Linguistic Probability Theory

Joe Halliwell

Doctor of Philosophy

School of Informatics

University of Edinburgh

2007

Abstract

In recent years probabilistic knowledge-based systems such as Bayesian networks and influence diagrams have come to the fore as a means of representing and reasoning about complex real-world situations. Although some of the probabilities used in these models may be obtained statistically, where this is impossible or simply inconvenient, modellers rely on expert knowledge. Experts, however, typically find it difficult to specify exact probabilities and conventional representations cannot reflect any uncertainty they may have. In this way, the use of conventional point probabilities can damage the accuracy, robustness and interpretability of acquired models. With these concerns in mind, psychometric researchers have demonstrated that fuzzy numbers are good candidates for representing the inherent vagueness of probability estimates, and the fuzzy community has responded with two distinct theories of *fuzzy probabilities*.

This thesis, however, identifies formal and presentational problems with these theories which render them unable to represent even very simple scenarios. This analysis leads to the development of a novel and intuitively appealing alternative - a theory of *linguistic probabilities* patterned after the standard Kolmogorov axioms of probability theory. Since fuzzy numbers lack algebraic inverses, the resulting theory is weaker than, but generalises its classical counterpart. Nevertheless, it is demonstrated that analogues for classical probabilistic concepts such as conditional probability and random variables can be constructed. In the classical theory, representation theorems mean that most of the time the distinction between mass/density distributions and probability measures can be ignored. Similar results are proven for linguistic probabilities.

From these results it is shown that directed acyclic graphs annotated with linguistic probabilities (under certain identified conditions) represent systems of linguistic random variables. It is then demonstrated these *linguistic Bayesian networks* can utilise adapted best-of-breed Bayesian network algorithms (junction tree based inference and Bayes' ball irrelevancy calculation). These algorithms are implemented in ARBOR, an interactive design, editing and querying tool for linguistic Bayesian networks.

To explore the applications of these techniques, a realistic example drawn from the domain of forensic statistics is developed. In this domain the knowledge engineering problems cited above are especially pronounced and expert estimates are commonplace. Moreover, robust conclusions are of unusually critical importance. An analysis of the resulting linguistic Bayesian network for assessing evidential support in glass-transfer scenarios highlights the potential utility of the approach.

Acknowledgements

The greatest debt is owed to my endlessly enthusiastic, insightful and patient supervisor, Qiang Shen. Speaking plainly, this thesis would not have been possible without his careful advice and stalwart support. Thank you, Qiang.

My second supervisor, Alan Smaill and my colleague Jeroen Keppens have both made invaluable contributions to the present work. Thank you, Alan and Jeroen.

I am also indebted to the former members of Qiang's research group at the University of Edinburgh for their interest in and useful feedback on my ideas. So, I'd like to thank Alexios Chouchoulas, Ronan Daly, Michelle Galea, Zhiheng Huang and Richard Jensen.

My time at the University of Edinburgh has been very happy, but would not have been so without the dear friends and intellectual comrades I have been lucky enough to find in Paul Crook, Colin Fraser, Annabel Harrison, Sebastian Mhatre, Fiona McNeill, Alison Pease and Dan Winterstein. Thanks guys.

Outwith this circle, I have the great privilege to be friends with Hamish Allan, Andrew Back, Stephen Blythe, Sam Collier, Aaron Crane, Harry Day, Lucas Dixon, Stephan van Erp, Helen Foot, Terry Grayshon, Alex Heneveld, Chris Hinds, Veronika Holtzmann, Jethro Green, Max MacAndrews, Jarred and Sarah McGinnis, Ewen Maclean, Laura Meikle, Jelena Meznaric, Miranda Millward, Hannu Rajaniemi, Nils Roeder, Rebecca Smith, Viktor Tron, Chris Scott, Graham Steel, Matthew Williams-Grey, Ika Willis and Susie Wilson. There have been many late nights, long days, pots of tea, pints of beer, lunches, dinners, games, books, crosswords, dreams, ideas, projects, plans, and conversa-

tions. For these, thank you all.

Finally, I'd like to thank my Mum and Dad for starting me on this path, and sticking by me ever since. This work, such as it is, is for you.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Joe Halliwell)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The research documented in this thesis represents an attempt to answer a set of linked questions about the mathematical and computational aspects of modelling probabilities as fuzzy quantities. The motivations for this study are sketched in the following section. This feeds into the set of specific research questions that follows. The chapter closes with an outline of the remainder of the thesis.

1.1 Motivation

Many different types of uncertainty can be found in the sorts of information intelligent systems must process and it is not uncommon to find several of these represented in a single item. So, for example, in constructing a knowledge-based system for industrial fault diagnosis, an expert might supply that if the water pressure is well above tolerance levels, then it is extremely likely that the output valve will fail. Here, the meaning of the term "extremely likely" combines elements of both fuzzy and probabilistic uncertainty.

The use of fuzzy sets to model every-day descriptions such as "John is tall" is no doubt familiar. The idea behind the set of theories sharing the name "fuzzy probability" is that imprecise linguistic characterisations of probabilistic uncertainty can be treated in an analogous way. The goal then, put simply, is to develop a principled approach to statements such as

$$\text{It is quite likely to rain tomorrow.} \tag{1.1}$$

A possible objection at this stage is that (1.1) is hopelessly uninformative. If (probabilistic) information about the next day's weather is crucial to a system's successful operation there are surely better ways to obtain it. In short, why bother attempting to utilise such woefully low-grade information? The answer, of course, is the standard argument for "computing with words" (Zadeh, 1996): whilst gold-standard numerical information may be available about tomorrow's weather, there are probabilistic assessments which are too difficult, expensive or simply impossible to obtain with such precision.

For example, consider the questions: Will there be artificial intelligence in 10 years? 100 years? 1000 years? Consultation with an expert is unlikely to yield much beyond vague probabilistic statements like "It is extremely unlikely that we will have (true) artificial intelligence in ten years time." But if such information is to be used within the framework of classical probability theory, numerical estimates of the probabilities of interest are required.

In such cases and indeed many that are less speculative, the difficulty of obtaining point estimates of probability has been widely reported (Kahneman et al., 1985; Zimmer, 1983). Whilst an expert may be willing to assert that it is extremely likely that there will be intelligent constructs this time next millennium it would seem odd, a loss of academic integrity even, to state that the probability of that occurrence is 0.93. Indeed, a committee of the U.S. National Research Council (National Research Council Governing Board Commitee on the Assessment of Risk, 1981; Wallsten et al., 1986) has written that there is "an important responsibility not to use numbers, which convey the impression of precision, when the understanding of relationships is indeed less secure. Thus whilst quantitative risk assessment facilitates comparison, such comparison may be illusory or misleading if the use of precise numbers is unjustified." Subjective probability assessments are often the product of countless barely articulate intuitions and are often best expressed in words. It is misleading to seek to express them with numerical precision.

1.2 The case for fuzzy probabilities

Responding to these difficulties, researchers have attempted to obtain point values for probabilistic terms experimentally. The general form of these investigations is to present subjects with probabilistic terms requesting a numerical translation. It is hardly surprising that studies such as Budescu and Wallsten (1985) have concluded that point estimates of probability terms vary too greatly between subjects and exhibit too great an overlap to be useful for many problems.

Attempts to model probabilistic terms using fuzzy sets, however, have proven more successful. For example, a relatively sophisticated experimental method for eliciting fuzzy models of probabilistic terms has been developed by Wallsten et al. (1986) and the inter-subjective stability of generated terms has been examined with promising results. In addition, Zimmer (1986) has reported that verbal expressions of probabilistic uncertainty were "more accurate" than numerical values in estimating the frequency of multiple attributes by experimental studies. Whilst there are outstanding problems such as context sensitivity with the fuzzy approach to modelling probabilistic terms, these psychometric studies are unanimous in preferring it to numerical estimates.

1.3 Problem statement

The research presented in this thesis represents an attempt to answer the question: what form should a theory of fuzzy probabilities take?

It does not seek to offer a view on the philosophical aspects, viability or utility of representing vague concepts in general using fuzzy sets, nor on the application of fuzzy arithmetic in particular to modelling vague probabilities. Instead it is hoped that the reader will evaluate the present work in the light of the psychometric studies outlined above, and on the assumption that these are sensible and useful things to do.

This is not, then, an experimental study – although suggestions for supporting experiments will be made from time to time. Instead it has the form of a systematic exploration of the mathematical, computational and aspects of a particular combination of probability theory and fuzzy logic.

This perhaps complicates evaluation. Artificial Intelligence – understood as the attempt to coax machines into exhibiting intelligent behaviour – is fundamentally an experimental discipline. Nevertheless, there is a long tradition of more analytic work in AI that proposes an abstract theory developed for a set of identified reasons and where the immanent experimental programme is assumed or deferred. Indeed, Zadeh's initial forays into fuzzy logic which provide significant inspiration for the present study, have this character.

The test of quality for such work must be whether the concepts and methods (and their computational counterparts of representations and algorithms) that are introduced are sufficiently novel, coherent, interesting and potentially useful. The critique of existing work demonstrates that the theory of linguistic probabilities is novel and non-trivial – others have walked a similar path but have stumbled on it.

Similarly, it is hoped, the coherence and richness of the theory should be evident from the relative ease with which sophisticated concepts such as random variables and conditional probability can be adapted. Finally, the potential utility is suggested by the development of linguistic probability networks as a scheme for representing and reasoning about multivariate linguistic probability distributions and their application to a realistic case study.

1.4 Overview

The chapter following this brief introduction presents a summary of the key concepts and notation that constitute the technical background of the present work. Although much of the material here will be familiar to the reader the notation and manner of presentation are significant: some of the notation, particularly that associated with fuzzy numbers and their arithmetic, where there are not yet universally accepted conventions, has been invented or adapted for the present purposes; and the way in which classical probability theory is introduced intended to prefigure the sequence and structure of the theory developed here.

Chapter 3 examines a number of related pieces of research which also seek to hybridize fuzzy logic and probability theory. Because of the breadth and somewhat inter-disciplinary nature of this area of research there have been few systematic reviews. Chapter 3 attempts to address this deficit. Most of the text, however, is given over to a discussion and original critique of what are identified as the two existing theories of "fuzzy probabilities".

In response to these criticisms, Chapter 4 develops the core theory of linguistic probabilities. Unlike its two predecessors in interest, the theory is explicitly patterned after the standard measure-theoretic axioms of contemporary probability theory. It is shown that linguistic probability measures, like their classical counterparts, are monotonic and continuous. Analogues for the classical concepts of conditional probability, independence and discrete and continuous random variables are also introduced and discussed.

Chapter 5 develops the theory into a computational application by examining how linguistic probabilities could be used in a Bayesian-network-like graphical knowledge representation. It is shown that the resulting representation is consistent, if not (in a certain sense, to be defined) complete.

Classical Bayesian networks are increasingly used to assess the strength of evidential support provided by forensic evidence. Chapter 6, introduces this application area and then, through a detailed case study, shows how linguistic Bayesian networks might be used to address some deficiencies in the current approach.

Finally, chapter 7 draws matters to a close. The chief claims and achievements of the work are summarised and directions for future research are sketched in some detail.

Chapter 2

Background

This chapter introduces and discusses the basic mathematical concepts that underlie the remainder of the work. Naturally, it is anticipated that some of the technical material will be familiar to the reader. In these cases, however a brief rehearsal serves both to introduce the requisite notation, and where there may be several alternatives in the literature, to clarify precisely which definitions are in operation.

The chapter is divided into three main sections. First the basic principles of fuzzy logic and fuzzy set theory are introduced against a background of classical and non-standard logic. These elementary fuzzy concepts are then used to define the concept of a fuzzy number, the chosen model for fuzzy probabilities. Some auxiliary results regarding the (natural) algebra and topology of fuzzy numbers are presented which are central to the development of linguistic probability theory.

The second section presents a brief overview of the standard Kolmogorov axiomatization of (classical) probability theory and sets the familiar entities of probability theory, in particular conditional probability and random variables into this context.

The third and final section presents a brief overview of the theoretical underpinnings of (classical) Bayesian networks. The structure of these section last two sections is mirrored by chapters 4 and 5 respectively, which seek to emulate the constructions they present in the revised axiomatization.

2.1 Fuzzy logic

One key section of Aristotle's Metaphysics, a work that has profoundly influenced the modern (realist) concept of truth, formal logic and linguistic philosophy, states that "...the understanding either affirms or denies every object of understanding or thought... whenever it is right or wrong. When, in asserting or denying, it combines the predicates in one way, it is right; when in the other, it is wrong."

Interpretative difficulties aside, this short excerpt illustrates two key features of Aristotle's thought. First, all statements are (understood as) either true or false. Second, this understanding is either correct or incorrect in virtue of the world. Setting aside the metaphysical/epistemological aspects this can be glossed as

$$\text{Every statement is either true or false.} \tag{2.1}$$

In contemporary philosophy of logic this has become known as the Principle of Bivalence.

The principle is meta-logical in that it cannot be formulated within logic itself. However two formulations are available within logic namely the law of non-contradiction

$$\models \neg(\phi \wedge \neg\phi) \tag{2.2}$$

and the law of the excluded middle (or tertium non datur)

$$\models \phi \vee \neg\phi \tag{2.3}$$

These differ from 2.1 in that they do not necessarily assert that P has a truth value.

In classical logic these two laws hold and it is therefore usually said to satisfy 2.1. But this has been the subject of controversy. So, for example, from a philosophical point of view it is not clear that a prospective statement, such as "Mary will go to the shops tomorrow" has (at this moment) any definite truth value. A similar difficulty is presented by intensional contexts. So, for example, one may believe "either John is having an affair or John is not having an affair" without believing either disjunct separately. Such modal statements have suggested to some researchers that 2.1 may be rejected while 2.3 and 2.2 are retained.

An opposite area of difficulty is highlighted by intuitionistic (constructive) logic, which equates truth and provenness. Gödel's Theorem (Gödel, 1931) demonstrates that for any sufficiently strong, consistent axiomatic system, there is a simple mathematical statement, G, that is true, but neither provable nor refutable. Thus intuitionistic logic rejects 2.3, because $G \vee \neg G$ can only be proven by the (impossible) proof or refutation of G. Interestingly, however, 2.2 is retained as a trivial consequence of the semantics of negation, namely that $\neg P$ can only be derived from a proof that P leads to a contradiction, but in the case of 2.2, the P in question is itself the very definition of a contradiction!

