Principle of Least commitment* [16, 48] allows us to select the belief function $Bel^* \in \mathcal{B}$ such that $Bel^*(B) \le Bel(B)\ \forall\ B \in \mathcal{R}_2\ \forall\ Bel \in \mathcal{B}$. This principle says that You should not give more support to a proposition than justified by Bel_1. It implies that Bel_2 (and its basic belief mass m_2) is related to Bel_1 (and its basic belief mass m_1) by:

$$\forall\ B \in \mathcal{R}_2 \quad m_2\left(B\right) = \sum_{A \in \mathcal{R}_1 : \Lambda(A) = B} m_1\left(A\right),$$

where the sum is 0 when no A satisfies the constraint. bel_2 is called the *vacuous extension* of bel_1 on $\mathcal{R}_2$ [33].

3 The Pignistic Probability Derived from a Belief Function

3.1 The Generalized Insufficient Reason Principle

Let us give a context. Given the evidence available to You, the TBM claims the existence of a belief function that describes Your credal state on the frame of discernment.

Suppose a *decision* must be made based on this credal state. As is well known, decisions will be coherent if the underlying uncertainties can be described by a probability distribution defined on 2^{Ω} [5]. Therefore, one must find a rule that allows for the construction of a probability distribution from a belief function in the case of forced decision. We only consider forced bets, forced decisions, as is done classically by Bayesians. (The unforced decisions considered in [12, 18, 50] concern ULP contexts.) The solution that will satisfy behavior requirements introduced in Sect. 3.2 happens to be a generalization of the Insufficient Reason Principle. Another justification can be found in [42]. This solution already appeared in [7, 51] but with no justification.

Let us consider a credibility space $(\Omega,\ \mathcal{R},\ bel)$ that describes Your beliefs on $\mathcal{R}$. Let $A \in \mathcal{R}$ and $A = A_1 \cup A_2 \cup \cdots \cup A_n$, where the A_i's are distinct atoms of $\mathcal{R}$. The mass $m(A)$ corresponds to that part of Your belief that is restricted to A and that, due to lack of further information, cannot be allocated to a proper subset of A. In order to build the probability distribution needed for decision making (hence qualified as pignistic) on $\mathcal{R}$, You distribute $m(A)$ equally among the atoms of A. Therefore, $m(A)/n$ is given to each $A_i,\ i = 1, \ldots, n$. This procedure corresponds to the *Insufficient Reason Principle*: if You must build a probability distribution on n elements, given a lack of information, give a probability $1/n$ to each element. This procedure is repeated for each mass m. Let $BetP$ be the pignistic probability distribution so derived. For all atoms $x \in \mathcal{R}$,

$$BetP\,(x) = \sum_{x \subseteq A \in \mathcal{R}} \frac{m\,(A)}{|A|} = \sum_{A \in \mathcal{R}} m\,(A)\, \frac{|x \cap A|}{|A|},$$

where $|A|$ is the number of atoms of $\mathcal{R}$ in A, and for $B \in \mathcal{R}$,

$$BetP\,(B) = \sum_{A \in \mathcal{R}} m\,(A)\, \frac{|B \cap A|}{|A|}.$$

Of course, *BetP* is a probability function, but we call it a pignistic probability function to stress the fact that it is the probability function in a decision context. The principle underlying this procedure is called the Generalized Insufficient Reason Principle since the Insufficient Reason Principle has been used at the level of each focal proposition of the belief function. As described up to here it is only an *ad hoc* principle, but it can be justified by natural behavior requirements.

3.2 Derivation of the Generalized Insufficient Reason Principle

Let us give a credibility space $(\Omega,\ \mathcal{R},\ Bel)$. Let m be the basic belief assignment corresponding to *bel.* Let $BetP(\cdot;\ m)$ be the pignistic probability defined on $\mathcal{R}$. The parameter "m" is added in order to enhance the basic belief assignment from which *BetP* is derived.

Assumption 1 *$\forall x$ atom of $\mathcal{R}$, $BetP(x;\ m)$ depends only on $m(X)$ for $x \subseteq X \in \mathcal{R}$.*

Assumption 2 *BetP(x; m) is continuous (or bounded) for each $m(X)$, $x \subseteq X \in \mathcal{R}$.*

Assumption 3 *Let G be a permutation defined on Ω. For $X \subseteq \Omega$, let $G(X) = \{G(x) :\ x \in X\}$. Let $m' = G(m)$ be the basic belief assignment given to the propositions of Ω after the permutation has been performed, that is for $X \in \mathcal{R}$, $m'(G(X)) = m(X)$. Then for any atom x of $\mathcal{R}$, $BetP(x;\ m) = BetP(G(x);\ G(m))$.*

In other terms, *BetP* is invariant by permutations of Ω.

Assumption 4 *Let $(\Omega, \mathcal{R},\ Bel)$ be the credibility space that describes Your beliefs on $\mathcal{R}$, such that it is known by You that ϖ is not an element of the atom $X \in \mathcal{R}$ (so $\forall A \in \mathcal{R}, A \cong A \cup X$ and $Bel(A) = Bel(A \cup X)$ by the Consistency Axiom). Let us consider the credibility space $(\Omega', \mathcal{R}', Bel')$, where $\Omega' = \Omega - X, \mathcal{R}'$ is the boolean algebra built from the atoms of $\mathcal{R}$ that are not subset of X (so every element A of $\mathcal{R}'$ is also an element of $\mathcal{R}$, and $\forall A \in \mathcal{R}'$, $Bel'(A) = Bel(A)$ by the Consistency Axiom). Let $BetP(x;\ m)$ and $BetP'(x;\ m')$ be the pignistic probabilities derived from $Bel(m)$ and $Bel'(m')$, respectively. Then for every atom $x \in \mathcal{R}'$,*

$$BetP\,(x;\ m) = BetP'\,(x;\ m')\,,$$
$$BetP\,(X;\ m) = 0.$$

The major assumption A1 says that $BetP(x;\ m)$ may depend on $Bel(\bar{x})$ but not on the way the basic belief masses used to compute $Bel(\bar{x})$ are distributed among themselves.

The three other assumptions are classical requirements. Assumption A2 could be weakened as one only needs that, for each $m(X),\ x \in X, BetP(x;\ m)$ is continuous in a point, or bounded, or measurable, or majorizable by a measurable function on a set of positive measure (see [1, p. 142]).

Assumption A3 is the classical *Anonymity Requirement*: renaming the elements of Ω does not modify the probabilities. That $m'(G(X)) = m(X)$ results from the Consistency Axiom as $G(X) \cong X$.

Assumption A4 only states that impossible atoms do not change the pignistic probabilities.

Theorem 1. *Let $(\Omega, \mathcal{R})$ be a propositional space and m be a basic belief assignment on $\mathcal{R}$. Let $|A|$ be the number of atoms of $\mathcal{R}$ in A. Under assumptions A1 to A4, for any atom x of $\mathcal{R}$*

$$BetP\,(x; m) = \sum_{x \subseteq A \in \mathcal{R}} \frac{m\,(A)}{|A|}. \tag{3}$$

Proof. Given in Appendix A.

The transformation defined by (3) is called the pignistic transformation.

Corollary 1. *If Bel is a probability distribution P, then BetP is equal to P.*

Example 1. The same pignistic transformation was derived in [42] by assuming different requirements whose overall idea follows the next scenario. Let us consider two friends of Yours, G and J. You know they will toss a fair coin and the winner will visit You tonight. You want to buy the drink Your friend would like to receive tonight: coke, wine, or beer. You can only buy one drink. Let $D = \{\text{coke, wine, beer}\}$.

Let us suppose that $Bel_G(d)\ \forall\ d \subseteq D$ quantifies Your belief about the drink G will ask for, should G come. Given Bel_G, You build Your pignistic probability $BetP_G$ about the drink G will ask by applying the (still to be deduced) pignistic transformation. You identically build the pignistic probability $BetP_J$ based on Bel_J, Your belief about the drink J will ask, should J come. The two pignistic probability distributions $BetP_G$ and $BetP_J$ are in fact the conditional probability distributions about the drink that will be drunk given that G respectively J comes. The pignistic probability distributions $BetP_{GJ}$ about the drink that Your visitor will ask is then

$$BetP_{GJ}\,(d) = 0.5\ BetP_G\,(d) + 0.5\ BetP_J\,(d) \quad \forall\ d \subseteq D.$$

You will use these pignistic probabilities $BetP_{GJ}(d)$ to decide which drink to buy.

But You could as well reconsider the whole problem and compute first Your belief about the drink Your visitor (V) would like to receive. In [42], we show that such a belief is the average of Bel_G and Bel_J:

$$Bel_V\,(d) = 0.5\ Bel_G\,(d) + 0.5\ Bel_J\,(d) \quad \forall\ d \subseteq D.$$

Given Bel_V, You could then build the pignistic probability $BetP_V$ You should use to decide which drink to buy. We proved that if $BetP_V$ and $BetP_{GJ}$ must be equal, then the pignistic transformation must be the one given by the Generalized Insufficient Reason Principle (relation (3)).

3.3 Betting Frames

Let us consider a credibility space $(\Omega, \mathcal{R}_0,\ Bel_0)$. Before betting, one must define a betting frame $\mathcal{R}$ on Ω, i.e. the set of atoms on which stakes will be allocated. The *granularity* of this frame $\mathcal{R}$ is defined so that a stake could be given to each atom independently of the stakes given to the other atoms. For instance, if the stakes given to atoms A and B of $\mathcal{R}_0$ must necessarily be always equal, both A and B belong to the same granule of $\mathcal{R}$. The *betting frame* $\mathcal{R}$ is organized so that the granules are the atoms of $\mathcal{R}$, and $\mathcal{R}$ is the result obtained by applying a sequence of coarsenings and/or refinements on $\mathcal{R}_0$. Let us suppose the initial belief Bel_0 is defined on $\mathcal{R}_0$. Then the belief function Bel induced by Bel_0 on $\mathcal{R}$ is (see Sect. 2.3):

$$\forall\ A \in \mathcal{R}, \quad Bel\,(A) = Bel_0\left(\Lambda^{-1}\,(A)\right),$$

where Λ is the transformation from $\mathcal{R}_0$ to $\mathcal{R}$, and

$$\forall\ A \in \mathcal{R}, \quad \Lambda^{-1}\,(A) = \cup\,\{X : X \in \mathcal{R}_0,\ \Lambda\,(X) \subseteq A\}\,.$$

(See also [33, pp. 146–147].)

The pignistic probability $BetP$ is then built from the belief function Bel so derived on $\mathcal{R}$.

Betting under Total Ignorance

To show the power of the TBM approach, let us consider one of those disturbing examples based on total ignorance.

Let us consider two propositions denoted A_1 and A_2. You know that one and only one proposition is true. But You don't know what the two propositions are. You just know their number and their labels. You must bet on A_1 versus A_2. In the TBM, Your belief about the truth of A_1 and A_2 is described by a vacuous belief function and the pignistic probabilities on the betting frame $\{A_1,\ A_2\}$ are

$$BetP\,(A_1) - BetP\,(A_2) = \frac{1}{2}.$$

Let us now consider three propositions, denoted $B_1,\ B_2$, and B_3. You know that one and only one proposition is true. But You don't know what the three propositions are. You just know their number and their labels. You must bet on B_1 versus B_2 versus B_3. In the TBM, Your belief about the truth of $B_1,\ B_2$, and B_3 is described by a vacuous belief function and the pignistic probabilities on the betting frame $\{B_1,\ B_2,\ B_3\}$ are

$$BetP'\left(B_1\right) = BetP'\left(B_2\right) = BetP'\left(B_3\right) = \frac{1}{3}.$$

Now You learn that A_1 is logically (hence doxastically) equivalent to B_1 and A_2 is logically (doxastically) equivalent to $B_2 \cup B_3$. Within the TBM, this information will not modify Your beliefs and Your pignistic probabilities. If You were a Bayesian, You must adapt Your probabilities as they must give the same probabilities to A_1 and B_1. Which set of probabilities are You going to update, and why, especially since it must be remembered that You have no knowledge whatsoever about what the propositions are.

In a Bayesian approach, the problem raised by this type of example results from the requirement that doxastically equivalent propositions should receive identical beliefs, and therefore identical probabilities. Within the TBM, the only requirement is that doxastically equivalent propositions should receive equal beliefs (it is satisfied as $Bel(A_1) = Bel'(B_1) = 0$). Pignistic probabilities depend not only on these beliefs but also on the structure of the betting frame, hence $BetP(A_1) \neq BetP'(B_1)$ is acceptable as the two betting frames are different.

In a betting context, the set of alternatives and their degrees of refinement are relevant to the way Your bets are organized. Of course, if $BetP(A_1) = \frac{1}{2}$ had been a well-justified probability, then $BetP'(B_1)$ would also have been $\frac{1}{2}$. But here $BetP(A_1) = \frac{1}{2}$ is based only on the knowledge of the number of alternatives on which You can bet and *nothing else*. The difference between $BetP(A_1)$ and $BetP'(B_1)$ reflects the difference between the two betting contexts. Of course, as required, both A_1 and B_1 share the same degrees of belief.

3.4 Assessing Degrees of Belief

Given a propositional space $(\Omega, \mathcal{R}_0)$, the assessment of $Bel_0(A)\ \forall\ A \in \mathcal{R}_0$ can be obtained from the betting behavior established on other algebras $\mathcal{R}$ defined on Ω (or any refinement of Ω). Given such a betting frame $\mathcal{R}$ and its corresponding pignistic probability $BetP$ on $\mathcal{R}$, one can determine the set of belief functions S on 2^{Ω} that would lead to $BetP$ through (3) when the betting frame is $\mathcal{R}$. Construct various betting frames $\mathcal{R}_i$ on Ω and assess the corresponding $BetP_i$ and S_i. Note that the same evidence underlies all bets and that the difference between the $BetP_i$ results only from the difference between the structure of the betting frames $\mathcal{R}_i$. Let us consider a refining Λ

from $\mathcal{R}$ to $\mathcal{R}'$. Then, given the Consistency Axiom, the relation between *Bel* defined on $\mathcal{R}$ and Bel' defined on $\mathcal{R}' = 2^{\Omega'}$ is such that:

$$\begin{aligned} m'(\Lambda(A)) &= m(A) \quad &\forall\, A \in \mathcal{R}, \\ m'(B) &= 0, &\text{otherwise,} \end{aligned}$$

where $\Lambda(A) = \{\Lambda(x) : x \in A\}$. Bel' is the vacuous extension of *Bel* from $\mathcal{R}_0$ to $2^{\Omega'}$ [33, p. 146]. The strategy for defining various betting frames $\mathcal{R}_i$ allows for the construction of a family of S_i whose intersection contains only one element, Bel_0. An empty intersection would imply inconsistency between the pignistic probabilities. The number of potential betting frames is large enough to guarantee that a unique solution be obtained.

Example 2. Let us suppose that $\Omega_0 = \{a, b\}$ where $\{a\}$ = "John will come tonight" and $\{b\}$ = "John will not come tonight". Let us consider the betting frame $\mathcal{R}$ with atoms $\{a\}$ and $\{b\}$, and Your pignistic probabilities on frame $\mathcal{R}$:

$$BetP(\{a\}) = \frac{4}{9}, \qquad BetP(\{b\}) = \frac{5}{9}.$$

Suppose ψ and $\bar{\psi}$ are two complementary but otherwise unknown propositions. $\{a\} \cap \psi$ will occur if John comes tonight and proposition ψ is true. $\{a\} \cap \bar{\psi}$ will occur if John comes tonight and proposition $\bar{\psi}$ is true. Let us consider the betting frame $\mathcal{R}'$ with atoms $\{a\} \cap \psi$, $\{a\} \cap \bar{\psi}$, and $\{b\}$ and Your pignistic probabilities on it:

$$BetP'(\{a\} \cap \psi) = BetP'\left(\{a\} \cap \bar{\psi}\right) = \frac{7}{27}, \qquad BetP'(\{b\}) = \frac{13}{27}.$$

Then the unique solution for m_0 is:

$$m_0(\{a\}) = \frac{2}{9}, \qquad m_0(\{b\}) = \frac{3}{9}, \qquad m_0(\{a, b\}) = \frac{4}{9}.$$

It solves the two linear equations derived from (3):

$$\begin{aligned} \frac{4}{9} &= m_0(\{a\}) + \frac{1}{2} m_0(\{a, b\}), \\ \frac{7}{27} &= \frac{1}{2} m_0(\{a\}) + \frac{1}{3} m_0(\{a, b\}). \end{aligned}$$

It might seem odd that $\{b\}$ receives pignistic probabilities of $\frac{5}{9}$ and $\frac{13}{27}$ according to the betting context. It reflects the fact that a large amount $(\frac{4}{9})$ of Your initial belief was left unassigned (i.e. given to $\{a, b\}$). This example corresponds to a state in which You have very weak support for $\{a\}$ and for $\{b\}$. You are not totally ignorant as in Sect. 3.3.1, but still in a state of "strong" ignorance. Part of $BetP(\{b\}) = \frac{5}{9}$ is due to justified beliefs $(\frac{3}{9})$ but the remainder results from a completely unassigned part of belief that You distribute equally among the alternatives of Your betting frame.

Wilson [52] showed that the set of pignistic probabilities that can be obtained from a given belief function *Bel* on a frame $\mathcal{R}$ is equal to the set of probability functions "compatible" with *Bel* and its associated plausibility function *Pl*, i.e. the set of probability functions P on $\mathcal{R}$ such that $Bel(A) \leq P(A) \leq pl(A)\ \forall\ A \in \mathcal{R}$. So whatever the betting frame, $BetP(A) \geq Bel(A)\ \forall A \in \mathcal{R}$. Suppose You ignore what the appropriate betting frame is, You nevertheless know that, $\forall\ A \in \mathcal{R}$, the lowest bound of $BetP(A)$ is $Bel(A)$. Therefore $Bel(A)$ can then be understood as the lowest pignistic probability one could give to A when the betting frame is not fixed [12].

3.5 Conditional Betting Behaviors

Let us consider a credibility space $(\Omega, \mathcal{R}, Bel)$ and let us suppose You learn that proposition $A \in \mathcal{R}$ is true. Then *Bel* must be conditioned on A by Dempster's rule of conditioning and *BetP* is built from this conditional belief function.

But a distinction must be made between the following two cases:

- suppose You know that A is true, then You condition *Bel* on A before deriving *BetP*.
- Suppose You know that the bet will be cancelled if A is false, then You derive *BetP* from the unconditioned *Bel* and condition *BetP* on A using the classical probabilistic conditioning.

The first case corresponds to "factual" conditioning, the second to "hypothetical" conditioning. In the factual case, A is true for every bet that can be created. In the hypothetical case, A can be true in some bets, and false in others. Pignistic probabilities obtained in these two contexts usually reflect the difference between the contexts: (A always true; A possibly false). This distinction was already considered in 1931 by Ramsey who noted that:

> the degree of belief in P given $Q\ldots$ roughly expresses the odds at which the subject would now bet on P, the bet only to be valid if Q is true.... This is not the same as the degree to which he would believe P, if he believed Q for certain: for knowledge of Q might for psychological reasons profoundly alter his whole system of beliefs. [30, p. 79]

Ramsey distinguished between a bet on $P \cap Q$ versus $\bar{P} \cap Q$ in a context $\{P{\cap}Q, \bar{P}{\cap}Q, \bar{Q}\}$ and a bet on P versus $\bar{P}$ when Q is known to be true, hence in the context $\{P{\cap}Q, \bar{P}{\cap}Q\}$. In the TBM, Ramsey's allusion to "psychological reasons" applies at the credal level: learning Q modifies our credal state, hence of course our pignistic probabilities.

Note: Recent work by Dubois and Prade [8] has shown the difference between two forms of conditioning: focusing and updating (which might better be called "revision"). Our factual conditioning seems to correspond to their updating. Focusing is not considered here.

3.6 The Avoidance of Dutch Books

The pignistic probability we build on the betting frame from the underlying belief guarantees that no static Dutch Book can be constructed. To construct a Dutch Book, one implicitly defines the betting frame, i.e. the set of atoms of the Boolean algebra built from all the options. The pignistic probabilities are built using this betting frame and no Dutch Book can be constructed as far as the bets are established according to a probability measure.

In order to show how Dutch Books are avoided, we reconsider the two bets under total ignorance considered in Sect. 3.3.1. One could think of building the following Dutch Book.[3]

> Before knowing $A_1 \cong B_1$, You would accept to pay \$0.45 for winning \$1 if A_1 were true (as $BetP(A_1) = 0.5$). (For any bet, You would accept to pay up to $\$x$ with $x = BetP(X)$ if You won \$1 when X is true.) You would also accept to pay \$0.60 for winning \$1 if $B_2 \cup B_3$ were true (as $BetP'(B_2 \cup B_3) = 0.66$). Given that You don't know what A_1, A_2, B_1, B_2, and B_3 say, the two bets are acceptable together. Now You learn that $B_2 \cup B_3$ is true iff A_2 is true. Therefore, by accepting the two bets together, You commit Yourself to pay $\$(0.45 + 0.60) = \$1.05$ for the certainty of winning \$1. Hence a Dutch Book has been built against You, as You will surely loose \$0.05.

The argument is wrong because it does not take into due consideration the problem of the betting frame. Once $A_1 \cong B_1$ is known, You will not accept both bets simultaneously. Before accepting a bet, You must always build the betting frame, i.e. You must establish the granularity, i.e. the list of elements on which stakes can freely be allocated.

In the present case, once You know $A_1 \cong B_1$, You must decide if stakes on B_2 and B_3 will always be the same or might vary. If they must always be the same, then You use the betting frame $\{A_1, A_2\}$ and reject the second bet. If they might be different, then You use the betting frame $\{B_1, B_2, B_3\}$ and reject the first bet. Dutch Books are thus avoided.

The existence of two types of conditioning does not permit the construction of a dynamic Dutch Book. If bets are based on "hypothetical" facts, conditioning must then be performed by applying classical probability conditioning. If bets are based on "factual" facts, then *every* bet must be organized accordingly. Some atoms are definitively eliminated since they are impossible, conditioning is performed at the level of the belief function by applying Dempster's rule of conditioning, and $BetP$ is derived from the conditional belief function. Dutch Books can still be avoided as one cannot build a set

[3] This example was suggested by P. Garbolino in [3].

of bets where the "factual" fact is treated as unknown in some cases and accepted in others. A "factual" fact is either known or unknown, but once it is known, it is known for every bet. The difference between "hypothetical" facts and "factual" facts is to be found in the fact that "factual" facts are true for every bet, whereas hypothetical facts can be denied in some bets.

4 The Murder of Mr. Jones

4.1 The Problem

Big Boss has decided that Mr. Jones must be murdered by one of the three people present in his waiting room whose names are Peter, Paul, and Mary. Big Boss has decided that the killer on duty will be selected by a throw of a dice: if it is an even number, the killer will be female; if it is an odd number, the killer will be male. You, the judge, know that Mr. Jones has been murdered, who was in the waiting room and about the dice throwing, but You do not know what the outcome was and who was selected. *You are also ignorant as to how Big Boss would have decided between Peter and Paul in the case of an odd number being observed.* Given the available information, Your odds for betting on the sex of the killer would be 1 to 1 for male versus female.

You then learn that should Peter not be selected by Big Boss, he would necessarily have gone to the police station at the time of the killing in order to have a perfect alibi. Peter indeed went to the police station, so he is not the killer. The question is how You would bet now on male versus female: should Your odds be 1 to 1 (as in the TBM) or 1 to 2 (as in the Bayesian model).

Note that the alibi evidence makes "Peter is not the killer" and "Peter has a perfect alibi" equivalent. The more classical evidence "Peter has a perfect alibi" would only imply P("Peter is not the killer" | "Peter has a perfect alibi") = 1. But P("Peter has a perfect alibi" | "Peter is not the killer") would stay undefined and would then give rise to further discussion, which for our purpose would be useless. In this presentation, the latter probability is also 1.

4.2 The TBM Approach

Let k be the killer. The waiting room evidence E_0 and its resulting basic belief assignment m_0 are:

$$\begin{aligned} E_0:\ & k \in \Omega = \{\text{Peter},\ \text{Paul},\ \text{Mary}\}, \quad \mathcal{R}_0 = 2^{\Omega}, \\ & m_0(\{\text{Peter},\ \text{Paul},\ \text{Mary}\}) = 1. \end{aligned}$$

The dice-throwing pattern (evidence E_1) induces the following basic belief assignment:

$$E_1\text{: dice experiment}, \quad \mathcal{R}_1 = \{\text{Male},\ \text{Female}\},$$

$$m_1(\text{Female}) = 0.5,$$
$$m_1(\text{Male}) = 0.5.$$

Combining E_0 and E_1 induces the basic belief assignment m_{01}:

$$E_{01}\text{: } E_0 \text{ and } E_1, \quad \mathcal{R}_{01} = 2^{\Omega},$$
$$m_{01}(\{\text{Mary}\}) = 0.5,$$
$$m_{01}(\{\text{Peter, Paul}\}) = 0.5.$$

The 0.5 belief mass given to {Peter, Paul} corresponds to the part of belief that supports "Peter or Paul", could possibly support each of them, but given the lack of further information, cannot be divided more specifically between Peter and Paul.

Suppose You had to bet on the killer's sex. You would obviously bet on Male = {Peter, Paul} versus Female = {Mary} at odds 1 to 1.

Peter's alibi pattern (evidence E_2) induces the basic belief assignment m_2.

$$E_2\text{: } A = \text{"Peter went to the police station"}$$
$$= \text{"Peter has a perfect alibi"},$$
$$E_2\text{: } k \in \{\text{Paul, Mary}\}, \quad \mathcal{R}_2 = 2^{\Omega},$$
$$m_2(\{\text{Paul, Mary}\}) = 1.$$

Conditioning m_{01} on E_2 by Dempster's rule of conditioning leads to m_{012}:

$$E_{012}\text{: } E_{01} \text{ and } E_2, \quad \mathcal{R}_{012} = 2^{\Omega},$$
$$m_{012}(\{\text{Mary}\}) = m_{012}(\{\text{Paul}\}) = 0.5.$$

The basic belief mass that was given to "Peter or Paul" is transferred to Paul. Your odds for betting on Male versus Female would now still be 1 to 1, as before.

4.3 The Bayesian Solution

Suppose You were a Bayesian. Therefore Your degrees of belief are quantified by probability distributions and all pieces of evidence are taken in consideration by adequately revising Your probability distributions through the Bayesian conditioning processes.

Given E_1, You build a probability distribution P_1 on $\Omega = \{\text{Peter, Paul, Mary}\}$:

$$P_1\,(k \in \{\text{Mary}\}) = 0.5, \qquad P_1\,(k \in \{\text{Peter, Paul}\}) = 0.5.$$

You would also bet on male versus female the odds being 1 to 1.

When You learn E_2, i.e. that Peter went to the police station, You condition P_1 on {Paul, Mary} in order to compute P_{12}, where

$$P_{12}\,(k \in \{\text{Mary}\}) = P_1\,(k \in \{Mary\}\,|k \in \{\text{Mary, Paul}\})$$
$$\frac{P_1\,(k \in \{\text{Mary}\})}{P_1\,(k \in \{\text{Mary}\}) + P_1\,(k \in \{\text{Paul}\})} = \frac{0.5}{0.5 + x}$$

with $x = P_1(k \in \{\text{Paul}\})$. But x is unknown. No evidence whatsoever has been given about x.

Usually Bayesians encountering this problem will assume that $x = 0.25$ leading to a 1 to 2 odds. They obtain $x = 0.25$ by either applying the Insufficient Reason Principle or a symmetry argument or a minimum entropy argument on P_1 to evaluate $P_1(k \in \{\text{Paul}\})$. It is of course the most natural assumption…but it is still an assumption extraneous to the available evidence, and any other value in [0, 0.5] could as well be assumed. Any such value would correspond to some *a priori* probability on Peter versus Paul, which is not justified by any of the available pieces of evidence. All that is known to You is that there were two men whose names were Peter and Paul, and *nothing* else.

Another justification for $x = 0.25$ could be obtained as follows. Suppose evidence E_2': "if Paul were not the killer, he would go to the police station to have a perfect alibi and Paul went to the police station". E_2' is E_2 where Peter and Paul interchange their role. A bet on male versus female should be the same after evidence E_2 where Peter and Paul interchange their role. A bet on male versus female should be the same after evidence E_2 and after evidence E_2'. This symmetry requirement is satisfied only with $x = 0.25$. Therefore Bayesians can hardly avoid the 1 to 2 odds. In the TBM, the requirement that bets after E_2 and after E_2' should be the same is automatically satisfied: the 0.5 mass that was given by m_{01} to "Peter or Paul" is transferred to "Paul" under E_2 and to "Peter" under E_2' and the bets of male versus female remain unchanged.

Our analysis of Mr. Jones' case could be rephrased by saying that Big Boss used a deck of 52 cards instead of a dice. Mary is the killer if the card is red, the killer is male if the card is black. Peter is not the killer. In how many ways could Big Boss have selected a card so that Paul is the killer? The answer is not "any number between 0 and 26" as none of the cards had Paul written on them. All black cards are identical, they all mean "male". To introduce the idea that some black cards could point to Paul, the others to Peter, would lead to a ULP analysis as we would be in a context in which there is a probability that Paul is the killer (the proportion of cards marked Paul) but we do not know the value of such a proportion. This a another problem, different from the one we have analyzed. The two problems should not be confused. The difference between such ULP approach and the TBM is detailed in [8, 38, 43].

4.4 Conditional Bets

The example can be used to illustrate the difference between bets according to the context in which the conditioning information E_2 is taken into account (see Sect. 3.5). Before learning evidence E_2, if You want to bet on Paul versus Mary, the betting frame is {Paul, Mary, Peter} and $BetP(\text{Paul}) = BetP(\text{Peter}) = 0.25$, $BetP(\text{Mary}) = 0.5$. To bet on Paul versus Mary corresponds then to conditioning the pignistic probabilities $BetP$ on $\neg$Peter, hence the resulting pignistic probabilities, $BetP'(\text{Paul}) = \frac{1}{3}$ and $BetP'(\text{Mary}) = \frac{2}{3}$, and the 1 to 2 odds. After learning evidence E_2, the betting frame is {Paul, Mary} and You condition Bel_{01} on ¬Peter from which You derive Bel_{012}, the pignistic probabilities $BetP(\text{Paul}) = 0.5$ and $BetP(\text{Mary}) = 0.5$, and the 1 to 1 odds.

The difference results from Your openness to the fact that Peter might be the killer before learning E_2 and Your knowledge that Peter is not the killer after learning E_2.

5 The Guards and Posts Paradigm

We will present three paradigms: the guards and posts, the translators, and the unreliable sensor paradigms. The first paradigm helps to explain the conditioning process. The second paradigm shows that the TBM solution is fundamentally different from the ULP solution, but might lead to the mistaken idea that the TBM is somehow related to likelihood theory. The third paradigm shows that the TBM leads to a solution different from the Bayesian, the likelihood, and the fiducial solutions.

In each paradigm, the Bookean algebra $\mathcal{R}$ on which beliefs are defined is the power set of Ω, the frame of discernment.

The Paradigm[4]

Suppose a military officer must organize guard duty in his camp. There are three possible posts (π_1, π_2, and π_3) but only one is to be occupied. There are three soldiers who could be appointed for guard duty (S_1, S_2, and S_3). The officer will randomly select one of the three soldiers by tossing a dice. Soldier S_1 is selected if the dice outcome is 1 or 2, soldier S_2 is selected if the dice outcome is 3 or 4, and soldier S_3 is selected if the dice outcome is 5 or 6. Each soldier has a habit in that

- if selected, soldier S_1 will always go to post π_1 or π_2.
- if selected, soldier S_2 will always go to post π_1 or π_2 or π_3.
- if selected, soldier S_3 will always go to post π_2.

Before the officer selects the soldier, each of them writes down on a piece of paper where he will go if he is selected. As a result, there are six possible worlds w_i, $i = 1, \ldots, 6$, where each world corresponds to one particular selection of the posts (see left-hand column of Table 1). After the officer selects the guard, there are 18 worlds (referred to as worlds w_{ij} if soldier S_j is selected in world w_i). You are about to attack the camp and You want to know which post is occupied in order to avoid it. You know all the facts described up to here, but You do not know which soldier was selected. What is Your belief about which post is occupied? The frame of discernment $\Omega = \{\pi_1, \pi_2, \pi_3\}$ and $\mathcal{R} = 2^{\Omega}$. Table 2 presents the degrees of belief for each set of posts. Initially the basic

[4] This paradigm was suggested by Yen-Teh Hsia.

Table 1. The set of six worlds that represent the six possible ways posts could be selected by each soldier, and the post occupied in the eighteen possible worlds after the soldier has been selected by the officer

	Post selected by each soldier			Occupied post according to the selected soldier			Remaining worlds after Case 1 conditioning			Remaining worlds after Case 2 conditioning		
World	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3
w_1	π_1	π_1	π_2	π_1	π_1	π_2	π_1	π_1	π_2	π_1	π_1	
w_2	π_1	π_2	π_2	π_1	π_2	π_2				π_1		
w_3	π_1	π_3	π_2	π_1	π_3	π_2	π_1	π_3	π_2	π_1	π_3	
w_4	π_2	π_1	π_2	π_2	π_1	π_2					π_1	
w_5	π_2	π_2	π_2	π_2	π_2	π_2						
w_6	π_2	π_3	π_2	π_2	π_3	π_2					π_3	
Worlds w_i	1/3 probabilities of being selected	1/3	1/3	18 worlds w_{ij} where soldier S_j is selected in w_i								

belief assignment on $\mathcal{R}$ is such that $m(\{\pi_1,\ \pi_2\}) = m(\{\pi_1,\ \pi_2,\ \pi_3\}) = m(\{\pi_2\}) = \frac{1}{3}$.

Two cases of conditioning can then be considered.

Case 1. The soldiers and You learn that post π_2 is so unpleasant to occupy that the soldiers will not select it if they can go elsewhere (it applies thus to soldiers S_1 and S_2, but not S_3). Hence the worlds w_2, w_4, w_5, and w_6 become impossible (Table 1). Table 2 presents Your beliefs about which post is occupied. The 0.33 basic belief masses given initially to $\{\pi_2\}$, $\{\pi_1,\ \pi_2\}$, and $\{\pi_1,\ \pi_2,\ \pi_3\}$ are transferred to $\{\pi_2\}$, $\{\pi_1\}$, and $\{\pi_1,\ \pi_3\}$, respectively. Suppose You decide to bet on which post is occupied. The betting frame is $\{\pi_1,\ \pi_2,\ \pi_3\}$. The pignistic probability is $BetP(\pi_1) = \frac{3}{6}$, $BetP(\pi_2) = \frac{2}{6}$, and $BetP(\pi_3) = \frac{1}{6}$.

Table 2. Degrees of belief and their related basic belief masses for the paradigm of Table 1. Before conditioning and after Case 1 and 2 conditionings

	Initial state		Case 1		Case 2	
Posts	m	Bel	m	Bel	m	Bel
π_1	0	0	0.33	0.33	0.5	0.5
π_2	0.33	0.33	0.33	0.33	0	0
π_3	0	0	0	0	0	0
$\{\pi_1,\ \pi_2\}$	0.33	0.66	0	0.66	0	0.5
$\{\pi_1,\ \pi_3\}$	0	0	0.33	0.66	0.5	1.0
$\{\pi_2,\ \pi_3\}$	0	0.33	0	0.33	0	0
$\{\pi_1,\ \pi_2,\ \pi_3\}$	0.33	1.0	0	1.0	0	1.0

Case 2. You are able to observe post π_2 and it is empty. Hence, the selected soldier had not selected π_2 before being assigned guard duty. Thus the actual world is not one of the worlds $w_{2,2}$, $w_{4,1}$, $w_{5,1}$, $w_{5,2}$, $w_{6,1}$, $w_{1,3}$, $w_{2,3}$, $w_{3,3}$, $w_{4,3}, w_{5,3}$, and $w_{6,3}$. After renormalization, basic belief masses of 0.5 are given to $\{\pi_1\}$ and $\{\pi_1,\ \pi_3\}$. The betting frame is $\{\pi_1,\ \pi_3\}$. The pignistic probability is $BetP(\pi_1) = \frac{3}{4}$ and $BetP(\pi_3) = \frac{1}{4}$.

A probability solution could be derived if You accept that the 0.33 masses given to each soldier are somehow distributed among the six possible worlds. Suppose You accept an equal distribution. So each of the eighteen worlds receives a probability of $\frac{1}{18}$. Case 1 conditioning would leave a probability of $\frac{1}{6}$ to each of the six remaining worlds. The derived probabilities on $\{\pi_1,\ \pi_2,\ \pi_3\}$ are $P(\pi_1) = \frac{3}{6}$, $P(\pi_2) = \frac{2}{6}$, and $P(\pi_3) = \frac{1}{6}$, as in the TBM analysis. In Case 2 conditioning, the solutions differ: $P(\pi_1) = \frac{5}{7}$ and $P(\pi_3) = \frac{2}{7}$.

Even without postulating the equi-distribution of the 0.33 basic belief masses among the eighteen worlds w_{ij}, probabilists might be tempted to defend the idea that the probabilities 0.33 used for the soldier selection do not apply once the conditioning information of Case 2 is known. Indeed, they could defend that the fact that π_2 is not occupied somehow supports the hypothesis that soldier S_2 was selected. Hence, the updated probability P' should be such that $P'(S_2) > P'(S_1)$. This is the basis of Levi's criticisms (see Sect. 8). The answer from the TBM point of view is that no probability whatsoever is built on the w_{ij} space, only on the S_j space. So the fact that, in Case 2, there are fewer remaining possible worlds for S_1 than for S_2 (3 versus 4) is irrelevant. Case 2 really is the case in which the TBM acquires its originality when compared with the probability approach.

6 The Translator Paradigm

The Paradigm

Shafer and Tversky [36] have described a *translator experiment* to explain Shafer's theory. Let $T = \{t_i :\ i = 1, 2, \ldots, n\}$ be a set of translators and $\Omega = \{c_j :\ j = 1, 2, 3\}$ be a set of messages that can be generated by a given device. For each message $c_j \in \Omega$, the translator t_i translates it into an element of some given space Θ. Let $f_i(c_j)$ denote the element of Θ obtained by the translation performed by translator t_i of the message c_j. Table 3 presents an example where $\Omega = \{c_1, c_2, c_3\}$, $\Theta = \{\theta,\ \bar{\theta}\}$, and $T = \{t_0, \ldots, t_7\}$. The crosses in the last three columns indicate the elements of Ω that are translated into θ for each translator. So translator t_1 translates c_1 into θ and c_2 and c_3 into $\bar{\theta}$: $f_1(c_1) = \theta$, $f_1(c_2) = f_1(c_3) = \bar{\theta}$. Given $\theta \in \Theta$, let $A_i \subseteq \Omega$ be the set of messages that are translated as θ by translator t_i. In Table 3, $A_1 = \{c_1\}$, $A_4 = \{c_1, c_2\}, \ldots$. Note that it was not said that $A_i \cap A_j = \emptyset$ for $i \neq j$. Suppose that a message is selected in Ω (we do not say randomly selected, see Sect. 8). Suppose that a translator is selected *by a chance process*

Table 3. Translator paradigm with 8 translators t_i, 3 messages c_j, and 2 observations θ and $\bar{\theta}$. Columns 3–5 present the values of $f_i(c_j)$. The last three columns present the elements of Ω that are translated into θ for each translator

		$f_i(c_j)$			Given θ		
T	$P(t_i)$	c_1	c_2	c_3	c_1	c_2	c_3
t_0	p_0	$\bar{\theta}$	$\bar{\theta}$	$\bar{\theta}$	·	·	·
t_1	p_1	θ	$\bar{\theta}$	$\bar{\theta}$	×	·	·
t_2	p_2	$\bar{\theta}$	θ	$\bar{\theta}$	·	×	·
t_3	p_3	$\bar{\theta}$	$\bar{\theta}$	θ	·	·	×
t_4	p_4	θ	θ	$\bar{\theta}$	×	×	·
t_5	p_5	θ	$\bar{\theta}$	θ	×	·	×
t_6	p_6	$\bar{\theta}$	θ	θ	·	×	×
t_7	p_7	θ	θ	θ	×	×	×

among the set T of translators and independently of the selected message. Let p_i be the probability that translator t_i is selected. You observe only θ, the result of the transformation of the unknown message, but You ignore which translator was selected. Furthermore You are totally ignorant of how the message was selected. What can be said about Your beliefs $Bel(c)$ for $c \subseteq \Omega$ given that θ was observed?

TBM Analysis

With the TBM, the following basic belief masses are assumed on $T \times \Omega \times \Theta$:

$$m\left(\bigcup_{t \in \Theta} \bigcup_{j \in J_\tau} \{(t_i, \tau, c_j)\}\right) = p_i \quad \text{where } J_\tau = \{j : f_i\,(c_j) = \tau\}\,. \tag{4}$$

So p_2 is allocated to $\{(t_2, \bar{\theta}, c_1),\ (t_2, \theta, c_2),\ (t_2, \bar{\theta}, c_3)\}$, p_5 is allocated to $\{(t_5, \theta, c_1),\ (t_5, \bar{\theta}, c_2),\ (t_5, \theta, c_3)\}, \ldots$. The origin of such an assignment is to be found in Sect. 8.

Learning that θ is true, the transfer of the basic belief masses m by Dempster's rule of conditioning leads to the basic belief masses m^* on $T \times \Omega \times \{\theta\}$:

$$m^*\left(\bigcup_{j \in J_\theta} \{(t_i, \theta, c_j)\}\right) = p_i,$$

$$m'\left(\bigcup_{j \in J_\theta} \{(t_i, \theta, c_j)\}\right) = \frac{p_i}{1 - p_0} = p_i',$$

i.e. the weight p_i is given to the messages that are indicated by the crosses on the line t_i in the columns "Given θ" (Table 3, right part). The knowledge that there are only eight translators (closed-world assumption) justifies the normalization of the basic belief assignment m^* into m' by dividing m^* by $1 - p_0$.

By marginalization of m' (i.e. Bel') on Ω, one computes for $c \subseteq \Omega$

$$Bel_\theta\left(c\right) = Bel'\left(\bigcup_i \bigcup_{c_j \in c} \{(t_i, \theta, c_j)\}\right) = \sum_{i \in I} p'_i$$

with $I = \{i : \; f_i^{-1}(\theta) \subseteq c\}$ and $p'_0 = 0$.

For example: $m_\theta(\{c_1\}) = p'_1$ and $m_\theta(\{c_1, c_2\}) = p'_4$. Table 4 presents the values of $Bel_\theta(c)$ and $pl_\theta(c)$ for some $c \subseteq \Omega$.

Bayesian Analysis

For a Bayesian analysis, one claims the existence of some $P(c_j), j = 1, 2, 3,$ but their values are missing. One must compute:

$$P\left(c_j|\theta\right) = \frac{P\left(\theta|c_j\right)P\left(c_j\right)}{P\left(\theta\right)}.$$

One has:

$$P\left(\theta|c_j\right) = \sum_i P\left(\theta|t_i, c_j\right)P\left(t_i|c_j\right) = \sum_{i \in I} p_i$$
$$\text{where } I = \{i : f_i\left(c_j\right) = \theta\}$$

because

$$P\left(\theta|t_i, c_j\right) = \begin{cases} 0, & \text{if } f_i\left(c_j\right) = \bar{\theta}, \\ 1, & \text{if } f_i\left(c_j\right) = \theta, \end{cases}$$
$$P\left(t_i|c_j\right) = p_i$$

as the translators are selected independently of the message.

The ULPs for $P(c_j|\theta)$ are computed by evaluating its extremes where the $P(c_j)$ are allowed to vary on their domain. Let Δ be the set of all vectors (x, y, z) where $x, y, z \in [0, 1]$ and $x + y + z = 1$. The vectors

Table 4. Translator paradigm: TBM analysis

Ω	$Bel_\theta(c)$	$pl_\theta(c)$
$\{c_1\}$	p'_1	$p'_1 + p'_4 + p'_5 + p'_7$
$\{c_1, c_2\}$	$p'_1 + p'_2 + p'_4$	$p'_1 + p'_2 + p'_4 + p'_5 + p'_6 + p'_7$
$\{c_1, c_2, c_3\}$	$p'_1 + p'_2 + p'_3 + p'_4 + p'_5 + p'_6 + p'_7$	$p'_1 + p'_2 + p'_3 + p'_4 + p'_5 + p'_6 + p'_7$

$(P(c_1),\ P(c_2), P(c_3))$ are the elements of Δ. The upper and lower conditional probabilities P^* and P_* for c_j, given θ, are

$$P^*\left(c_j|\theta\right) = \sup\nolimits_\Delta P\left(c_j|\theta\right) = \sup\nolimits_\Delta \frac{P\left(\theta|c_j\right)P\left(c_j\right)}{\sum_\nu P\left(\theta|c_\nu\right)P\left(c_\nu\right)} = 1,$$

$$P^*\left(c_j|\theta\right) = \inf\nolimits_\Delta P\left(c_j|\theta\right) = \inf\nolimits_\Delta \frac{P\left(\theta|c_j\right)P\left(c_j\right)}{\sum_\nu P\left(\theta|c_\nu\right)P\left(c_\nu\right)} = 0,$$

This ULP solution provides no information and is different from the TBM solution.

Dempster–Shafer Analysis

The Dempster–Shafer analysis of the paradigm leads to the same solution as the TBM. But the origin of this solution is connected to probability theory and open to criticisms about the appropriateness of the use of Dempster's rule of conditioning.

In Dempster–Shafer theory, it is postulated that there is a mapping M from T to $T\times\Omega\times\Theta$, that there is a probability distribution on T, and $Bel(\omega)$ for $\omega \subseteq T\times\Omega\times\Theta$ is defined as the probability of $M_*^{-1}(\omega) = \{t_i :\ M(t_i) \subseteq \omega\}$:

$$Bel\left(\omega\right) = P\left(M_*^{-1}\left(\omega\right)\right).$$

The knowledge that θ is true induces the adaptation of M into $M_\theta = M \cap (T \times \Omega \times \{\theta\})$ and Bel is updated to:

$$Bel_\theta\left(\omega\right) = P\left(M_{\theta^*}^{-1}\left(\omega\right)\right) = P\left(\left\{t_i : M_\theta\left(t_i\right) \subseteq \omega\right\}\right).$$

In the present paradigm, one has a.o. $M(t_4) = \{(t_4, c_1, \theta),\ (t_4, c_2, \theta), (t_4, c_3, \bar{\theta})\}$, and $P(M(t_4)) = p_4$ is correct. Once θ is known, the mapping M becomes $M_\theta = M \cap (T \times \Omega \times \{\theta\})$; a.o. $M_\theta(t_4) = \{(t_4, c_1, \theta),\ (t_4, c_2, \theta)\})$, and $Bel(M_\theta(t_4)) = p'_4$ where the denominator is the normalization factor related to the closed-world assumption (given θ, we know that t_0 was not the translator).

The problem with such a model is: why do we use p'_i and not $P(t_i|\theta)$ as it should according to probability theory [24]? One has

$$P\left(t_i|\theta\right) \approx P\left(\theta|t_i\right)p_i = P\left(\left\{c_j : f_i\left(c_j\right) = \theta\right\}|t_i\right)p_i,$$

e.g. $P(t_1|\theta) \approx P(c_1)p_1,\ P(t_4|\theta) \approx P(\{c_1, c_2\})p_4, \ldots$. It is impossible to get $P(t_i|\theta) = P(t_i)$ for all i. So the Dempster–Shafer solution cannot be obtained.

The difficulty with the paradigm lies in the fact that the values of $P(t_i|\theta)$ are unknown as the $P(c)$'s are unknown. Note that if one could assess $P(t_i|\theta)$, then $P(c)$ could be deduced. In that case all problems would disappear. Each analysis, be it the TBM, the Dempster–Shafer, or any upper and lower probabilities analysis, would lead to the Bayesian solution. But we are considering the case where $P(t_i|\theta)$ cannot be assessed since $P(c)$ is completely unknown. In the TBM, such a probability measure on $T \times \Omega \times \Theta$ and the concept of $P(t_i|\theta)$ are neither assumed nor even defined. Once θ is known, the TBM conditions the initial belief on $T \times \Omega \times \{\theta\}$ by transferring the basic belief masses.

Levi's criticism of the Dempster–Shafer analysis is based on the assumption that this last analysis represents a generalized Bayesian analysis, in which case the concept of a probability on $T \times \Omega \times \Theta$ is claimed. Once all relations with probability theory are set aside, such grounds for criticism disappear (see Remark 8.3).

The *difference between the TBM and the Dempster–Shafer solutions* resides in the fact that the TBM is free from any underlying probability theory. The probability information relative to the translators explained the origin of the basic belief masses at the credal level. But apart from that, any concept of probability is useless. A statement like

$$Bel_\theta\left(c\right) \leq P\left(c|\theta\right) \leq Pl_\theta\left(c\right)$$

is meaningless since we never built any probability measure on the frame of discernment Ω at the credal level, so the symbol P is undefined. ($P(c|\theta)$ should not be confused with $BetP_\theta(c)$ derived at the pignistic level.) Note that given any belief function, one can build a set of compatible probability functions such that

$$Bel\left(A\right) \leq P\left(A\right) \leq pl\left(A\right) \quad \forall\ A \in \mathcal{R}.$$

This is just a mathematical property without any interpretation relevant to the model (except for the comments at the end of Sect. 3.4).

The TBM could be viewed as a “purified” Dempster–Shafer model, i.e. purified from any probabilistic connotation. Hence, it forestalls criticisms aimed at the strange conditioning process encountered in the Dempster–Shafer solution which is at odds with plain probability approaches.

It is interesting to note that

$$pl_\theta\left(c_j\right) = P\left(\theta|c_j\right) = l\left(c_j|\theta\right),$$

where $l(c_j|\theta)$ is the likelihood of c_j given θ. There are some analogies between the TBM and *likelihood theory.* On the singletons of Ω, the two solutions are equivalent here. The analogy between the likelihood solution and the TBM solution is not always present as will be seen in the unreliable sensor paradigm.

7 The Unreliable Sensor Paradigm

The Paradigm

Let us consider a sensor with which You must check the temperature of a preparation: either the temperature is “cold” (Cold) or “hot” (Hot). Under correct working conditions, the sensor answers are given by a lamp that is "blue" (B) if the temperature is cold, and "red" (R) if the temperature is hot. Unfortunately, the sensor is not reliable as its thermometer is sometimes broken, in which case the sensor status can be B or R. In such a context, the sensor answer (B or R) is unrelated to the real temperature (Cold or Hot).

The only information known by You is what is indicated on the box containing the sensor: "Warning: the thermometer included in this sensor can be broken. *The probability that it is broken is 20%.* When the thermometer is *not broken*, the sensor is a perfectly reliable detector of the temperature situation. When the thermometer is not broken: a blue light means the temperature is cold, a red light means that the temperature is hot. When the thermometer is *broken*, the sensor answer is unrelated to the temperature."

You use the sensor and the light is red. What is Your degree of belief $Bel(\mathrm{Hot}|R)$ that the temperature is hot given the red light is on?

Let $\Theta = \{R,\ B\},\ \Omega = \{\mathrm{Cold},\ \mathrm{Hot}\},\ T = \{\mathrm{ThW},\ \mathrm{ThB}\}$ where ThW and ThB mean "thermometer-sensor in working conditions" and "thermometer-sensor broken".

TBM Solution

The TBM solution consists in assuming that the masses 0.8 and 0.2 are allocated on $\Theta \times \Omega \times T$ such that (see Fig. 2)

$$\begin{aligned} &m\left(\{(\mathrm{R},\ \mathrm{Hot},\ \mathrm{ThW}),(\mathrm{B},\ \mathrm{Cold},\ \mathrm{ThW})\}\right) = 0.8,\\ &m\left(\{(\mathrm{R},\ \mathrm{Cold},\ \mathrm{ThB}),(\mathrm{R},\ \mathrm{Hot},\ \mathrm{ThB})\ ,\right.\\ &\quad\left.(\mathrm{B},\ \mathrm{Cold},\ \mathrm{ThB}),(\mathrm{B},\ \mathrm{Hot},\ \mathrm{ThB})\}\right) = 0.2. \end{aligned}$$

When You learn that the light is red (R), the masses are transferred such that

$$\begin{aligned} m'\left(\{(\mathrm{R},\ \mathrm{Hot},\ \mathrm{ThW})\}\right) &= 0.8,\\ m'\left(\{(\mathrm{R},\ \mathrm{Cold},\ \mathrm{ThB}),(\mathrm{R},\ \mathrm{Hot},\ \mathrm{ThB})\}\right) &= 0.2. \end{aligned}$$

Marginalization on Ω provides:

$$\begin{aligned} Bel_{\mathrm{R}}\left(\mathrm{Hot}\right) &= 0.8, & bel_{\mathrm{R}}\left(\mathrm{Cold}\right) &= 0.0,\\ Pl_{\mathrm{R}}\left(\mathrm{Hot}\right) &= 1.0, & pl_{\mathrm{R}}\left(\mathrm{Cold}\right) &= 0.2. \end{aligned}$$

Should You have any *a priori* knowledge about the risk that the temperature is hot or cold, the credal equivalent of $P(\mathrm{Hot})$ and $P(\mathrm{Cold})$, it should be combined with the present results by Dempster's rule of combination.

8 Bayesian Solution

The Bayesian solution assumes $P(\mathrm{ThW}) = 0.8$ and $P(\mathrm{ThB}) = 0.2$. One also has $P(\mathrm{Hot}|R,\ \mathrm{ThW}) = 1$ and $P(\mathrm{Cold}|\mathrm{B},\ \mathrm{ThW}) = 1$. Note that we do not

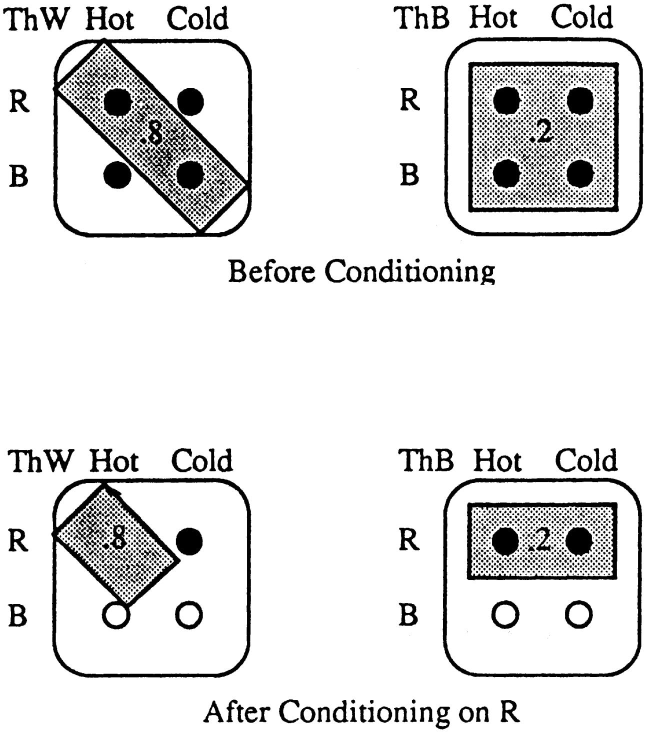


Fig. 2. The unreliable sensor paradigm, basic belief assignment and impact of conditioning on R

have $P(\text{R}|\text{ThB}) = 0$, the light can be red when the thermometer is broken. Furthermore, when the thermometer is broken, the probability that the system is Hot is independent from the sensor answer, and the thermometer is broken independently of the status of the system: $P(\text{Hot}|\text{R, ThB}) = P(\text{Hot})$. Also when the thermometer is in working condition (ThW), the probability that the light is red is the probability that the temperature is hot (Hot): $P(\text{R}|\text{ThW}) = P(\text{Hot})$. Then:

$$\begin{aligned}
P\left(\text{Hot}|\text{R}\right) &= P\left(\text{Hot}|\text{R},\ \text{ThW}\right) P\left(\text{ThW}|\text{R}\right) \\
&\quad + P\left(\text{Hot}|\text{R}, \text{ThB}\right) P\left(\text{ThB}|\text{R}\right) \\
&= \frac{1 \cdot P\left(\text{R}|\text{ThW}\right) P\left(\text{ThW}\right)}{P\left(\text{R}\right)} + \frac{P\left(\text{Hot}\right) P\left(\text{R}|\text{ThB}\right) P\left(\text{ThB}\right)}{P\left(\text{R}\right)} \\
&= \frac{0.8 \cdot P\left(\text{Hot}\right) + 0.2 \cdot P\left(\text{Hot}\right) P\left(\text{R}|\text{ThB}\right)}{P\left(\text{R}|\text{ThW}\right) P\left(\text{ThW}\right) + P\left(\text{R}|\text{ThB}\right) P\left(\text{ThB}\right)} \\
&= \frac{P\left(\text{Hot}\right)\left(0.8 + 0.2 \cdot P\left(\text{R}|\text{ThB}\right)\right)}{0.8 \cdot P\left(\text{Hot}\right) + 0.2 \cdot P\left(R|\text{ThB}\right)} .
\end{aligned}$$

The real problem encountered by the Bayesians is not so much in assessing $P(\mathrm{Hot})$, which could be known in practice but $P(\mathrm{R}|\mathrm{ThB})$, i.e. the probability that the light is red when the thermometer is broken. It is hard to imagine a serious hypothesis for such an ill-defined probability. There are so many ways for a thermometer to be broken that any particular value seems hardly justified. Bayesians could go on by assuming such a value... but of course the quality of their conclusions is strictly related to the quality of their assumptions.

Likelihood Solution

The likelihood solution $l(\mathrm{Hot}|\mathrm{R}) = P(\mathrm{R}|\mathrm{Hot})$ cannot be derived in this example as we cannot assess $P(\mathrm{R}|\mathrm{Hot})$

$$\begin{aligned} P\left(\mathrm{R}|\mathrm{Hot}\right) &= P\left(\mathrm{R}|\mathrm{ThW},\ \mathrm{Hot}\right) P\left(\mathrm{ThW}|\mathrm{Hot}\right) \\ &\quad + P\left(\mathrm{R}|\mathrm{ThB}, \mathrm{Hot}\right) P\left(\mathrm{ThB}|\mathrm{Hot}\right) \\ &= 1 \cdot 0.8 + 0.2 \cdot P\left(\mathrm{R}|\mathrm{ThB}\right), \end{aligned}$$

and we are faced with the same problem as the Bayesians: what is $P(\mathrm{R}|\mathrm{ThB})$? In this example $pl_{\mathrm{R}}(\mathrm{Hot}) = 1$: it is different from $l(\mathrm{Hot}|\mathrm{R})$ (except if You can defend $P(\mathrm{R}|\mathrm{ThB}) = 1$). Hence the TBM solution is not the likelihood solution.

Fiducial Solution

The TBM solution is also different from the fiducial solution. A *fiducial analysis* might consist in assuming:

$$\begin{aligned} P\left(\mathrm{R}|\mathrm{Hot}\right) &= P\left(\mathrm{B}|\mathrm{Cold}\right) = 0.8, \\ P\left(\mathrm{B}|\mathrm{Hot}\right) &= P\left(\mathrm{R}|\mathrm{Cold}\right) = 0.2, \end{aligned}$$

in which case, whatever $P(\mathrm{Hot})$,

$$\begin{aligned} P\left(\left\{\left(\mathrm{R},\ \mathrm{Hot}\right),\ \left(\left(\mathrm{B},\ \mathrm{Cold}\right)\right\}\right) &= 0.8, \\ P\left(\left\{\left(\mathrm{B},\ \mathrm{Hot}\right), \left(\mathrm{R},\ \mathrm{Cold}\right)\right\}\right) &= 0.2. \end{aligned}$$

As we know that R is true, the 0.8 mass is transferred to (R, Hot) and the mass 0.2 to (R, Cold). Marginalization on Ω gives $P(\mathrm{Hot}|\mathrm{R}) = 0.8$ and $P(\mathrm{Cold}|\mathrm{R}) = 0.2$. The solution is similar to the TBM as far as $Bel(\mathrm{Hot}|\mathrm{R})$ is concerned, but not as far as $Bel(\mathrm{Cold}|\mathrm{R})$ is concerned. The TBM does not provide any support to Cold, whereas the fiducial model gives it a 0.2 support, hence the difference between them.