Another strand in the philosophy of logic has been to reject 2.1 by the addition of one or more intermediate truth values. So, for example, in order to address epistemic concerns one might add an "undecided" truth value. The classical truth functional operators are then extended to encompass this new value. Very many different trivalent logics have been proposed, prompting the MathWorld encyclopaedia to remark (with jocular precision) that "there are 3072 such logics". Nevertheless, such distinguished philosophers such as Emile Post and Charles Peirce may be numbered amongst this authors of this vast body of research.

Nowadays, trivalent logics are viewed as just a particular class amongst many multi-valued logics. Much pioneering work in this area was undertaken by the Polish mathematician and philosopher, Jan Łukasiewicz and however this work largely remained within a logicomathematical niche until, in the mid 1960s Lotfi Zadeh proposed a systematic study of what he termed "fuzzy logic" as a means to model the vagueness of natural language.

2.1.1 Fuzzy logic in AI

Zadeh was extremely successful in demonstrating the practical utility of the fuzzy approaches to knowledge representation and (equally important) in persuading other researchers to back the programme of work he outlined.

Control techniques utilising meaningful natural language rule sets to implicitly generate sophisticated control surfaces were an early and continuing area of success. Unusually for an Artificial Intelligence technique, fuzzy logic saw rapid adoption outside academia and a host of commercial and domestic applications.

Nevertheless, Zadeh has, from the outset, maintained that the fuzzy approach is a generic knowledge representation strategy and that "fuzzification" is a process which can be applied to any mathematical theory. This idea has been taken up with gusto by a host of researchers and has resulted in a wide-ranging but systematic effort to fuzzify various areas of mathematics. The fruit of these labours is reflected in Section 2.1.4 and following.

2.1.2 Fuzzy sets

This section introduces the basic formal concepts and notation associated with fuzzy sets.

Definition 2.1.1 (Fuzzy set)**.** Given a universe of discourse, D, a *fuzzy set*, A, in D is determined by its membership function

$$\mu_A : D \to [0,1] \tag{2.4}$$

For any $d \in D$, $\mu_A(d)$ is termed d's *degree of membership* in A.

The force of this definition is that two fuzzy sets are equal if they have the same membership function. An alternative representation of fuzzy sets is through their alphacuts. If a fuzzy set is thought of as a landscape (in a three dimensional Venn diagram) then its alphacuts correspond to its contours.

Definition 2.1.2 (Alphacut)**.** Given a fuzzy set, A, the *alphacut* of A at $\alpha \in [0,1]$

is the set of elements with membership degree at least α i.e.

$$A_{\lfloor\alpha} = \{x \in D \,:\, \mu_A(x) \geq \alpha\} \tag{2.5}$$

That alphacuts and membership functions are equivalent and mutually-determining representations is a trivial consequence of their definitions.

Theorem 2.1.3 (Decomposition). *Given fuzzy sets A and B with universe of discourse D, $A = B$ if and only if $A_{\lfloor\alpha} = B_{\lfloor\alpha}$ for all $\alpha \in [0,1]$*

Proof. If $\mu_A = \mu_B$ then for all $\alpha \in [0,1]$

$$A_{\lfloor\alpha} = \{x \in D \,:\, \mu_A(x) \geq \alpha\} = \{x \in D \,:\, \mu_B(x) \geq \alpha\} = B_{\lfloor\alpha} \tag{2.6}$$

Now suppose $A_{\lfloor\alpha} = B_{\lfloor\alpha}$ for all $\alpha \in [0,1]$. If $\mu_A(d) < \mu_B(d)$ for some $d \in D$ then $d \in B_{\lfloor\mu_B(d)}$, but $d \notin A_{\lfloor\mu_B(d)}$, which is a contradiction. By symmetry, $\mu_A = \mu_B$. □

2.1.3 Triangular norms

Triangular norms were introduced by Schwiezer and Sklar (1963) as a model of distances in probabilistic spaces. In fuzzy logic they are analogues for the Boolean "and" truth functor.

Definition 2.1.4 (Triangular norm). A mapping $T : [0,1]^2 \rightarrow [0,1]$ is a *triangular norm* if it satisfies the following properties,

a) (Symmetry) for all $x, y \in [0,1]$, $T(x,y) = T(y,x)$

b) (Associativity) for all $x, y, z \in [0,1]$, $T(x, T(y,z)) = T(T(x,y),z)$

c) (Monotonicity) if $x \leq x', y \leq y'$, $T(x,y) \leq T(x',y')$

d) (Identity) for all $x \in [0,1]$, $T(x,1) = x$

Triangular norms are frequently termed "t-norms".

T-norms generalise the Boolean "and" operation. The Boolean "or" operation is generalised by t-conorms.

Definition 2.1.5 (Triangular co-norm)**.** A mapping $S : [0,1]^2 \to [0,1]$ is a *triangular co-norm* if and only if

$$T(x,y) = 1 - S(1-x, 1-y) \tag{2.7}$$

is a triangular norm. Triangular co-norms are frequently termed "s-norms".

There are an infinite number of t- and s-norm pairs. Some examples are presented in Table 2.1. The choice of t-norm is a somewhat vexed issue. For practical applications one approach has been to treat the selection of truth-functional operator as an optimization problem in its own right. For example, Song et al. (2003) describe a parameterised family of operators and a learning algorithm for choosing the optimal parameter given a set of training data.

Historically, min and max were Zadeh's choice of truth functors and these remain, for the author at least, the most intuitively appealing. In addition generalised t-norm notation is rather cumbersome. Finally, min and max (and their infinitistic counterparts inf and sup) are uniquely well-suited to the development of fuzzy sets of numbers as they generalise immediately to arbitrary numbers of operands. For these reasons, the remainder of this text will utilise them as concrete truth-functional operators, but the reader should bear in mind that most of what follows hold regardless of the choice of t-norm.

	t-norm		s-norm		
Zadeh	$T_Z(x,y)$	$= \min(x,y)$	$S_Z(x,y)$	$= \max(x,y)$	
Łukasicwixz	$T_L(x,y)$	$= \max(x+y-1,0)$	$S_L(x,y)$	$= \min(x+y,1)$	
Product/probabilistic	$T_P(x,y)$	$= xy$	$S_P(x,y)$	$= x+y-xy$	
Weak	$T_W(x,y)$	$= \begin{cases} \min(x,y) & \max(x,y)=1, \\ 0 & \text{otherwise} \end{cases}$	$S(x,y)$	$= \begin{cases} \max(x,y) & \min(x,y)=0, \\ 1 & \text{otherwise} \end{cases}$	
Yager family	$Y(x,y,p)$	$= 1-\min(1, \sqrt[p]{(1-x)^p+(1-y)^p})$	$S_Y(x,y,p)$	$= \min(1, \sqrt[p]{x^p+y^p})$	for $p > 0$
Dubois and Prade family	$T_D(x,y,\alpha)$	$= \frac{xy}{\max(x,y,\alpha)}$	$S_D(x,y,\alpha)$	$= 1-T_D(1-x,1-y,\alpha)$	for $\alpha \in (0,1)$
Frank family	$T_F(x,y,\lambda)$	$= \begin{cases} T_Z(x,y) & \lambda=0, \\ T_P(x,y) & \lambda=1, \\ T_L(x,y) & \lambda=\infty, \\ 1-\log_\lambda\{1+\frac{(\lambda^a-1)(\lambda^b-1)}{(\lambda-1)}\} & \text{otherwise} \end{cases}$	$S_F(x,y,\lambda)$	$= 1-T_F(1-x,1-y,\lambda)$	for $\lambda >= 0$

Table 2.1: Examples of t-norms and their associated s-norms

2.1.4 The Extension Principle

As indicated earlier, Zadeh has conceived of fuzzy theory not as a single theory per se, but rather as a process of "fuzzification" – a method for generalizing any specific representation scheme or theory from crisp (roughly speaking, first-order predicate logic) to fuzzy. One of the central results that supports this contention is the Extension Principle. This identifies a natural way to extend maps between classical sets to maps on fuzzy sets defined over them (as a universe of discourse).

Definition 2.1.6 (Extension Principle)**.** Given a map,

$$f : A_1 \times A_2 \times \ldots \times A_n \to B \tag{2.8}$$

the natural fuzzy extension, $\tilde{f}$, is the map determined by:

$$\mu_{\tilde{f}(a_1,a_2,\ldots,a_n)}(y) = \sup_{f(x_1,x_2,\ldots,x_n)=y} T(\mu_{a_1}(x_1), \mu_{a_2}(x_2), \ldots \mu_{a_n}(x_n)) \tag{2.9}$$

for all fuzzy sets $a_1, a_2, \ldots, a_n$ defined on $A_1, A_2, \ldots, A_n$ respectively. $\tilde{f}$ is sometimes referred to as the *sup-t convolution* of f.

In other words, the *possibility* of a particular element being in the image of a fuzzy set under an extended function is simply the possibility of the disjunction of each element that maps to it belonging to that set.

A theorem first presented by Nguyen (1978) for classical fuzzy truth operators and extended to arbitrary sup t-norm convolutions by Fullér and Keresztfalvi (1990) connects alpha cuts and extended functions. In plain language it states that the alphacuts of the image of a fuzzy set (under an extended operator) are just the images of the alphacuts of that set.

Theorem 2.1.7. *Given a map $f : X \to Y$ and a fuzzy set, A with universe of discourse X then for all $\alpha \in [0,1]$,*

$$\tilde{f}(A)_{\lfloor\alpha} = \{f(x) \ : \ x \in A_{\lfloor\alpha}\} \tag{2.10}$$

Proof. The proof follows immediately from the definitions. For all $y \in Y$,

$$y \in \tilde{f}(A)_{\lfloor\alpha} \iff \sup_{f(a)=y} \mu_A(a) \geq \alpha \tag{2.11}$$

$$\iff \exists a' \in A \text{ with } f(a') = y \text{ such that } \mu_A(a') \geq \alpha \tag{2.12}$$

$$\iff y \in \{f(x) \ : \ x \in A_{\lfloor\alpha}\} \tag{2.13}$$

□

This result has also been extended to arbitrary L-fuzzy systems by Bertoluzza and Bodini (1998).

2.2 Fuzzy numbers

Fuzzy numbers are simply fuzzy sets of real numbers whose membership functions have the "right sort" of properties. Although the precise conditions remain the subject of some debate, for example, it is sometimes required that the membership function be "unimodal". Alternatively the convexity condition may be removed. the following definition is by far the most commonplace:

Definition 2.2.1 (Fuzzy number)**.** A fuzzy number, a, is a fuzzy set of real numbers with the following properties

a) μ_a is normal, i.e. there is $x \in \mathbb{R}$ such that $\mu_a(x) = 1$

b) μ_a is convex, i.e. for all $x, y, z \in \mathbb{R}$ if $x \leq y \leq z$ then $\mu_a(y) \geq \min(\mu_a(x), \mu_a(z))$

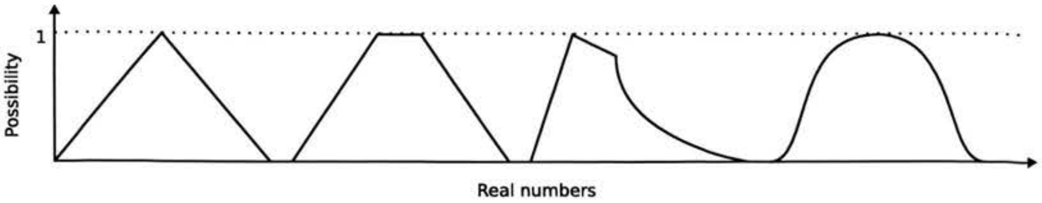


Figure 2.1: Examples of fuzzy numbers

c) μ_a is upper semi-continuous, i.e. for all $\alpha \in [0,1]$ $\{x \in \mathbb{R} : \mu_a(x) < \alpha\}$ is open

d) the support of a, $a_{\lfloor 0}$, is bounded.

Note that this definition also covers what might be termed "fuzzy intervals". Examples of fuzzy numbers can be found in Figure 2.1. The set of all fuzzy numbers are termed the "fuzzy numbers" and denoted, $\mathbb{R}^{\mathcal{F}}$.

Because the weak alphacut at zero of a fuzzy number is always the entire real line, it is notationally convenient to consider instead its level sets.

Definition 2.2.2 (Level set). Given $a \in \mathbb{R}^{\mathcal{F}}$ and $\alpha \in [0,1]$ define the α level set of a as

$$L_\alpha(a) = \begin{cases} a_{\lfloor \alpha} & \alpha \in (0,1] \\ \mathrm{Cl}(\{x \in \mathbb{R} : \mu_a(x) > 0\}) & \alpha = 0 \end{cases} \tag{2.14}$$

where $\mathrm{Cl}(X)$ is the closure (in $\mathbb{R}$) of $X \subset \mathbb{R}$.

An immediate and attractive consequence of these definitions is the following theorem.

Theorem 2.2.3. *The level sets of a fuzzy number are intervals of the form* $[x,y]$.

Proof. Since the support of a is bounded each $L_\alpha(a)$ is bounded. Let $x = \inf(L_\alpha(a))$ and $y = \sup(L_\alpha(a))$. Clearly, $L_\alpha(a) \subseteq [x,y]$. By upper semi-continuity $x, y \in L_\alpha(a)$

and hence by convexity, $[x,y] \subseteq L_\alpha(a)$. □

It will occasionally be useful to refer to the upper and lower bounds of these alphacut intervals. For these quantities the following notation is adopted,

$$\lfloor A \rfloor_\alpha = \inf(A_{\lfloor\alpha}) \tag{2.15}$$

$$\lceil A \rceil_\alpha = \sup(A_{\lfloor\alpha}) \tag{2.16}$$

There is a natural embedding of the real numbers and their closed intervals into $\mathbb{R}^{\mathcal{F}}$. This will be denoted by a χ subscript. So, for example, the membership function of 1_χ (the embedding of 1) is simply the characteristic function

$$\mu_{1_\chi}(x) = \begin{cases} 1 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.17}$$

A similar notation will be adopted for embedded intervals. So, for example,

$$\mu_{[0,1]_\chi}(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases} \tag{2.18}$$

Lemma 2.2.4. *If $a_\chi = a'_\chi$ then $a = a'$. Similarly if $[a,b]_\chi = [a',b']_\chi$ then $[a,b] = [a',b']$.*

Proof. These results follow immediately from the definitions. □

2.2.1 Arithmetic operators

The Extension Principle may also be used to define fuzzy counterparts to the standard arithmetic operators of addition, multiplication, subtraction and division. If the standard arithmetic operators are considered as (continuous) maps from $\mathbb{R}^2 \to \mathbb{R}$ the straightforward application of the principle yields the following definitions for the extended operators. As is conventional, the extension of a real arithmetic operator will be denoted by circling its usual symbol. In the context of the fuzzy numbers it is also possible to derive these operators by examining the effects of performing interval-based calculations at each alphacut.