9 Origin of the Basic Belief Masses

In each paradigm, a certain probability is associated with a certain basic belief mass. Such underlying probabilities are not necessary, as shown in Example 3.3, but they simplify our presentation. Should they have been omitted, the origin of the basic belief masses might have been felt to be somehow mysterious. We explain now the link between the basic belief masses and the probabilities when they exist.

Let $(\Omega, \mathcal{R})$ be a propositional space. Let us suppose that Your evidential corpus $\mathrm{EC}_t^{\mathrm{Y}}$ induces a belief on $\mathcal{R}$ such that there is a coarsening $\mathcal{R}'$ of $\mathcal{R}$ on which Your beliefs are described by a probability distribution P'. To obtain Your degrees of belief Bel' on $\mathcal{R}'$, only the frame $\mathcal{R}'$ needs to be considered. One gets: $\forall\ A \in \mathcal{R}',\ Bel'(A) = P'(A)$, and

$$\begin{aligned} m'(x) &= P'(x) && \text{for all atoms } x \text{ of } \mathcal{R}', \\ m'(A) &= 0 && \text{for non atomic } A \text{ of } \mathcal{R}'. \end{aligned}$$

The numerical equality between Bel' and P' can be justified by generalizing Hacking's Frequency Principle [13] to belief functions. The original principle is: when the objective probability of an event A is p, then the subjective probability of A is p. We just generalize it by requiring that the belief in A is p when the probability of A is p (whatever the nature of the probability).

In our paradigms, the atoms of the coarsenings $\mathcal{R}'$ are:

- in the murder of Mr. Jones: {Mary} and {Peter, Paul},
- in the guards and posts paradigm: the three soldiers,
- in the translator paradigm: the eight translators,
- in the unreliable sensor: the states ThW and ThB.

Problems appear once $\mathcal{R}$ is considered. Probabilists claim that the probability $P'(x)$ given to atom $x \in \mathcal{R}'$ is the sum of the probabilities $P(y)$ given to the atoms y of $\mathcal{R}$ that belong to $\Lambda(x)$, where Λ is the refining from $\mathcal{R}'$ to $\mathcal{R}$ corresponding to the coarsening from $\mathcal{R}$ to $\mathcal{R}'$:

$$P'(x) = \sum_{y \in \Lambda(x)} P(y)\,.$$

Given $\mathrm{EC}_t^{\mathrm{Y}}$, this tells nothing about the value of P, You can only compute the ULPs for the $P(A)$'s for $A \in \mathcal{R}$ or create P by using some general principles (like the Insufficient Reason principle or Maximum Entropy Principle). The major point about probability analyses is that the probabilists postulate the existence of a probability distribution P on $\mathcal{R}$—an item of information in fact not included in $\mathrm{EC}_t^{\mathrm{Y}}$.

The TBM considers only the information in $\mathrm{EC}_t^{\mathrm{Y}}$, hence it *does not postulate any probability distribution P on $\mathcal{R}$*. Nowhere in the paradigms are such functions P on $\mathcal{R}$ claimed. In practice to build Bel on $\mathcal{R}$, we only allocate the masses $m'(x)$ to $\Lambda(x) \in \mathcal{R}$:

$$m\left(\Lambda\left(x\right)\right) = \begin{cases} m'(x) & \text{for all atoms } x \in \mathcal{R}', \\ 0 & \text{otherwise.} \end{cases}$$

Such allocation is justified by the Principle of Least Commitment [48]. This principle translates the idea that You should not give more support to a proposition than justified. The least committed belief function Bel induced by Bel' on $\mathcal{R}$ is the vacuous extension of Bel' on $\mathcal{R}$ [33, p. 146]. Let $\mathcal{B}$ be the set of belief functions Bel^* defined on $\mathcal{R}$ such that $Bel^*(X) = Bel'(X)\ \forall\ X \in \mathcal{R}'$. The vacuous extension Bel of Bel' is the minimal element of $\mathcal{B}$ such that $Bel(A) \leq Bel^*(A)\ \forall\ A \in \mathcal{R},\ \forall Bel^* \in \mathcal{B}$.

Remark 3. We selected paradigms for which there exists a probability distribution on a coarsening because at least the numerical values given initially to the basic belief masses can be explained. The evaluation of the basic belief masses when there is no coarsening $\mathcal{R}'$ on which a probability distribution can be defined is discussed in Sect. 3.4.

Remark 4. The major difference between the TBM and probabilistic approaches obviously lies in the way we create the beliefs on $\mathcal{R}$ knowing the belief on $\mathcal{R}'$. The TBM is based on what is available and nothing else, whereas the probability analysis requires the existence of a probability distribution on $\mathcal{R}$. Consider the murder of Mr. Jones: in the case of a male killer (odd number thrown) the TBM accepts that Peter is *arbitrarily* selected by Big Boss whereas the probabilists claim that Peter is *randomly* selected by Big Boss.

Remark 5. The non-existence of a probability distribution on $\mathcal{R}$ resolves the problem raised by Levi [24]. Let us consider the translator paradigm. Once θ is learnt, why don't we condition the $p_i = p(t_i)$ on θ and thus use $p(t_i|\theta)$ as should be the case in a bona fide probability analysis? The concept of $p(t_i|\theta)$ is valid iff one can describe a probability distribution at least on the space $T \times \Theta$, which is not claimed in the TBM analysis. Hence, Levi's criticism does not apply to our model, but it does apply to some interpretations of the Dempster–Shafer model (those with a ULP connotation).

10 Handling Evidence at the Credal Level

We shall now detail how evidence is handled in the TBM and Bayesian models.

The *TBM* is based on the following underlying model:

- Credal states that represent the impact of the evidential corpus $\mathrm{EC}_t^{\mathrm{Y}}$ on a boolean algebra $\mathcal{R}$ are described by belief functions on $\mathcal{R}$.

- Pieces of evidence are to be taken into consideration at the credal level, conditioning being obtained via Dempster's rule of conditioning, the rule that underlies the TBM.
- Whenever a bet has to be established, the betting frame must be defined and the credal state prevailing at that time t induces a pignistic probability distribution on the elements of the bet.

In the case of the murder of Mr. Jones this schema becomes:

$$\begin{tikzcd}
\text{Credal state} & m_0 \arrow[r, "E_1", "\text{throwing the dice}"'] & m_{01} \arrow[r, "E_2", "\text{Peter's alibi}"'] \arrow[d] & m_{012} \arrow[d] \\
\text{Pignistic probability} & & P_{01} & P_{012}
\end{tikzcd}$$

In the *Bayesian model*, the same pattern reduces to:

$$\begin{tikzcd}
\text{Probability} & P_0 \arrow[r, "E_1", "\text{throwing the dice}"'] & P_1 \arrow[r, "E_2", "\text{Peter's alibi}"'] & P_2
\end{tikzcd}$$

The difference between the credal and pignistic levels is reminiscent of the difference between thought and action, between "inference" (how belief is affected by evidence) and "action" (which of several possible courses of action seems best) [50, p. 1].

Which model fits "reality"? Justifications of the Bayesian model are based on betting and decision arguments through the introduction of some requirements that lead to additive measures. But at the pignistic level, we also represent Your beliefs by a probability distribution, therefore we satisfy those requirements. This does not mean that additivity also pervades the credal level. No justifications are given by Bayesians for such requirements except that they just do not distinguish between a credal and a pignistic level. Their axioms always center on forced decisions (or preferences) but not on belief itself. They relate to observable behaviors that reflect an underlying credal state, not to the credal state itself.

To understand the difference between the two models, let us re-analyze in detail the Bayesian solution for the case of Mr. Jones. Re-consider the formula:

$$P\,(\text{killer is Mary}|\text{Peter's alibi}) = \frac{1}{1+x}$$

where

$$x = P\,(\text{killer is Paul}|\text{dice} = \text{odd})\,.$$

The only way to arrive at a Bayesian solution which is identical to the TBM solution is by postulating $x = 1$, i.e. that Big Boss will select Paul if an odd number is thrown. This case does not fit in with the available evidence (those in $\mathrm{EC}_t^{\mathrm{Y}}$). The case $x = 1$ fits in with the case "Peter was not and could not have been the killer", whereas the case $x < 1$ fits with the available information: "Peter was not the killer but could have been".

Intuitively, the real conceptual problem is to decide if, given Peter was not the killer, the knowledge that Peter might or might not have been the killer is relevant to the bet "Male versus Female". Bayesians say it is, the TBM says it is not.

Mathematically, the difference between the two solutions results from the necessity, in the context of probability, to split the 0.5 probability given to the males by the throwing of dice among the two males. Then later, the mass given to Peter cannot be given back to Paul once the alibi for Peter becomes available. Instead, in the TBM, the mass is not split, and is later transferred as a whole to Paul.

11 Conclusions

An argument we encountered when comparing the respective merits of the Bayesian model and the TBM runs as follows. Let us consider the case of Mr. Jones. Let

M = "the male chosen is always Paul".
B = "the TBM is true".
P = "the Bayesian model is true".
C = "the odds on male versus female are 1 to 1 once Peter's alibi is available".

One has to be careful not to use the following deceptive reasoning:

B implies C.
Assumption M is necessary in order to get C.
I dislike assumption M.
Therefore I dislike B.

The second proposition is wrong. The correct reasoning is:

B implies C.
If P, then assumption M is necessary in order to get C.
I dislike assumption M.
Therefore, if P, then I dislike B.

This is not sufficient to conclude that "I dislike B".

For the case of Mr. Jones, the Bayesian approach leads to a bet on Male versus Female with the odds at 1 to 2 whereas the belief function approach leads to a bet with the odds at 1 to 1. Which of the two is adequate is a matter of personal opinion? We feel that 1 to 1 is adequate. Others might prefer 1 to 2.

The argument that the Bayesian approach is correct because it complies with the probability theory is circular, hence useless. Description of credal states can be done by at least two normative models: the classical Bayesian and the TBM. Which of the two is correct cannot be established. It can only be tested. As Smith [50, p. 1] stated: "beliefs are insubstantial mental processes and it is not easy to lay down generally acceptable principles according to which our belief is ‘better’ than another.”

The interest of the “Mr. Jones” example lies in the fact that there is a case where both theories lead to different results. As a result, this particular example can be used as a test, a discriminating tool to distinguish between the two models. That it would convince the Bayesians is not sure but we hope here to have suggested some answers to Lindley’s challenge [25].

In summary, we have presented the TBM through the analysis of some paradigms and the comparison of the TBM solutions with the classical Bayesian solutions. The TBM aims at quantifying our degree of belief that a given proposition is true (where “belief” could be renamed “support”, “assurance”, “commitment”,...). We use belief as it is the most natural word even though one could argue about the value of such a choice.

The principal assumption on which the TBM depends is the concept of “parts of belief” supporting propositions and that due to a lack of further information cannot support a more specific proposition. We showed how betting behaviors can be established by constructing pignistic probabilities, and explained why Dutch Books cannot be constructed to disprove the TBM. The semantics of our model are provided by its betting behavior. It is essentially identical to the “exchangeable bets” semantics of the Bayesians. The difference lies in the way bets are adapted when the betting frames are changed. The paradigms illustrate the model and allow us to enhance its originality in comparison to the classical probability models. The two-level structure of our belief (credal and pignistic) is detailed. The missing element of our presentation, a clear axiomatic justification of the TBM, is presented in a forthcoming paper (see also [53]). The present paper concentrates on the presentation of the model rather than its foundations.

Uncertainty is a polymorphous phenomenon [44]. There is a different mathematical model for each of its varieties. No single model fits all cases. The real problems when quantifying uncertainty is to recognize its nature and to select the appropriate model. The Bayesian model is only one of them. The TBM is also only one of them. Each has its own field of applicability. Neither is always better than the other [32]. As Fisher once put it:

> La seule direction pratique qui nous est ouverte, est de concevoir clairement le processus intellectuel exact d’une méthode et ensuite de peser, considérer, critiquer et finalement décider si la méthode est ou non acceptable, si j’ose dire, pour notre conscience scientifique.. . .[10, p. 193][5]

[5] The only practical direction open to us is to conceive clearly the exact intellectual process of a method and then to weight, consider, criticize and finally decide whether or not the method is acceptable, if I dare say it, to our scientific conscience. . ..

Appendix A. Proof of Theorem 3.1

Let (Ω, $\mathcal{R}$, Bel) be a credibility space and m the basic belief assignment associated to Bel. Let $\mathcal{A}$ denote the set of atoms of $\mathcal{R}$. Given Assumption A1, there is a function f such that, for $x \in \mathcal{A}$, $BetP(x;m) = f(\{m(X) : x \subseteq X\})$.

Suppose Bel is such that there is an $A \in \mathcal{R}$ with $m(A) > 0$ and a pair of atoms y and z of $\mathcal{R}$ with $y \neq z$ and $y \cap A = \emptyset$, $z \cap A \neq \emptyset$. Let $0 \leq \delta \leq \varepsilon \leq m(A)$.

Let m' be the basic belief assignment on $\mathcal{R}$ such that $m'(A) = m(A) - \varepsilon$, $m'(y) = m(y) + \varepsilon - \delta$, $m'(z) = m(z) + \delta$, and $m'(B) = m(B)$ for all $B \in \mathcal{R}$, $B \neq A$, $B \neq y$, $B \neq z$. Let

$$g\left(x, A, \varepsilon\right) = \begin{cases} f\left(m(x), \ldots, m\left(A\right) - \varepsilon, \ldots\right) - f\left(m\left(x\right), \ldots, m\left(A\right), \ldots\right), \\ \text{if } x \subseteq A, \\ 0, \qquad \text{otherwise;} \end{cases}$$

$$h\left(x, y, \varepsilon - \delta\right) = \begin{cases} -f\left(m\left(y\right) + \varepsilon - \delta, \ldots\right) + f\left(m\left(y\right), \ldots\right), & \text{if } x = y, \\ 0, & \text{if} x \neq y. \end{cases}$$

$$h\left(x, z, \delta\right) = \begin{cases} -f\left(m\left(z\right) + \delta, \ldots\right) + f\left(m\left(z\right), \ldots\right), & \text{if } x = z, \\ 0, & \text{if} x \neq z. \end{cases}$$

Since

$$\sum_{x \in \mathcal{A}} BetP\left(x; m\right) = \sum_{x \in \mathcal{A}} BetP\left(x; m'\right) = 1,$$

we have

$$\sum_{x \in \mathcal{A}} \left(BetP\left(x; m'\right) - BetP\left(x; m\right)\right) = 0.$$

Therefore, for the given $A \in \mathcal{R}$,

$$\sum_{x \in \mathcal{A}, x \subseteq A} g(x, A, \varepsilon) = h\left(y, y, \varepsilon - \delta\right) + h\left(z, z, \delta\right). \tag{A.1}$$

Since $g(x,\ A,\ \varepsilon)$ is independent of δ for all $x \subseteq A$, $h(y,\ y, \varepsilon - \delta) + h(z,\ z,\ \delta)$ is also independent of δ. Let

$$\begin{aligned} &h\left(y, y, \varepsilon - \delta\right) + h\left(z, z, \delta\right) - f\left(m\left(y\right), \ldots\right) - f\left(m\left(z\right), \ldots\right) = H\left(\varepsilon\right), \\ &K\left(\varepsilon - \delta\right) = -f\left(m\left(y\right) + \varepsilon - \delta, \cdots\right), \\ &L\left(\delta\right) - f\left(m\left(z\right) + \delta, \ldots\right). \end{aligned}$$

The relation

$$\begin{aligned} &h\left(y, y, \varepsilon - \delta\right) + h\left(z, z, \delta\right) \\ &\quad = -f\left(m\left(y\right) + \varepsilon - \delta, \ldots\right) + f\left(m\left(y\right), \ldots\right) - f\left(m\left(z\right) + \delta, \ldots\right) \\ &\qquad + f\left(m\left(z\right), \ldots\right) \end{aligned}$$

can be written as

$$H\left(\varepsilon\right)=K\left(\varepsilon-\delta\right)+L\left(\delta\right).$$

It is a Pexider's equation whose solutions for H, K, and L, given Assumption A2, are linear in their argument [1, Theorem 1, p. 142]. Hence

$$f\left(m\left(z\right),\ldots\right)=\alpha+\beta m\left(z\right),$$

where α and β may depend on the basic belief masses given to the strict supersets of z.

The important point up to here is that both $h(y,\ y,\ \varepsilon-\delta)+h(z,\ z,\ \delta)$ is linear in ε and does not depend on δ. Let $h(y,\ y,\ \varepsilon-\delta)+h(z,\ z,\ \delta)=c\varepsilon+d$.

The proof that $g(x,\ A,\ \varepsilon)$ in (A.1) is linear in all its arguments $m(\cdot)$ is based on the following procedure given for the case that A is the union of four atoms $y_1,\ y_2,\ y_3$, and y_4. Let $m(A)=a$. For $i,\ j,\ k=1,\ 2,\ 3,\ 4,\ i\neq j\neq k\neq i$, let

$$\begin{aligned}
x_i&=\left\{m\left(y_i\cup B\right):B\subseteq\bar{A}\right\},\\
x_{ij}&=\left\{m\left(y_i\cup y_j\cup B\right):B\subseteq\bar{A}\right\},\\
x_{ijk}&=\left\{m\left(y_i\cup y_j\cup y_k\cup B\right):B\subseteq\bar{A}\right\},\\
x_{1234}&=\left\{m\left(y_1\cup y_2\cup y_3\cup y_4\cup B\right):B\subseteq\bar{A},B\neq\emptyset\right\},
\end{aligned}$$

Then (A.1) becomes (the $m(A)$-term is put as first element of f and is not included in x_{1234}):

$$\begin{aligned}
&f\left(a-\varepsilon,x_1,x_{12},x_{13},x_{14},x_{123},x_{124},x_{134},x_{1234}\right)\\
&\quad-f\left(a,x_1,x_{12},x_{13},x_{14},x_{123},x_{124},x_{134},x_{1234}\right)+\\
&f\left(a-\varepsilon,x_2,x_{12},x_{23},x_{24},x_{123},x_{124},x_{234},x_{1234}\right)\\
&\quad-f\left(a,x_2,x_{12},x_{23},x_{24},x_{123},x_{124},x_{234},x_{1234}\right)+\\
&f\left(a-\varepsilon,x_3,x_{13},x_{23},x_{34},x_{123},x_{134},x_{234},x_{1234}\right)\\
&\quad-f\left(a,x_3,x_{13},x_{23},x_{34},x_{123},x_{134},x_{234},x_{1234}\right)+\\
&f\left(a-\varepsilon,x_4,x_{14},x_{24},x_{34},x_{124},x_{134},x_{234},x_{1234}\right)\\
&\quad-f\left(a,x_4,x_{14},x_{24},x_{34},x_{124},x_{134},x_{234},x_{1234}\right)\\
&=c\varepsilon+d.
\end{aligned}$$

Let $x_i=u,\ x_{ij}=v,\ x_{ijk}=w$, and $x_{1234}=t$ for all $i,\ j,\ k=1,2,3,4$. One gets:

$$\begin{aligned}
&4\left(f\left(a-\varepsilon,u,v,v,v,w,w,w,t\right)-f\left(a,u,v,v,v,w,w,w,t\right)\right)\\
&\quad=c\varepsilon+d;
\end{aligned}$$

hence $f(a,u,v,v,v,w,w,w,t)$ is linear in a.

Keep all equalities as before except for $x_{123} \neq w$, then:

$$\begin{aligned}&3\left(f\left(a-\varepsilon, u, v, v, v, x_{123}, w, w, t\right)-f\left(a, u, v, v, v, x_{123}, w, w, t\right)\right)+\\&\left(f\left(a-\varepsilon, u, v, v, v, w, w, w, t\right)-f\left(a, u, v, v, v, w, w, w, t\right)\right)\\&\quad=c\varepsilon+d.\end{aligned}$$

The second term in the left-hand side is linear in a, so the first term in the left-hand side is also linear in a.

Suppose now $x_{124} \neq w$, then

$$\begin{aligned}&2\left(f\left(a-\varepsilon, u, v, v, v, x_{123}, x_{124}, w, t\right)-f\left(a, u, v, v, v, x_{123}, x_{124}, w, t\right)\right)+\\&\left(f\left(a-\varepsilon, u, v, v, v, x_{124}, w, w, t\right)-f\left(a, u, v, v, v, x_{124}, w, w, t\right)\right)+\\&\left(f\left(a-\varepsilon, u, v, v, v, w, w, w, t\right)-f\left(a, u, v, v, v, w, w, w, t\right)\right)\\&\quad=c\varepsilon+d.\end{aligned}$$

The second and third terms in the left-hand side are linear in a, hence so is the first. Therefore, f is linear in a whatever the x_{ijk}-terms and x_{1234}. We drop them in the following relations about f.

Suppose $x_{12} \neq v$, then

$$\begin{aligned}&2\left(f\left(a-\varepsilon, u, x_{12}, v, v\right)-f\left(a, u, x_{12}, v, v\right)\right)+\\&2\left(f\left(a-\varepsilon, u, v, v, v\right)-f\left(a, u, v, v, v\right)\right)\\&\quad=c\varepsilon+d.\end{aligned}$$

The second term in the left-hand side is linear in a, hence so is the first.

Suppose $x_{13} \neq v$, then

$$\begin{aligned}&\left(f\left(a-\varepsilon, u, x_{12}, x_{13}, v\right)-f\left(a, u, x_{12}, x_{13}, v\right)\right)+\\&\left(f\left(a-\varepsilon, u, x_{12}, v, v\right)-f\left(a, u, x_{12}, v, v\right)\right)+\\&\left(f\left(a-\varepsilon, u, x_{13}, v, v\right)-f\left(a, u, x_{13}, v, v\right)\right)+\\&\left(f\left(a-\varepsilon, u, v, v, v\right)-f\left(a, u, v, v, v\right)\right)\\&\quad=c\varepsilon+d.\end{aligned}$$

The second, third, and fourth terms in the left-hand side are linear in a, hence so is the first. Therefore f is linear in its first argument a whatever its other arguments.

The general proof of the linearity of f in its arguments $m(\cdot)$ is obtained by tediously generalizing this reasoning for any A. Let n_X be the number of atoms of $\mathcal{A}$ in $X \in \mathcal{R}$. The proof is valid if $n_A < n_\Omega - 1$, since we need at least two atoms of $\mathcal{R}$ not in the set A used in the derivation.

Let $F = \{X : x \subseteq X\}$. The general solution can be written as:

$$f\left(\left\{m\left(X\right) : x \subseteq X\right\}\right)=\sum_{G \subseteq F} \beta\left(G\right) \prod_{Y \in G} m\left(Y\right), \tag{A.2}$$

where the $\beta(G)$ might depend on the $m(Y)$ with $n_Y \geq n_\Omega - 1$.

Suppose a belief function Bel with $m(X) = 1 - \omega$ and $m(\Omega) = \omega$ for $X \in \mathcal{R}$ and $n_X < n_\Omega - 1$. Then for all atoms x in X,

$$\begin{aligned} BetP\left(x;m\right) = \beta\left(\{\}\right) + \beta\left(\{X\}\right)\left(1-\omega\right) + \beta\left(\{\Omega\}\right)\omega \\ + \beta\left(\{X,\Omega\}\right)\omega\left(1-\omega\right) \end{aligned}$$

and for y not an atom of X,

$$BetP\left(y;m\right) = \beta\left(\{\}\right) + \beta\left(\{\Omega\}\right)\omega.$$

By adding these terms on the atoms of $\mathcal{R}$, one gets:

$$\begin{aligned} 1 = n_\Omega\beta\left(\{\}\right) + n_X\beta\left(\{X\}\right)\left(1-\omega\right) + n_\Omega\beta\left(\{\Omega\}\right)\omega \\ + n_X\beta\left(\{X,\Omega\}\right)\omega\left(1-\omega\right). \end{aligned}$$

This being true for all ω in [0, 1], the coefficients of the terms in ω and ω^2 must be zero. So

$$\begin{aligned} &\beta\left(\{X,\Omega\}\right) = 0, \\ &n_\Omega\beta\left(\{\Omega\}\right) = n_X\beta\left(\{X\}\right), \\ &1 = n_\Omega\beta\left(\{\}\right) + n_X\beta\left(\{X\}\right). \end{aligned}$$

The same argument can be repeated in order to show that every coefficient $\beta(G) = 0$, whenever there is more than one element in G. Relation (A.2) becomes:

$$BetP\left(x;m\right) = \beta\left(\{\}\right) + \sum_{x\subseteq X} \beta\left(\{X\}\right) m\left(X\right),$$

where the β may depend on the basic belief masses given to those elements of $\mathcal{R}$ with $n_\Omega - 1$ or n_Ω atoms. We show that β does not depend on those basic belief masses.

Let $(\Omega, \mathcal{R}, Bel)$ be a credibility space. Suppose it is known (by You) that the actual world ϖ is not an element of the set B that contains two atoms b_1 and b_2 of $\mathcal{R}$. So $\forall\ A \in \mathcal{R},\ A \cong A \cup X$ where X may be $b_1,\ b_2$, or $b_1 \cup b_2$, and $Bel(A) = Bel(A \cup X)$ by the Consistency Axiom. Let $(\Omega', \mathcal{R}', Bel')$ be the credibility space where $\Omega' = \Omega - B$, the set $\mathcal{A}'$ of atoms of $\mathcal{R}'$ is equal to $\mathcal{A} - (b_1 \cup b_2)$, and $Bel'(A) = Bel(A)$ for all $A \in \mathcal{R}'$. By construction, for all $Y \in \mathcal{R},\ n_Y \geq n_\Omega - 1$ and $m(Y) = 0$. Let $BetP(x;\ m)$ and $BetP'(x;\ m')$, x atom of $\mathcal{R}'$, be the pignistic probabilities derived from $Bel(m)$ and $Bel'(m')$. The basic belief masses involved in the coefficients β are explicitly written. One has:

$$\begin{aligned} BetP\left(x;m\right) &= \beta\left(\{\};\{m\left(Y\right):n_Y = n_\Omega - 1\},m\left(\Omega\right)\right) \\ &\qquad + \sum_{x\subseteq X\in\mathcal{R}} \beta\left(\{X\};\{m\left(Y\right):n_Y = n_\Omega - 1\},m\left(\Omega\right)\right)m\left(X\right) \\ &= \beta\left(\{\};\{0:n_Y = n_\Omega - 1\},0\right) \\ &\qquad + \sum_{x\subseteq X\in\mathcal{R}'} \beta\left(\{X\};\{0:n_Y = n_\Omega - 1\},0\right)m\left(X\right), \end{aligned}$$

where all terms $m(X)$ are zero if $n_X \geq n_\Omega - 1$, or equivalently $m(X) = 0$ if $X \notin \mathcal{R}'$.

One also has:

$$\begin{aligned} BetP'(x; m') =& \beta(\{\}; \{m(Z) : n_Z = n_{\Omega'} - 1\}, m(\Omega')) \\ &+ \sum_{x \subseteq X \in \mathcal{R}'} \beta(\{X\}; \{m(Z) : n_Z = n_{\Omega'} - 1\}, m(\Omega'))\, m(X). \end{aligned}$$

By Assumption A4, $BetP(x;\ m) = BetP'(x;\ m')$ for all atoms x of $\mathcal{R}'$. Hence

$$\begin{aligned} &\beta(\{X\}; \{m(Z) : n_Z = n_{\Omega'} - 1\}, m(\Omega')) \\ &\quad = \beta(\{X\}; \{0 : n_Y = n_\Omega - 1\}, 0). \end{aligned}$$

So β does not depend on the basic belief masses $m(Z)$ for $n_Z \geq n_{\Omega'} - 1$.

Furthermore, by Assumption A3, the coefficients β depend only on the number of atoms in their arguments. Hence

$$BetP(x; m) = 1/n_\Omega - \beta(\{Q\}) + n_\Omega \beta(\{\Omega\}) \sum_{x \subseteq X} m(X)/n_X.$$

Let $m(A) = 1$ for $A \in \mathcal{R}$ and $n_A < n_\Omega - 1$. By Assumption A3, $BetP(x;\ m) = 1/n_A$ for every atom x that is a subset of A. It implies $\beta(\{\Omega\}) = 1/n_\Omega$. Hence

$$BetP(x; m) = \sum_{x \subseteq X} m(X)/n_X.$$

Acknowledgements

The authors are indebted to P. Besnard, M. Clarke, D. Dubois, P. Garbolino, Y.T. Hsia, H. Kyburg, B. Marchal, V. Poznanski, H. Prade, E. Ruspini, T. Seidenfelt, D. Spiegelhalter, L. Simar and N. Wilson for their useful and most stimulating comments, especially when they disagree.

Bibliography

[1] J. Aczel, *Lectures on Functional Equations and Their Applications* (Academic Press, New York, 1966).

[2] P.K. Black, Is Shafer general Bayes? in: *Proceedings Third Workshop on Uncertainty in Artificial Intelligence*, Seattle, WA (1987) 2–9.

[3] M.R.B. Clarke, C. Froidevaux, E. Gregoire and P. Smets, eds., Special Issue on Uncertainty, Conditional and Non Monotonicity: Positions and Debates in Non-Standard Logics, *J. Appl. Non-Classical Logics* **1** (2) (1991) 103–310.

[4] R.T. Cox, Probability, frequency and reasonable expectation, *Amer. J. Phys.* **14** (1946) 1–13.

[5] M.H. DeGroot, *Optimal statistical decisions* (McGraw-Hill, New York, 1970).

[6] D. Dubois, P. Garbolino, H.E. Kyburg, H. Prade and P. Smets, Quantified uncertainty, *J. Appl. Non-Classical Logics* **1** (1991) 105–197.

[7] D. Dubois and H. Prade, On several representations of an uncertain body of evidence, in: M.M. Gupta and E. Sanchez, eds., *Fuzzy Information and Decision Processes* (North-Holland, Amsterdam, 1982) 167–181.

[8] D. Dubois and H. Prade, Focusing versus updating in belief function theory, Internal Report IRIT/91–94/R, IRIT, Université P. Sabatier, Toulouse, France (1991).

[9] P.O. Ekelof, *Rättegång IV* (Stockholm, 5th ed., 1982).

[10] R.A. Fisher, Conclusions fiduciaires, *Ann. l'Inst. Henri Poincaré* **10** (1948) 191–213.