Definition 2.2.5 (Fuzzy-arithmetic operators). For all $a, b \in \mathbb{R}^{\mathcal{F}}$ the extended operators $\oplus, \otimes, \ominus, \oslash$ are determined by

$$\mu_{a\oplus b}(z) = \sup_{x+y=z} \min(\mu_a(x), \mu_b(y)) \tag{2.19}$$

$$\mu_{a\otimes b}(z) = \sup_{xy=z} \min(\mu_a(x), \mu_b(y)) \tag{2.20}$$

$$\mu_{a\ominus b}(z) = \sup_{x-y=z} \min(\mu_a(x), \mu_b(y)) \tag{2.21}$$

$$\mu_{a\oslash b}(z) = \sup_{\frac{x}{y}=z} \min(\mu_a(x), \mu_b(y)) \tag{2.22}$$

2.2.2 Partial orderings

For real numbers, a, b, $a \le b$ if and only if $a = \min(a,b)$. Following this observation, the Extension Principle may be used to induce a natural partial ordering, $\preccurlyeq$, on the fuzzy numbers

$$a \preccurlyeq b \iff a = \tilde{\min}(a,b) \tag{2.23}$$

$$\iff \mu_a(z) = \sup_{min(x,y)=z} \min(\mu_a(x), \mu_b(y)) \; \forall z \in \mathbb{R} \tag{2.24}$$

This ordering extends the standard (total) ordering of the reals in the sense that agrees with the standard ordering on the set embedded reals.

Since the set of fuzzy numbers whose membership functions are zero outside some given real interval $[a,b]$ can be characterised as

$$\{x \in \mathbb{R}^{\mathcal{F}} \ : \ a_\chi \preccurlyeq x \wedge x \preccurlyeq b_\chi\} \tag{2.25}$$

It is therefore natural to denote such an interval of fuzzy numbers by $[a_\chi, b_\chi]$.

Another partial order on the fuzzy numbers is generated by the fuzzy subset relation.

Definition 2.2.6 (Subsumption). Given $a, b \in \mathbb{R}^{\mathcal{F}}$ b is said to *subsume* a (written $a \subseteq b$) if and only if for all $x \in \mathbb{R}$,

$$\mu_a(x) \leq \mu_b(x) \tag{2.26}$$

One number subsumes another if, loosely speaking, it is a less precise version of it. Subsumption has some useful properties under the fuzzy arithmetic operators which will be discussed in a following section.

A key property of the subsumption ordering, that has not been widely observed, is that it "carries over" extended operators in the sense of the following Lemma.

Lemma 2.2.7. *Given an operator* $* : \mathbb{R}^n \to \mathbb{R}$ *and fuzzy numbers,* $a_1, a_2, \ldots a_n, b_1, b_2, \ldots b_n$ *such that* $a_i \subseteq b_i$ *for all* $1 \leq i \leq n$ *then*

$$\circledast(a_1, a_2 \ldots a_n) \subseteq \circledast(b_1, b_2 \ldots b_n) \tag{2.27}$$

Proof. By the Extension Principle

$$\begin{aligned} \circledast(a_1, a_2 \ldots a_n)(x) &= \sup_{x=*(x_1,x_2 \ldots x_n)} \{ \min_{1 \le i \le n} a_i(x_i) \} && (2.28) \\ &\le \sup_{x=*(x_1,x_2 \ldots x_n)} \{ \min_{1 \le i \le n} b_i(x_i) \} && (2.29) \\ &= \circledast(b_1, b_2 \ldots b_n)(x) && \square \end{aligned}$$

This result allows complex calculations (such as the Bayesian sum of products expression for joint probability distribution considered later) to be rearranged and computed from partial results just as in the classical case.

2.2.3 Functions

Extending a real operator (i.e. a mapping from $\mathbb{R}^n$ to *reals*) does not necessarily yield a fuzzy real operator. However if the operator maps closed intervals to closed intervals, then its extension will map fuzzy reals to fuzzy reals. All continuous functions have this property. Furthermore if the function is injective (or, equivalently, monotonic) then whenever

Lemma 2.2.8. *If $f : \mathbb{R} \to \mathbb{R}$ is continuous then the map $\tilde{f} : \mathbb{R}^{\mathcal{F}} \to \mathbb{R}^{\mathcal{F}}$ determined by*

$$\mu_{\tilde{f}(a)}(y) = \sup\{\mu_a(x) \ : \ y = f(x)\} \qquad (2.30)$$

is well-defined.

Proof. By Theorem 2.1.7 and 2.2.3 it suffices to consider the level sets. By the Extreme Value Theorem, since f is continuous the image of level sets (i.e. closed intervals) are also closed intervals. Thus $\tilde{f}$ is well-defined. □

2.2.4 Algebraic properties

With the exception of $\oslash$, the basic arithmetical operators described above are closed with respect to the fuzzy numbers. As with their classical analogues, fuzzy addition and multiplication are commutative and associative with identities 0_χ and 1χ respectively. There are however, some important differences between classical and fuzzy arithmetic. First, fuzzy numbers are not distributive in the classical sense. Instead they are *subdistributive* i.e. for all $a, b, c \in \mathbb{R}^{\mathcal{F}}$,

$$a \otimes (b \oplus c) \subseteq (a \otimes b) \oplus (a \otimes c) \tag{2.31}$$

It is important to note however, that where the fuzzy numbers involved are strictly positive (or negative) full distributivity is retained.

Second, in general, fuzzy numbers have neither additive nor multiplicative inverses, although there are (non-unique) pseudo-inverses. In particular, for all $a \in \mathbb{R}^{\mathcal{F}}$, $0_\chi \subseteq a \oplus (0_\chi \ominus a)$ and with the usual cautions about 0, $1_\chi \subseteq a \otimes (1_\chi \oslash a)$.

In short, $\mathbb{R}^{\mathcal{F}}$ equipped with $\oplus, \otimes$ and their respective identities $0_\chi, 1_\chi$ is a subdistributive, commutative semi-ring.

2.2.5 Solving fuzzy equations

The lack of true multiplicative and additive inverses presents a problem for solving systems of fuzzy arithmetical equations. In classical linear algebra solutions are found by "doing the same" to both sides of an equation. To take a trivial example, consider solving $y = x + 3$ for x. In Peano arithmetic, the (additive) inverse of 3 is added to both sides yielding

$$y+(-3)=x+3+(-3)=x+0=x \tag{2.32}$$

In fuzzy arithmetic, however, since the constant term ("3" above) might not have an inverse this simple strategy fails and indeed the equation might not have a solution (depending, roughly speaking, on the value of y).

However it is possible to obtain traction under various conditions.

Lemma 2.2.9. *If f is a continuous, monotonic function then its fuzzy extension $\tilde{f}$ has the property that for all $a,b \in \mathbb{R}^{\mathcal{F}}$*

$$\tilde{f}(a) \subseteq \tilde{f}(b) \iff a \subseteq b \tag{2.33}$$

Proof. By Lemma 2.2.8 $\tilde{f}$ is well-defined. Since f is both continuous and monotonic, it is injective at each level set and so by Theorem 2.1.7 the desired result is obtained. □

2.2.6 Topology

This section presents the standard metric, d_∞ on the space of fuzzy numbers which is essential in providing substance to the to the infinite sums required for the development of probability theory.

The metric, d_∞, is derived from the standard Hausdorff metric.

Definition 2.2.10 (Hausdorff metric). Given a metric space (X,d) let F_X be the set of all closed, bounded subsets of X. Given $A \in F_X$ and $r > 0$ let $N_r(A,r)$ denote the neighbourhood of A with radius r i.e. $\bigcup_{x\in A} B(x,r)$. The *Hausdorff* *metric* is given by

$$d_H(A,B) = \inf\{r > 0 \,:\, A \subset N_r(B,r) \wedge B \subset N_r(A,r)\} \qquad (2.34)$$

If a metric space is complete (i.e. every Cauchy sequence has a limit in that space) then the induced Hausdorff metric space inherits this property.

Theorem 2.2.11. *Hausdorff metric spaces inherit completeness*

If (X,d) is a complete metric space, then the Hausdorff metric induced by d is also complete.

Proof. Suppose (A_n) is a Cauchy sequence with respect to the Hausdorff metric. By selecting a sub-sequence if necessary, we may assume that A_n and A_{n+1} are within 2^{-n} of each other, that is, that $A_n \subset N_r(A_{n+1}, 2^{-n})$ and $A_{n+1} \subset N_r(A_n, 2^{-n})$.

Now for any natural number N, there is a sequence $(d_n)_{n \geq N}$ in X such that $x_n \in A_n$ and $d(x_n, x_{n+1}) < 2^{-n}$. Any such sequence is Cauchy with respect to d and thus converges to some $x \in X$. Now, by applying the triangle inequality, for any $n \geq N$, $d(x_n, x) < 2^{-n+1}$.

Define A to be the set of all x such that x is the limit of a sequence $(x_n)_{n \geq 0}$ with $n \in A_n$ and $d(x_n, x_{n+1}) < 2^{-n}$. Then A is nonempty.

Furthermore, for any n, if $x \in A$, then there is some $x_n \in A_n$ such that $d(x_n, x) < 2^{-n+1}$, and so $A \subset N_r(A_n, 2^{-n+1})$. Consequently, the set $\overline{A}$ is nonempty, closed and bounded.

Now, suppose $\varepsilon > 0$. Thus $\varepsilon > 2^{-N} > 0$ for some N. Let $n \geq N+1$. Then by applying the claim in the first paragraph, for any $x_n \in A_n$, there is some $x \in X$ with $d(x_n, x) < 2^{-n+1}$. Hence $A_n \subset N_r(\overline{A}, 2^{-n+1})$. Hence the sequence (A_n)

converges to A in the Hausdorff metric. □

In this way it is possible to produce a complete metric space over the set of closed bounded intervals of $\mathbb{R}$. Recalling that the level sets of a fuzzy number are such closed bounded intervals this in turn can be extended to the fuzzy numbers. Note that whilst various other Hausdorff-like metrics have been proposed for fuzzy sets in general, these have quite serious problems (Brass, 2002).

Definition 2.2.12 (Extended Hausdorff metric d_∞). The extended Hausdorff metric, $d_\infty : \mathbb{R}^{\mathcal{F}} \times \mathbb{R}^{\mathcal{F}} \to \mathbb{R}$ is determined by

$$d_\infty(a,b) = \sup_{\alpha \in (0,1]} d_H(L_\alpha a, L_\alpha b) \tag{2.35}$$

Under this metric the distance between embedded real numbers is simply the standard Euclidean distance, in the sense of the following Lemma.

Lemma 2.2.13. *If $a,b \in \mathbb{R}$ then $d_\infty(a_\chi, b_\chi) = |a-b|$.*

Proof. This is an immediate consequence of the definition of the extended Hausdorff metric. □

It is well-known that $(\mathbb{R}^{\mathcal{F}}, d_\infty)$ is a complete metric space.

Theorem 2.2.14. *$(\mathbb{R}^{\mathcal{F}}, d_\infty)$ is a complete metric space.*

Proof. Suppose $a_1, a_2, \ldots \in \mathbb{R}^{\mathcal{F}}$ is a Cauchy sequence then at each alphacut there is a Cauchy sequence with respect to d_H. By completeness there is a limit interval associated with that alpha. □

2.2.7 Convergence results

With this metric in place it is possible to define the convergence of sequences and sums in the usual manner.

Definition 2.2.15 (Convergence of fuzzy numbers). A sequence $a_1, a_2, \ldots \in \mathbb{R}^{\mathcal{F}}$ is said to *converge* to $a \in \mathbb{R}^{\mathcal{F}}$ if for all $\varepsilon > 0$ there is N such that for all $n \geq N$

$$d_\infty(a_n, a) \leq \varepsilon \tag{2.36}$$

Such a convergent sequence and its limit will be denoted $a_n \to a$.

Definition 2.2.16 (Convergent series). Given a sequence $a_1, a_2, \ldots \in \mathbb{R}^{\mathcal{F}}$, consider the sequence of partial sums $b_n = \sum_{i=1}^{n} a_i$. If b_n is a convergent sequence then a_n is said to be a convergent series with limit $\sum^{\infty} a_n$.

Lemma 2.2.17. *If $a_n \to a \in \mathbb{R}$ then $(a_n)_\chi \to a_\chi \in \mathbb{R}^{\mathcal{F}}$.*

Proof. By Lemma 2.2.13 for all n, $d_\infty((a_n)_\chi, a_\chi) = |a_n - a|$. Now, given $\varepsilon > 0$, since a_n is a convergent sequence, there exists an N such that

$$d_\infty((a_n)_\chi, a_\chi) = |a_n - a| \leq \varepsilon \tag{2.37}$$

for all $n \geq N$. □

2.2.8 Computational issues

It has often been observed that commonly used classes of fuzzy number are not closed under the standard arithmetic operators. So, for example, the product of two polygonal fuzzy numbers is not polygonal. This has lead some to conclude that it is not possible to have a correct (i.e. accurate) and computationally tractable calculus of fuzzy numbers.

If however, fuzzy numbers are represented by a pair of finite series of finite polynomial shoulder functions determining the upper and lower boundaries of their alphacuts, then their arithmetic combinations also fall into this class and can be computed exactly with relative ease. As a simple example, consider the fuzzy number, *a*, determined by the membership function,

$$\mu_a(x) = \begin{cases} x & \text{if } x \in [0,1] \\ 2-x & \text{if } x \in (1,2] \\ 0 & \text{otherwise} \end{cases} \tag{2.38}$$

Then a has an equivalent representation as

$$a_{\lfloor\alpha} = [\alpha, 2-\alpha] \tag{2.39}$$

Now, $a \otimes a$ has a membership function which can be obtained through solving quadratic equations and paying careful attention to boundaries

$$\mu_{a\otimes a}(x) = \begin{cases} \sqrt{x} & \text{if } x \in [0,1] \\ 2-\sqrt{x} & \text{if } x \in (1,4] \\ 0 & \text{otherwise} \end{cases} \tag{2.40}$$

But this is hardly easy to represent and performing further computations with it will be increasingly complicated. On the other hand, the alphacut representation is trivially calculated as

$$(a \otimes a)_{\lfloor\alpha} = [\alpha^2, \alpha^2 - 4\alpha + 4] \tag{2.41}$$

Further computations can be performed with similar ease. Naturally, "zero-crossing" fuzzy numbers (numbers whose membership at 0 is non-zero) require some caution and introduce precision errors (as it becomes necessary to "split" the polynomials at their roots).

2.3 Probability theory

This section constitutes a brisk introduction to the formal underpinnings of contemporary probability theory.

2.3.1 Fuzziness and probability

Both fuzzy logic and probability theory are considered theories of uncertainty, however they deal with very different types of uncertainty and in very different ways. Where fuzzy logic represents an attempt to model the uncertainty arising from the vagueness of everyday concepts, probability theory models uncertainty about what has happened (or will happen).

Unlike fuzzy logic, classical probability assumes that any proposition is either absolutely true or absolutely false, it's just that one might not be certain which. So, an apple drawn at random from a bag must be either red or not red and certainly not (as a fuzzy account might have it) a little of each.

Another key difference between the two is that fuzzy logic is truth functional: the truth value of a compound proposition is fully determined by the truth values of its constituents. So, the truth value of a conjunction is conventionally the minimum of the truth values of the conjuncts. This differs radically from probability theory where the probability of a conjunction is not determined by the probabilities of the conjuncts unless further information, such as an assertion of independence, is provided.

2.3.2 Probability Theory

The predominant formalisation of probability theory is that provided by Kolmogorov. These standard definitions may be found in any introductory text on probability theory e.g. Grimmet and Welsh (1986). Given an experiment or trial, such as rolling a die, the set of all possible outcomes or *sample space* will be denoted Ω. So, in the die example $\Omega = \{1,2,3,4,5,6\}$. Clearly, various questions may be asked about the outcome of a trial. Some of these will be elementary, of the form "Was the outcome ω?", but others will be about groups of states. Returning to the die example, one might enquire "Was the outcome an odd number?" Moreover, it is often convenient to specify the probability of propositions modelled as such groups of atomic outcomes. The notion of an event space is used to capture the idea that the relevant propositions should be closed under logical operators.

Definition 2.3.1 (Event space). A set $\mathcal{E}$ is termed an event space on a set Ω of possible outcomes if and only if

a) $\mathcal{E} \subseteq \mathbb{P}(\Omega)$

b) $\mathcal{E}$ is non-empty.

c) If $A \in \mathcal{E}$ then $A^c = \Omega \setminus A \in \mathcal{E}$

d) If $A_1, A_2, \ldots \in \mathcal{E}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{E}$

Events spaces are sometimes also referred to as "sigma algebras" and are said to be closed under complementation and countable union. Observe that these conditions entail that both Ω and $\varnothing = \{\}$ are elements of $\mathcal{E}$. With the notion of an event space in place it is possible to define the central concept of a probability measure.

Definition 2.3.2 (Classical probability measure). A mapping $\mathrm{P} : \mathcal{E} \to \mathbb{R}$ is termed a probability measure on $(\Omega, \mathcal{E})$ if and only if for all $E_1, E_2 \in \mathcal{E}$

(CP1) $\mathrm{P}(E_1) \geq 0$

(CP2) $\mathrm{P}(\Omega) = 1$

(CP3) If E_1 and E_2 are disjoint (i.e. $E_1 \cap E_2 = \varnothing$) then $\mathrm{P}(E_1) + \mathrm{P}(E_2) = \mathrm{P}(E_1 \cup E_2)$

Where P is such a probability measure, the tuple $(\Omega, \mathcal{E}, \mathrm{P})$ is termed a probability space.

2.3.3 Conditional probability

Definition 2.3.3 (Conditional probability). Given a probability space $(\Omega, \mathcal{E}, \mathrm{P})$ and $E_1, E_2 \in \mathcal{E}$ the *conditional probability* of E_1 given E_2, is defined as

$$\mathrm{P}(E_1 \mid E_2) = \frac{\mathrm{P}(E_1 \cap E_2)}{\mathrm{P}(E_2)} \tag{2.42}$$

An immediate consequence of this definition is that conditional probability with respect to a given event (E) is itself a probability measure i.e. $(\Omega, \mathcal{E}, \mathrm{P}(x \mid E))$ forms a probability space.

Although the bar notation ($|$) is standard, on closer inspection it is somewhat misleading since bar is not a set-theoretic operation. The P of $\mathrm{P}(E_1 \mid E_2)$, is therefore *not the same as* the P of $\mathrm{P}(E_1)$. For this reason, some probabilists prefer the notation $P_{E_2}(E_1)$ as this more clearly expresses the relationship with, but difference from the original measure.

To put things pragmatically for a moment, conditional probabilities have two chief uses.

The first is to capture the effect of new information, that is understanding $\mathrm{P}(A \mid B)$ and the probability that A *will* occur, given that B has already.

The second is in what artificial intelligence practitioners might term knowledge engineering. Here conditional probabilities may be used to construct a probabilistic model from a sequence of hypothetical questions. In this case $\mathrm{P}(A \mid B)$ may be understood as the probability that A *would* occur, should B.

Although these uses coincide in the classical case, there is a difference in emphasis. In the first the unknown is the conditional, while in the second it is the measure as a whole.

2.3.4 Random variables

Definition 2.3.4 (Random variable). Given a probability space $(\Omega, \mathcal{E}, \mathrm{P})$, and a domain D_X, a function $X : \Omega \rightarrow D_X$ is termed a *random variable* on $(\Omega, \mathcal{E}, \mathrm{P})$ if and only if for all $x \in D_X$,

$$X^{-1}(x) = \{\omega \in \Omega \,:\, X(\omega) = x\} \in \mathcal{E} \tag{2.43}$$

Note that where two or more random variables are defined with respect to the same probability space it is trivial to construct a random variable that represents their joint distribution. So, suppose X and Y are random variables, consider $Z : \Omega \rightarrow D_X \times D_Y$ given by

$$Z : \omega \mapsto (X(\omega), Y(\omega)) \tag{2.44}$$

Now for any $z = (x,y) \in D_X \times D_Y$

$$Z^{-1}(z) = X^{-1}(x) \cap Y^{-1}(y) \tag{2.45}$$

which as a finite intersection of elements is in $\mathcal{E}$.

2.3.5 Discrete random variables

Given a random variable X if D_X is finite, X is termed a *discrete random variable*. Discrete random variables have a useful representational property through the concept of a mass function.

Definition 2.3.5 (Mass function). Given a discrete random variable, X, defined on a probability space $(\omega, \mathcal{E}, \mathrm{P})$, the function $\mathrm{p} : D_X \to [0,1]$ determined by

$$\mathrm{p}_X(x) = \mathrm{P}(X^{-1}(x)) \tag{2.46}$$

is termed the *mass function* of X.

Theorem 2.3.6 (Representation Theorem). *Given a domain, $D = \{d_1, d_2, \ldots, d_N\}$, and $\pi_1, \pi_2, \ldots, \pi_N \in [0,1]$ such that $\sum_{i=1}^{N} \pi_i = 1$ there is a random variable, X with domain such that for all $d_i \in D$,*

$$\mathrm{p}_X(d_i) = \pi_i \tag{2.47}$$

Proof. Since D is finite, define $P : \mathbb{P}(D) \to [0,1]$ by

$$\mathrm{P}(E) = \sum_{i \,:\, d_i \in E} \pi_i \tag{2.48}$$

By construction, $\mathrm{P}(D) = 1$. Moreover, for all $E_1, E_2 \in \mathbb{P}(D)$, $\mathrm{P}(E_1) \geq 0$ and if $E_1 \cap E2 = \varnothing$, $\mathrm{P}(E_1) + \mathrm{P}(E_2) = \mathrm{P}(E_1 \cup E_2)$. So, $(D, \mathbb{P}(D), \mathrm{P})$ is a probability space. Finally, take X to be the identity map (mapping elements of D to themselves). Clearly, $\mathrm{p}_X(d_i) = \pi_i$ as required. □

This theorem allows the formal differences between probability measures and mass functions to be elided most of the time.

2.3.6 Continuous random variables

Continuous random variables are simply random variables whose domain is the Borel sigma algebra of $\mathbb{R}$ (or $\mathbb{R}^n$ in the multivariate case). This is usually expressed through the following equivalent definition.

Definition 2.3.7 (Continuous Random Variable)**.** Given a probability space $(\Omega, \mathcal{E}, \mathrm{P})$ a function $X : \Omega \to \mathbb{R}$ is termed a continuous random variable if and only if for all $\alpha \in \mathbb{R}$ the preimage of $(-\infty, \alpha]$ is in $\mathcal{E}$ i.e.

$$\{x \in \Omega \, : \, X(x) \leq \alpha\} \in \mathcal{E} \tag{2.49}$$

2.4 Bayesian networks

Bayesian networks are a knowledge representation technique. As such they have applications in knowledge engineering and machine-learning contexts. This section describes the basic principles and advantages of BNs

2.4.1 Basic graph concepts

Bayesian networks are graphical models, for the purposes of conceptual and notational clarity I will rehearse some basic concepts and defintions.

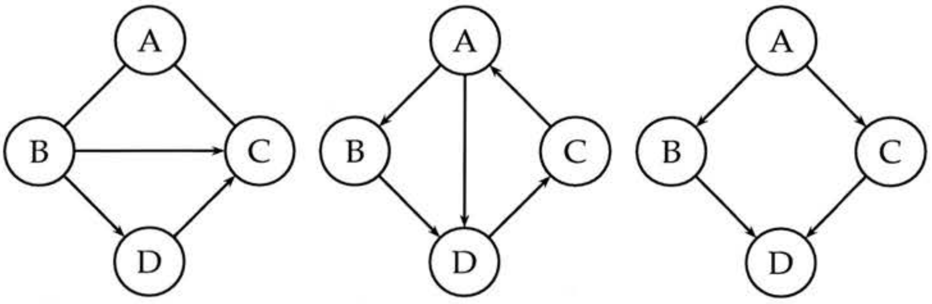
Definition 2.4.1 (Simple graph). A *simple graph* is a pair, $\mathfrak{G} = (V, \mathfrak{P})$, where V is a finite set of vertices (or nodes) and $\mathfrak{P}$ a function

$$\mathfrak{P} : V \to \mathbb{P}(V \setminus \{v\}) \tag{2.50}$$

The nodes in $\mathfrak{P}(v)$ are said to be the *parents* of v. The set $\{v\} \cup \mathfrak{P}(v)$ is termed the *family* of v and denoted $\mathfrak{F}(v)$

Note that this definition excludes both reflexive loops (nodes that are their own parents) and "multiple edges", which are sometimes considered in broader graph theory. Graphs are depicted as circled nodes with arcs (or edges) connecting them to their parents.

Definition 2.4.2 (Directed graph). The arc between two nodes, v_1, v_2 is said to be *undirected* if $v_1 \in \mathfrak{P}(v_2)$ and $v_2 \in \mathfrak{P}(v_1)$. If this is not the case, the arc is said to be *directed*. Graphs containing only (un)directed arcs are themselves termed (un)directed. If a graph contains a mixture of both directed and undirected arcs, it is termed a *chain graph*.



(a) A simple graph (b) Directed graph (c) Directed acyclic graph

Figure 2.2: Some examples of different types of graph

Undirected arcs are commonly depicted as plain connecting lines, whereas directed arcs have arrow pointing from parent nodes to their children. Figure 2.2(a) displays a sample chain graph with $V = \{A,B,C,D\}$, $\mathfrak{P}(A) = \{B,C\}$, $\mathfrak{P}(B) = \{A\}$, $\mathfrak{P}(C) = \{A,B,D\}$ and $\mathfrak{P}(D) = \{B\}$.

Definition 2.4.3 (Connected graph)**.** An undirected graph, $\mathfrak{G} = (V,\mathfrak{P})$ is said to be *connected* if there is a "path" between any two nodes i.e. for all $v_1, v_2 \in V$ there is a sequence $l_1, l_2, \ldots, l_N \in V$ such that $l_1 = v_1$, $l_N = v_2$ and $l_{i+1} \in \mathfrak{P}(l_i)$ for all $i \in 1,2,\ldots,N-1$.

Lemma 2.4.4. *A directed acyclic graph, $(V,\mathfrak{P})$ has at least one node, l such that for all $v \in V$, $l \notin \mathfrak{P}(v)$. Such a node is termed a* leaf.

Proof. Suppose for a contradiction that every node has a child. In this case there exists a sequence $v_1, v_2, \ldots, v_N$ such that $v_{i+1} \in \mathfrak{P}(v_i)$ for all i. However if $N > |V|$ by the pigeon hole principle $v_i = v_j$ for some $i \neq j$. This contradicts the acylicity of the graph. □

2.4.2 Representation

A Bayesian network represents a joint probability mass function as a directed acyclic graph whose nodes correspond to discrete random variables. Associated with each node is the conditional probability table for that variable *given* its parent nodes. However, these associations are a matter of interpretation whereas the network itself is a formal construct:

Definition 2.4.5 (Bayesian network)**.** A *Bayesian network* consists of three elements:

a) A finite directed acyclic graph, $\mathfrak{G} = (V,\mathfrak{P})$

b) An association between each of node of $\mathfrak{G}$, v_i and a domain or set

D_{v_i}

c) An association between each node of $\mathfrak{G}$, v_i and a function

$$\mathrm{p}_i : D_{v_i} \times \prod_{v \in \mathfrak{P}(v_i)} D_v \to [0,1] \tag{2.51}$$

with the property that for all $(d_{i1}, d_{i2}, \ldots d_{im}) \in \prod_{v \in \mathfrak{P}(v_i)} D_v$

$$\sum_{d \in D_{v_i}} \mathrm{p}_i(d, d_{i1}, d_{i2}, \ldots d_{im}) = 1 \tag{2.52}$$

Now, to flesh out the claim that such a graphical structure is an "appropriate" representation for joint probability mass functions it will be useful to examine two properties of the approach.