[11] P. Gärdenfors, B. Hansson and N.E. Sahlin, eds., *Evidentiary Value: Philosophical, Judicial and Psychological Aspects of a Theory* (C.W.K. Gleerups, Lund, Sweden, 1983).

[12] R. Giles (1982) Foundation for a possibility theory, in: M.M. Gupta and E. Sanchez, eds., *Fuzzy Information and Decision Processes* (North-Holland, Amsterdam, 1982) 183–195.

[13] I. Hacking, *Logic of Statistical Inference* (Cambridge University Press, Cambridge, England, 1965).

[14] P. Hájek, Deriving Dempster's rule, in: *Proceedings IPMU'92*, Palma de Mallorca, Spain (1992) 73–75.

[15] J.Y. Halpern and R. Fagin, Two views of belief: belief as generalized probability and belief as evidence, *Artif. Intell.* **54** (1992) 275–318.

[16] Y.-T. Hsia, Characterizing belief with minimum commitment, in: *Proceedings IJCAI-91*, Sydney, Australia (1991) 1184–1189.

[17] D. Hunter, Dempster-Shafer versus probabilistic logic, in: *Proceedings Third Workshop on Uncertainty in Artificial Intelligence*, Seattle, WA (1987) 22–29.

[18] J.Y. Jaffray, Application of linear utility theory for belief functions, in: B. Bouchon, L. Saitta and R.R. Yager, eds., *Uncertainty and Intelligent Systems* (Springer, Berlin, 1988) 1–8.

[19] F. Klawonn and E. Schwecke, On the axiomatic justification of Dempster's rule of combination, *Int. J. Intell. Syst.* **7** (1990) 469–478.

[20] F. Klawonn and P. Smets, The dynamic of belief in the transferable belief model and specialization-generalization matrices, in: D. Dubois, M.P. Wellman, B. d'Ambrosio and P. Smets, eds., *Uncertainty in AI 92* (Morgan Kaufmann, San Mateo, CA, 1992) 130–137.

[21] H.E. Kyburg Jr, Objectives probabilities, in: *Proceedings IJCAI-87*, Milan, Italy (1987) 902–904.

[22] H.E. Kyburg Jr, Bayesian and non-Bayesian evidential updating, *Artif. Intell.* **31** (1987) 271–293.

[23] K.B. Laskey, Beliefs in belief functions: an examination of Shafer's canonical examples, in: *Proceedings Third Workshop on Uncertainty in Artificial Intelligence*, Seattle, WA (1987) 39–46.

[24] I. Levi, Consonance, dissonance and evidentiary mechanisms, in: P. Gärdenfors, B. Hansson and N.E. Sahlin, eds., *Evidentiary Value: Philosophical, Judicial*

and Psychological Aspects of a Theory (C.W.K. Gleerups, Lund, Sweden, 1983) 27–43.

[25] D.V. Lindley, The probability approach to the treatment of uncertainty in artificial intelligence and expert systems, *Stat. Sci.* **2** (1987) 17–24.

[26] H.T. Nguyen, On random sets and belief functions, *J. Math. Anal. Appl.* **65** (1978) 531–542.

[27] H.T. Nguyen and P. Smets, On dynamics of cautious belief and conditional objects, *Int. J. Approx. Reasoning* **8** (1993) 89–104.

[28] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, San Mateo, CA, 1988).

[29] J. Pearl, Reasoning with belief functions: an analysis of compatibility, *Int. J. Approx. Reasoning* **4** (1990) 363–390.

[30] F.P. Ramsey, Truth and probability, in: H.E. Kyburg and H.E. Smokler, eds., *Studies in Subjective Probability* (Wiley, New York, 1931) 61–92.

[31] E.H. Ruspini, The logical foundations of evidential reasoning, Tech. Note 408, SRI International, Menlo Park, CA (1986).

[32] A. Saffiotti, An AI view of the treatment of uncertainty, *Knowl. Eng. Rev.* **2** (1988) 75–98.

[33] G. Shafer, *A Mathematical Theory of Evidence* (Princeton University Press, Princeton, NJ, 1976).

[34] G. Shafer, Perspectives in the theory and practice of belief functions, *Int. J. Approx. Reasoning* **4** (1990) 323–362.

[35] G. Shafer, P.P. Shenoy and K. Mellouli, Propagating belief functions in qualitative Markov trees, *Int. J. Approx. Reasoning* **1** (1987) 349–400.

[36] G. Shafer and A. Tversky, Languages and designs for probability, *Cogn. Sci.* **9** (1985) 309–339.

[37] P. Smets, Un modèle mathématico-statistique simulant le processus du diagnostic médical, Doctoral Dissertation, Université Libre de Bruxelles, Bruxelles, Belgium (1978). Available through University Microfilm International, 30–32 Mortimer Street, London WIN 7RA, Thesis 80–70,003.

[38] P. Smets, Upper and lower probability functions versus belief functions, in: *Proceedings International Symposium on Fuzzy Systems and Knowledge Engineering*, Guangzhou, China (1987) 17–21.

[39] P. Smets, Belief functions, in: P. Smets, A. Mamdani, D. Dubois and H. Prade, eds., *Non-standard Logics for Automated Reasoning* (Academic Press, London, 1988) 253–286.

[40] P. Smets, The combination of evidence in the transferable belief model, *IEEE Trans. Pattern Anal. Mach. Intell.* **12** (1990) 447–458.

[41] P. Smets, The transferable belief model and possibility theory, in: *Proceedings NAFIPS-90* (1990) 215–218.

[42] P. Smets, Constructing the pignistic probability function in a context of uncertainty, in: M. Henrion, R.D. Shachter, L.N. Kanal and J.F. Lemmer, eds., *Uncertainty in Artificial Intelligence* **5** (North-Holland, Amsterdam, 1990) 29–40.

[43] P. Smets, The transferable belief model and other interpretations of Dempster-Shafer's model, in: *Proceedings 6th Conference on Uncertainty in Artificial Intelligence*, Cambridge, MA (1990).

[44] P. Smets, Varieties of ignorance, *Inf. Sci.* **57–58** (1991) 135–144.

[45] P. Smets, The nature of the unnormalized beliefs encountered in the transferable belief model, in: D. Dubois, M.P. Wellman, B. d'Ambrosio and P. Smets, eds., *Uncertainty in AI 92* (Morgan Kaufmann, San Mateo, CA, 1992) 292–297.

[46] P. Smets, The transferable belief model and random sets, *Int. J. Intell. Syst.* **7** (1992) 37–46.

[47] P. Smets, Resolving misunderstandings about belief functions: a response to the many criticisms raised by J. Pearl, *Int. J. Approx. Reasoning* **6** (1992) 321–344.

[48] P. Smets, Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem, *Int. J. Approx. Reasoning* **9** (1993) 1–35.

[49] P. Smets, An axiomatic justification for the use of belief function to quantify beliefs, in: *Proceedings IJCAI-93*, Chambery, France (1993) 598–603.

[50] C.A.B. Smith, Consistency in statistical inference and decision, *J. Roy. Stat. Soc. B* **23** (1961) 1–37.

[51] P.M. Williams, Discussion, in: G. Shafer, ed., Belief functions and parametric models, *J. Roy. Stat. Soc. B* **44** (1982) 322–352.

[52] N. Wilson, Decision making with belief functions and pignistic probabilities, in: M. Clarke, R. Kruse and S. Moral, eds., *Symbolic and Quantitative Approaches to Reasoning Under Uncertainty* (Springer, Berlin, 1993) 364–371.

[53] S.K.M. Wong, Y.Y. Yao, P. Bollmann and H.C. Bürger, Axiomatization of qualitative belief structure, *IEEE Trans. Syst. Man Cybern.* **21** (1992) 726–734.

29

A k-Nearest Neighbor Classification Rule Based on Dempster-Shafer Theory

Thierry Denœux

Abstract. In this paper, the problem of classifying an unseen pattern on the basis of its nearest neighbors in a recorded data set is addressed from the point of view of Dempster-Shafer theory. Each neighbor of a sample to be classified is considered as an item of evidence that supports certain hypotheses regarding the class membership of that pattern. The degree of support is defined as a function of the distance between the two vectors. The evidence of the k nearest neighbors is then pooled by means of Dempster's rule of combination. This approach provides a global treatment of such issues as ambiguity and distance rejection, and imperfect knowledge regarding the class membership of training patterns. The effectiveness of this classification scheme as compared to the voting and distance-weighted k-NN procedures is demonstrated using several sets of simulated and real-world data.

1 Introduction

In classification problems, complete statistical knowledge regarding the conditional density functions of each class is rarely available, which precludes application of the optimal Bayes classification procedure. When no evidence supports one form of the density functions rather than another, a good solution is often to build up a collection of correctly classified samples, called the *training set*, and to classify each new pattern using the evidence of nearby sample observation. One such non-parametric procedure has been introduced by Fix and Hodges [11], and has since become well-known in the Pattern Recognition literature as the voting k-nearest neighbor (k-NN) rule. According to this rule, an unclassified sample is assigned to the class represented by a majority of its k nearest neighbors in the training set. Cover and Hart [4] have provided a statistical justification of this procedure by showing that,

This work was partially supported by EEC funded Esprit Project 6757 EMS (Environmental Monitoring System).

as the number N of samples and k both tend to infinity in such a manner that $k/N \to 0$, the error rate of the k-NN rule approaches the optimal Bayes error rate. Beyond this remarkable property, the k-NN rule owes much of its popularity in the Pattern Recognition community to its good performance in practical applications. However, in the finite sample case, the voting k-NN rule is not guaranteed to be the optimal way of using the information contained in the neighborhood of unclassified patterns. This is the reason why the improvement of this rule has remained an active research topic in the past 40 years.

The main drawback of the voting k-NN rule is that it implicitly assumes the k nearest neighbors of a data point x to be contained in a region of relatively small volume, so that sufficiently good resolution in the estimates of the different conditional densities can be obtained. In practice, however, the distance between x and one of its closest neighbors is not always negligible, and can even become very large outside the regions of high density. This has several consequences. First, it can be questioned whether it is still reasonable in that case to give all the neighbors an equal weight in the decision, regardless of their distances to the point x to be classified. In fact, given the k nearest neighbors $x^{(1)}, \cdots, x^{(k)}$ of x, and $d^{(1)}, \cdots, d^{(k)}$ the corresponding distances arranged in increasing order, it is intuitively appealing to give the label of $x^{(i)}$ a greater importance than to the label of $x^{(j)}$ whenever $d^{(i)} < d^{(j)}$. Dudani [10] has proposed to assign to the ith nearest neighbor $x^{(i)}$ a weight $w^{(i)}$ defined as:

$$w^{(i)} = \frac{d^{(k)} - d^{(i)}}{d^{(k)} - d^{(1)}} \qquad d^{(k)} \neq d^{(1)} \tag{1}$$

$$= 1 \qquad d^{(k)} = d^{(1)\cdot} \tag{2}$$

The unknown pattern x is then assigned to the class for which the weights of the representatives among the k nearest neighbors sum to the greatest value. This rule was shown by Dudani to be admissible, i.e. to yield lower error rates than those obtained using the voting k-NN procedure for at least one particular data set. However, several researchers, repeating Dudani's experiments, reached less optimistic conclusions [1], [16], [6]. In particular, Baily and Jain [1] showed that the distance-weighted k-NN rule is not necessarily better than the majority rule for small sample size if ties are broken in a judicious manner. These authors also showed that, in the infinite sample case ($N \to \infty$), the error rate of the traditional unweighted k-NN rule is better than that of any weighted k-NN rule. However, Macleod et al. [15] presented arguments showing that this conclusion may not apply if the training set is finite. They also proposed a simple extension of Dudani's rule allowing for a more effective use of the kth neighbor in the classification.

Apart from this discussion, it can also be argued that, because the weights are constrained to span the interval [0, 1], the distance-weighted k-NN procedure can still give considerable importance to observations that are very dissimilar to the pattern to be classified. This represents a serious drawback when all the classes cannot be assumed to be represented in the training set, as is often the case in some application areas as target recognition in noncooperative environments [5] or diagnostic problems [9]. In such situations, it may be wise to consider that a point that is far away from any previously observed pattern most probably belongs to an unknown class for which no information has been gathered in the training set, and should therefore be rejected. Dubuisson and Masson [9] call *distance reject* this decision, as opposed to the *ambiguity reject* introduced by Chow [3] and for which an implementation in a k-NN rule has been propoposed by Hellman [12]. Dasarathy [5] has proposed a k-NN rule where a distance reject option is made possible by the introduction of the concept of an *acceptable neighbor*, defined as a neighbor whose distance to the pattern to be classified is smaller than some threshold learnt from the training set. If there is less than some predefined number of acceptable neighbors of one class, the pattern is rejected and later considered for assignment to a new class using a clustering procedure.

Another limitation of the voting k-NN procedure is that it offers no obvious way to cope with uncertainty or imprecision in the labelling of the training data. This may be a major problem in some practical applications, as in the diagnostic domain, where the true identity of training patterns is not always known, or even defined, unambiguously, and has to be determined by an expert or via an automatic procedure that is itself subject to uncertainty. From a slightly different point of view, it may also be argued that patterns, even correctly labelled, have some degree of "typicality" depending on their distance to class centers, and that atypical vectors should be given less weight in the decision than those that are truly representative of the clusters [14]. Fuzzy sets theory offers a convenient formalism for handling imprecision and uncertainty in a decision process, and several fuzzy k-NN procedures have been proposed [13], [14]. In this approach, the degree of membership of a training vector x to each of M classes is specified by a number of u_i, with the following properties:

$$u_i \in [0,1] \tag{3}$$

$$\sum_{i=1}^{M} u_i = 1. \tag{4}$$

The membership coefficients u_i can be given (e.g. by experts) or computed using the neighbors of each vector in the training set [14]. The membership of an unseen pattern in each class is then determined by combining the memberships of its neighbors. Keller et al. [14] have proposed a rule in which membership assignment is a function of both the vector's distance from its k nearest neighbors, and those neighbors' memberships in the possible classes. Beyond an improvement in classification performance over the crisp k-NN procedure, this approach allows a richer information content of the classifier's output by providing membership values that can serve as a confidence measure in the classification.

In this paper, a new classification procedure using the nearest neighbors in a data set is introduced. This procedure provides a global treatment of important issues that are only selectively addressed in the aforementioned methods, namely: the consideration of the distances from the neighbors in the decision, ambiguity and distance rejection, and the consideration of uncertainty and imprecision in class labels. This is achieved by setting the problem of combining the evidence provided by nearest neighbors in the conceptual framework of Dempster-Shafer (D-S) theory. As will be seen, this formalism presents the advantage of permitting a clear distinction between the presence of conflicting information—as happens when a pattern is close to several training vectors from different classes—and the scarcity of information—when a pattern is far away from any pattern in the training set, or close to patterns whose class memberships are not defined precisely. In the following section, the basics of D-S theory are recalled. The application to a new k-NN procedure is then described, and experimental results are presented.

2 Dempster-Shafer Theory

Let $\varTheta$ be a finite set of mutually exclusive and exhaustive hypotheses about some problem domain, called the *frame of discernment* [19]. It is assumed that one's total belief induced by a body of evidence concerning $\varTheta$ can be partitioned into various portions, each one assigned to a subset of $\varTheta$. A *basic probability assignment* (BPA) is a function m from $2^{\varTheta}$, the power set of $\varTheta$, to [0, 1], verifying:

$$m\left(\emptyset\right) = 0 \tag{5}$$

$$\sum_{A \subseteq \varTheta} m\left(A\right) = 1. \tag{6}$$

The quantity $m(A)$, called a *basic probability number*, can be interpreted as a measure of the belief that one is willing to commit *exactly* to A, and not to any of its subsets, given a certain piece of evidence. A situation of total ignorance is characterized by $m(\varTheta) = 1$.

Intuitively, a portion of belief committed to a hypothesis A must also be committed to any hypothesis it implies. To obtain the total belief in A, one must therefore add to $m(A)$ the quantities $m(B)$ for all subsets B of A. The function assigning to each subset A of $\varTheta$ the sum of all basic probability numbers for subsets of A is called a *belief function*:

$$Bel\left(A\right) = \sum_{B \subseteq A} m\left(B\right). \tag{7}$$

Bel(A), also called the *credibility* of A, is interpreted as a measure of the total belief committed to A. The subsets A of Θ such that $m(A) > 0$ are called the *focal elements* of the belief function, and their union is called its *core*. The *vacuous* belief function has Θ for only focal element, and corresponds to complete ignorance. Other noticeable types of belief functions are *Bayesian* belief functions, whose focal elements are singletons, and *simple support* functions, that have only one focal element in addition of Θ.

It can easily be verified that the belief in some hypothesis A and the belief in its negation $\bar{A}$ do not necessarily sum to 1, which is a major difference with probability theory. Consequently, *Bel(A)* does not reveal to what extent one believes in $\bar{A}$, i.e. to what extent one doubts A, which is described by $Bel(\bar{A})$. The quantity $Pl(A) = 1 - Bel(\bar{A})$, called the *plausibility* of A, defines to what extent one fails to doubt in A, i.e. to what extent one finds A *plausible*. It is straightforward to show that:

$$Pl\,(A) = \sum_{B \cap A \neq \emptyset} m\,(B)\,. \tag{8}$$

As demonstrated by Shafer [19], any one of the three functions *m, Bel* and *Pl* is sufficient to recover the other two. This follows from the definition of $Pl(A)$ as $1 - Bel(\bar{A})$, and:

$$m\,(A) = \sum_{B \subseteq A} (-1)^{|A \backslash B|}\,\mathrm{Bel}\,(B)\,. \tag{9}$$

A BPA can also be viewed as determining a set of probability distributions P over 2^{Θ} satisfying:

$$Bel\,(A) \leq P\,(A) \leq Pl\,(A) \tag{10}$$

for all $A \subseteq \Theta$. For that reason, *Bel* and *Pl* are also called *lower* and *upper* probabilities, respectively. This fundamental imprecision in the determination of the probabilities reflects the “weakness”, or incompleteness of the available information. The above inequalities reduce to equalities in the case of a Bayesian belief function.

Given two belief functions Bel_1 and Bel_2 over the same frame of discernment, but induced by two independent sources of information, we must define a way by which, under some conditions, these belief functions can be combined into a single one. Dempster's rule of combination is a convenient method for doing such pooling of evidence. First, Bel_1 and Bel_2 have to be *combinable*, i.e. their cores must not be disjoint. If m_1 and m_2 are the BPAs associated with Bel_1 and Bel_2, respectively, this condition can also be expressed as:

$$\sum_{A \cap B = \emptyset} m_1\,(A)\,m_2\,(B) < 1. \tag{11}$$

If Bel_1 and Bel_2 are combinable, then the function $m : 2^{\Theta} \mapsto [0{,}1]$, defined by:

$$m(\emptyset) = 0 \tag{12}$$

$$m(\theta) = \frac{\sum_{A \cap B = \theta} m_1(A)\, m_2(B)}{1 - \sum_{A \cap B = \emptyset} m_1(A)\, m_2(B)} \qquad \theta \neq \emptyset \tag{13}$$

is a BPA. The belief function *Bel* given by m is called the orthogonal sum of Bel_1 and Bel_2, and is denoted $Bel_1 \oplus Bel_2$. For convenience, m will also be denoted $m_1 \oplus m_2$. The core of *Bel* equals the intersection of the cores of Bel_1 and Bel_2.

Although Dempster's rule is hard to justify theoretically, it has some attractive features, such as the following: it is commutative and associative; given two belief functions Bel_1 and Bel_2, if Bel_1 is vacuous, then $Bel_1 \oplus Bel_2 = Bel_2$; if Bel_1 is Bayesian, and if $Bel_1 \oplus Bel_2$ exists, then it is also Bayesian.

The D-S formalism must also be considered in the perspective of decision analysis [2]. As explained above, under D-S theory, a body of evidence about some set of hypotheses Θ does not in general provide a unique probability distribution, but only a set of *compatible* probabilities bounded by a belief function *Bel* and a plausibility function *Pl*. An immediate consequence is that simple hypotheses can no longer be ranked according to their probability: in general, the rankings produced by *Bel* and *Pl* will be different. This means that, as a result of lack of information, the decision is, to some extent, indeterminate. The theory does not make a choice between two distinct strategies: select the hypothesis with the greatest degree of belief—the most *credible*, or select the hypothesis with the lowest degree of doubt—the most *plausible*.

This analysis can be extended to decision with costs. In the framework of D-S theory, there is nothing strictly equivalent to Bayesian expected costs, leading unambiguously to a single decision. It is however possible to define lower and upper bounds for these costs, in the following way [7], [2]. Let M be the number of hypotheses, and U be an $M \times M$ matrix such that $U_{i,j}$ is the cost of selecting hypothesis θ_i if hypothesis θ_j is true. Then, for each simple hypothesis $\theta_i \in \Theta$, a *lower* expected cost $E_*[\theta_i]$ and an *upper* expected cost $E^*[\theta_i]$ can be defined:

$$E_*[\theta_i] = \sum_{A \subseteq \Theta} m(A) \min_{\theta_j \in A} U_{i,j} \tag{14}$$

$$E^*[\theta_i] = \sum_{A \subseteq \Theta} m(A) \max_{\theta_j \in A} U_{i,j}. \tag{15}$$

The lower (respectively: upper) expected cost can be seen as being generated by a probability distribution compatible with m, and such that the density of $m(A)$ is concentrated at the element of A with the lowest (respectively: highest) cost. Here again, the choice is left open as to which criterion should be used for the decision. Maximizing the upper expected cost amounts to minimizing the worst possible consequence, and therefore generally leads to more conservative decisions. Note that, when U verifies:

$$U_{i,j} = 1 - \delta_{i,j} \tag{16}$$

where $\delta_{i,j}$ is the Kronecker symbol, the following equalities hold:

$$E_* [\theta_i] = 1 - Pl(\{\theta_i\}) \tag{17}$$

$$E^* [\theta_i] = 1 - Bel(\{\theta_i\}) . \tag{18}$$

In the case of $\{0, 1\}$ costs, minimizing the lower (respectively: upper) expected cost is thus equivalent to selecting the hypothesis with the highest plausibility (respectively: credibility).

3 The Method

3.1 The Decision Rule

Let $\mathcal{X} = \{x^i = (x^i_1, \cdots, x^i_P) | i = 1, \cdots, N\}$ be a collection on N P-dimensional training samples, and $\mathcal{C} = \{C_1, \cdots, C_M\}$ be a set of M classes. Each sample x^i will first be assumed to possess a class label $L^i \in \{1, \cdots, M\}$ indicating with certainty its membership to one class in $\mathcal{C}$. The pair $(\mathcal{X}, \mathcal{L})$, where $\mathcal{L}$ is the set of labels, constitutes a training set that can be used to classify new patterns.

Let x^s be an incoming sample to be classified using the information contained in the training set. Classifying x^s means assigning it to one class in $\mathcal{C}$, i.e. deciding among a set of M hypotheses: $x^s \in C_q, q = 1, \ldots, M$. Using the vocabulary of D-S theory, $\mathcal{C}$ can be called the *frame of discernment* of the problem.

Let us denote by Φ^s the set of the k-nearest neighbors of x^s in $\mathcal{X}$, according to some distance measure (e.g. the Euclidian one). For any $x^i \in \Phi^s$, the knowledge that $L^i = q$ can be regarded as a piece of evidence that increases our belief that x^s also belongs to C_q. However, this piece of evidence does not by itself provide 100% certainty. In D-S formalism, this can be expressed by saying that only some part of our belief is committed to C_q. Since the fact that $L^i = q$ does not point to any other particular hypothesis, the rest of our belief cannot be distributed to anything else than $\mathcal{C}$, the whole frame of discernment. This item of evidence can therefore be represented by a BPA $m^{s,i}$ verifying:

$$m^{s,i}(\{C_q\}) = \alpha \tag{19}$$

$$m^{s,i}(\mathcal{C}) = 1 - \alpha \tag{20}$$

$$m^{s,i}(A) = 0 \quad \forall A \in 2^{\Theta} \setminus \{\mathcal{C}, \{C_q\}\} \tag{21}$$

with $0 < \alpha < 1$.

If x^i is far from x^s, as compared to distances between neighboring points in C_q, the class of x^i will be considered as providing very little information regarding the class of x^s; in that case, α must therefore take on a small value. On the contrary, if x^i is close to x^s, one will be much more inclined to believe that x^i and x^s belong to the same class. As a consequence, it seems reasonable to postulate that α should be a decreasing function of $d^{s,i}$, the distance between x^s and x^i. Furthermore, if we note:

$$\alpha = \alpha_0 \phi_q \left(d^{s,i}\right) \tag{22}$$

where the index q indicates that the influence of $d^{s,i}$ may depend on the class of x^s, the following additional conditions must be imposed on α_0 and ϕ_q:

$$0 < \alpha_0 < 1 \tag{23}$$

$$\phi_q (0) = 1 \tag{24}$$

$$\lim_{d \to \infty} \phi_q (d) = 0. \tag{25}$$

The first two conditions indicate that, even if the case of a zero distance between x^i and x^s, one still does not have certainty that they belong to the same class. This results from the fact that several classes can, in general, simultaneously have non zero probability densities in some regions of the feature space. The third condition insures that, in the limit, as the distance between x^s and x^i gets infinitely large, the belief function given by $m^{s,i}$ becomes vacuous, which means that one's belief concerning the class of x^s is no longer affected by one's knowledge of the class of x^i.

There is obviously an infinitely large number of decreasing functions ϕ verifying (24) and (25), and it is very difficult to find any a priori argument in favor of one particular function or another. We suggest to choose ϕ_q as:

$$\phi_q (d) = e^{-\gamma_q d^\beta} \tag{26}$$

with $\gamma_q > 0$ and $\beta \in \{1, 2, \cdots\}$. β can be arbitrarily fixed to a small value (1 or 2). Simple heuristics for the choice of α_0 and γ_q will be presented later.

For each of the k-nearest neighbors of x^s, a BPA depending on both its class label and its distance to x^s can therefore be defined. In order to make a decision regarding the class assignment of x^s, these BPAs can be combined using Dempster's rule. Note that this is always possible, since all the associated belief functions have $\mathcal{C}$ as a focal element.

Let us first consider two elements x^i and x^j of Φ^s belonging to the same class C_q. The BPA $m^{s,(i,j)} = m^{s,i} \oplus m^{s,j}$ resulting from the combination of $m^{s,i}$ and $m^{s,j}$ is given by:

$$m^{s,(i,j)} (\{C_q\}) = 1 - \left(1 - \alpha_0 \phi_q \left(d^{s,i}\right)\right) \left(1 - \alpha_0 \phi_q \left(d^{s,j}\right)\right) \tag{27}$$

$$m^{s,(i,j)} (\mathcal{C}) = \left(1 - \alpha_0 \phi_q \left(d^{s,i}\right)\right) \left(1 - \alpha_0 \phi_q \left(d^{s,j}\right)\right) \tag{28}$$

If we denote by Φ_q^s the set of the k-nearest neighbors of x^s belonging to C_q, and assuming that $\Phi_q^s \neq \emptyset$, the result of the combination of the corresponding BPAs $m_q^s = \oplus_{x^i \in \Phi_q^s} m^{s,i}$ is given by:

$$m_q^s\left(\{C_q\}\right) = 1 - \prod_{x^i \in \Phi_q^s} \left(1 - \alpha_0 \phi_q\left(d^{s,i}\right)\right) \tag{29}$$

$$m_q^s\left(\mathcal{C}\right) = \prod_{x^i \in \Phi_q^s} \left(1 - \alpha_0 \phi_q\left(d^{s,i}\right)\right). \tag{30}$$

If $\phi_q^s = \emptyset$, then m_q^s is simply the BPA associated with the vacuous belief function: $m_q^s(\mathcal{C}) = 1$.

Combining all the BPAs m_q^s for each class, a global BPA $m^s = \oplus_{q=1}^{M} m_q^s$ is obtained as:

$$m^s\left(\{C_q\}\right) = \frac{m_q^s\left(\{C_q\}\right) \prod\limits_{r \neq q} m_r^s(\mathcal{C})}{K} \qquad q = 1 \cdots, M \tag{31}$$

$$m^s\left(\mathcal{C}\right) = \frac{\prod\limits_{q=1}^{M} m_q^s(\mathcal{C})}{K} \tag{32}$$

where K is a normalizing factor:

$$K = \sum_{q=1}^{M} m_q^s\left(\{C_q\}\right) \prod_{r \neq q} m_r^s\left(\mathcal{C}\right) + \prod_{q=1}^{M} m_q^s\left(\mathcal{C}\right) \tag{33}$$

$$= \sum_{q=1}^{M} \prod_{r \neq q} m_r^s\left(\mathcal{C}\right) + (1 - M) \prod_{q=1}^{M} m_q^s\left(\mathcal{C}\right). \tag{34}$$

The focal elements of the belief function associated with m^s are the classes represented among the k-nearest neighbors of x^s, and $\mathcal{C}$. The credibility and plausibility of a given class C_q are:

$$Bel^s\left(\{C_q\}\right) = m^s\left(\{C_q\}\right) \tag{35}$$

$$Pl^s\left(\{C_q\}\right) = m^s\left(\{C_q\}\right) + m^s\left(\mathcal{C}\right) \tag{36}$$

Therefore, both criteria produce the same ranking of hypotheses concerning x^s.

If an $M \times M$ cost matrix U is given, where $U_{i,j}$ is the cost of assigning an incoming pattern to class i, if it actually belongs to class j, then lower and upper expected costs are defined for each possible decision:

$$E_*\left[C_q\right] = \sum_{A \subseteq \mathcal{C}} m^s\left(A\right) \min_{C_r \in A} U_{q,r} \tag{37}$$

$$= \sum_{r=1}^{M} m^s\left(\{C_r\}\right) U_{q,r} + m^s\left(\mathcal{C}\right) \min_{C_r \in \mathcal{C}} U_{q,r} \tag{38}$$

$$E^*\left[C_q\right] = \sum_{A \subseteq \mathcal{C}} m^s\left(A\right) \max_{C_r \in A} U_{q,r} \tag{39}$$

$$= \sum_{r=1}^{M} m^s\left(\left\{C_r\right\}\right) U_{q,r} + m^s\left(\mathcal{C}\right) \max_{\mathcal{C}_r \in \mathcal{C}} U_{q,r}. \tag{40}$$

Note that minimizing the lower or upper expected cost do not necessarily lead to the same decision, as can be seen from the following example. Let us consider the problem of assigning an incoming sample x^s to one of three classes ($M = 3$). It is assumed that the consideration of the k-nearest neighbors of x^s has produced a BPA m^s such that $m^s(\{C_1\}) = 0.2,\ m^s(\{C_2\}) = 0,\ m^s(\{C_3\}) = 0.4$ and $m^s(\mathcal{C}) = 0.4$. The cost matrix is:

$$U = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 2 & 0 \end{pmatrix}.$$

The lower and upper expected costs are, in that case:

$$\begin{array}{lll} E_*\left[C_1\right] = 0.4 & E_*\left[C_2\right] = 0.6 & E_*\left[C_3\right] = 0.2 \\ E^*\left[C_1\right] = 0.8 & E^*\left[C_2\right] = 1.0 & E^*\left[C_3\right] = 1.0. \end{array}$$

Thus, C_3 minimizes E_*, while C_1 minimizes E^*.