The first property is what Charniak (1991) calls *consistency*: that the product of the node functions may be construed as a joint probability mass function.

Theorem 2.4.6 (Consistency). *Given a Bayesian network as above, the product of the node functions where (as before) $v_i j$ denotes the jth parent of node i,*

$$\lambda(v_1, v_2, \ldots, v_n) = \prod_{i=1}^{|V|} \mathrm{p}(v_i, v_{i1}, v_{i2}, \ldots v_{im}) \tag{2.53}$$

determines a joint probability mass function on $\prod_{v \in V} D_v$.

Proof. By 2.3.6 it is sufficient to show that these terms are positive and sum to unity. The proof proceeds by induction on N. Where $N = 1$ the result holds trivially. Now suppose the result holds for all Bayesian networks with n nodes and consider a Bayesian network with $N + 1$ nodes. By 2.4.4, $\mathfrak{G}$has a leaf node. Removing this node yields a graph of n nodes which (by inductive hypothesis) is consistent. Finally, by re-arranging the sum

$$\begin{aligned}
& \sum_{d_i \in D_{v_i}} \prod_{i=1}^{N+1} \mathrm{p}_i(d_i, d_{i1}, d_{i2}, \ldots, d_{im}) && (2.54) \\
&= \sum_{d_i \in D_{v_i}, i \neq l} \prod_{i \neq l}^{N+1} \mathrm{p}_i(d_i, d_{i1}, d_{i2}, \ldots, d_{im}) \sum_{d_l \in D_{v_l}} \mathrm{p}_l(d_l, d_{l1}, d_{l2}, \ldots d_{lm}) && (2.55) \\
&= \sum_{d_i \in D_{v_i} \forall i \neq l} \prod \mathrm{p}_i(d_i, d_{i1}, d_{i2}, \ldots, d_{im}) && (2.56) \\
&= 1 && (2.57)
\end{aligned}$$

the desired result is obtained. □

Consistency guarantees that the function encoded by a Bayesian network structure has the right form to be a joint distribution. It does not, however, demonstrate any connection between the abstract node functions or graphical structure and that distribution.

The second property – what will be termed *completeness* – shows that Bayesian networks are a sufficient representation i.e. that every joint distribution can be represented in that way.

Theorem 2.4.7 (Completeness). *Any probability mass function can be represented as a Bayesian network.*

Proof. Given a joint mass function $\mathrm{p}(X_1, X_2, \ldots, X_N)$, let $\mathfrak{G} = (V, \mathfrak{P})$ where

$$V = \{X_1, X_2, \ldots, X_N\} \tag{2.58}$$

and

$$\mathfrak{P} : X_i \mapsto \{X_j \in V \ : \ j < i\} \tag{2.59}$$

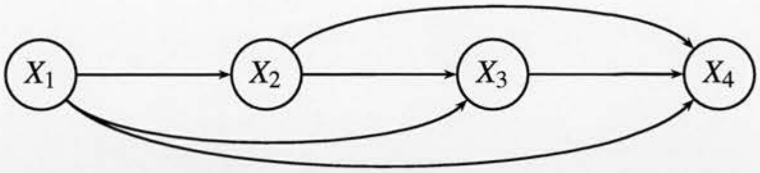


Figure 2.3: A sufficient graph for a joint distribution with four random variables.

Note that these constructions are well-founded since N is finite. An example of such a graph for a joint mass function with $N = 4$ is shown in Figure 2.3. Clearly this is a DAG. Moreover by Bayes' Theorem the product of node functions is

$$\prod_i^n \mathrm{p}(X_i \mid \mathfrak{P}(X_i)) = p(X_1, X_2, \ldots, X_N) \tag{2.60}$$

as required. □

2.4.3 Inference in Bayesian networks

Bayesian networks are not only an efficient representation of a joint probability distribution. The conditional independence relations encoded in the graphical structure support a variety of algorithms for computing probabilities of interest. This process is commonly termed inference in Bayesian networks.

It is a well-established result (Cooper, 1990) that exact inference in Bayesian networks is NP-hard (Garey and Johnson, 1979). Indeed Dagum and Luby (1993) have proven the same of approximate inference within a given error bound.

Nevertheless, outside the worst-case scenarios envisaged by complexity theory, the additional structure of conditional independence encoded by Bayesian networks, can be utilised to efficiently compute conditional probability values.

From the first-generation message-passing (Pearl, 1986, 1988) techniques to the current state-of-the-art join tree-based schemes (Lauritzen and Speigelhalter, 1988; Jensen, 1989; Huang and Darwiche, 1994) efficient inference algorithms have a common objective: to re-arrange the directed factorisation of a joint probability mass function in order to minimize the number of addition and multiplication operations required.

By way of a concrete example suppose the task was to marginalise the Markov chain

$$\lambda(x,y,z) = f(x,y)g(y,z)h(z) \tag{2.61}$$

with respect to y and z. The simplest approach is just to sum

$$\sum_{y \in D_y, z \in D_z} f(x,y)g(y,z)h(z) \tag{2.62}$$

This calculation involve $|D_y||D_z|$ additions and $2|D_y||D_z|$ multiplications. However this sum of products may be re-arranged as

$$\sum_{y \in D_y} f(x,y) \sum_{z \in D_z} g(y,z)h(z) \tag{2.63}$$

which requires only $|D_y| + |D_z|$ additions and $|D_y||D_z|$ multiplications. Although the asymptotic time complexity remains exponential, this calculation can be much more efficient in practice.

Algebraically, all that these transformations require is that the underlying ring (or indeed semi-ring) be distributive.

2.5 Summary

This chapter has sought to present in condensed form the mathematical notation and concepts that underpin the work presented in this thesis.

The material presented in the first two sections, describing fuzzy logic and probability theory, is an essential pre-requisites for understanding the following chapter. There a survey is made of the ways in which other fuzzy logic researchers have attempted to hybridize fuzzy logic and probability theory.

But the detail provided above comes into play most crucially in Chapter 4 where the novel theory of Linguistic probabilities is developed. The various concepts, theorems and lemmas marshaled above will feed into both the specification of theory and the pendant analysis of its properties. Roughly speaking, the aim will be to "carry over" as much of classical probability theory as possible, but also to illuminate the differences between the two.

The final section above, with its brisk sketch of Bayesian network research may be largely disregarded until Chapter 5 where the theoretical application of the Linguistic Bayesian network is presented.

Chapter 3

Vague probabilities

Over the years, various hybrid fuzzy-probabilistic theories have been presented in the literature. It is convenient to think of these in context of the hierarchy of classical probability theory a diagram of which may be found in Figure 3.1. The ovals in this diagram represent different "spaces", understood conventionally as classical sets, and the arrows represent the mappings whose definitions essentially characterise the theory. The various hybrid theories described here may be seen as replacing one or more of the spaces, and the relevant functions with appropriate fuzzy generalisations.

3.1 Fuzzy random variables

In practice there are many situations where the factors of interest are not real numbers (or real vectors in the more general multivariate case) but linguistic terms. So, for example, one might be interested in whether a particular

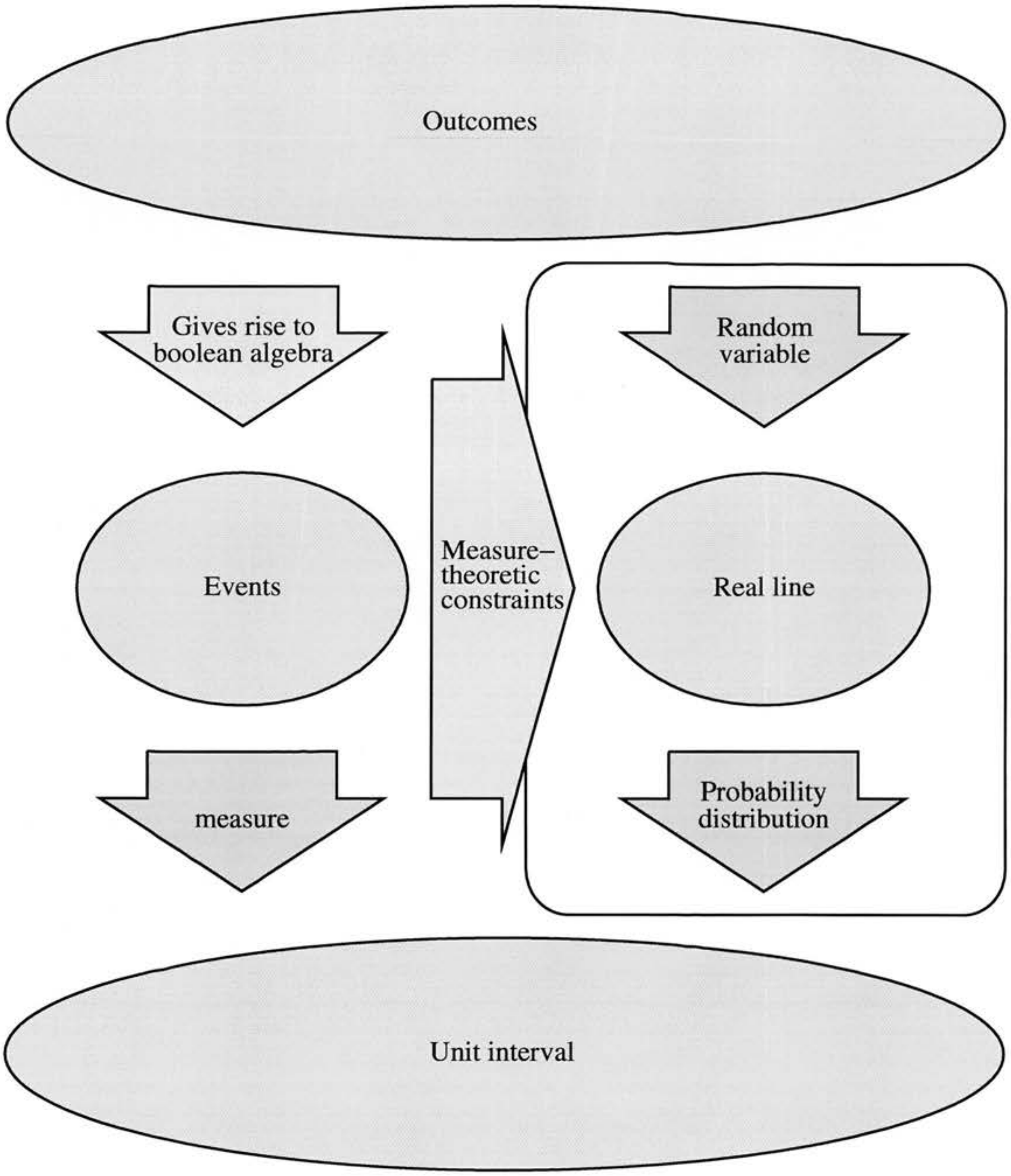


Figure 3.1: A schematic of conventional probability theory. The oval shapes represent classical sets. In hybrid fuzzy-probabilistic theories, one or more of these is replaced by an appropriate fuzzy generalisation.

measurable is *much greater than zero*. Alternatively it may be the case that the outcomes of an experiment are expressed in inexact linguistic terms. As an example (derived from Puri and Ralescu (1986)), consider a survey whereby people are asked to give their opinion about the weather. Typical responses might include "cold", "not bad for this time of year" and so on. In this context is reasonable to ask: What is the average opinion about the weather?

The term "fuzzy random variable" was first introduced in the literature by Kwakernaak (1978a,b) and refined by Kruse and Meyer (1987). In this theory (KKM) fuzzy random variables are interpreted as fuzzy perceptions of classical random variables (referred to as the *original* of its fuzzy counterpart), much as a linguistic label can be interpreted as fuzzy perceptions of its base variable.

Of the two, the second framework is more general but their equivalence under reasonable assumptions has been demonstrated by Zhong and Zhou (1987). The main difference seems to be stylistic - Kwakernaak's framework is more real-analysis oriented, Puri and Ralescu's more topological. In both settings it is possible to give well-founded fuzzy counterparts of statistics such as expected value and variance (Kruse and Meyer, 1987; Puri and Ralescu, 1986). There are also proofs of major limit theorems such versions of the Law of Large Numbers and the Central Limit Theorem. Summaries of these results can be found in Corral et al. (1998) and Ralescu (1995b).

3.2 Fuzzy information systems

An alternative type of approach concerns what conclusions may be drawn about a classical probabilistic system on the basis of available fuzzy information. Okuda et al. (1978) and Tanaka et al. (1979) have developed techniques centring on the concept of a fuzzy information system.

Given a probability space $(\Omega, \mathcal{E}, \mathrm{P})$ and a classical random variable, X a Borel-measurable fuzzy set of $X(\Omega)$ is termed fuzzy information. A partition (in the sense of Ruspini (1970)) is then defined to model all available observations.

3.3 Zadeh's fuzzy probabilities

The term "fuzzy probability" first appears, albeit incidentally, in the second part of Lotfi Zadeh's seminal paper on linguistic variables (Zadeh, 1975b). Here, Zadeh mentions in passing that fuzzy quantifiers, which are introduced to capture the sense of vague quantifiers such as "most" or "few" can be thought of as being like fuzzy probabilities. It was not until the mid-eighties, however, that Zadeh revisited the idea. In "Fuzzy probabilities" (Zadeh, 1984) he presents two quite separate answers to the question "What is the probability of a fuzzy event?".