However, in the case of {0,1} costs, that will exclusively be considered henceforth, minimizing the lower (resp. upper) expected cost amounts to maximizing the plausibility (resp. credibility). In that case, and under the assumption that the true class membership of each training pattern is known, both criteria therefore lead to the same decision rule D:

$$q^s_{\max} = \arg\ \max_p\ m^s\left(\left\{C_p\right\}\right) \Rightarrow D\left(x^s\right) = q^s_{\max} \tag{41}$$

where $D(x^s)$ is the class label assigned to x^s.

Note that the consideration of the distances makes the probability of a tie taking place much smaller than in the simple majority rule, whose relationship with D can also be described by the following theorem:

Theorem 1. *If the k nearest neighbors of a data point x^s are located at the same distance of x^s, and if $\phi_1 = \phi_2 = \cdots = \phi_M = \phi$, then the decision rule D produces the same decision as the majority rule.*

Proof. Let us denote by ℓ the distance between x^s and all of its k nearest neighbors $x^i \in \Phi^s$. For all $q \in \{1, \ldots, M\}$, m^s_q is defined by:

$$m^s_q\left(\left\{C_q\right\}\right) = 1 - \left(1 - \alpha_0 \phi\left(\ell\right)\right)^{\left|\Phi^s_q\right|} \tag{42}$$

$$m^s_q\left(\mathcal{C}\right) = \left(1 - \alpha_0 \phi\left(\ell\right)\right)^{\left|\Phi^s_q\right|}. \tag{43}$$

Thus:

$$m^s\left(\{C_q\}\right) = \frac{\left(1-\left(1-\alpha_0\phi\left(\ell\right)\right)^{\left|\Phi_q^s\right|}\right)\left(1-\alpha_0\phi\left(\ell\right)\right)^{k-\left|\Phi_q^s\right|}}{K} \quad q \in \{1,\cdots,M\} \tag{44}$$

$$m^s\left(\mathcal{C}\right) = \frac{\left(1-\alpha_0\phi\left(\ell\right)\right)^k}{K}. \tag{45}$$

For any p and q in $\{1,\cdots,M\}$ such that $m^s(\{C_q\}) > 0$, we have:

$$\frac{m^s\left(\{C_p\}\right)}{m^s\left(\{C_q\}\right)} = \frac{\left(1-\alpha_0\phi\left(\ell\right)\right)^{k-\left|\Phi_p^s\right|} - \left(1-\alpha_0\phi\left(\ell\right)\right)^k}{\left(1-\alpha_0\phi\left(\ell\right)\right)^{k-\left|\Phi_q^s\right|} - \left(1-\alpha_0\phi\left(\ell\right)\right)^k}. \tag{46}$$

Therefore:

$$m^s\left(\{C_p\}\right) > m^s\left(\{C_q\}\right) \Leftrightarrow k - \left|\Phi_p^s\right| < k - \left|\Phi_q^s\right| \tag{47}$$

$$\Leftrightarrow \left|\Phi_p^s\right| > \left|\phi_q^s\right|. \tag{48}$$

3.2 Reject Options

The decision rule D can easily be modified so as to include ambiguity and distance reject options. The ambiguity reject option, as introduced by Chow [3] consists in postponing decision-making when the conditional error of making a decision given x^s is high. This situation typically arises in regions of the feature space where there is a strong overlap between classes. In that case, an incoming sample x^s to be classified will generally be close to several training vectors belonging to different classes. Hence, this can be viewed as a problem of conflicting information.

The distance reject option discussed in [9] corresponds to a different situation, where the point x^s to be classified is far away from any previously recorded sample, and is therefore suspected of belonging to a class that is not represented in the training set. The problem here no longer arises from conflict in the data, but from the weakness or scarcity of available information.

In our framework, the first case will be characterized by a BPA m^s that will be uniformly distributed among several classes. As a consequence, both the maximum plausibility $Pl^s(\{C_{q^s_{\max}}\})$ and the maximum credibility $Bel^s(\{C_{q^s_{\max}}\})$ will take on relatively low values. In the second case, most of the probability mass will be concentrated on the whole frame of discernment $\mathcal{C}$. As a consequence, only $Bel^s(\{C_{q^s_{\max}}\})$ will take on a small value; as the distance between x^s and its closest neighbor goes to infinity, $Bel^s(\{C_{q^s_{\max}}\})$ actually goes to zero, while $Pl^s(\{C_{q_{\max}}\})$ goes to one.

As a result, it is possible to introduce ambiguity and distance reject options by imposing thresholds $Pl_{\min}$ and $Bel_{\min}$ on the plausibility and credibility, respectively. The sample x^s will be ambiguity rejected if $Pl^s(\{C_{q^s_{\max}}\})$

$< Pl_{\min}$, and it will be distance rejected if $Bel^s(\{C_{q^s_{\max}}\}) < Bel_{\min}$. Note that, in the case of $\{0.1\}$ costs, these thresholds correspond to thresholds $E_{*\max}$ and $E^*_{\max}$ on the lower and upper expected costs, respectively:

$$E_{*\max} = 1 - Pl_{\min} \tag{49}$$

$$E^*_{\max} = 1 - Bel_{\min}. \tag{50}$$

The determination of $Pl_{\min}$ has to be based on a tradeoff between the probabilities of error and reject, and must therefore be left to the designer of the system. The choice of $Bel_{\min}$ is more problematic, since no unknown class is, by definition, initially included in the training set. A reasonable approach is to compute $Bel^i(\{C_{q^i_{\max}}\})$ for each x^i in the training set using the leave-one-out method, and define a distinct threshold $Bel^q_{\min}$ for each class C_q as:

$$Bel^q_{\min} = \min_{x^i \in \mathcal{X}, L^i = q} Bel^i\left(\left\{C_{q^i_{\max}}\right\}\right). \tag{51}$$

3.3 Imperfect Labelling

In some applications, it may happen that one only has imperfect knowledge concerning the class membership of some training patterns. For example, in a three class problem, an expert may have some degree of belief that a sample x^i belongs to a class C_3, but still consider as possible that it might rather belong to C_1 or C_2. Or, he may be sure that x^i does not belong to C_3, while being totally incapable of deciding between C_1 and C_2. In D-S formalism, one's belief in the class membership of each training pattern x^i can be represented by a BPA m^i over the frame of discernment $\mathcal{C}$. For example, if the expert is sure that x^i does not belong to C_3, has no element to decide between C_1 and C_2, and evaluates the chance of his assessment being correct at 80%, then his belief can be represented in the form of a BPA as:

$$m^i\left(\{C_1, C_2\}\right) = 0.8 \tag{52}$$

$$m^i\left(\mathcal{C}\right) = 0.2 \tag{53}$$

with all remaining $m^i(A)$ values equal to zero.

The approach described in above can easily be generalized so as to make use of training patterns whose class membership is represented by a BPA. If x^s is a sample to be classified, one's belief about the class of x^s induced by the knowledge that $x^i \in \Phi^s$ can be represented by a BPA $m^{s,\,i}$ deduced from m^i and $d^{s,\,i}$:

$$m^{s,i}\left(A\right) = \alpha_0 \phi\left(d^{s,i}\right) m^i\left(A\right) \quad \forall A \in 2^{\mathcal{C}} \backslash \mathcal{C} \tag{54}$$

$$m^{s,i}\left(\mathcal{C}\right) = 1 - \sum_{A \in 2^{\mathcal{C}} \backslash \mathcal{C}} m^{s,i}\left(A\right) \tag{55}$$

where ϕ is a monotonically decreasing function verifying (24) and (25).

As before, the $m^{s,\,i}$ can then be combined using Dempster's rule to form a global BPA:

$$m^s = \bigoplus_{x^i \in \Phi^s} m^{s,i} \tag{56}$$

Note that, while the amount of computation needed to implement Dempster's rule increases only linearly with the number of classes when the belief functions given by the $m^{s,\,i}$ are simple support functions as considered in Sect. 3.1, the increase is exponential is the worst general case. However, more computationally efficient approximation methods such as proposed in [21] could be used for very larger numbers of classes.

4 Experiments

The approach described in this paper has been successfully tested on several classification problems. Before presenting the results of some of these experiments, practical issues related to the implementation of the procedure need to be addressed.

Leaving alone the rejection thresholds, for which a determination method has already been proposed, and assuming an exponential form for ϕ_q as described in (26), the following parameters have to be fixed in order to allow the pratical use of the method: $k,\ \alpha_0,\ \gamma_q,\ q = 1, \cdots, M$ and β.

As in the standard k-NN procedure, the choice of k is difficult to make *a priori*. Although our method seems to be far less sensitive to this parameter than the majority rule, a systematic search for the best value of k may be necessary in order to obtain optimal results.

For the choice of α_0 and γ_q, several heuristics have been tested. Good results on average have been obtained with $\alpha_0 = 0.95$ and γ_q determined seperately for each class as $1/d_q^{\beta}$, where d_q is the mean distance between two training vectors belonging to class C_q.[1] The value of β has been found to have very little influence on the performance of the method. A value of $\beta = 1$ has been adopted in our simulations.

The following examples are intended to illustrate various aspects of our method, namely: the shape of the decision boundaries and reject regions for simple two-dimensional data sets, the relative performance as compared to the voting and distance-weighted k-NN rules for different values of k, and the effect of imperfect labelling.

4.1 Experiment 1

The purpose of this experiment is to visualize the decision boundary and the regions of ambiguity and distance reject for two different two-dimensional

[1] This heuristic was suggested to me by Lalla Meriem Zouhal.

data sets of moderate size. The first data set is taken from two Gaussian distributions with the following characteristics:

$$\mu_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \mu_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$
$$\Sigma_1 = 0.25 I \quad \Sigma_2 = I$$

where I is the identity matrix. There are 40 training samples in each class.

The second data set consists of two non-Gaussian classes of 40 samples each separated by a non-linear boundary. Both data sets are represented in Figs. 1–4, together with the lines of equal maximum credibility $Bel^s(\{C_{q^s_{\max}}\})$ and plausibility $Pl^s(\{C_{q^s_{\max}}\})$, for $k = 9$. As expected, the region of low plausibility is concentrated in each case around the class boundary, allowing for ambiguity reject, whereas small credibility values are obtained in the regions of low probability density. The distance reject regions, as defined in Sect. 3.2, are delimited by dotted lines.

For the first data set, the estimated error rate obtained using an independent test set of 1000 samples is 0.084, against 0.089 for the voting 9-NN rule. The corresponding results for the second data set and leave-one-out error estimation are 0.075 for both methods.

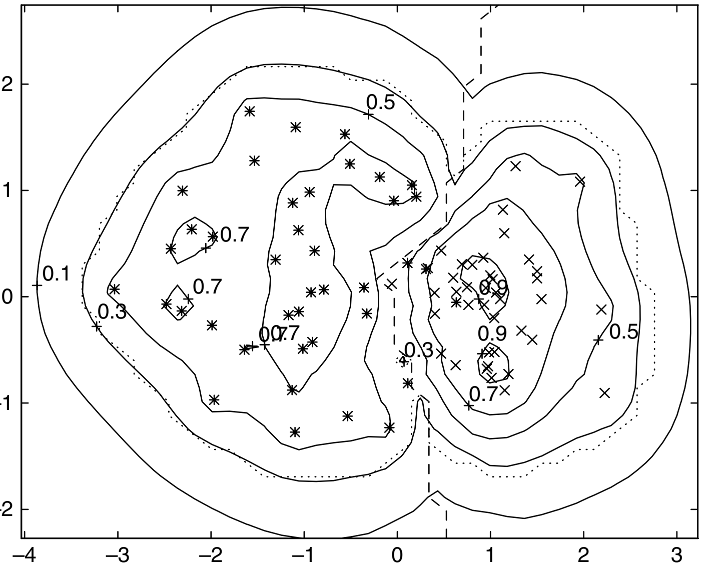


Fig. 1. Lines of equal maximum credibility ($Bel^s(\{C_{q^s_{\max}}\})$) for $k = 9$ (Gaussian data). Samples falling outside the region delimited by the dotted line are distance rejected

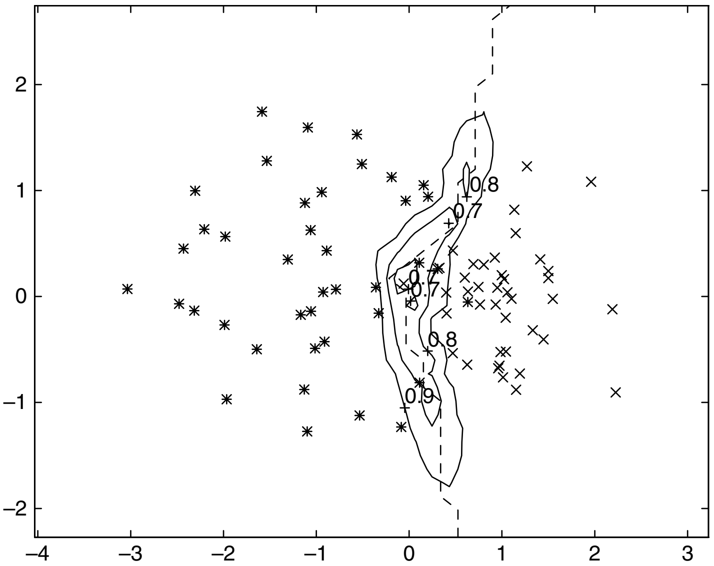


Fig. 2. Lines of equal maximum plausibility $(Pl^s(\{C_{q^s_{\max}}\}))$ for $k = 9$ (Gaussian data)

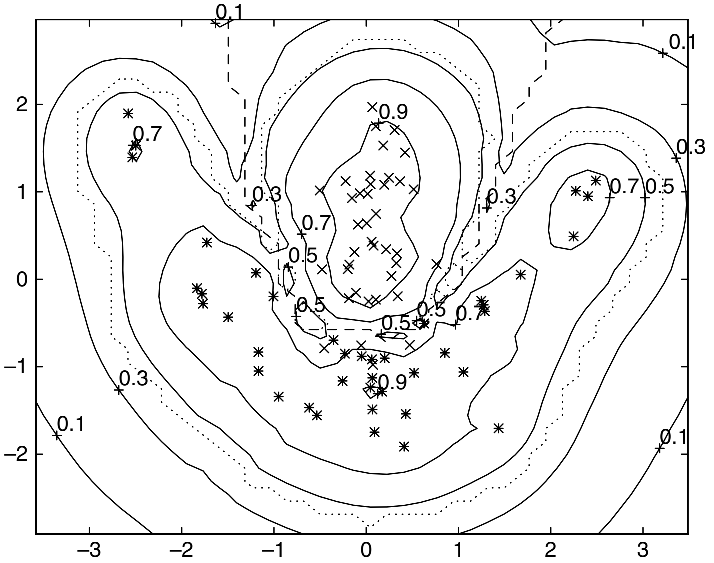


Fig. 3. Lines of equal maximum credibility $(Bel^s(\{C_{q^s_{\max}}\}))$ for $k = 9$ (non-Gaussian data). Samples falling outside the region delimited by the dotted line are distance rejected

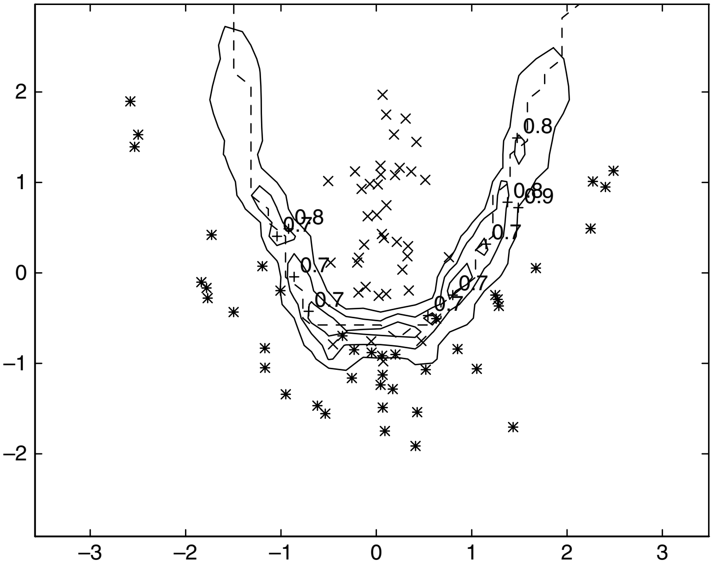


Fig. 4. Lines of equal maximum plausibility $(Pl^s(\{C_{q^s_{\max}}\}))$ for $k = 9$ (non-Gaussian data)

4.2 Experiment 2

A comparison between the performances of the voting k-NN procedure, the distance-weighted k-NN rule and our method was performed using one artificial and two real-world classification problems. In the majority rule, ties were resolved by randomly selecting one of the tied pattern classes.

The first problem implies three Gaussian distributions in a three-dimensional space, with the following characteristics:

$$\mu_1 = \begin{pmatrix}1\\1\\1\end{pmatrix} \qquad \mu_2 = \begin{pmatrix}-1\\1\\0\end{pmatrix} \quad \mu_2 = \begin{pmatrix}0\\-1\\1\end{pmatrix} .$$
$$\Sigma_1 = I \qquad\qquad \Sigma_2 = I \qquad\qquad \Sigma_3 = 2I$$

Training sets of 30, 60, 120 and 180 samples have been generated using prior probabilities (1/3, 1/3, 1/3). Values of k ranging from 1 to 25 have been investigated. A test set of 1000 samples has been used for error estimation. For each pair (N, k), the reported error rates are averages over 5 trials performed with 5 independent training sets. The results are presented in Table 1 and Figs. 5–8.

The second data set is composed of real-world data obtained by recording examples of the eleven steady state vowels of English spoken by fifteen speakers [8], [18]. Words containing each of these vowels were uttered once by the fifteen speakers. Four male and four female speakers were used to build

Table 1. Results of the Second Experiment (Gaussian Data, 1000 Test Samples) for the Voting k-NN Rule (k-NN), The Distance-Weighted k-NN Rule (Weighted k-NN) and Our Method (D-S): Best Error Rates (Means Over 5 Runs) with Corresponding Values of k (Upper Numbers) and Average Error Rates Integrated Over the Different Values of k (Lower Number)

	Classification rule		
	k-NN	weighted k-NN	Dempster-Shafer
$N = 30$	0.326 (5)	0.299 (16)	0.267 (15)
	0.397	0.338	0.306
$N = 60$	0.309 (8)	0.293 (21)	0.260 (23)
	0.335	0.314	0.284
$N = 120$	0.296 (7)	0.277 (25)	0.254 (22)
	0.306	0.300	0.280
$N = 180$	0.280 (18)	0.267 (14)	0.249 (23)
	0.296	0.293	0.273

a training set, and the other four male and three female speakers were used for building a test set. After suitable preprocessing, 568 training patterns and 462 test patterns in a 10 dimensional input space were collected. Figure 9 shows the test error rates for the three methods with k ranging from 1 to 30.

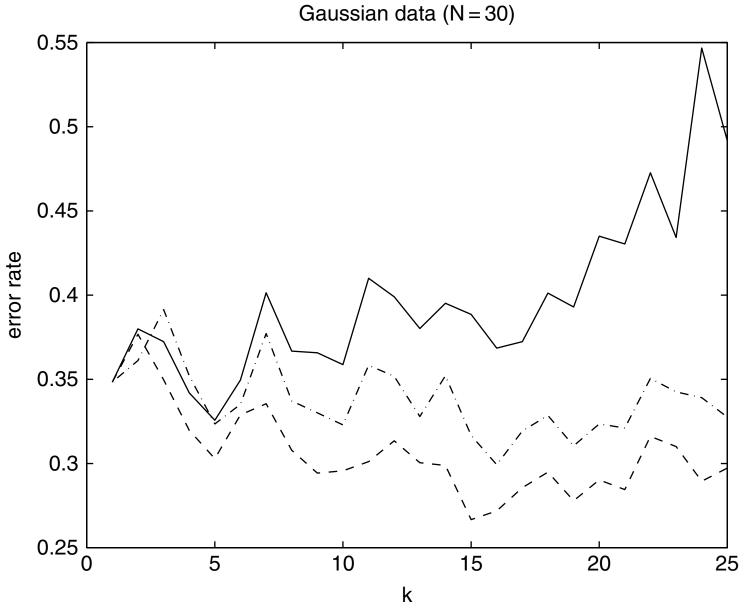


Fig. 5. Mean classification error rates for the voting k-NN rule (-), the distance-weighted k-NN rule (-.) and our method (- -) as a function of k (Gaussian data, $N = 30$)

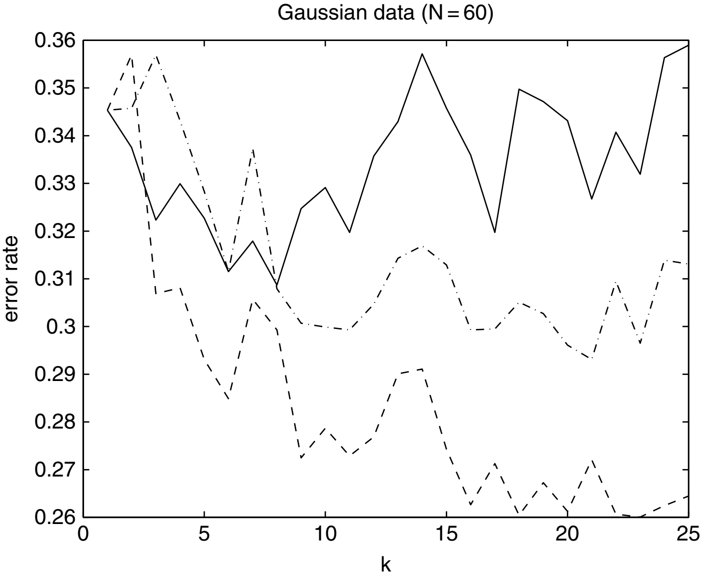


Fig. 6. Mean classification error rates for the voting k-NN rule (-), the distance-weighted k-NN rule (-.) and our method (- -) as a function of k (Gaussian data, $N = 60$)

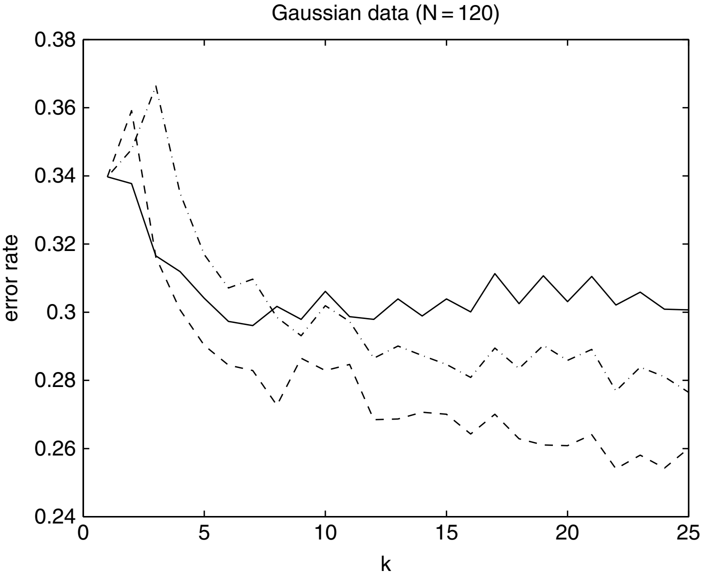


Fig. 7. Mean classification error rates for the voting k-NN rule (-), the distance-weighted k-NN rule (-.) and our method (- -) as a function of k (Gaussian data, $N = 120$)

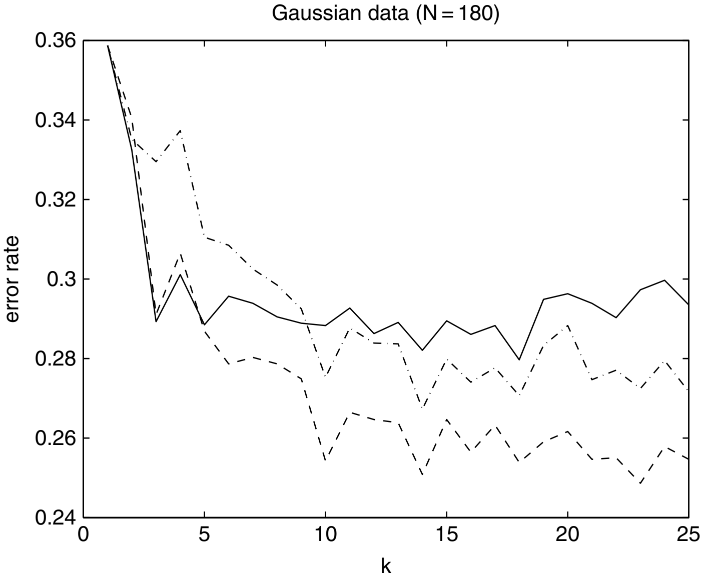


Fig. 8. Mean classification error rates for the voting k-NN rule (-), the distance-weighted k-NN rule (-.) and our method (- -) as a function of k (Gaussian data, $N = 180$)

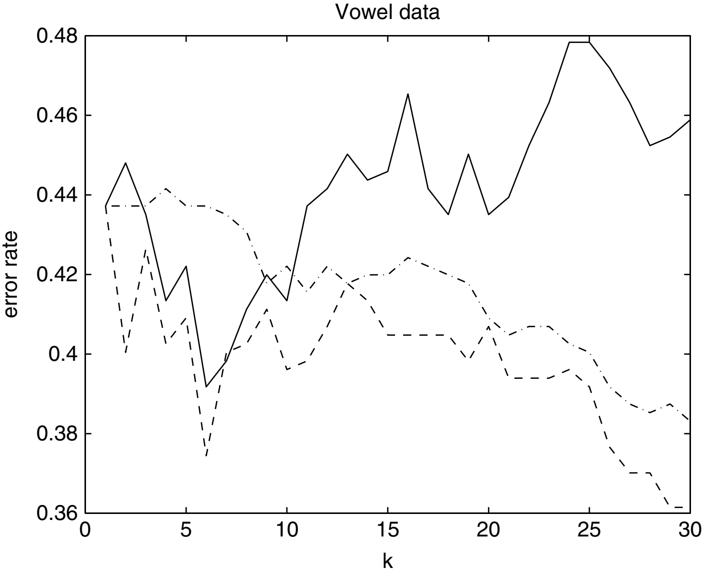


Fig. 9. Mean classification error rates for the voting k-NN rule (-), the distance-weighted k-NN rule (-.) and our method (- -) as a function of k (Vowel data)

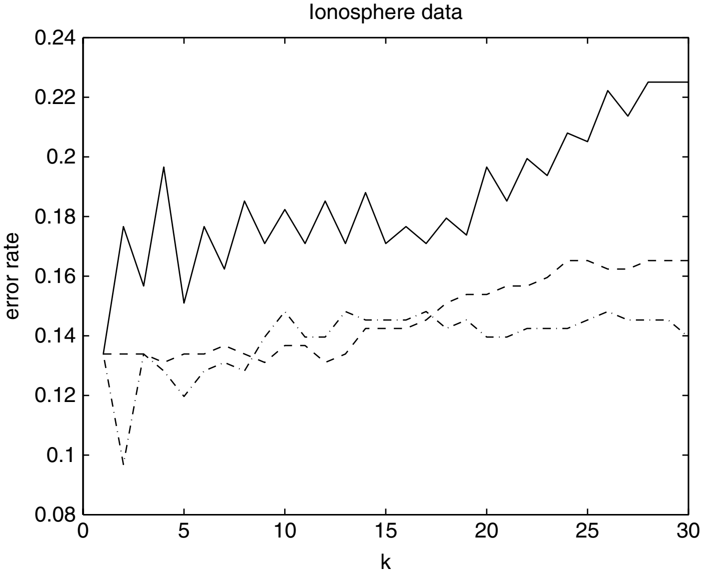


Fig. 10. Mean classification error rates for the voting k-NN rule (-), the distance-weighted k-NN rule (-.) and our method (- -) as a function of k (Ionosphere data)

The third task investigated concerns the classification of radar returns from the ionosphere obtained by a radar system consisting of a phased array of 16 high-frequency antennas [17], [20]. The targets were free electrons in the ionosphere. Radar returns were manually classified as "good" or "bad" depending on whether or not they showed evidence of some type of structure in the ionosphere. Received signals were processed using an autocorrelation function whose arguments are the time of a pulse and the pulse number. This processing yielded 34 continuous attributes for each of the 351 training instances collected. The classification results for different values of k are described in Fig. 10. The figures shown are leave-one-out estimates of the error rates, computed using the training data.

Not surprisingly, the performances of the two methods taking into account distance information are better than that of the voting k-NN rule, for the three classification problems investigated. Whereas the error rate of the voting k-NN rule passes by a minimum for some problem-dependent number of neighbors, the results obtained by the two other methods appear to be much less sensitive to the value of k, provided k is chosen large enough. Our method clearly outperforms the distance-weighted approach on the Gaussian data sets and the vowel recognition task. Both methods are almost equivalent on the ionosphere data.

4.3 Experiment 3

In order to study the behavior of our method in case of imperfect labelling, the following simulation has been performed. A data set of 120 training samples has been generated using the three Gaussian distributions of the previous experiment. For each training vector x^i, a number p^i has been generated using a uniform distribution on [0, 1]. With probability p^i, the label of x^i has been changed (to any of the other two classes with equal probabilities). Denoting by L^i the new class label of x^i, and assuming that $L^i = q$, then the BPA m^i describing the class membership of x^i has been defined as:

$$m^i\left(\{C_q\}\right) = 1 - p^i \tag{57}$$

$$m^i\left(\mathcal{C}\right) = p^i \tag{58}$$

and $m^i(A) = 0$ for all other $A \subseteq \mathcal{C}$. Hence, $m^i(\mathcal{C})$ is an indication of the reliability of the class label of x^i. Using the D-S formalism, it is possible to make use of this information, by giving less importance to those training vectors whose class membership is uncertain. This property can be expected to result in a distinctive advantage over the majority rule in a situation of this kind.

As can be seen from Fig. 11, our results support this assumption. The two curves correspond to the voting k-NN rule and our method with consideration

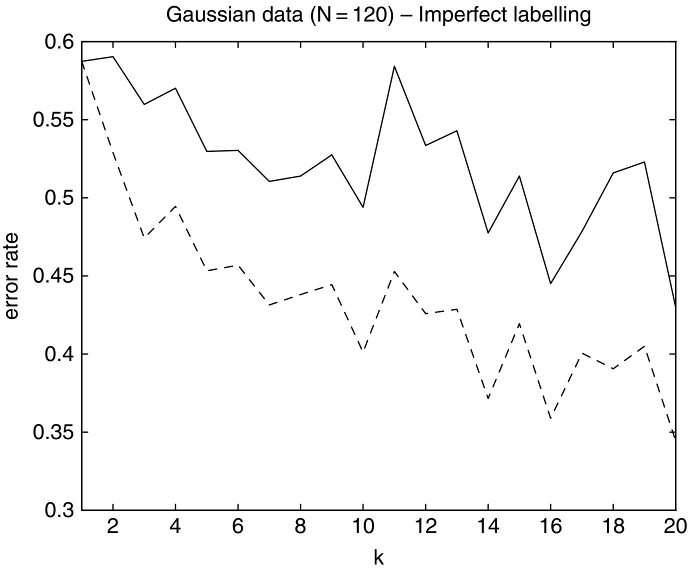


Fig. 11. Mean classification error rates for the voting k-NN rule (-) and our method with consideration of uncertainty in class labels (- -), as a function of k (Gaussian data, $N = 120$)

of uncertainty in class labels. As before, the indicated error rates are averages over 5 trials. The lowest rates achieved, as estimated on an independent test set of 1000 samples, are 0.43 and 0.34, respectively. The percentages of performance degradation resulting from the introduction of uncertainty in the class labels are respectively 54% and 21%. These results indicate that the consideration of the distances to the nearest neighbors *and* of the BPAs of these neighbors both bring an improvement over the majority rule in that case.

5 Conclusion

Based on the conceptual framework of D-S theory, a new non parametric technique for pattern classification has been proposed. This technique essentially consists in considering each of the k nearest neighbors of a pattern to be classified as an item of evidence that modifies one's belief concerning the class membership of that pattern. D-S theory then provides a simple mechanism for pooling this evidence in order to quantify the uncertainty attached to each simple or compound hypothesis. This approach has been shown to present several advantages. It provides a natural way of modulating the importance of training samples in the decision, depending on their nearness to the point to be classified. It allows for the introduction of ambiguity and distance reject options, that receive a unified interpretation using the concepts of lower and upper expected costs. Situations in which only imperfect knowledge is available concerning the class membership of some training patterns are easily dealt with by labelling each recorded sample using basic probability numbers attached to each subset of classes. Simulations using artificial and real-world data sets of moderate sizes have illustrated these different aspects, and have revealed in each case a superiority of the proposed scheme over the voting k-NN procedure in terms of classification performance. In two cases, the results obtained with our method were also better than those obtained with the distance-weighted k-NN rule, while both methods yielded similar results in a third experiment. It should be noted that these results are obviously not sufficient to claim the superiority of our approach for all possible data sets, although no counterexample has been encountered up to now. The comparison with the weighted or unweighted k-NN rules in the infinite sample case is also an interesting, but so far unanswered question.

Another particularity of the technique described in this paper is the quantification of the uncertainty attached to the decisions, in a form that permits combination with the outputs of complementary classifiers, possibly operating at different levels of abstraction. For example, given three classes C_1, C_2 and C_3, one classifier may discriminate between class C_1 and the other two, while another one may help to discern C_2 and C_3. By combining the BPAs produced by each of these classifiers, Dempster's rule offers a way to assess the reliability of the resulting classification. This approach is expected to be particularly useful in data fusion applications, where decentralized decisions based on data coming from disparate sensor sources need to be merged in order to achieve a final decision.

Acknowledgment

The vowel data were obtained from the Carnegie Mellon University collection of neural net benchmarks maintained by Matt White, under the supervision of Scott E. Fahlman. The ionosphere data were obtained from the UCI Repository of machine learning databases maintained by Patrick M. Murphy and David W. Aha. The author wishes to express his thanks to the anonymous referees for their helpful comments and suggestions during the revision of this paper.

Bibliography

[1] T. Baily and A. K. Jain, "A note on distance-weighted k-nearest neighbor rules," *IEEE Trans. Syst. Man Cyber.*, vol. 8, no. 4, pp. 311–313, 1978.

[2] W. F. Caselton and W. Luo, "Decision making with imprecise probabilities: Dempster-Shafer theory and application," *Water Resources Research*, vol. 28, no. 12, pp. 3071–3081, 1992.

[3] C. K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 41–46, 1970.

[4] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-13, no. 1, pp. 21–27, 1967.

[5] B. V. Dasarathy, "Nosing around the neighborhood: A new system structure and classification rule for recognition in partially exposed environments," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, no. 1, pp. 67–71, 1980.

[6] ——, "Nearest neighbor norms: NN pattern classification techniques," *IEEE Computer Society Press*, Los Alamitos, CA, 1991.

[7] A. P. Dempster and A. Kong, "Comment," *Stat. Sci.*, vol. 2, no. 1, pp. 32–36, 1987.

[8] D. H. Deterding, "Speaker normalization for automatic speech recognition," Ph.D. thesis, University of Cambridge, 1989.

[9] B Dubuisson and M. Masson, "A statistical decision rule with incomplete knowledge about classes," *Pattern Recognition*, vol. 26, no. 1, pp. 155–165, 1993.

[10] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *IEEE Trans. Syst. Man Cyber.*, vol. 6, pp. 325–327, 1976.

[11] E. Fix and J. L. Hodges, "Discriminatory analysis, nonparametric discrimination: Consistency properties," Technical Report 4, USAF School of Aviation Medicine, Randolph Field, TX, 1951.

[12] M. E. Hellman, "The nearest neighbor classification rule with a reject option," *IEEE Trans. Syst. Man Cyber.*, vol. 3, pp. 179–185, 1970.

[13] A. Jozwik, “A learning scheme for a fuzzy k-NN rule,” *Pattern Recognition Letters*, vol. 1, pp. 287–289, 1983.

[14] J. M. Keller, M. R. Gray, and J. A. Givens, “A fuzzy k-NN neighbor algorithm,” *IEEE Trans. Syst. Man Cyber.*, vol. 15, no. 4, pp. 580–585, 1985.

[15] J. E. Macleod, A. Luk, and D. M. Titterington, “A re-examination of the distance-weighted k-nearest neighbor classification rule,” *IEEE Trans. Syst. Man Cyber.*, vol. 17, no. 4, pp. 689–696, 1987.

[16] R. L. Morin and D. E. Raeside, “A reappraisal of distance-weighted k-nearest-neighbor classification for pattern recognition with missing data,” *IEEE Trans. Syst. Man Cyber.*, vol. 11, no. 3, pp. 241–243, 1981.

[17] P. M. Murphy and D. W. Aha, “UCI Repository of machine learning databases [Machine-readable data repository],” University of California, Department of Information and Computer Science., Irvine, CA, 1994.

[18] A. J. Robinson, “Dynamic error propagation networks,” Ph.D. thesis, Cambridge University Engineering Department, 1989.

[19] G. Shafer, *A Mathematical Theory of Evidence.* Princeton, NJ: Princeton University Press, 1976.

[20] V. G. Sigillito, S. P. Wing. L. V. Hutton, and K. B. Baker, “Classification of radar returns from the ionosphere using neural networks,” in *Johns Hopkins APL Technical Digest*, vol. 10, pp. 262–266, 1989.

[21] B. Tessem, “Approximations for efficient computation in the theory of evidence,” *Artificial Intelligence*, vol. 61, pp. 315–329, 1993.

30

Logicist Statistics II: Inference*

Arthur P. Dempster

Abstract. A perspective on statistical inference is proposed that is broad enough to encompass modern Bayesian and traditional Fisherian thinking, and interprets frequentist theory in a way that gives appropriate weights to both science and mathematics, and to both objective and subjective elements. The aim is to inject new thinking into a field held back by a longstanding lack of consensus.

1 Introduction

Was there ever a Fisherian cult in statistics? In the IMS presidential address that preceded my Fisher Memorial Lecture by two days at the 1998 Joint Statistical Meetings in Dallas, Persi Diaconis conferred "guru" status on Fisher, and used related terms such as "mysticism" and "cult". I beg to differ. Certainly Fisher was held in awe by many, but his statistical following was never as large, or organized, as those of present day advocates of "frequentist" and "Bayesian" inferential theories. Nor are these conventional categories as monolithic as the prominence of the labels suggests. Important ideas are involved, but statisticians are scarcely divided into two schools that compete for hearts and minds. For science, each methodology has obvious limitations, ambiguities, and deficiencies. I argue that their associated prescriptions should be absorbed into a broader and more complex outlook. Statistics needs a bigger toolbox of inferential methods, in fact a toolbox consistent with Fisher's thought.

Whereas research statisticians tend to call Fisher's writing obscure, my longheld sense has been that his ideas cohere easily with a simple perspective that I am calling "logicist". Fisher believed that the primary task of the practicing statistician is to get inside the mind of a research scientist facing a set of challenging questions. Only then should he or she turn to mathematical idealizations that represent aspects of a question under study and lead directly

* The 1998 COPSS R. A. Fisher Memorial Lecture, revised April 2002. Partial support through NSF Grant 94-04396 is gratefully acknowledged.

to answering carefully formulated and situation-specific questions involving uncertainties. It was largely because so much statistical research came to be dominated by exclusively mathematical discussions, with scarcely any reference to backgrounds in science, that Fisher became highly critical of the direction of mathematical statistics. This circumstance has changed little. A basic thesis of my presentation is that the logic of applied statistics is a logic of scientific practice, with mathematics in a supporting role.

The world of scientific reasoning is fundamentally different from the world of mathematical reasoning, being concerned with knowing and understanding empirical phenomena. Science depends on both informal and supporting formal reasoning about phenomena, including reasoning about uncertain aspects. There is a major distinction between the roles of formal reasoning in science and in mathematics, since the former rests on omnipresent connections with specific features of the objective world, and continually uses these connections to motivate choices of what to record and what to compute. The worlds of mathematics and science interact, often and successfully, but nevertheless represent different cultures, speaking languages that overlap only in part.

In mathematics, an element or operation is understood to have a precise place in a defined abstract structure. Not being consciously aware of mathematical abstraction, however, most nonmathematical scientists use informal language when referring to mathematical entities such as stochastic models or statistical inference procedures. In science, it is normal and necessary to define terms and concepts through ordinary language, without recourse to precise mathematics. In my presentation, the terms “logic”, “probability”, and “independence” connote technical concepts that are partly scientific and partly mathematical. Nonmathematical scientific semantics are acquired, as we acquire most language, through repeated exposures that gradually build understanding of subtle distinctions and contexts. In parallel, certain generally understood mathematical representations are essential to a minimally complete picture of formal statistical methods in daily use. Statistical sciences are quantitative and formal, and hence inseparable from mathematics, at least to the point of appreciating basic mathematics of data structures and probability, including familiarity with related computations. It is equally essential to internalize the scientific understanding that connects with formal statistical modeling and inference.

Fisher’s career illustrates the complexities of bridging the two worlds. He understood the substance and value of mathematics, and used mathematics to make pathbreaking contributions to agricultural science and genetics, as well as to core mathematical statistics. While drawing freely on his own formidable mathematical talents, Fisher also warned against overemphasis on mathematical theory, for example in teaching, where “there should be a nucleus of teachers with practical experience in all departments teaching statistical methods” (Fisher, 1960). His repeated statements on the practical limitations of both frequentist and Bayesian theories were too easily dismissed, however, due in part to his penchant for rude and dismissive polemics. Fisher’s scientific successes and communication failures hold lessons for anyone wishing to bring science and mathematics into improved and mutually beneficial harmony.

Logicism, as I use the term, signifies principled and explicit reasoning about specific scientific unknowns and uncertainties, whether the reasoning is informal or formal. Nonignorable uncertainties are traditionally approached in scientific discourse through informal and somewhat vague language applied to premises, arguments, and conclusions. Fisher in his scientific writing used informal language with a logical flavor, see for example Fisher (1955, 1956, 1960), but as a theoretician he understood that informal judgments can and should often be supplemented with formal analysis. For example, rather than tolerate sole reliance on expert perceptions that differential yields observed in agricultural studies are or are not large enough to justify concluding that experimental practices are effective, he proposed, implemented, and disseminated the methodology of randomized experimentation, together with formal significance tests based on ANOVA models. As argued in Part I (Dempster, 1998b), the force of Fisher's logic depends on statistical models being taken seriously as tentative and idealized conceptions that link specific external (objective) realities with incomplete and uncertain (subjective) knowledge of those realities.

In contrast to logicism, the frequentist idealization of practice holds that a statistician chooses and applies "procedures", such as a testing or confidence procedure, and reports numerical outputs. Interpretation is left to informal judgment by a user assumed to have mathematical and contextual knowledge and understanding of the meaning of long run properties associated with sampling models. Fisher thought differently. In parallel with reporting numerical or graphical outputs, such as a p-value from a significance test, or a plotted likelihood, or a fiducial interval, he sought to convey a logical interpretation constituting a formal inference about the specific situation under study. Such interpretations are of two basic types, described and analyzed below using the labels "postdictive" and "predictive". While not logic in the sense of a highly structured mathematical system, the Fisherian system is formal because its operations consist of precise algorithms applied to formal input data, and depend on an assumed formal mathematical model of a specific situation, including formal probabilistic representation of uncertainties. Inferential statistical computation implements data-specific processes of formal artificial reasoning from premises to conclusions.

In featuring Fisher's attitudes, my goal is not to reestablish a Fisherian school of inference, but rather to relate issues raised by Fisher to a broader sweep of scientific thinking. One important necessity in my view is recognition that complementarity of subjective and objective aspects of formal models and modeling is as much about the deterministic "equations" of traditional science as about the probabilistic models that dominate contemporary statistics. I argue here for recognition of the subjective "you" of Good (1950) (or "thou" of Savage (1962)) as an everpresent concominant of all scientific activity that seeks to make use of formal models. For example, if I use Newtonian mechanics to calculate the progress of an asteroid, it is implied that the formal "you" in the mathematical model translates into advice for the real me, that I may accept or question, and ultimately build upon for further study or action.

I believe that the opposition to anything subjective that roils academic statistics comes mainly from a tradition of mathematics teaching that was largely conditioned by objectivist attitudes to the deterministic models that have impressive credentials in physical and engineering sciences. This philosophy of objectivity is, however, viable only in limited areas of hard science, and only then by suppressing the ubiquitous "you". A more catholic attitude permits extended use of formal models in wider realms of science. Especially important in contemporary practice are models, such as linear and nonlinear Kalman filters and other "hidden Markov models" that capture practically important uncertainties through integration of deterministic and probabilistic formal analysis.

The mathematical side of probability is marvellous to behold, and remains under active development after three centuries, but it is not an immediate concern in this discussion. On the science side, what I am calling formal subjective probability was taken seriously by leading 19th Century mathematicians and scientists, such as Boole and Maxwell. As reviewed by Keynes (1921) and Porter (1986) the subject was actively debated by leading British academics in the latter half of the 19th Century. One position that emerged from the controversies of that time is the now familiar stance that probabilities meriting serious scientific consideration should have objective bases in long run frequencies. Many scientifically important formal probabilities are indeed approximately determined from observed quantities in well identified formal data structures, and vastly more are indirectly obtained by computation from such empirical inputs. When interpreted informally, however, the term probability is inseparable from logical interpretation as a measure of someone's uncertainty. Much is lost when probability is wholly converted from an instrument of logic into an empirical concept little different from a physical dimension or a population count. The richness of probabilistic thinking depends on understanding the interaction of mathematical theory with informal perceptions of uncertainties associated with precisely definable outcomes.

In a mathematical framework, probabilistic thinking can be expressed as applying to hypothetical examples, but in scientific applications it involves quite specific objective phenomena and questions, such as whether the surface climate of our planet will warm substantially in the next century (Houghton *et al*, 2001). A key to the bridging problem is to build credible formal probabilistic representations, adding a hard skeleton to informal judgments that are too often malleable and influenced by poorly analyzed assumptions. Statistical applications of broad sense probability have been too long held back by dated controversies over the obvious logical content of probability.

A simple schematic of logicist statistical inference identifies three aspects:

1. Two distinct modes of statistical inference exist, here referred to as S-inference and P-inference, my abbreviations for “inference based on sampling distributions”, and “inference based on posterior distributions”. Both S-inference and P-inference are routinely used in daily practice, and are only partially separable in their domains of application.
2. Both S-inference and P-inference rest on a common scientific concept of probability abbreviated here to FSP for “Formal Subjective Probability”. FSP was described and defended at length in Dempster (1998b). A numerical instance of FSP is linked to the unknown truth of a defined objective fact. “Formal” is used because every FSP is an element of a formal mathematical model of a specified objective situation. “Subjective” refers to the circumstance that scientific interpretation of a numerical probability is simply a degree of certainty of a subjectively imagined “you”. Along with this interpretation goes understanding that a specific FSP represents uncertainty tied to specific limited evidence or uncertain information concerning the unknown truth of the defined objective fact. Good probabilistic science, like any good science, requires credible and cogent evidence to support an assumption, and a readiness to rethink in the face of conscious and balanced arguments and data.
3. The motivating interpretation of FSP is forward-looking or “predictive”. But when new information is judged to alter a previously understood state of evidence sufficiently to invalidate a probability model, an original predictive probability may no longer be “live” in the sense of retaining validity for prediction. In particular, if the truth of a fact in question is established or disproved with certainty by new evidence, as for example by a precise and accurate observation, then the original probability is completely “dead”, and the term “postdictive” becomes a propos for its new interpretation. S-inference is inference based on postdictive interpretation, while P-inference is based on predictive interpretation.

My perception of statistical inference has developed over 40 plus years, mostly in ways consistent with today’s presentation. A tour can be had from Dempster (1964, 1968a, 1968b, 1971, 1974, 1976, 1983, 1990, 1997, 1998a, 1998b, 2002). My thinking about predictive inference is colored by the theory of probabilistic reasoning variously referred in a spectrum of literatures as “theory of belief functions”, or “Dempster-Shafer theory” that is fundamentally logicist (Shafer, 1976, Kohlas amd Monney, 1995). While resting on a concept of information combination from independent sources that generalizes the conditioning mechanism of Bayesian logic, the D-S formulation further extends Bayesian logic by admitting the common sense feature of allowing formal predictive statements to express degrees of ignorance beyond probabilistic uncertainty (Dempster 1966, 1967,1968a, 1968b, 1988). Just as Bayesian logic is increasingly being pressed into service to cope with complexities of contemporary science, the natural D-S extension is likely to find application in a widening range of scientific situations.

The remainder of the paper expands on the themes introduced above. I continue in Sect. 2 with further comments on Fisher, followed by a discussion of FSP in Sect. 3.The key concept of a parametric family of stochastic models is analyzed in Sect. 4. Fisher's terminology "forms of quantitative inference" is reflected in the titles of Sects. 5 and 6 that describe the concepts of postdictive and predictive reasoning, respectively. Some diverse commentary is collected in Sect. 7, including a plea for consensus.

2 More on Fisher

Fisher was a distinguished scientist, of wide learning and a broad talents. In statistics, he led through identifying, naming, and communicating important inferential ideas, forged in direct contact with his own scientific work and that of colleagues. His methods were often new at the time. Many remain central to contemporary applied statistical practice. While Fisher left an unpolished set of norms and principles for uncertain inference, I find that his understanding easily encompasses frequentism and Bayesianism, and can evolve to meet present and future problem-solving needs of statistical sciences.

My task here is to address aspects of Fisher inadequately represented in critiques and evaluations that continue many decades after his death (Aldrich, 1997, Barnard, 1992, Dempster, 1998a, 1998b, Edwards, 1997, Efron, 1998, Fienberg and Hinkley, 1980, Neyman, 1967, Rao, 1992, Savage, 1976, Zabell, 1989), in addition to a full length biography by his daughter Joan (Box 1978), and scholarly collections of writings (Bennett 1971, 1990). Many appreciations have described his ingenious and insightful derivations of sampling distributions, and his major contributions to the corpus of now standard design and analysis tools. Instead, I focus on the way he understood probabilistic reasoning under uncertainty, because it defines the intellectual matrix in which his better known innovative methods developed.

The inferential logic that is an integral part of statistical science was jump-started by two major papers on "estimation" (Fisher, 1922, 1925). By isolating and naming concepts, including "specification", "likelihood", "efficiency", "sufficiency", and "maximum likelihood", Fisher gave substance and generality, and ultimately wide recognition, to fundamental ideas surrounding statistical information. By hindsight, a scattered and unfocused history of related concepts can be traced over several earlier centuries, but Fisher put what I am calling S-inference on the scientific map. Although he attempted to do the same for P-inference with his "fiducial" argument (e.g., Fisher, 1930, 1935b, 1939, 1956, 1960), he did not achieve the coup that he may have hoped for. The progress of P-inference has been slower and more controversial at every step.

Fisher's innovations derived from natural genius interacting with Cantabridgian perceptions of mathematics as a servant of science. In particular, his attitude to "probability" was more pragmatic than rigorous in the sense of 20th Century mathematics. His mathematical education occurred in an objectivist environment, whence especially his early writing reflects prejudices of his teachers, for example, in opposition to "inverse probability", as Bayesian posterior inference was often called at the time. But the understanding of inference developed by the mature Fisher was much broader. Late in life, at a time when he had mostly lost his audience, he wrote discussions of his conception of probability (Fisher, 1958, 1959). But even his admirers pay little heed to his career-long repetition of terms such as "inductive inference". Statisticians should adopt Fisher's embrace of probabilistic logic. First, however, it is necessary that FSP be recognized and developed as an essential foundation of statistical science.

3 Formal Subjective Probability

Like the mountain, FSP is there, and begs mastering before the practice of statistical inference can be fully understood and made operational. Whereas in a conventional view of the statistical scene different perspectives on probability are adopted by frequentists and Bayesians, almost as defining characteristics of the respective sects, I argue on the contrary that FSP identifies a common foundation supporting viewpoints that are hard to reconcile only in the narrow sense that they pose and answer different questions, but essentially within a common framework.

In particular, FSP is a common support for both S-inference and P-inference. A sampling distribution is a property of a formal stochastic model, and describes the uncertainty of a formal user, the ubiquitous "you", looking forward to a "random" outcome not yet observed. Because the probabilities specified by a typical sampling model are determined only after the values of parameters are also specified, there are mysteries about where such models come from, and more mysteries about the usefulness of an idea whose fundamental interpretation depends on parameter values that are assumed to be unknown. I return to these mysteries in Sect. 4. There is, however, no mystery about the interpretation of a numerically determined sampling probability as characterizing uncertainty about a prospective sampling process. On the other hand, P-inference involves similar prospective characterizations of uncertainty referring to remaining unknowns after sample values are observed. A switch from S-inference to P-inference does not imply a change in the meaning of probabilities, only a change in the circumstances of application, sometimes accompanied by a switch from predictive to postdictive interpretation, and sometimes indicating a need to recompute probabilities for predictive interpretation. These are matters for professional judgment, not philosophical dispute.

Since S-inference and P-inference are the two principal branches of operational statistical inference, recognition of their dependence on the common conception called here FSP is enormously important as a major step toward eliminating roadblocks to consensus among statisticians. Once FSP is perceived as the sine qua non of statistical inference, its invisibility as an explicit element of most statistical writing seems almost surreal. Because Fisher implicitly saw FSP as a natural piece of the statistician's technical equipment, whereas Neyman steadfastly rejected it, the stage was set for protracted controversies and polemics that rumbled back and forth from about 1934 to Fisher's death in 1962 (for example Fisher, 1934, 1935a, 1939, and Neyman, 1935b, 1941, 1961), and beyond as Neyman continued to describe and defend his position (Neyman, 1977). Neyman resorted to a combination of frequentism and behaviorism as his preferred way to retain part of what FSP entails, while avoiding reference to mental processes or subjectivity that he and his followers believed to be inconsistent with science. In Sect. 7, I review Neyman's "inductive behavior" philosophy from the perspective of current realities, and argue that it is too narrow to support a full system of inference, hence is unsuited to underlie scientific practice. Behavior can be observed without reference to rational thought, but an explanation of behavior that makes no reference to the agent drawing logical inferences from evidence eliminates essential meaning. Logicism can rationalize behavior, but not vice versa.

Here are salient features of the unifying concept of FSP:

1. Objective sources for FSP in repetitions are basic, but so are subjective sources, including judgments about what to count to obtain a relevant frequency probability, and about connections of FSP to formal reasoning about specific unknowns. By maintaining rigorous formal structures and rules, along with informal professional standards of practice, formal subjective reasoning transcends criticisms of being pseudoscience coming from doctrinaire believers in purely objective science.
2. The important purely formal side of FSP is especially prominent in statistical literature, including much accumulated knowledge about practical computation as well as justly admired abstract mathematics, but the connections of FSP to informal reasoning and judgment are equally necessary for uses in science and for decision-making.
3. Specificity is key to relating FSP to inference. To have meaning outside an abstract mathematical structure, a specific numerical probability must be associated with a unique real world situation whose uncertainty is or was in play. A live FSP is predictively interpretable, and should be a reflection of specific formalized evidence and information. A live FSP must be revised when evidence and information change. A dead FSP can only be interpreted postdictively.
4. Inferential computation fuses together empirical data and constructed models, with both types contributing knowledge about an objective

> situation. Such computation, involving formal subjective probabilities, is in effect a process in a formal logic of uncertainty.

Too many academics fixate on narrow aspects of FSP, such as "subjective belief" FSP, or repeated copies of an FSP that make up a long run frequency, or Fisherian FSP derived from a "hypothetical population" with "no recognizable subsets" (Fisher, 1956, Cox, 1998), or for that matter a Dempster-Shafer FSP based on a "frame of discernment" with an associated "basic probability assignment" (Shafer, 1976). These are not separate or indeed separable concepts.

4 What Is a "Family" of Stochastic Models?

In statistics the term "random variable" is by widespread convention interpreted to signify a quantity with an objective real world basis, or more precisely with a value determined by a real world "random" process. The assumption of an objective foundation implies that the value of a specific random quantity is either directly obtainable from observations, at least in principle, or is indirectly obtainable by recursive computation from values of other quantities, both random and nonrandom, that can be traced back to such observable-in-principle quantities. A random quantity may also be called "stochastic", a term more often applied to an evolving or spatial system. A stochastic model typically has many random quantities assumed jointly "distributed" according to a "family" of multivariate probability measures, where a family member is identified by fixing a point in a defined "parameter" space, whence the term "parametric family". In modern practice, modeling a situation of realistic complexity generally requires multiple varieties of randomness, such as "hierarchical" randomness among and within populations, or random trajectories and random fields over physical time and space. Several parametric families then appear simultaneously within a statistical model of a complex system.

Theoretical statistics has been centered around the notion of parametric families of probability measures for at least the 75 years since Fisher coined the term "specification" for them, and put specifications at the center of his theory of estimation. In what sense does a specification "specify" an applied situation? A common but inadequate answer follows the rhetoric of random processes. To wit, analysts are assumed to "know" that an "unknown" member of an agreed family of random processes "generates" the empirical system under study. Two questions dog this attitude. First, since the assumed existence of objective "randomness" in the data-generating process is dubious in most situations where parametric models are applied, can one sensibly characterize a basic task of scientific inference as being to "decide" among hypotheses about such a process? And second, since much rides on the choice of family, how does one come to knowledge of a working specification?

Jakob Bernoulli introduced (see Todhunter, 1865 or or Stigler, 1986) what is now called the binomial sampling model, or "Bernoulli trials", and implicitly with it the first precise example of S-inference. The parameter in his urn model was an objective quantity, namely, the proportion of red balls in an urn known to contain only red and black balls. But he was explicit that his motivation for the model was to address the uncertain length of life of an individual of a given age. The urn model represents a situation that has analogies to the length of life question, but differs in that the specific choice of a relevant underlying human population is ambiguous at best, unlike the urn model situation. Similarly, when a few decades later Bayes introduced P-inference the artificial illustration leading to his uniform prior distribution concerned the objective stopping place of a ball randomly set in motion on a table. But he too meant his method to address realistic inferential situations where the binomial parameter is not easily objectified. The objective character of the "binomial p" parameter in routine statistical applications is sometimes spelled out precisely, as in close analogs of the urn model where a physical population is identified. But in many other cases, such as in widespread applications of logistic regression models, it is rare to have available an objective population underlying random choices given each specified vector of values for predictor variables.

The difficulty of assuming belief in an objective randomness that does not exist can be resolved only by recognizing a different reality, namely, that inference rests on subjective choices creating a proxy for the external world, not an approximation to a nonexistent objective process. The examples of uncertain reasoning from proxies illustrated by Bernoulli and Bayes easily scale up from their artificial examples to more general real world contexts that constitute the basis of applied inferential practice. The first question raised above is thus partially resolved because there is no need to believe in objective random processes. It remains, however, to address what it might mean to estimate parameter values, when it is understood that they lack objective existence. The mystery surrounding the first question is only partially resolved.