The following section outlines Zadeh's approach with reference to a simple data-centred example. It is followed by critical examination of the two techniques which identifies a series of statistical, theoretical and technical ambiguities and deficiencies.

x	b_0	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9
$\mu_B(x)$	0.4	0	0.2	0.8	1	0	0.9	0.4	0.1	0.5

Table 3.1: A bag of balls and their blackness

The main example of Zadeh (1984) seeks to assess the probability that a car, c, will be stolen as the probability that a randomly selected car is (fuzzily) like c and has been stolen. This is inexplicable as it is clear that a conditional probability (what is the chance that a car will be stolen given that it is like c?) would be more appropriate.

As a basis for the following discussion then, consider instead a bag containing 10 balls of varying shade. The fuzzy set of black balls (with respect to the universe of discourse $U = \{b_0, b_1, \ldots b_9\}$) has a membership function μ_B listed in Table 3.1. Zadeh's approaches will then be assessed with respect to the question: What is the probability of selecting a black ball?

3.3.1 ZFP Type 1

Zadeh's first approach to this question calculates the desired probability in terms of the crisp "cardinality" (or "sigma-count") of B. This is obtained by simply summing up the membership function of the fuzzy set.

$$\Sigma\text{count}(B) = 0.4 + 0 + 0.2 + 0.8 + 1 + 0 + 0.9 + 0.4 + 0.1 + 0.5 = 4.3 \tag{3.1}$$

The probability is then calculated by dividing this through by the size of the universe of discourse i.e.

$$\text{ZP}(B) = \frac{\Sigma\text{count}(B)}{|U|} = \frac{4.2}{10} = 0.42 \tag{3.2}$$

3.3.2 ZFP Type 2

Zadeh's second definition relies on the concept of "fuzzy cardinality" which is rather loosely defined in terms of the "sigma representation" of a fuzzy set. Cleaning up some ambiguities and recasting in modern notation yields the following definition,

Definition 3.3.1 (Zadeh's fuzzy cardinality). Given D, a universe of discourse and A, a fuzzy set defined over D, the *fuzzy cardinality of* A,

$$\mathrm{FGcount}(A)_{\lfloor\alpha} = [|A_{\lfloor 1}|, |A_{\lfloor\alpha}|] \tag{3.3}$$

Or, equivalently,

$$\mu_{\mathrm{FGcount}(A)}(x) = \max\left\{ \max_{x' \in X}(\mu_A(x')) \;:\; X \in \mathbb{P}(D) \wedge |X| \leq x \right\} \tag{3.4}$$

Definition 3.3.2 (Zadeh's fuzzy probability). Given D, a (finite) universe of discourse and A, a fuzzy set defined on that universe, the fuzzy probability of A,

$$\mathrm{ZFP}(A) = \frac{\mathrm{FGcount}(A)}{|D|_\chi} \tag{3.5}$$

Returning then to the black ball example, this definition yields the fuzzy number with the membership function,

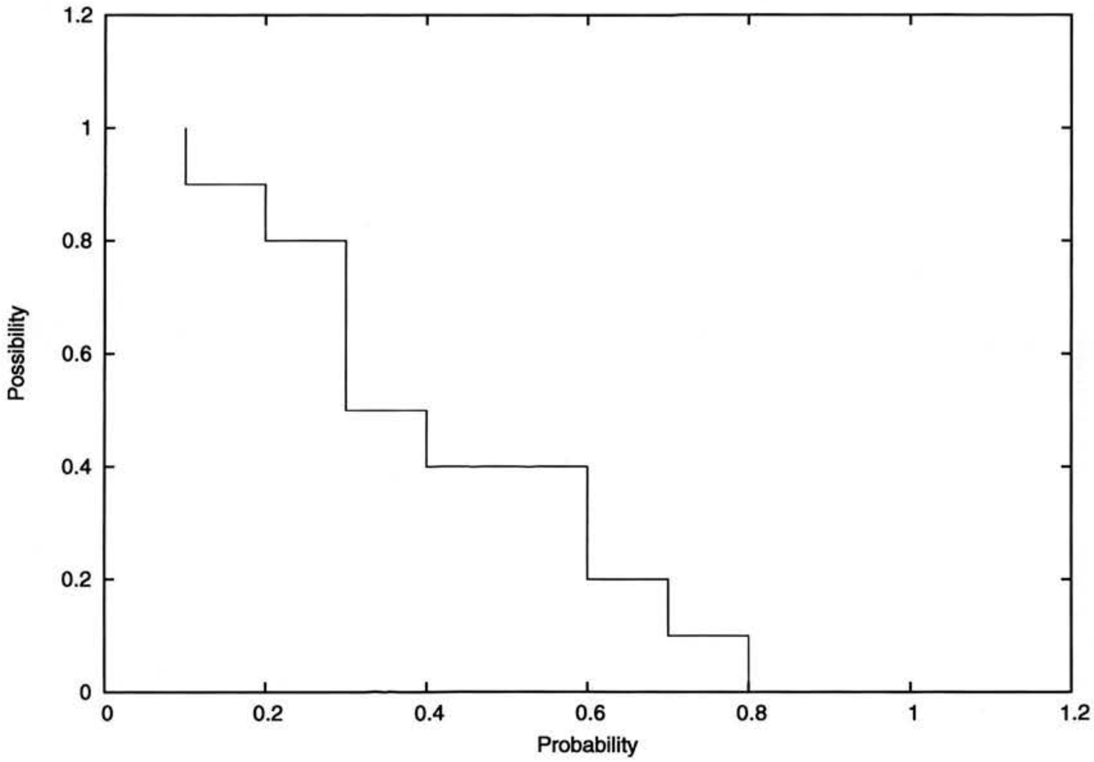


Figure 3.2: The fuzzy probability, $\mathrm{ZFP}(B)$. As is conventional the discontinuous membership function has been drawn with a single line.

$$\mu_{\mathrm{ZFP}(A)}(x) = \begin{cases} 0 & x < 0.1 \\ 1 & x = 0.1 \\ 0.9 & 0.1 < x \leq 0.2 \\ 0.8 & 0.2 < x \leq 0.3 \\ 0.5 & 0.3 < x \leq 0.4 \\ 0.4 & 0.4 < x \leq 0.6 \\ 0.2 & 0.6 < x \leq 0.7 \\ 0.1 & 0.7 < x \leq 0.8 \\ 0 & x > 0.8 \end{cases} \tag{3.6}$$

This is graphed in Figure 3.2.

3.3.3 Critical evaluation

Although these approaches have a certain intuitive appeal, they share a number of serious problems.

First, the definitions involved tacitly assume a finite set of outcomes with equiprobable elements. Such a space, although commonplace in didactic examples involving cards and coins, does not provide an adequate framework for infinite sets of outcomes (as required for continuous random variables). The assumption that outcomes are equiprobable is also problematic – consider, to take a trivial example, an unfair coin.

This said, it is relatively easy to see how to begin to remedy these deficiencies.

Definition 3.3.3 (Measurable fuzzy set). Given a probability space $(\Omega, \mathcal{E}, \mathrm{P})$ a fuzzy set A (defined over Ω) is said to be *measurable* if and only if for all $\alpha \in [0,1], A_{\lfloor\alpha} \in \mathcal{E}$.

By definition a fuzzy set is measurable if and only if its membership function is measurable (in the usual sense). Note also that the measurable sets form a closed class under (fuzzy) intersection, union and complementation.

Now, supposing that A is a measurable fuzzy set the natural extension of Zadeh's first measure, ZP, is the Lebesgue integral,

$$\mathrm{ZP}'(A) = \int_{\Omega} \mu_A \, d\mathrm{P} \tag{3.7}$$

An analogue for ZFP is similarly well-defined:

$$\mathrm{ZFP}'(A)_{\lfloor\alpha} = [\mathrm{P}(A_{\lfloor 1}), \mathrm{P}(A_{\lfloor\alpha})] \tag{3.8}$$

However, even if it is accepted that these formulations resolve the technical issues with Zadeh's approaches the resulting theory of probability has undesirable consequences.

First, the sort of fuzzy probabilities envisaged by ZFPall have the same sort of "left-crisp" shape – speaking loosely that of a right angle triangle facing to the right. Thus ZFP is out of the running when it comes to grounding subjective probability assessments such as "about an even chance".

Another concern, invited perhaps by the concentration of two quite separate ideas into a single paper, is that there are no guidelines for applicability i.e.

when to use the crisp probability of a fuzzy event and when the fuzzy. It is perhaps unfair to level this criticism at what is clearly a preliminary investigation, but on the other hand this lack of context makes it difficult to see how to extend the work.

Finally, whilst Zadeh's approach may be useful for such data-centred applications, from the point of view of the probability theorist this approach is somewhat dubious as it rests on the assumption that the set out outcomes is both finite and that individual outcomes are equiprobable.

Zadeh's theory then, as a finite theory, cannot be expected to provide analogues for classical random variables. But worse still, the equiprobability assumption means that it is restricted to analytic examples (such as lotteries and card games) where all information is available. Ideally a theory of fuzzy probabilities would support a clear calculus in their own terms rather than, as here, a frequentist estimation method.

3.4 Bayesian Fuzzy probabilities

On Zadeh's theory fuzziness in a probability is secondary – merely a shadow of the primary fuzziness of the event of interest itself. Although survey papers have tended to conflate the two, it is exactly this point that distinguishes the different approach to "fuzzy probabilities" taken by Jain and Agogino (1990). Arguably this paper has been the most influential publication in the area, however it will be demonstrated that for technical reasons the theory it presents cannot provide a satisfactory model for qualitative probability assessments.

Jain and Agogino call their version of fuzzy probabilities "Bayesian fuzzy probabilities". The presentation here differs substantially from the original formulation of these ideas. In particular, Jain and Agogino do not explicitly use the idea of a probability measure either to locate their belief measure, bf, in the context of an event space or to define the associcated "mean" function, m.

Definition 3.4.1 (Bayesian fuzzy probability measure). Given an event algebra, $\mathcal{E}$, defined over a set of outcomes, Ω, a function bf from the set of events to the set of "convex normalised fuzzy set[s] ... of $[0,1]$" is a *Bayesian fuzzy probability measure* if and only if for all $A, B \in \mathcal{E}$

(BF1) bf(A) has a unique "mean" i.e. there is a function $m : \mathcal{E} \to [0,1]$ such that for all $x \in [0,1]$, $\mu_{\mathrm{bf}(A)}(x) = 1$ if and only if $x = m(A)$ (in this case $\mathrm{bf}(A)$ is said to be *unimodal*)

(BF2) $\mu_{\mathrm{bf}(A)}$ is continuous on (0,1)

(BF3) m (as defined in BF1) is a probability measure

(BF4) $\mathrm{bf}(\Omega) = 1_\chi$

(BF5) If A and B are disjoint then $\mathrm{bf}(A) \oplus \mathrm{bf}(B) = \mathrm{bf}(A \cup B)$

At first sight this definition seems reasonable and indeed it can and has be used as an informal theory for reasoning with fuzzy probabilities, however as a formal theory it is seriously defective as a consequence of the following Lemma.

Lemma 3.4.2. *For any event $E \in \mathcal{E}$, $\mu_{\mathrm{bf}(E)}(x) = 0$ for all $x < m(E)$. Such a membership function is termed left-crisp.*

Proof. Consider an arbitrary event $E \in \mathcal{E}$. By definition $\mathrm{bf}(E)$ has a unique mode $m(E) \in [0,1]$ such that $\mu_{\mathrm{bf}(E)}(m(E)) = 1 = \mu_{\mathrm{bf}(E^c)}(1-m(E))$. Suppose, for a contradiction, that $0 < \mu_{\mathrm{bf}(E)}(x) \leq 1$ for some $0 \leq x < m(E)$. Clearly, $0 < 1 - m(E) + x < 1$ and by the definition of $\oplus$

$$\begin{aligned} 0 &= 1_{\chi}(1-m(E)+x) \\ &= \mu_{\mathrm{bf}(E\cup E^c)}(1-m(E)+x) \\ &= \max_{z+z'=1-m(E)+x} \min(\mu_{\mathrm{bf}(E)}(z), \mu_{\mathrm{bf}(E^c)}(z')) \\ &\geq \min(\mu_{\mathrm{bf}(E)}(x), \mu_{\mathrm{bf}(E^c)}(1-m(E))) \\ &> 0 \end{aligned}$$

which is the desired contradiction. Hence $\mu_{bfp(E)}(x) = 0$ for all $x < m(E)$. □

Thus every BFP is necessarily left-crisp and therefore the theory cannot act as a formal model for vague probability assessments such as "quite likely" which tail-off smoothly to the left of their peak. Note that this criticism applies also to Zadeh's fuzzy probabilities.

Worse still, there are two ways to strengthen this result to a proof that BFPs can only be embedded point probabilities (i.e. both left and right-crisp), both of which seem to reflect Jain and Agogino's intentions if not their precise formulation.

First, $\oplus$ is an operator defined on the fuzzy reals, not pairs of "convex normalised fuzzy set[s] …of $[0,1]$". Lemma 1 rests only on the assumption that Jain and Agogino tacitly intended some form of $\oplus$ restricted to the unit interval. Without such a restriction Bayesian fuzzy probabilities reduce immediately to embedded classical probabilities. To see this, suppose that for some

$m(E) < x \leq 1$, $\mu_{\mathrm{bf}(E)}(x) > 0$. Then, as before,

$$\begin{aligned} 0 &= \mu_{1_\chi}(1 - m(E) + x) \\ &= \mu_{\mathrm{bf}(E \cup E^c)}(1 - m(E) + x) \\ &\geq \min(\mu_{\mathrm{bf}(E)}(x), \mu_{\mathrm{bf}(E^c)}(1 - m(E))) \\ &> 0 \end{aligned}$$

And thus, $\mathrm{bf}(E) = m(E)_\chi$.

Second, it seems clear from their examples, that Jain and Agogino intend that for all $E \in \mathcal{E}$, $\mathrm{bf}(E) = 1_\chi \ominus \mathrm{bf}(E^c)$. Indeed this principle has considerable intuitive appeal, since it is roughly equivalent to the assertion that if you know something (however imprecise) about the probability of an event, then you know "just as much" about the probability of that event's complement. This will be elaborated further in the following Section.

But in this case, since every event is the complement of some left-crisp event, all events are also right-crisp. Again, the theory reduces to an embedding of classical point probabilities. Similar criticisms of the existing theories, albeit couched in very different language and developed independently can be found in Gert de Cooman's most recent work (de Cooman, 2003a) on possibilistic previsions.

3.5 Relationship with Type-2 fuzzy sets

There is a formal resemblence between the two theories of fuzzy probability described above and Type-2 fuzzy sets (Mendel and John, 2002; Mendel et al., 2007). First, like fuzzy probabilities, Type-II fuzzy sets may be derived from the process of fuzzification, applied in this case to fuzzy sets themselves. So, the membership function of a Type-2 fuzzy set maps from some universe of discourse to fuzzy sets of the unit interval (usually fuzzy numbers). Like any derived fuzzy theory, the Extension Principle may then be used to develop analogues for associated functions and relations. In the case of Type-2 fuzzy systems this means the fundamental truth-functional operators and inference procedures of conventional fuzzy logic.

The second similarity is that, as the name suggests Type-2 fuzzy theory is a second-order theory of uncertainty, which is to say that it represents an attempt to represent and process uncertainty about uncertainty. The primary uncertainty of Type-2 approaches is however that of vagueness or conceptual fit.

Although there are these two methodological and formal similarities between the Type-2 fuzzy modelling strategy and the approaches advanced by advocate of fuzzy probabilies there are equally significant differences. The crucial point is that fuzzy logic (like classical logic) is a truth-functional calculus – the truth values of individual propositions fully determine the truth values of compound propositions. At risk of labouring the point, this is simply not the case with probability theory where the probabilities of individual events do (in themselves) determine the probability of a compound event.

These differences mean that the representations, techniques, questions and answers associated with the growing body of literature on Type-2 fuzzy sets are different from those clustered around the idea of fuzzfying probabilities.

Thus whilst acknowledging a certain commonality of interest and approach, the present work is quite independent of that area of study.

3.6 Summary

This chapter has reviewed a number of hybrid fuzzy-probabilistic theories in an attempt to locate the research that will be presented in the following chapters with respect to existing work in the area. The discussion centred on the contributions from Zadeh and Jain and Agogino which are the most widely cited theories of "fuzzy probability".

A close analysis of these theories showed that they have significant weaknesses. Zadeh's work is practically-oriented and because it lacks a formalisation, not immediately applicable outside data-centric context and (relatedly) ill-at-ease with contemporary subjectivist accounts of probabilities. Jain and Agogino's work is therefore to be praised for its attempt to address these issues by providing a formalisation of the concepts. However, as shown above, this formalisation is fundamentally flawed, with their Bayesian fuzzy probabilities reducing immediately to embedded classical (or "point") probabilities.

Nevertheless, the core ideas behind these two bodies of work are sound. In this way, and as a matter of biography, they are the motive and cause of what follows.

Chapter 4

Linguistic Probability Theory

This chapter introduces the theory of linguistic probabilities. The presentation here is self-consciously patterned after the exposition of classical probability theory in Chapter 2 to emphasise the parallels between the proposed and the classical constructions.

The first section presents the basic characterisation of linguistic probability measures. The axioms are discussed and some simple consequences of the theory examined. The second section builds upon these measure-theoretic foundations by developing a theory of linguistic random variables. The discrete case is examined in some detail and – crucially – an analogue for the classical representation theorem for discrete random variables is provided. The chapter closes with a sketch of how the theory might be developed to cover continuous linguistic random variables.

4.1 Linguistic probability measures

As with the classical theory, Linguistic probabilities are defined in terms of a measurable function.

Definition 4.1.1 (Linguistic probability measure). Given an event algebra $\mathcal{E}$ defined over a set of outcomes Ω, a function $\mathrm{LP} : \mathcal{E} \to \mathbb{R}^{\mathcal{F}}$ is termed a *linguistic probability measure* if and only if for all $A, B \in \mathcal{E}$

(LP1) $0_\chi \preccurlyeq \mathrm{LP}(A) \preccurlyeq 1_\chi$

(LP2) $\mathrm{LP}(\Omega) = 1_\chi$

(LP3) If $A_1, A_2 \dots$ are a sequence of pairwise disjoint events then

$$\sum_{i}^{\infty}{}^{\circ}\, \mathrm{LP}(A_i) \supseteq \mathrm{LP}(\bigcup_{i}^{\infty} A_i) \tag{4.1}$$

(LP4) $\mathrm{LP}(A) = 1_\chi \ominus \mathrm{LP}(A^c)$

where LP is a linguistic probability measure on $(\Omega, \mathcal{E})$, the tuple $(\Omega, \mathcal{E}, \mathrm{LP})$ is termed a linguistic probability space. Note that the $\sum^{\circ}$ notation represents an ininite sum of the sort discussed in 2.2.7.

Like the first two axioms of classical probability theory LP1 and LP2 simply specify the quantity space in which probabilities will be assessed. Note that LP1 entails that linguistic probabilities have zero membership outside the unit interval and (together with LP4) that $\mathrm{LP}(\varnothing) = 0_\chi$. Again, although the choice of unit interval is somewhat arbitrary it simplifies the definition of expectation. The most significant parts therefore are LP3 and LP4. The underlying intuition is that vagueness in a probability acts as a soft constraint on all probabilities that are logically linked to it.

LP3 – countable subadditivity – is intended to capture the intuition one might know the probability of the union of (say) two disjoint events more precisely than the probabilities of either individually. Consider, for example, tossing a coin which one is told is almost unbiased. Here, knowledge about the probability of the result being heads (or tails) is uncertain, but the probability that result will be either heads or tails is certain (and equal to 1). Equally, the probability that the result will *not* be both heads and tails is certain (and equal to 0). Note that LP3 is asserting that $\mathrm{LP}(A_i)$ is a convergent series.

In a similar vein, LP4 expresses that knowing something about the probability of an event translates into equally precise knowledge about the probability of its complement. Or, to put it another way, that it is unthinkable that one might have more knowledge about the probability of an event than it's complement.

4.1.1 Properties of linguistic probability measures

Classical probability measures are monotonic and this property is shared by linguistic probability measures as the following theorem demonstrates.

Theorem 4.1.2. $(\Omega, \mathcal{E}, \mathrm{LP})$ *and* $A, B \in \mathcal{E}$ *if* $A \subseteq B$, *then* $\mathrm{LP}(A) \preccurlyeq \mathrm{LP}(B)$

Proof. By LP3, for any $\alpha \in [0,1]$ $\mathrm{LP}(B) \subseteq \mathrm{LP}(A) \oplus \mathrm{LP}(B \setminus A)$,

$$\lfloor \mathrm{LP}(B) \rfloor_\alpha \geq \lfloor \mathrm{LP}(A) \rfloor_\alpha + \lfloor \mathrm{LP}(B \setminus A) \rfloor_\alpha \tag{4.2}$$

Now, $\lfloor \mathrm{LP}(B \setminus A) \rfloor_\alpha \geq 0$ and hence $\lfloor \mathrm{LP}(B) \rfloor_\alpha \geq \lfloor \mathrm{LP}(A) \rfloor_\alpha$. On the other hand, since $\mathrm{LP}(A^c) \subseteq \mathrm{LP}(B^c) \oplus \mathrm{LP}(A^c \setminus B^c)$,

$$\lfloor \mathrm{LP}(A^c) \rfloor_\alpha \geq \lfloor \mathrm{LP}(B^c) \rfloor_\alpha + \lfloor \mathrm{LP}(A^c \setminus B^c) \rfloor_\alpha \tag{4.3}$$

As before, $\lfloor \mathrm{LP}(A^c \setminus B^c) \rfloor_\alpha \geq 0$ and hence

$$\lceil \mathrm{LP}(A) \rceil_\alpha = 1 - \lfloor \mathrm{LP}((^{c}A)) \rfloor_\alpha \leq 1 - \lfloor \mathrm{LP}(B^c) \rfloor_\alpha = \lceil \mathrm{LP}(B) \rceil_\alpha \qquad \square$$

In classical probability theory there is an important result asserting that probability measure are continuous. This result paves the way for continuous random variables and density functions. Linguistic probability measures also share this property.

Theorem 4.1.3. *Given a linguistic probability space* $(\Omega, \mathcal{E}, \mathrm{LP})$*, and an increasing sequence* $A_1 \subseteq A_2 \subseteq \ldots$ *with limit* $A = \bigcup_i^\infty A_i$ *then*

$$\mathrm{LP}(A) = \lim_{n \to \infty} \mathrm{LP}(A_n) \tag{4.4}$$

Proof. Define a sequence B_i, such that

$$B_i = A_{i+1} \setminus A_i \tag{4.5}$$

Clearly the B_is are disjoint from one another and from A_i (for a given i). Hence, by LP3, for all N,

$$\mathrm{LP}(A) \subseteq \mathrm{LP}(A_N) \oplus \mathop{\sum\!\!\!\!\!\circ}_{N}^{\infty} \mathrm{LP}(B_i) \tag{4.6}$$

By Theorem 4.1.2 the sequence of partial sums is decreasing and bounded below (by 0_χ) and therefore converges. Furthermore,

$$\lim_{N \to \infty} \mathop{\sum\!\!\!\!\!\circ}_{N}^{\infty} \mathrm{LP}(B_i) = 0_\chi \tag{4.7}$$

Now since all limits exist,

$$\mathrm{LP}(A) \subseteq \lim_{N \to \infty} \mathrm{LP}(A_N) \tag{4.8}$$

Similarly

$$\mathrm{LP}({A_N}^c) \subseteq \mathrm{LP}(A^c) \oplus \mathop{\sum\!\!\!\!\!\circ}_{N}^{\infty} \mathrm{LP}(B_i) \tag{4.9}$$

and hence

$$1_\chi \ominus \lim_{N\to\infty} \mathrm{LP}(A_N) = \lim_{N\to\infty} \mathrm{LP}({A_N}^c) \subseteq \mathrm{LP}(A^c) = 1_\chi \ominus \mathrm{LP}(A) \tag{4.10}$$

thus, by Lemma 2.2.9,

$$\lim_{N\to\infty} \mathrm{LP}(A_N) \subseteq \mathrm{LP}(A) \tag{4.11}$$

Taken together 4.8 and 4.11 yield the desired result. □

4.1.2 Relation with classical probabilities

Linguistic probability measures generalise classical probability measures in the sense of the following Lemmas.

Lemma 4.1.4. *Given a classical probability measure,* P, *the map* $\mathrm{LP} : \mathcal{E} \to \mathbb{R}^{\mathcal{F}}$ *determined by* $\mathrm{LP}(A) = (\mathrm{P}(A))_\chi$ *is a linguistic probability measure.*

Proof. Clearly, $\mathrm{LP}(\Omega) = 1_\chi$, $\mathrm{LP}(\varnothing) = 0_\chi$ and $0_\chi \preccurlyeq \mathrm{LP}(A) \preccurlyeq 1_\chi$ for all $A \in \mathcal{E}$ as required. Now for pairwise disjoint $A_1, A_2 \ldots \in \mathcal{E}$,

$$\mathrm{LP}(\bigcup_i^\infty A_i) = (\mathrm{P}(\bigcup_i^\infty A_i))_\chi = (\sum_i^\infty P(A_i))_\chi \tag{4.12}$$

which by Lemma 2.2.17,

$$= \sum_i^\infty{}_{\circ}\, \mathrm{P}(A_i)_\chi = \sum_i^\infty{}_{\circ}\, \mathrm{LP}(A_i) \tag{4.13}$$

Finally, for all $A \in \mathcal{E}$. $\mathrm{LP}(A) = (\mathrm{P}(A))_\chi = (1 - \mathrm{P}(A^c))_\chi = 1_\chi \ominus \mathrm{LP}(A^c)$. □

Thus any classical probability measure has an (unique embedding as an) equivalent linguistic probability measure. Similarly, any linguistic probability measure assigning only embedded point probabilities uniquely determines a classical probability measure.

Lemma 4.1.5. *Given a linguistic probability measure,* LP, *such that for all $A \in \mathcal{E}$, $\mathrm{LP}(A) = (p_A)_\chi$ for some $p_A \in \mathbb{R}$ the map, $\mathrm{P} : \mathcal{E} \to [0,1]$, determined by $\mathrm{P}(A) = p_A$ is a probability measure.*

Proof. Clearly, $\mathrm{P}(\varnothing) = 0$ and $\mathrm{P}(A) \geq 0$ for all $A \in \mathcal{E}$ as required. Now, given disjoint $A_1, A_2 \ldots \in \mathcal{E}$ and letting $A = \bigcup_i^\infty A_i$,

$$(p_A)_\chi = \mathrm{LP}(A) \subseteq \mathop{{\sum}\!\!\!\!\!\circ}_{i}^{\infty} \mathrm{LP}(A_i) = \sum_i^\infty (p_{A_i})_\chi = (\sum_i^\infty p_{A_i})_\chi \tag{4.14}$$

By Lemma 2.2.4, $p_A = \sum_i^\infty p_{A_i}$ as required. □

4.2 Conditional probability

As discussed in Chapter 2, conditional probabilities have two aspects. First they provide a means of updating a probability measure on receipt of new information. Second they allow the construction of complex probability measures through the consideration of simple what-if scenarios.

Given a linguistic probability space $(\Omega, \mathcal{E}, \mathrm{LP})$, suppose one discovers that $A \in \mathcal{E}$ has occurred. How should this affect ones knowledge about the linguistic probability of some other event B? As in the classical case, the idea will be to restrict the measure to A and normalise yielding a new probability measure. In classical probability theory the quantity space is a field (the real line) so this "normalization" is unproblematic. However the case of linguistic probabilities is not so clear cut.

To frame the issue more precisely: given a linguistic probability space $(\Omega, \mathcal{E}, \mathrm{LP})$ and $E \in \mathcal{E}$ (such that $\mathrm{LP}(E) \in (0_\chi, 1_\chi]$) is there a "conditional" linguistic probability measure on the sigma algebra $(E, \{E \cap A \,:\, A \in \mathcal{E}\})$ such that

$$\mathrm{LP}(E) \otimes \mathrm{LP}_E(A) = \mathrm{LP}(E \cap A) \tag{4.15}$$

for all $A \in \mathcal{E}$?