The second question is also fundamental, and has no simple answer. The truism that most acceptable formal probabilistic assumptions are at least loosely traceable to empirical frequencies is only a start. Consider random sampling from a normally distributed population, a parametric assumption that is a close competitor to binomial sampling in the statistical modeling sweepstakes. Unlike the binomial case, the normal model places a strong condition on the contents of the hypothetical urn, namely that the numerical values attached to each of the large number of balls in the urn are distributed so that a normal plot of their values is effectively a straight line. The question here concerns the prior assumption of such a specific parametric form. My sense is that most statisticians rather casually believe that they can often adopt parametric assumptions with little risk after plotting "the" data in ways that create "checks" on the specification. There is, however, a basic issue to consider. Data analysis can only address whether the particular "realized" family member associated with the sample data under analysis has a normal or other assumed form, not whether inferential computations involving a large cast of family members can be supported. Any parametric family assumption that happens to include a normal population that fits the data well is equally supported by a diagnostic such as an empirical normal plot.

So the second question remains: how to construct an empirical basis for an adopted specification, such as that of sampling a normal population? An acceptable answer, as far as it goes, is "use past experience with similar situations". Ideally, the statistician identifies and studies a collection of similar situations, hoping to find empirical shapes lying within plausible model-checking tolerances of some member of a hypothesized family. In practice, this prescription starts from a subjective judgment that certain situations encountered in the past are similar enough to the present situation that it is reasonable to assume a common family, and proceeds to an informal integration of what is known empirically about these situations to obtain a manageable and adequately inclusive family. In the end, the working statistician informally assesses the available evidence and takes a chance when pairing a precise parametric model with a particular new situation. Another possible use of a collection of past situations is to aid construction of a Bayesian prior distribution that reflects variation of parameter values across situations. Note, however, that using the collection to motivate a prior distribution assumes that the particular situation under analysis is effectively randomly sampled from the collection, a strong assumption going a step beyond motivating an assumed parametric family.

Parameters appearing in practice are often described as having "fixed but unknown" values. The "unknown" characterization is similarly understood by Bayesian and non-Bayesian statisticians alike. The label "fixed" suggests, however, that parameter values were determined by some objective nonrandom process. This harks back to the attitude that parameters are objective scientific quantities, which attitude was characterized above as holding only in limited special circumstances. When the concept of parameter is widened to include its subjective role in a mixed subjective-objective theory of inference, the assumption of an objective "fixing" process evaporates. It is therefore important in practice to separate situations where parameters can be assumed to have an objective basis from situations where by contrast they are no more than logical constructs, since in the former case they are legitimate final targets for P-inference, whereas in the latter case they have technical roles only.

If it cannot be maintained that parameter values are objective properties of data-generating mechanisms, then "chance" is no more than a label applied to FSP in certain familiar situations like "games of chance" or "randomized" statistical designs. The proposition that chance is a subtype of FSP brings with it the corollary that parameters determining chances are likewise instruments of logic, and resemble probabilities in that while sometimes identified with objective quantities they are not required to be such. Specified parametric families are in any case hypotheses constructed on the way to making inferential assertions.

5 Forms and Principles of Postdictive Inference

When an outcome that was predicted to have occurred with specific probability p, is in fact observed to have occurred, and consequently has posterior probability unity, then the "expired" original probability p can only be interpreted postdictively. Unless p is small, a postdictive interpretation has little or no impact on the interpreter, but if p is small, such as $p = .01$, then one may be entitled to register some surprise at the implied occurrence of of an improbable event, and go from there to draw tentative conclusions, perhaps about the effectiveness of a treatment in a randomized trial, or perhaps about the inappropriateness of a model assumption. The logic of such an inference is the essence of what Fisher calls a "test of significance".

The logic of postdictive interpretation is notoriously slippery. Something highly improbable always occurs in all but the most simple of observational complexes. The odds against particular observed outcomes are typically astronomical, so postdictive interpretation is typically restricted to "tail areas". But even here the question of what to select as a potential "extraordinary" (Laplace, 1814) outcome is critical, especially if the outcome is selected after preliminary inspection of data, as in Laplace's example of randomly ordered letters forming CONSTANTINOPLE. Fisher sometimes used the term "proof" to describe the finding that a null hypothesis is postdictively implausible, but proof in this sense is distinct from the mathematician's concept of formal proof. In particular, there is nothing in the logic that directly supports any alternative hypothesis that is often implicitly accepted to replace the null. Worse, there are countless examples where untrained users confuse postdictive interpretation of a small tail area p with a posterior probability that a null hypothesis holds, fallaciously implying that $1 - p$ is the posterior probability of some implicit alternative hypothesis.

Many good articles have warned against careless interpretation of Fisherian tests (e.g., Cox, 1977). In the end, however, the idea of postdictive interpretation is simply the FSP-based version of the fundamental principle that science progresses through the use of data to falsify hypotheses so is an inevitable element of the treacherous processes of improving scientific knowledge through new hypotheses with extended life cycles. In this spirit, Neyman and Egon Pearson invented a theory directed at aiding choices of suitable Fisherian tests (Neyman and Pearson,1928). Fisher was hostile to the Neyman-Pearson theory, unwisely in my opinion. He would have been better advised to see it for what it is, namely, a theory that evaluates procedures, not a replacement for the logic of significance testing. See also Sect. 7 below.

A concept that Fisher isolated, and thereby opened up for theoretical study and direct and indirect roles in applications, is the fundamental concept of "likelihood" that he originally defined (Fisher, 1922) essentially as the probability of what was observed (i.e., "the data") as a function of undetermined parameter values. There are many theoretical reasons, obvious ones from a Bayesian perspective, and more subtle ones from the frequentist decision-analytic perspective of Wald (1950), for recognizing the information-bearing quality of the likelihood of observed data. Once the FSP interpretation of stochastic models is recognized, however, it is evident that the direct interpretation of an observed likelihood is postdictive.

As Fisher made clear, likelihoods permit sensible interpretation only as ratios among the different parameter sets that are being compared. Specifically, observed likelihood provides relative probabilities of a set of outcomes that are known to have occurred. If two hypotheses are compared this way, and one is relatively unlikely, say only .01 as likely, as the other, then the relatively unlikely outcome should be regarded with suspicion. Fisher gave examples where likelihood functions were used to rule out parameter values regarded as too unlikely relative to the maximum of the likelihood, the implicit logic being postdictive as I use the term. Long ago I proposed extensions of the postdictive logic of likelihood testing (Dempster, 1974). Similar procedures were independently proposed more recently by Aitkin (1997). The appeal of these procedures depends in an essential way on recognizing the connection of likelihood to postdictive interpretation.

From the FSP perspective Neymanian confidence statements amount to a separation of parameter value sets into those accepted and those rejected by an associated family of significance tests. Fisher explicitly makes this point already in 1935 correspondence (Fisher, 1935, p.187). Neyman of course had something quite different in mind. When FSP is rejected, as Neyman advocated, neither predictive nor postdictive interpretation has any meaning. A consequence of frequentist dominance in academic statistics is that few if any current textbooks make the predictive/postdictive distinction clear. They cannot because FSP is not explicitly recognized. Many, perhaps most, users mentally interpret confidence statements as predictive, when the formal model supports only postdictive. While the frequentist "long run" interpretation preferred by Neyman and his followers is mathematically sound, it translates into a logical statement about something in the real world only insofar as a specific real world long run is identified. The typical user of a confidence statement makes no effort to define a real world long run, because he or she is rarely interested in a logical statement about any specific long run that might be implicit in the frequentist justification of a reported confidence statement. There is a further logical point to be made here, namely, that a "law of large numbers" that underlies a deterministic inference about a long run is a mathematical statement with premises and conclusions. The conclusions cannot have logical force unless the premises do also. Thus a claimed logical inference about a long run is simply void unless the input random variables are allowed logical interpretation, something Neyman is at pains to deny. Surely, however there is no magic whereby the consequences of assumptions acquire meaning while the assumptions have none. It appears therefore that FSP can justify inference about a long run, while frequentism alone cannot.

Bayesian adherents are also typically unenthusiastic about postdictive interpretation, but since P-inference in practice rests on predictive interpretation of FSP, it is possible at least to raise and debate the question in FSP terms, as I often did with Jimmie Savage in the 1960s. He staunchly defended his Bayesian decision-theoretic outlook, referring to the kinds of slipperiness of postdictive logic alluded to above. In the more ecumenical attitudes of the 1990s, however, my sense is that some Bayesians at least are willing to think about postdictive logic in Fisher's terms, and even to see Fisher as a friend. After all FSP is the basic thing, and postdictive logic is a commonplace form of reasoning with FSP in the context of a specific uncertain judgment.

6 Forms and Principles of Predictive Inference

The term predictive applies to statements about a factual assertion whose truth or falseness is objectively meaningful but is currently unknown to “you”. Hence usage is not limited to predicting an unfolding future that cannot be currently observed because defining events have not yet occurred. In my logicist formulation, the formal “you” is assumed to be working with an FSP-based model at the time predictive interpretations are made. The following discussion contrasts three approaches to prediction. The main logicist option available today for applied statistical practice is Bayesian inference, often credited in its contemporary form to Bruno de Finetti. See Bernardo and Smith (1994). I believe, however, that a more suitable paradigm is embodied in the D-S theory introduced briefly in my introduction. D-S theory is logicist in a fundamental way because it integrates nonprobabilistic “propositional” logic with probabilistic reasoning. The third approach to prediction discussed below is nonlogicist, and in some versions nonprobabilistic.

Historically, Bayesian predictive logic was well understood by many 19th Century scientists, especially astronomers familiar with the writing of Laplace and Gauss. Early advocates of nonlogicist frequency probability were also active from about 1850 onward, engaging in controversy with proponents of Bayesian “inverse” probability. In the 1920s and 1930s, Fisher and Neyman attempted in different ways to address the fundamental puzzle of inference from data “drawn” from parametric stochastic models. Their focus was on S-inference, deliberately avoiding the controversial topic of Bayesian priors. Fisher led with his theory of estimation, and then direct inference from likelihood, and finally, most controversially, his fiducial method introduced in Fisher (1930). Neyman soon followed with “confidence” statements that he originally conceived as making sense of Fisher's fiducial method. Most statisticians felt they could follow Neyman's reasoning, but not Fisher's, and in due course Fisher's thinking on the subject was largely dismissed, and ignored in teaching and practice. The fiducial method was, however, a source of D-S theory.

My own view has long been that Fisher was asking the right question by insisting on a predictive interpretation for predictive intervals. The theory behind confidence regions provides ingenious and mathematically correct answers to precisely formulated questions, but insofar as these answers have logicist interpretations, these interpretations can only be postdictive. If the task is defined as prediction, Fisher (1939) was on the right path, and Neyman (1941) had indeed taken a wrong turn. I suggest that future generations need to take a hard look with an open mind.

The Bayesian paradigm, as illustrated by many studies of complex phenomena in the 1990s (e.g., Gilks *et al*, 1996) directs the user to set up a model with a data structure that is detailed enough to adequately capture a set of objective phenomena, and then to proceed to develop relations that allow probability judgments about any and all objective unknowns specified in the model. Although the data structure often reflects deterministic relations, as in hierarchical classification structures, these are generally left implicit without drawing attention to obvious expressions in terms of formal propositional logic that parallels the predictive probabilistic logic. In Bayesian statistics, the explicitly considered nonprobabilistic facts are "the data" and these are combined with "prior" probabilistic relations via computation of conditional probabilities, as laid down by Bayes's rule. The habit of hiding propositional relations while exposing probabilistic relations, is presumably due to the suppressed recognition of logicism in the thinking of most statisticians, including enthusiasts for modern Bayesian tools.

The core of Bayes's remarkable contribution to the history of probability lies in his perceptive double use of the concept of conditional probability, especially noteworthy because he had first to craft the rule in terms that made his double application transparent. In the first application, the sampling density of observables given a parameter is written down as a function of the undetermined parameter value. This conditional density is then multiplied by the prior density of the parameter to form the joint marginal density of parameter and observables. The second application reverses the role of parameter and observables, but applies the same abstract rule to form the conditional density of the parameter given values for the observables. My reason for laying out this familiar story is to make plain that the concept of parametric family is just as central to Bayes as it had been to Bernoulli 50 years earlier. Fisher in the 1920s gave explicit names to concepts that Bayes and others through the 18th and19th Centuries had left implicit, calling the assumed sampling model the "specification", and renaming the density of the specification to be the "likelihood" when the values of observables are fixed and substituted into the sampling densities of the family.

It was part of the mindset of Fisher and many of his contemporaries, as it had been for a preceding generation (e.g., Edgeworth, 1884) that a precondition for Bayesian analysis is an empirical basis for the requisite prior distribution, an implication being that the usefulness of Bayes for science is thereby strongly limited. The post Fisher era disputes between Bayesians and Neymanians that reached a peak around 1960 turned mainly on the same issue (Pearson, 1962). The issue here is real, but risks trivialization. The issue is not that some priors are founded in data, while others are "made up" subjective judgments. To accept a frequency basis for a specific prior, is to make a subjective judgment of exchangeability of the specific situation with those making up a population on which prior frequency probabilities are based. When one such population presents, many others including superpopulations and subpopulations generally come in its train, whence subjective judgmental choices among populations are required. Frequency experience may also range from hard data to softer memories, thus posing choices between soft experience that is judged more relevant to the specific situation and hard data that is judged less relevant. Finally, as noted already in Sect. 4, the empirical foundations of parametric likelihood factors that are implicitly assumed by both Fisher and frequentists to be more sound than that of many Bayesian prior factors cannot be based only on the specific situation under analysis but also entail acceptance of adequate model fit to a range of other situations. The lesson is that all approaches to formal statistical reasoning implicitly assume a big and somewhat vague family of unrealized situations that includes the specific situation under analysis. As a matter of practice, the issue here may be less important than it was several decades ago because serious Bayesian applications are generally based on models that are stitched together from layers of parametric and hyperparametric models that often appear supportable scientifically, while ultimate hyperpriors are often arguably unimportant within plausible limits. There is little reason in any of this to question logicism.

The D-S perspective, as seen through my eyes, is based on the construction of a model with a formal data structure representing a defined piece of an objective world of facts. Above the data structure sits a set of probabilistic relations about which rather strong "independence" assumptions are made. Although my original motivation was to put a coherent foundation under the models and reasoning of Fisher's fiducial method, it soon became clear that the umbrella thus created includes Bayesian inference and propositional logic, and various mixtures, within a logicist stew. Shafer (1976) was responsible for lifting the theory beyond its narrow source in applications to parametric sampling theory, and may have inadvertently put up a barrier against probabilists and statisticians by introducing wholesale new terminology such as "frame of discernment", "basic probability assignment" and "commonality function" for what are essentially a sample space, a probability measure , and a characteristic function. Clearly, however, statistics was not ready for these ideas 30 or 20 years ago, or even today. There have all along, however, been pockets of interest in fields open to the "artificial intelligence" implicit in logicist thinking. Among the jungle of proposals that accompany attempts to reason formally about uncertainty, the D-S theory survives and I believe will grow and prosper. Detailed sorting out of the history, present state, and prospects requires a further lengthy paper.

Probability models of all kinds are built up in most examples by hypothesizing independent stochastic relations operating on different subsets of the variables of a multivariate system. A typical big and complex probability model is arrived at via "independence" assumptions. Independence has a precise meaning in the mathematics of probability theory, in terms of joint probabilities being products of marginal probabilities. Most models, especially in the hard sciences also have many nonprobabilistic relations, such as differential equations governing physical processes of motion or energy flow. In addition there are generally other nonprobabilistic logical statements, such as an assertion that a subsystem "I" is part of a larger system "K". These deterministic relations possess another kind of "independence" in that they are generally assumed to hold simultaneously, implying that they do not interfere with each other, or indeed interfere with the assumed probability laws. The key idea driving D-S theory is that there is one overarching independence concept at work both within and between the probabilistic and nonprobabilistic components of the full model. This is mathematically represented by the simple ("Dempster") rule of combination central to my first papers on the subject (Dempster, 1966, 1967) which essentially unify propositional and probabilistic logic in a coherent way.

D-S independence has a special role in relation to observations that is often insufficiently heeded by users who focus mechanically on applying formal rules of conditioning to data. The issue arises in various special manifestations of the theory, including Fisher's fiducial theory and standard Bayesian theory. The essence of Fisher's independence assumption is plain to see in the simplest example of a fiducial argument. Suppose that an observable X is a measurement of a quantity U, where measurement error $E = X - U$ is assumed distributed with a known density $f(e)$. The mechanics of the argument is to regard X as known, and to use the relation $U = X - E$ together with the density $f(e)$ to deduce the fiducial distribution of U.

The independence assumption here is evidently that E is independent of X. Most statisticians are programmed to think this is wrong, even a grievous error, based I believe on the random mechanism attitude to probability, namely that "nature" somehow first fixed U and then independently "drew" E at random. It is certainly a very fundamental part of informal scientific thought to delve into the operational, causal nature of error mechanisms. But this important scientific activity is quite different from depending on belief in a nonexistent objective random mechanism. From a logicist perspective, it is a deliberate judgment call in the course of model construction to decide on a particular type of independence. The example is too simple to convey the flavor of real modeling. In complex examples, it may not seem at all unreasonable to assume a Fisherian "pivotal quantity" (illustrated by E in the simple example) to be independent of observables.

D-S theory casts an unfamiliar light on Bayesian inference by noting that when the latter is viewed as a special case (Dempster, 1968a) there is an implicit assumption that the observables are independent of the other relations assumed in the model. Bayes's original definition of conditional probability has long seemed to me too casual in that users are not cautioned on the implicit independence assumption in Bayesian conditioning. It is not just observed facts that matter, but also whether these facts interact with the evidence that was used to construct a prior model. Data selection processes, for example, can easily bias naïve Bayesian inferences. Before Bayesian conditioning can be used, D-S independence of data and prior should be assessed and judged credible.

Logicism can provide an anchor for procedures based on nonprobabilistic heuristics. While logic and procedure are complementary (Dempster, 1968b), and logic is an important source of procedures, in most of modern statistics procedure is in effect used as a substitute for logic. For example, an area of active research concerns "classification", now often called "supervised machine learning" in terms popularized by computer scientists. The task is to learn from "training data" how to classify a statistical unit (object, person, etc.) into one of a defined finite set of classes on the basis of an observed multivariate vector of characteristic. In pure form, the learning is done by analyzing a set of samples of observed vectors of units from each of the classes. A "validation" data set is typically withheld from the training set, to be used for assessing the properties of the estimated classification rule, free from biases that would affect unadjusted assessment from the training data. Current interest centers around an approach to improving classifier performance called "boosting."

My point in raising classification and boosting is to contrast nonlogicist prediction with probabilistic logicism. It is premature to assess overall technical merit and importance of a field under rapid development, but the mindset especially among computer scientists appears to be that perfect classification is the goal. Failing that, maximizing correct classification rates on a range of trial examples is the prime objective, recognizing that "overfitting" may bias rates developed from training data alone unless adjusted from a validation set.. The objective when classifying a new unit is a discrete "decision" that is either right or wrong, evaluated through percentages of correct choices averaged over populations of assorted types. Nonprobabilistic boosting theory seeks bounds on the performance of a classifier (Schapire *et al*, 1998), much as bounds can be placed on the accuracy of a numerical algorithm. A mindset more congenial to statisticians is to postulate an intrinsic probability limit to accuracy of classification given stated data. Frequentist statisticians regard such probabilities as objective quantities, defined as hypothetical population frequencies. As is often the case, mysterious heuristics that work well are understandable as approximations to frequentist Bayesian decision procedures (Freeman *et al*, 2000). Carrying the argument one further step, full logicist justification asks for the Bayesian assumption to be credible, in the sense of being "your"

FSP representation of uncertainty, whence Bayesian posterior computation automatically takes care of overfitting without recourse to a validation set.

My criticism here is basically that the applied statistician's duty is not primarily knowledge derived from mathematical assessment of procedures. This is important work for academic theoreticians, but in actual practice the need is for case by case evaluation of the science-based assumptions that underlie the specifics of each inference and subsequent policy recommendation. The machine learning literature, for example, is typically silent on the sources of training and validation data sets, largely ignoring the detailed knowledge and understanding of scientific context that are central to applied predictive inference. While frequentist statistics provides tools to evaluate procedures, as far as I can see only FSP provides a direct entree to situations where specific uncertainties merit quantification. Denying FSP and associated probabilistic inference in effect pulls the rug out from under the science.

7 End Notes: Models, Behavior, and Decisions

Formal models and related formal inferences are thin constructed overlays on rich complexes of informal knowledge and understanding. They are tools that should not be confused with the realities that they are designed to illuminate. Discussion about whether models and inferences are right or wrong is misplaced. The issue is rather whether formal constructs accomplish the goals for which they are intended, namely to support and quantify informal judgments, always accompanied by specific concerns and safeguards against overdependence on fallible and limited idealizations. Nevertheless, since real people have only vague thoughts about uncertainty, they can derive major benefits from carefully crafted crutches that draw on FSP. Ultimate scientific reports necessarily use informal natural language to describe uncertainty, but when accompanied by accurate perceptions of the implied meaning of formal probabilistic inferences, including postdictive and predictive interpretations of FSP, formal statistical inferences add valuable quantitative support to assessments of uncertainty.

The approach described as proceduralism in Sect. 1 is quite different. The proceduralist approach to inferential analysis consists of selecting and applying a procedure, then reporting results for interpretation by the user. Selecting a procedure can be described as behavior, and hence can be rationalized as having occurred as the result of a choice among alternative possible behaviors. The decision to select and use a particular procedure is governed by demonstrating desirable properties of the procedure. Neyman (1955, 1957) promoted the term "inductive behavior" for this attitude to statistical analysis, where his use of the term "inductive" was evidently designed to provide a counterweight to Fisher's repeated use of the same term in relation to various styles of probabilistic reasoning about data. It was also in tune with a widespread belief in behaviorism as a philosophy of science, namely, the thesis that only phenomena that are objectively observed can be studied scientifically, and hence we should attempt only to construct a statistical theory to guide decisions that are a matter of objective record, not purely mental processes that cannot be directly measured.

I believe that behaviorist philosophy is less heard from at present, now that human brains can be observed more directly, and computer models move closer to mimicking human reasoning capabilities, thus providing formal models of such reasoning. Historically, however, the concept of decision-theoretic optimization of choices among procedures is a powerful metaphor for practice that characterizes much of the fundamental development in mathematical statistics that peaked about 1950. It has its roots in 18th Century discussions of what constitutes the "best mean" (Todhunter, 1865). Fisher's (1922, 1925) theory of estimation put point estimation on a modern track, especially through its identification of "maximum likelihood estimation" and the associated asymptotic "efficiency" concept. The Neyman-Pearson (1928) theory of hypothesis testing put an analogous choice-of-procedure foundation under Fisher's significance testing methodology. Wald's (1950) frequentist decision theory created an inclusive framework for both estimation and hypothesis testing, and included real world decisions, such as choices among different rules to underlie investment decisions. Wald's recognition of the "game against nature" analogy, and his fundamental theorems about "minimax" and "admissible" procedures that tied frequentist decision theory to "Bayesian decision rules", represent the high point of frequentist development.

There is a fundamental confusion at the root of statistical decision theory. Does the term "decision" refer to making a choice among different procedures for a class of applications, such as different procedures for testing a mathematically defined null hypothesis, as envisaged by inductive behavior? Or does it refer to statistical "decisions" in the sense of outputs of procedures, such as deciding to "accept" or "reject" a null hypothesis? Parallel questions concern the role of stochastic model assumptions. Is an assumed model primarily used to compute operating characteristics of different candidate procedures? Or is it used to compute inferential statistics such as confidence coefficients or p-values in a particular application of a specific procedure? The first of each of these pairs of questions is much closer to development of mathematical theory than it is to practice. As always in frequentist theory, there is no clear connection between what the theory usefully and legitimately means concerning long run outcomes, and how it helps convey a clear message about a specific data analysis at hand.

In considering when decision theory is applicable, a distinction is appropriate between the action of choosing and implementing one statistical procedure over another, and the action of choosing one formula over another and actually using it in daily life, for example, to determine what one decides is appropriate to pay for a security. A cost function is required for decision-theoretic analysis in both types of situation, but in the former case net cost is typically represented as a "loss function" of convenience, such as an expected squared error or an expected misclassification rate, while actions carried out in the real world have consequences that relate directly to costs with measurable economic, social, or medical effects. It is arguable that only the latter type constitutes an operationally meaningful application of decision theory. Either way, there are strong arguments, ranging from Wald's frequentist admissibility theorem, to axiomatic arguments such as those of Savage (1954), suggesting that if decisions must be made they ought to be made Bayesianly.

A difficulty with proceduralist decision theory is that users of statistical methods do not always understand the arms length separation of proceduralist advisors from in situ needs of scientists and professionals to draw inferences and make choices in specific situations. Although Neyman and Pearson and many distinguished theoreticians following them have been very clear in their expositions about the frequentist meaning of Type 1 and Type 2 errors, the statement that a hypothesis H is rejected at the .05 level is very widely misinterpreted by scientific users as meaning that there is probability .05 that the null hypothesis is true and .95 that it is false. Tukey (1961) refers to this as "badmandment number 100" and Lindley (1997) calls it "the fallacy of the transposed conditional". From the logicist perspective, the problem is that an FSP that can legitimately be interpreted postdictively is being illegitimately interpreted predictively, in large part because several generations of statistics teaching have not explained the natural postdictive logic of S-inference.

The Bayesian school would prefer to get rid of S-inference completely. This effort is futile, however, for we cannot replace Fisherian significance testing by a purely Bayesian approach without compromising the entrenched principle that science often advances by comparing observations against prior predictions of those observations, which is after all exactly what Fisherian testing does. Bayes factors (Kass and Raftery, 1995) offer one way to "select" a model, by comparing marginal likelihoods averaged over prior distributions of parameters. An alternative when the focus is on comparing alternative parametric forms is to use likelihoods in their original unaveraged form to compare distributions of parametric likelihoods in nonaveraged original forms (Aitkin, 1997, Dempster, 1974, 1997). Either or both of these methods may be used to assess fit of data to prior predictions, depending on the question asked. My preference is generally not to extend a Bayes factor ratio to a full Bayes posterior of a pair of competing hypotheses, since this assumes that choice of a mixture model with priors on the components has already been made. Direct postdictive interpretation of a Bayes factor ratio is on its own capable of supporting a modeling decision to drop a relatively "unlikely" component model. Professional judgments can differ here, without much practical effect.

Fisher (1935b) remarked that his pretentious title "The logic of inductive inference" might equally well have been "On making sense of figures". I believe that wider recognition of logicist inference offers a way forward for a statistics profession that is steadily losing ground to newer technology-based specialties strongly allied to computer science. We need to reassert leadership concerning quantitative assessment of uncertainty. I believe that the logicist thinking surveyed here offer a common sense approach that absorbs methods seen by too many colleagues as conflicting and contradictory. A healthy future for statistical professions may depend on developing a critical mass that can break down unnecessary roadblocks to unity.

Bibliography

AITKIN, M. (1997). The calibration of P-values, posterior Bayes factors and the AIC from the posterior distribution of likelihood (with discussion) *Statistics and Computing.* **7** 252–273.

ALDRICH, J. (1997) R. A. Fisher and the making of maximum likelihood (with discussion) *Statistical Science* **12** 162–176.

BARNARD, G. A. (1992). Review of statistical Inference and analysis: Selected correspondence of R. A. Fisher (ed. J. H. Bennett) *Statistical Science* **7** 5–12.

BENNETT, J. H. (1971). *Collected Papers of R. A. Fisher 1–5* Univ. of Adelaide. (Referred to below as JHB1, JHB2, etc.).

BENNETT, J. H. (ed.) (1990) *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher.* Oxford.

BERNARDO, J. M. and A. F. M. SMITH (1994) *Bayesian Theory.* Wiley.

BOOLE, G. (1854). *An Investigation into the Laws of Thought.* Walton and Maberly, London (reprinted 1951, Dover, NY).

BOX, FISHER, J. (1978). R. A. Fisher. *The Life of a Scientist.* Wiley.

COX, D. R. (1977), The role of significance tests (with discussion) *Scand. J. Statist.* **4** 49–63.

COX, D. R. (1998), Comment *Statistical Science* **13** 114–115.

DEMPSTER, A. P. (1964). On the difficulties inherent in Fisher's fiducial argument. *J. Amer. Statist. Assoc.* **59** 56–66.

DEMPSTER, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics* **37** 355–74.

DEMPSTER, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics* **38** 325–39.

DEMPSTER, A. P. (1968a). A generalisation of Bayesian inference (with discussion). *J.Roy. Statist. Soc.* B **30** 205–247.

DEMPSTER, A. P. (1968b). Crosscurrents in statistics. *Science* **160** 661–663.

DEMPSTER, A. P. (1971). Model searching and estimation in the logic of inference. In *Foundations of Statistical Inference* (eds, V. P. Godambe and D. A. Sprott) 56–78. Holt Rinehart and Winston.

DEMPSTER, A. P. (1974). The direct use of likelihood for significance testing. *Proceedings of the Conference on Foundational Issues in Statistical Inference (O. Barndorff-Nielson)* Reprinted in *Statist. Comput.* **7** (1997) 247–252.

DEMPSTER, A. P. (1975). A subjectivist look at robustness. *Bull. Int. Statist. Inst.* **46** Book 1 349–374.

DEMPSTER, A. P. (1983). Purposes and limitations of data analysis. *Scientific Inference, Data Analysis, and Robustness* (ed. G. E. P. Box and T. Leonard) Academic Press.

DEMPSTER, A. P. (1989). Construction and local computation aspects of network belief functions. *Influence Diagrams Belief Nets and Decision Analysis* (eds. R. M. Oliver and J. Q. Smith) 121–141 Wiley.

DEMPSTER, A. P. (1990). Causality and statistics. *J. Statist. Planning and Inference* **25** 261–278.

DEMPSTER, A. P. (1997). Commentary on a paper by Murray Aitkin and on discussion by Mervyn Stone. *Statistics and Computing.* **7** 265–269.

DEMPSTER, A. P. (1998a). Comment. *Statistical Science* **13** 120–121.

DEMPSTER, A. P. (1998b). Logicist Statistics I. Models and Modeling. *Statistical Science* **13** 248–276.

DEMPSTER, A. P. (2002). John Tukey as "Philosopher". *Annals of Statistics* **30** 1619–1628.

EDGEWORTH, F. Y. (1884). The philosophy of chance. *Mind* **9** 223–35.

EDWARDS, A. W. F. (1997) What did Fisher mean by inverse probability in 1912–1922? (with discussion) *Statistical Science* **12** 177–184.

EFRON, B. (1998) R. A. Fisher in the 21st Century (with discussion) *Statistical Science* **13** 95–122.

FIENBERG, S. E. and D. V. HINKLEY (1980) *R. A. Fisher: An Appreciation.* Lecture Notes in Statistics 1 Springer.

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc.* London **A 222** 309–368. (Reprinted in JHB1)

FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Camb. Phil. Soc.* **22** 700–725. (Reprinted in JHB2)

FISHER, R. A. (1930). Inverse probability *Proc. Camb. Phil. Soc.* **26** 154–57, 172–173. (Reprinted in JHB2)

FISHER, R. A. (1934). Discussion of a paper by Neyman. *J. Roy. Statist. Soc.* **97** 614–619.

FISHER, R. A. (1935a). Discussion of a paper by Neyman. *J. Roy. Statist. Soc.* **Supp. 2** 107–180.

FISHER, R. A. (1935b). The logic of inductive inference (with discussion). *J. Roy. Statist. Soc.* **98** 39–82. (Reprinted in JHB3)

FISHER, R. A. (1935c). Letter to T. Koopmans: 29 November 1935. In *Statistical Inference and Analysis: Selected correspondence of R. A. Fisher.* (ed. J. H. Bennett).

FISHER, R. A. (1939). A note on fiducial inference. *Ann. Math. Statist.* **10** 383–388. (Reprinted in JHB4)

FISHER, R. A. (1948). Conclusions fiduciaires. *Ann. Inst. H. Poincar* **10** 191–213. (Reprinted in JHB5)

FISHER, R. A. (1955). Statistical methods and scientific induction. *J. Roy. Statist. Soc.* **B 17** 69–78. (Reprinted in JHB5)

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh. (Slightly revised versions appeared in 1958 and 1960).

FISHER, R. A. (1958). The Nature of Probability. *Centennial Review* **B** 261–74.(Reprinted in JHB5)

FISHER, R. A. (1959). Mathematical probability in the natural sciences. *Technometrics* 121–29.

FISHER, R. A. (1960). Scientific thought and the refinement of human reasoning. *J. Op. Res. Soc. of Japan* **3** 1–10 . (Reprinted in JHB5)

FREEDMAN, J., TREVOR H., AND R. TIBSHIRANI (2000). Additive logistic regression (with discussion). *Ann. Statist.* **28** 337–407.

GILKS W. R., S. Richardson, and D. J. Spieglehalter (eds.) (1996) *Markov Chain Monte Carlo in Practice.* Chapman and Hall.

GOOD I. J. (1950). *Probability and the Weighing of Evidence.* Griffin, London.

HOUGHTON, J. T., et al (2001). *Climate Change 2002. The Scientific Basis.* Cambridge.

KASS, R. E. and A. E. RAFTERY (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

KEYNES, J. M. (1921). *A Treatise on Probability.* MacMillan, London. (Reprinted 1962 Harper Torchbooks)

KOHLAS, J. and P.-A. MONNEY (1995). *A Mathematical Theory of Hints. An Approach to the Dempster-Shafer Theory.of Evidence.* Lecture Notes in Economics and Mathematical Systems 425 Springer.

LAPLACE, P. S. (1951) *A Philosophical Essay on Probabilities*, Dover (Translated by F. W. Truscott and F. L. Emory from an 1840 French edition)

LINDLEY, D. V. (1997) Comment. *Statistical Science* **12** 149–152.

NEYMAN, J. (1934). On the two different aspects of the representative method of sampling (with discussion). *J. Roy. Statist. Soc.* 97 558–625.

NEYMAN, J. (1935). Statistical problems in agricultural experimentation (with discussion). *J. Roy. Statist. Soc.* Supp. **2** 107–180.

NEYMAN, J. (1935). On the problem of confidence intervals. *Ann. Math. Statist.* **6** 111–116.

NEYMAN, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika* **32** 128–150.

NEYMAN, J. (1955) The problem of inductive inference. *Commun. Pure and Appl. Math.* **8** 13–46.

NEYMAN, J. (1957) Inductive behavior as a basic concept of the philosophy of science. *Rev. Int. Statist. Inst.* **25** 22–35.

NEYMAN, J. (1961). Silver jubilee of my dispute with Fisher. *J. Op. Res. Soc. of Japan* **3** 145–154.

NEYMAN, J. (1967). R. A. Fisher (1890–1962): An appreciation. *Science* **156** 1456–62.

NEYMAN, J. (1977) Frequentist probability and frequentist statistics. *Synthese* **36** 97–131.

NEYMAN, J. and E. S. PEARSON (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* **20A Part I**, 175–240, **Part II**, 263–294.

PEARSON, E. S. (1962). Thoughts on statistical inference. *Ann. Math. Statist.* **33** 394–403.

PORTER, T. M. (1986). *The Rise of Statistical Thinking, 1820–1900.* Princeton University Press.

RAO, C. R. (1992) R. A. Fisher: the founder of modern statistics. *Statistical Science* **7** 34–48.

SAVAGE, L. J. (1954). *The Foundations of Statistics.* Wiley. (Reprinted 1972 Dover).

SAVAGE, L. J. (1962). *The Foundations of Statistical Inference.* Methuen, London.

SAVAGE, L. J. (1976). On rereading Fisher. *Annals of Statistics* **4** 441–500.

SHAFER, G. (1976) *A Mathematical Theory of Evidence.* Princeton University Press.

SHAFER, G. (1990). The unity and diversity of probability (with discussion). *Statistical Science* **5** 435–462.

SCHAPIRE, R. E., et al (1998). Boosting the margine: A new explanation for the effectiveness of voting methods. *Ann. Statist.* **26** 1651–1686.

STIGLER, S. M. (1986). *The History of Statistics. The Measurement of Uncertainty before 1900.* Harvard University Press.

TODHUNTER, I. (1865). *A History of the Mathematical Theory of Probability from the time of Pascal to that of Laplace.* Cambridge University Press. (Reprinted Chelsea 1949, 1965)

TUKEY, J. W. (1961). Data analysis and behavioral science, or learning to bear the quantitative mans burden by shunning badmandments. *Collected Works of J. W Tukey* 3 (1984) 187–390 (ed. Lyle V. Jones) Wadsworth.

WALD, A. (1950). *Statistical Decision Functions.* Wiley.

ZABELL, S. (1989). R. A. Fisher on the history of inverse probability (with discussion) *Statistical Science* **5** 435–462.

About Editors

Arthur P. Dempster (see pp. vii)

Glenn Shafer (see pp. viii)

Ronald R. Yager has worked in the area of fuzzy sets and related disciplines of uncertainty modeling for over twenty-five years. He has published over 500 papers and fifteen books. He was the recipient of the IEEE Computational Intelligence Society Pioneer award in Fuzzy Systems. Dr. Yager is a fellow of the IEEE, the New York Academy of Sciences and the Fuzzy Systems Association. He was recently given an award by the Polish Academy of Sciences for his contributions. He served at the National Science Foundation as program director in the Information Sciences program. He was a NASA/Stanford visiting fellow and a research associate at the University of California, Berkeley. He has been a lecturer at NATO Advanced Study Institutes. He received his undergraduate degree from the City College of New York and his Ph. D. from the Polytechnic University of New York. Currently, he is Director of the Machine Intelligence Institute and Professor of Information and Decision Technologies at Iona College. He is editor and chief of the International Journal of Intelligent Systems. He serves on the editorial board of a number of journals including the IEEE Transactions on Fuzzy Systems, Neural Networks, Data Mining and Knowledge Discovery, IEEE Intelligent Systems, Fuzzy Sets and Systems, the Journal of Approximate Reasoning and the International Journal of General Systems. In addition to his pioneering work in the area of fuzzy logic he has made fundamental contributions in decision making under uncertainty and the fusion of information.

Liping Liu is a Professor of Information Systems at University of Akron. He received his Ph.D. in Business from University of Kansas in 1995. His research interests are in the areas of Uncertainty Reasoning and Decision Making in Artificial Intelligence, Electronic Business, and Systems Development. His articles have appeared in *Decision Support Systems, European Journal of*

Operational Research, *IEEE Transactions on System, Man, and Cybernetics, International Journal of Approximate Reasoning, Information and Management, Journal of Risk and Uncertainty*, and others. He served on the program committees or as a track chair for INFORMS, AMCIS, IRMA, IIT, etc. He has strong practical and teaching interests in e-business systems design, development, and integration using advanced DBMS, CASE, and RAD tools. He has won two teaching awards. His recent consulting experience includes designing and developing a patient record management system, a payroll system, a course management system, and an e-travel agent, and administering Oracle databases for medium and large corporations.

About Authors

Jeffery Barnett did his undergraduate work in mathematics at UCLA and Indian University and his graduate work in computer science at the University of California, Irvine. His research interests include artificial intelligence, mathematics, and system and process architectures. He has been at the Northrop Grumman Corporation since 1984 where he was a cofounder of the Automation Sciences Laboratory. Prior to that time, he was on staff at USC/ISI and was visiting faculty at UCLA and the University of California, Irvine. He has a long standing interest in technology that will enable flexible robust artificially intelligent systems, particularly methods for representing and combining evidence and for reasoning about control of problem solvers.

Alain Chateauneuf, born in 1945 in Le Mans (France) is a graduate (ingénieur civil) of the Ecole Nationale Supérieure du Génie Maritime. He passed a Ph.D. in Mathematics at University Pierre et Marie Curie (Paris 6) in 1981. His academic career began in 1974 as assistant-professor in mathematics at University Paris 1 Panthéon-Sorbonne ; he became professor at University Paris 1 in 2000. His research areas are Mathematical Economics and Decision Theory, where he is particularly interested in the non-additive decision models.

Arthur P. Dempster (see pp. vii)

Thierry Denœux graduated in 1985 as an engineer from the Ecole Nationale des Ponts et Chaussées in Paris, and received a doctorate from the same institution in 1989. Currently, he is Full Professor with the Department of Information Processing Engineering at the Université de Technologie de Compiègne, France. His research interests concern belief functions theory, fuzzy data analysis and, more generally, the management of imprecision and uncertainty in data analysis, pattern recognition and information fusion. He is the Editor-in-Chief of the *International Journal of Approximate Reasoning,* and a member of the editorial board of *Fuzzy Sets and Systems.*

Didier Dubois graduated from the “Ecole Nationale Superieure de l’Aeronautique et de l’Espace”, Toulous, France, in 1975, and received the “Docteur-Ingenieur” degree from the same school in 1977. He also received the “Docteur es Science”degree from the University of Grenoble, France, in 1983. From 1980 till 1983, he worked as a research engineer at the Center d’Etudes et de Recherches de Toulouse, in the Production Research Area. He is presently a CNRS researcher at the University of Toulouse, Laboratoire Languages et Systemes Informatiques. He coauthored with Henri Prade two books on fuzzy sets and possibility theory, in 1980 and 1985 respectively. He is also co-editing an exchange bulletin, called BUSEFAL, about fuzzy sets and related topics. His main topics of interest are the modeling of impression and uncertainty, the representation of knowledge and approximate reasoning for expert systems, operations research and decision analysis.

Thomas D. Garvey is Associate Director of SRI’s AI Center and manages a group that focuses on continuous planning and execution, uncertainty, C2, and practical knowledge representation. He has applied these technologies to problems in electronic warfare, radar, sonar signal processing, antisubmarine warfare, helicopter mission planning, C3I, computer security, intelligence analysis and modeling biological signaling systems and the immune system. Dr. Garvey has served the Defense Advanced Research Projects Agency (DARPA) as assistant office director for Planning and C2. He has participated in high level advisory groups including the US Air Force Scientific Advisory Board, DARPA’s Information Science and Technology group, and advisory groups for specific military commands.

Jean Gordon received an A.B. in mathematics from Princeton University in 1972. She then received a Ph.D. in mathematics from Dartmouth College in 1977. Her thesis and subsequent mathematical papers were in the field of algebraic coding theory and abstract algebra. She was an assistant professor of mathematics at Williams College until 1981 when she decided to pursue a career in medicine. During her studies at Stanford University School of Medicine, she worked as a research assistant to Edward Shortliffe in the department of medical computer science. During this time, the paper published in this book was written as the culmination of their collaboration. After obtaining an M.D. degree in 1986, she completed a residency in dermatology and wrote an expert system for dermatopathology as a research interest. Alas, she is currently in private practice in Mountain View where she eradicates warts, wrinkles and rashes!

Jean-Yves Jaffray was born in Rennes (Brittany, France) in 1939. He received a degree in engineering from the Ecole des Mines de Paris in 1961 and became an associate professor in mathematics at Faculté des Sciences de Paris, later University Pierre et Marie Curie (Paris 6), in 1965. He passed his doctorat d’Etat in mathematics at Paris 6 in 1974 and obtained the following year a position of professor at University Paris 1 Panthéon-Sorbonne, where he remained three years. He moved then back to Paris 6, to the Computer Science department (LIP6) to which he still belongs. His research interest are in mathematical economics, and especially in utility theory and decision making under risk and uncertainty; part of his recent work deals with Bayesian networks.

Robert Kennes studied mathematics at the University of Brussels. After research in algebra (with Prof. G. Papy) and space aeronomy (with Prof. M. Nicolet), he joined, in 1988, Prof. Philippe Smets at IRIDIA to work, under his supervision, on belief functions.

Professor Kohlas studied mathematics and physics at the University of Zurich. He got there a PhD in Operations Research. Afterwards he worked as a senior scientist at the Asea Brown Boveri research center in the domain of automatic control of electrical power systems. 1973 he got appointed professor at the University of Fribourg, Switzerland. He was visiting scientist at the IBM Scientific Center, Paris, France and visiting professor at the University of Lausanne the Swiss Federal Institute of Technology, Lausanne, the University of Padova and the University of Kansas, Lawrence. He was member of program committees of various conferences in Operations Research and Artificial Intelligence, and he served as editor and referee for numerous scientific journals. He was appointed vice rector of the University of Fribourg from 1996–1999. His current research interest covers inference under uncertainty based on probabilistic argumentation systems, its relation to Dempster-Shafer theory of evidence, algebraic information theory and uncertain information. He published several books and numerous papers in these and related fields.

Liping Liu (see pp. 793).

Roger Logan was a graduate student in the Department of Mathematics at the University of Kansas. He joined, Professor Glenn Shafer, at School of Business, under his supervision, to study belief functions.

John D. Lowrance has been a member of SRI's Artificial Intelligence Center since 1980. He has led and participated in basic and applied research programs in perception, foundations for expert systems, uncertainty calculi for knowledge-based systems, knowledge-based planning methodologies, intelligent simulation, the integration of multi-source knowledge, representations of knowledge, link analysis, and the design and implementation of AI support tools and programming languages. Dr. Lowrance's Ph.D. dissertation introduced the AI community to the Dempster-Shafer theory of belief functions and laid the foundation for *evidential reasoning*, a methodology for representing and reasoning from evidence (i.e., information that is potentially uncertain, incomplete, and inaccurate). He is the former Assistant Director of SRI's AI Center and currently is the Director of that Center's Representation and Reasoning Program. His application-oriented research has developed approaches to multi-sensor integration, knowledge-based simulations, analysis of intelligence data, logistics planning, medical diagnosis, sonar data interpretation, vehicle tracking, forensic accounting, target systems analysis, and management decision aids. In addition, he was the principle architect of Grasper (a programming language that supports interactive graph-processing) and Gister (an evidential-reasoning and argument construction tool). Dr. Lowrance's most recent work is aimed at making evidential reasoning accessible to real world analysts and decision makers. As such, he has been the technical and managerial lead in the development of SEAS (a tool to aid intelligence analysts in recording, understanding, and comparing analytic arguments), along with Angler (a tool to promote divergent and convergent thinking) and LAW (a link analysis tool that finds close matches for graphically specified patterns). Dr. Lowrance received his A.B. in Computer Science and Mathematics from Indiana University, and M.S. and Ph.D. in Computer and Information Science from the University of Massachusetts. Dr. Lowrance has numerous publications in conference proceedings, journals, and books since 1974.

Paul-André Monney is currently working as an Independent Consultant in Statistics and Reasoning under Uncertainty. Previously, he was an Associate Professor of Statistics in the Department of Quantitative Economics at the University of Fribourg, Switzerland, and, more recently, he was for two years a Visiting Associate Professor of Statistics at Purdue University, USA. He received a Doctoral Degree in Mathematics and the Venia Legendi in Statistics from the University of Fribourg. He has done extensive research in Theoretical Computer Science and Statistics, in particular the Dempster-Shafer Theory of Evidence. Dr. Monney has numerous publications, including articles in scientific journals such as Artificial Intelligence, International Journal of Approximate Reasoning, Journal of Computational and Applied Mathematics and Zeitschrift für Operations Research. He is co-author of A Mathematical Theory of Hints - An Approach to the Dempster-Shafer Theory of Evidence, a book presenting a new perspective on the Dempster-Shafer Theory of Evidence. In 2003, he published a book entitled A Mathematical Theory of Arguments for Statistical Evidence, in which the theory of hints is applied to the field of statistics. Dr. Monney has served as a member of the program committee or as a referee for several international journals and conferences.

Hung Nguyen received his Doctorat d'Etat es Sciences Mathematiques (Ph.D.) at the University of Sciences and Technologies of Lille (France) in 1975. After spending several years at the University of California, Berkeley, and the University of Massachusetts, Amherst, he joined the faculty at New Mexico State University, Las Cruces, in 1981 where he is professor of Mathematical Sciences at present. His research contributions in the general field of decision-making under uncertainty includes statistical inference in diffusion processes, theory of conditional events, foundations of fuzzy logics and decision-making with random set data (statistical inference with set-valued observations, and uncertainty analysis in intelligent systems).He is co-author of several books, such as Fundamentals of Mathematical Statistics, Mathematics of Data Fusion, A First Course in Fuzzy Logics, and author of the upcoming book An Introduction to Random Sets. Hung Nguyen has been awarded the LIFE chair of fuzzy theory at Tokyo Institute of Technology, Japan, in 1992–93; the Westhafer award on research at New Mexico State University, in 2000; the distinguished fellow of the ASEE- Summer Faculty Research Programs ; and the distinguished visiting Lukacs professorship of statistics at Ohio State University, Bowling Green, in Spring 2002.

Henri Prade is a Charge de Recheche at the National Center for Scienfific Research (CNRS). He received the Engineer and the Doctor-Engineer degree from the Ecole Nationale Superieure de l'Aeronautique et de l'Espace respectively in 1975 and in 1977. he is the author (with Dr. D. Dubois) of tow books: "Fuzzy Sets and Systems: Theory and Applications" (Academic Press, New York, 1980) and "Theorie des Possibilites (Masson, Paris, 1985), and over eighty technical papers. He is co-editor of the fuzzy set bulletin, Busefal, which has appeared quarterly since 1979. He is also a member of the editorial organizations of the Inter. J. "Fuzzy Sets & Systems". His current research interests are in fuzzy set and possibility theory, approximate reasoning and non-classical logics, artificial intelligence and expert systems, natural language communication systems, incomplete and uncertain information data base management systems, and operations research.

Galina L. Rogova received both her MS and PhD in Moscow, Russia. Currently she is consulting on behalf of Encompass Consulting she founded in 2000. Galina Rogova is also an adjunct professor at the State University of New York at Buffalo where she is actively involved in the programs of the Center for Multisource Information Fusion. Her research interest is focused on information fusion, machine learning, pattern recognition, reasoning and decision making under uncertainty, and medical imaging. She authored numerous papers in these fields. She is a member of a number of professional societies including the International Society of Information Fusion and AAAS.

Enrique H. Ruspini received his degree of Licenciado en Ciencias Matemáticas from the University of Buenos Aires, Argentina, and his doctoral degree in System Science from the University of California at Los Angeles. Prior to joining SRI (formerly *Stanford Research Institute*), Dr. Ruspini held positions at the University of Buenos Aires, the University of Southern California, UCLA's Brain Research Institute, and Hewlett-Packard Laboratories. Dr. Ruspini is one the earliest contributors to the development of fuzzy-set theory and its applications, having introduced its use to the treatment of numerical classification and clustering problems. He has also made significant contributions to the understanding of the foundations of fuzzy logic and approximate-reasoning methods. His recent research has focused on the application of fuzzy-logic techniques to the development of systems for intelligent control of teams of autonomous robots, information retrieval, qualitative description of complex objects, and knowledge discovery and pattern matching in large databases. Dr. Ruspini, who has lectured extensively in the United States and abroad and is the author of over 100 original research papers, is a Fellow of the Institute of Electrical and Electronics Engineers, a First Fellow of the International Fuzzy Systems Association, a Fulbright Scholar, and a SRI Institute Fellow. Dr. Ruspini was the General Chairman of the Second IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'93) and of the 1993 IEEE International Conference on Neural Networks (ICNN'93). In 2004, Dr. Ruspini received the *Meritorious Service Award* of the IEEE Neural Networks Society for leading the transition of the Neural Networks Council into Society status. He is one of the founding members of the North American Fuzzy Information Processing Society and the recipient of that society's King-Sun Fu Award. Dr. Ruspini is a former member of the IEEE Board of Directors (Division X Director, 2003–2004), the Past-President (President-2001) of the IEEE Neural Networks Council and its past Vice-president of Conferences. Dr. Ruspini, who has led numerous IEEE technical, educational, and organizational activities, is also a member of the IEEE Nominations and Appointments Committee and of the Administrative Committee of the IEEE Computational Intelligence Society. Dr. Ruspini is also the Editor in Chief (together with P.P.Bonissone and W. Pedrycz) of the *Handbook of Fuzzy Computation*, published by Institute of Physics Publishing, a member of the Advisory Boards of the *IEEE Transactions on Fuzzy Systems*, and of the *International Journal of Fuzzy Systems*, and a member of the Editorial Board of the the *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, *Fuzzy Sets and Systems*, the *Journal of Applied Nonstandard Logics*, *Mathware and Soft Computing*, and the *Journal of Advanced Computational Intelligence and Intelligent Informatics.*

Glenn Shafer (see pp. viii)

Prakash P. Shenoy is the Ronald G. Harper Distinguished Professor of Artificial Intelligence in Business, University of Kansas at Lawrence. He received a B.Tech. in Mechanical Engineering from the Indian Institute of Technology, Bombay, India, in 1973, and an M.S. and a Ph.D. in Operations Research from Cornell University in 1975 and 1977, respectively. His research interests are in the areas of uncertain reasoning and decision analysis. He is the inventor of valuation-based systems, an abstract framework for knowledge representation and inference that includes Bayesian probabilities, Dempster-Shafer belief functions, Spohn's kappa calculus, Zadeh's possibility theory, propositional logic, optimization, solving systems of equations, database retrieval, and other domains. He serves as the North-American editor of *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, and as an Associate

Editor of *Management Science*, and the *International Journal of Approximate Reasoning*, and as an ad-hoc referee for over 30 journals and conferences in Artificial Intelligence and Management Science/Operations Research.

Edward H. Shortliffe is the founding dean of the Phoenix campus of the University of Arizona's College of Medicine, where he is also appointed Professor of Basic Medical Sciences, Professor of Medicine, and (at Arizona State University) Professor of Biomedical Informatics. After receiving an A.B. in Applied Mathematics from Harvard College in 1970, he moved to Stanford University where he was awarded a Ph.D. in Medical Information Sciences in 1975 and an M.D. in 1976. During the early-1970s, he was principal developer of the medical expert system known as MYCIN. In January 2000 he moved to Columbia University, where he was Chair of the Department of Biomedical Informatics, Deputy Vice President for Strategic Information Resources, Professor of Medicine, Professor of Computer Science, and Director of Medical Informatics Services for the New York-Presbyterian Health Care System. He continues to be closely involved with biomedical informatics graduate training and his research interests include the broad range of issues related to integrated decision-support systems, their effective implementation, and the role of the Internet in health care. He is an elected member of the Institute of Medicine and the American Society for Clinical Investigation, a fellow of the American College of Medical Informatics and the American Association for Artificial Intelligence, and a Master of the American College of Physicians. Editor-in-Chief of the *Journal of Biomedical Informatics*, he has authored over 230 articles and books in the fields of medical computing and artificial intelligence.

Philippe Smets (1938–2005) was born in Brussels (Belgium) in 1938. He received a medical doctor degree in 1963 from the Université Libre de Bruxelles (ULB), a Master degree in experimental statistics from North Carolina State University, and his PhD degree in medical statistics from ULB in 1978. He was the founder of the IRIDIA laboratory (Institut de Recherches Interdisciplinaires et de Développements en Intelligence Artificielle) at ULB. Under his leadership, IRIDIA became a major Belgian research institute in Artificial Intelligence and related topics, and an internationally renowned place. Philippe Smets contributed more than 100 papers to Dempster-Shafer theory, and to its comparison with alternative approaches to uncertainty reasoning. Philippe Smets was the father of the European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU), which has taken place every two years since 1991. He was also an active participant to the annual Uncertainty in Artificial Intelligence (UAI) Conference in the nineties and was the first European UAI co-program chair in 1991. He served on the editorial boards of many journals including *International Journal of Approximate Reasoning*, *Journal of Logic and Computation*, *Information Sciences*, *Fuzzy Sets and Systems*, *IEEE Transactions on Fuzzy Systems*, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *Journal of Applied Non Classical Logics*, and *Mathware and Soft Computing.*

Rajendra P. Srivastava is Ernst & Young Distinguished Professor of Accounting and Director of the Ernst & Young Center for Auditing Research and Advanced Technology at the School of Business, University of Kansas. He holds a Ph.D. in accounting from the University of Oklahoma, Norman (1982) and a Ph.D. in physics from Oregon State University, Corvallis (1972). Professor Srivastava's publications have appeared in *The Accounting Review, Journal of Accounting Research, Auditing: A Journal of Practice and Theory*, *Journal of Management Information Systems* and in many other accounting, AI and physics journals. Professor Srivastava received the 1996 Award for Notable Contribution to AI & Expert Systems Research in Accounting from the AI/Expert Systems Section of the American Accounting Association. He is currently Associate Editor of *Journal of Emerging Technologies in Accounting*, AI/ET Section Journal of the American Accounting Association, and has been a member of the Editorial and Review Board of several journals including: *The Accounting Review*, *Auditing: A Journal of Practice and Theory*, *Indian Accounting Review*, *International Journal of Auditing*, and *International Journal of Accounting and Information Systems.* Raj served as the Chairman of the AI/Emerging Technology Section of American Accounting Association during 1994–95.

Thomas M. Strat is a Program Manager at the Defense Advanced Research Projects Agency. His technical interests include computer vision, video understanding, and applications in robotics, reconnaissance, surveillance, and targeting. Previously, Dr. Strat was co-founder and Chief Technology Officer of ObjectVideo, Inc., a leading supplier of intelligent video surveillance systems. Dr. Strat spent 15 years as a research scientist and project leader at the SRI Artificial Intelligence Center, where he designed and implemented computer vision and evidential reasoning systems.Tom received the B.S., M.S., and E.E. degrees in Electrical Engineering and Computer Science from M.I.T., and the Ph.D. in Computer Science from Stanford University. He has published more than 100 scientific papers on image understanding and evidential reasoning and has authored or edited several books on computer vision.

Amos Tversky (1937–1996) was Davis Brack Professor of Behavioral Sciences and one of the world's most respected and influential psychologists. Amos earned a bachelor's degree from Hebrew University in 1961, and his doctorate in 1965 from the University of Michigan. After holding teaching positions at Michigan and Harvard, Amos returned to Hebrew University, where he began his long collaboration with Danny Kahneman. He remained at Hebrew University until joining the Stanford Faculty in 1978. A member of the faculty senate from 1991 on, and a key advisory board member, Amos' contributions to the Stanford community were extremely memorable. He also contributed to a number of interdisciplinary programs, and was a cofounder of the Stanford Center of Conflict and Negotiation. Amos' accomplishments were recognized with all the honors that academia can bestow. A fellow at the Center for Advanced Study in 1970, he was elected to the American Academy of Arts and Sciences in 1980, and the National Academy of Science in 1985. He also won (with Kahneman) the American Psychological Association's award for distinguished scientific contribution in 1982, and MacArthur and Guggenheim Fellowships in 1984, and was awarded honorary doctorates by the University of Chicago, Yale University, the University of Goteborg in Sweden and the State University of New York at Buffalo.

Ronald R. Yager (see pp. 793)

John Yen received his Ph.D. in Computer Science from University of California, Berkeley in 1986. His thesis advisor is Prof. Lotfi A. Zadeh, the father of fuzzy logic. Between 1986 and 1989, he was the main architect at USC Information Sciences Institute (ISI) for an AI architecture that pioneers a knowledge-level integration involving rules and semantic-Web knowledge representation technologies. From 1989 to 2001, he was a Professor of Computer Science and the Director of Center for Fuzzy Logic, Robotics, and Intelligent Systems at Texas A&M University. He joined Penn State's College of Information Sciences and Technology in 2001, and became Professor-in-Charge in 2003. He is currently the Associate Dean for Research and Graduate Programs of the college. Yen has published a textbook on fuzzy logic, two edited volumes, more than 50 journal articles, and more than 100 refereed conference papers on topics related to fuzzy logic, Dempster-Shafer Theory, AI, and intelligent agents. Yen was the Vice President of Publication for IEEE Neural Networks Council, which is the predecessor of IEEE Computational Intelligence Society. He received the NSF Young Investigator Award in 1992, and is a Fellow of IEEE.

Nevin L. Zhang received his BSc degree in Applied Mathematics from China University of Electrical Sciences and Technology in 1983, his MSc and PhD degrees in Applied Mathematics from Beijing Normal University in 1986 and 1989 respectively, and his PhD degree in Computer Science from University of British Columbia in 1994. Thereafter, he joined Computer Science Department, The Hong Kong University of Science and Technology, where he is now an Associate Professor.Nevin Lianwen Zhang has been on the programming committee of the Conference on Uncertainty in Artificial Intelligence (UAI) since 1994. He co-chaired the program committee of the Seventh European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU-2003). He served on the editorial board of Journal of Artificial Intelligence Research (JAIR) (1999–2002) and is currently an associate editor of JAIR. Nevin Lianwen Zhang publishes mainly on, Influence Diagrams, Bayesian Networks (inference and learning), Partially Observable Markov Decision Processes. His current research focuses on latent structure discovery and its applications.

Author Index

Subject Index