The following example shows that the answer to this question is negative at least in some cases. Consider flipping a perfectly fair coin. If the result is heads then the coin is flipped again. If not, then an *approximately* fair coin is flipped instead. The event that the first coin is heads will be denoted A, the event that the second coin – whichever coin is used – is heads, B. Table 4.1 represents a linguistic probability measure that might reasonably be used to model this scenario.

$\mathrm{LP}(X \cap Y)$	$Y = B$	$Y = B^c$
$X = A$	0.25_χ	0.25_χ
$X = A^c$	$[0.2, 0.3]_\chi$	$[0.2, 0.3]_\chi$

Table 4.1: The coins example that proves conditionalisations of linguistic probability examples do not always exist.

Now consider the composite event, D, that both coins – whichever coin is used second – show the same, be it heads or tails i.e.

$$D = ((A \cap B) \cup (A^c \cap B^c)) \tag{4.16}$$

The linguistic probability of D may be computed as $[0.45, 0.55]_\chi$, since

$$\mathrm{LP}(D) \oplus 0.25_\chi = \mathrm{LP}(D) \oplus \mathrm{LP}(A^c) \supseteq 1_\chi \ominus [0.2, 0.3]_\chi = [0.7, 0.8]_\chi \tag{4.17}$$

$\mathrm{LP}(X \cap Y)$	$X = A$	$X = A^c$
$Y = B$	$[0.03, 0.08]_\chi$	$[0.06, 0.14]_\chi$
$Y = B^c$	$[0.16, 0.27]_\chi$	$[0.56, 0.72]_\chi$

Table 4.2: A fully-factorable linguistic probability measure

and

$$\mathrm{LP}(D) \subseteq \mathrm{LP}(A \cap B) \oplus \mathrm{LP}(A^c \cap B^c) = 0.25_\chi \oplus [0.2, 0.3]_\chi = [0.45, 0.55]_\chi \tag{4.18}$$

This accords with intuition, since D's probability is, roughly speaking, dependent on the probability that the approximately fair coin yields heads. However, there can be no conditional linguistic probability measure with respect to D that satisfies equation 4.15 since no fuzzy number multiplied by $[0.45, 0.55]_\chi$ will yield $0.25_\chi = \mathrm{LP}(A \cap B)$.

But this example is also factorable, in the sense that conditioning on A yields a consistent linguistic probability measure with

$$\mathrm{LP}(A) = \mathrm{LP}(A^c) = 0.5_\chi \tag{4.19}$$

$$\mathrm{LP}_A(B) = \mathrm{LP}_A(B^c) = 0.5_\chi \tag{4.20}$$

$$\mathrm{LP}_{A^c}(B) = \mathrm{LP}_{A^c}(B) = [0.4, 0.6]_\chi \tag{4.21}$$

The linguistic probability measure tabulates in Table 4.2 shows that there are "non-trivial" fully-factorable measures.

4.3 Linguistic random variables

As in classical probability theory random variables are often more useful than events, although as we shall see, there is a simple connection between the two. The notation used in this section will generally suggest the univariate case. However, it should be noted that the results presented apply equally and immediately to the multivariate case by considering the domains that are Cartesian products of other sets.

The definition of a linguistic analogue to a classical random variable is straightforward:

Definition 4.3.1 (Linguistic random variable). Given a linguistic probability space $(\Omega, \mathcal{E}, \mathrm{LP})$, and a domain D_X, a function $X : \Omega \rightarrow D_X$ is termed a *linguistic random variable* on $(\Omega, \mathcal{E}, \mathrm{LP})$ if and only if for all $x \in D_X$,

$$X^{-1}(x) = \{\omega \in \Omega \; : \; X(\omega) = x\} \in \mathcal{E} \tag{4.22}$$

Where D_X is finite, X is termed a linguistic discrete random variable.

4.3.1 Linguistic discrete random variables

Definition 4.3.2 (Discrete linguistic random variable). Given a linguistic probability space $(\Omega, \mathcal{E}, \mathrm{LP})$, and a domain D_X, a function $X : \Omega \rightarrow D_X$ is termed a *discrete linguistic random variable* on $(\Omega, \mathcal{E}, \mathrm{LP})$ if and only if:

a) $\mathrm{Image}(X) = \{X(\omega) \; : \; \omega \in \Omega\}$ is countable

b) For all $x \in D_X$, $\{\omega \in \Omega \; : \; X(\omega) = x\} \in \mathcal{E}$

Definition 4.3.3 (Mass function). The *mass function* of a discrete linguistic random variable, X on $(\Omega, \mathcal{E}, \mathrm{LP})$ is the function, $\mathrm{lp}_X : D_X \to \mathbb{R}^{\mathcal{F}}$ determined by

$$\mathrm{lp}_X(x) = \mathrm{LP}(\{\omega \in \Omega \,:\, X(\omega) = x\}) \tag{4.23}$$

By definition, a linguistic mass function satisfies

$$0_\chi \preccurlyeq \mathrm{lp}_X(x) \preccurlyeq 1_\chi \tag{4.24}$$

and

$$\mathrm{lp}_X(x) \subseteq 1_\chi \ominus (\sum^{\circ}_{x' \neq x} \mathrm{lp}_X(x')) \tag{4.25}$$

for all $x \in D_X$. Note that b) also entails that $1_\chi \subseteq \sum^{\circ}_x \mathrm{lp}_X(x)$ since for any $x \in D_X$,

$$\begin{aligned} 1_\chi &\subseteq 1_\chi \ominus \mathrm{lp}_X(x) \oplus \mathrm{lp}_X(x) && (4.26)\\ &\subseteq 1_\chi \ominus \big(1_\chi \ominus \sum^{\circ}_{x' \neq x} \mathrm{lp}_X(x')\big) \oplus \mathrm{lp}_X(x) && (4.27)\\ &\subseteq \sum^{\circ}_{x'} \mathrm{lp}_X(x') && (4.28) \end{aligned}$$

Whilst these conditions are necessary they are also sufficient in the sense of the following theorem.

Theorem 4.3.4 (Representation Theorem). *If $D = \{d_i \,:\, i \in I\}$ is a non-empty finite set (indexed by I) and $\{\pi_i \,:\, i \in I\}$ is a set of fuzzy numbers such that for all $i \in I$*

$$0_\chi \preccurlyeq \pi_i \preccurlyeq 1_\chi \tag{4.29}$$

and

$$\pi_i \subseteq 1_\chi \ominus \sum^{\circ}_{j \neq i} \pi_j \tag{4.30}$$

then there exists a linguistic probability space, $(\Omega, \mathcal{E}, \mathrm{LP})$ *and a discrete linguistic random variable,* X*, on* $(\Omega, \mathcal{E}, \mathrm{LP})$ *with the mass function*

$$\mathrm{lp}_X(d) = \begin{cases} \pi_i & \text{if } d = d_i \text{ for some } i \in I \\ 0_\chi & \text{otherwise} \end{cases} \tag{4.31}$$

Proof. The proof proceeds by construction. Let $\Omega = D$, $\mathcal{E} = \mathbb{P}(\Omega)$ and define $\mathrm{LP} : \mathcal{E} \to \mathbb{R}^{\mathcal{F}}$ by

$$\mathrm{LP}(A) = (\sum^{\circ}_{i\,:\,d_i \in A} \pi_i) \cap (1_\chi \ominus \sum^{\circ}_{i\,:\,d_i \notin A} \pi_i) \cap [0,1]_\chi \tag{4.32}$$

By definition $\mathrm{LP}(\Omega) = 1_\chi$ and $\mathrm{LP}(\varnothing) = 0_\chi$ as required. Now, for any $A \in \mathcal{E}$ since and $0_\chi \preccurlyeq \pi_i$ for all $i \in I$ there is an $x \in [0,1]$ such that

$$\mu_{\sum_{i\,:\,d_i \in A} \pi_i}(x) = 1 \quad \text{and} \quad \mu_{1_\chi \ominus \sum_{i\,:\,d_i \notin A} \pi_i}(x) = 1 \tag{4.33}$$

Hence $\mathrm{LP}(A)$ as defined is in $[0_\chi, 1_\chi]$. Since S is finite it suffices to consider disjoint $A, B \in \mathcal{E}$. By definition,

$$\mathrm{LP}(A \cup B) \subseteq \sum^{\circ}_{i\,:\,d_i \in A \cup B} \pi_i = (\sum^{\circ}_{i\,:\,d_i \in A} \pi_i) \oplus (\sum^{\circ}_{i\,:\,d_i \in B} \pi_i) \subseteq \mathrm{LP}(A) \oplus \mathrm{LP}(B) \tag{4.34}$$

Similarly, by definition,

$$1_\chi \ominus \mathrm{LP}(A^c) = 1_\chi \ominus \{(\sum^{\circ}_{i\,:\,d_i \in A^c} \pi_i) \cap (1_\chi \ominus \sum^{\circ}_{i\,:\,d_i \notin A^c} \pi_i) \cap [0,1]_\chi\} \tag{4.35}$$

$$= (1_\chi \ominus \sum^{\circ}_{i\,:\,d_i \in A^c} \pi_i) \cap (1_\chi \ominus 1_\chi \oplus \sum^{\circ}_{i\,:\,d_i \notin A^c} \pi_i) \cap [0,1]_\chi \tag{4.36}$$

$$= \mathrm{LP}(A) \tag{4.37}$$

Thus $(\Omega, \mathcal{E}, \mathrm{LP})$ is a linguistic probability space.

Finally, define $X : \Omega \rightarrow \Omega$ as the identity i.e. such that $X(\omega) = \omega$. Now, given $d \in D$, if $d \neq d_i$ for all $i \in I$, $X^{-1}(s) = \varnothing$ and hence $\mathrm{lp}_X(d) = 0_\chi$. Otherwise, $X^{-1}(d) = \{d_i\}$ for some $i \in I$ and since $\pi_i \subseteq 1_\chi \ominus \sum^{\circ}_{j \neq i} \pi_j$

$$\mathrm{lp}_X(s) = \mathrm{LP}(\{d_i\}) = \pi_i \qquad \square$$

This theorem is important for practical applications of the theory as it allows probabilistic modelling to dispense with measure theory almost all the time and concentrate on random variables which are typically the entities of interest. It is also an essential component in the proof that linguistic analogues for Bayesian networks can be constructed.

Note that the full strength of condition 4.30 is only required to prove that the constructed linguistic random variable exactly coincides with the relevant π_i. If it were replaced by the significantly weaker condition

$$1_\chi \subseteq \sum^{\circ}_{i \in I} \pi_i \tag{4.38}$$

then $(\omega, \mathcal{E}, \mathrm{LP})$ as constructed above would still be a linguistic probability space. Thus 4.32 can also be seen as a procedure for correcting an improperly specified random variable.

4.3.2 Real-valued discrete linguistic random variables

A (fuzzy) real-valued linguistic random variable is simply a random variable whose domain is the (fuzzy) real line. Where the variable is also discrete i.e. having only a finite range, expectation can be calculated in the usual way.

Definition 4.3.5 (Expectation). Given a (fuzzy) real-valued discrete linguistic random variable $X : \Omega \to D_X$ on $(\Omega, \mathcal{E}, \mathrm{LP})$, the expectation of $X, E(X)$ is defined by

$$E(X) = \begin{cases} \sum_{x \in D_X} \mathrm{LP}(X^{-1}(x)) \otimes x_\chi & \text{if } D_X \subset \mathbb{R} \\ \sum_{x \in D_X} \mathrm{LP}(X^{-1}(x)) \otimes x & \text{if } D_X \subset \mathbb{R}^{\mathcal{F}} \end{cases} \tag{4.39}$$

Since the fuzzy numbers form a commutation semi-ring, this expectation shares many properties with its classical counterpart. So, for example, expectation is linear in the sense that for all $\alpha, \beta \in \mathbb{R}^{\mathcal{F}}$

$$E(\alpha X + \beta Y) = \alpha \otimes E(X) \oplus \beta \otimes E(Y) \tag{4.40}$$

4.3.3 Continuous linguistic random variables

As in the classical case, random variables are essentially integrable functions e.g.

Definition 4.3.6 (Continuous linguistic random Variable). Given a linguistic probability space $(\Omega, \mathcal{E}, \mathrm{LP})$, a map

$$X : \Omega \to \mathbb{R} \tag{4.41}$$

is termed a continuous linguistic random variable if and only if for all $x \in \mathbb{R}$

$$\{\omega \in \Omega \ : \ X(\omega) \leq x\} \in \mathcal{E} \tag{4.42}$$

4.4 Summary

This chapter introduced the axiomatic characterisation of linguistic probability theory. Following a brief discussion of these axioms which sought to ground them in a set of reasonable normative insights, some basic properties of linguistic probability measures were demonstrated to wit: monotonicity and continuity. These fundamental properties pave the way for the following developments.

Next, the relationship with classical probabilities was examined. It was demonstrated that the natural embedding of a classical probability measure is consistent with the axioms of linguistic probability theory, and that conversely a linguistic probability measure ranging over embedded real numbers corresponds exactly to a classical probability measure.

After this, the issue of conditional probability was examined. The key result of this section was that (unlike their classical counterparts) linguistic probability measures are not necessarily "factorable" – put simply, given such a measure, a decomposition into conditional and prior need not exist.

Finally, the theory was used to develop fuzzy analogues for random variables. The discrete case was explored in some detail and a sketch of the approach to continuous case was presented.

As explained in Chapter 2, these last two elements, conditionals and discrete random variables are the essential constituents of a Bayesian graphical representation. The following chapter explores this theoretical application in more detail.

Chapter 5

Linguistic Bayesian Networks

This chapter develops an analogue to classical Bayesian networks. A set of sufficient criteria for an annotated graph to represent a discrete linguistic joint probability mass function are presented. This can be understood as showing the *soundness* of the representation. The issue of representational completeness is then discussed. No conclusion is presented on this point although it is shown that the standard constructive proof strategy is not applicable in this context. Inference procedures are then examined and it is shown that best-of-breed classical procedures may be utilised with little modification. These points are illustrated by converting a classical Bayesian network to the proposed scheme.

5.1 Representation

In order to prove that it is possible to specify a linguistic joint probability distribution in the form of a Bayesian network, it is sufficient to show that multiplying the conditional probability table at a node by its priors yields a joint distribution. The following Lemma and Theorem present this result for discrete linguistic random variables. The Lemma is intuitively very obvious, but rather fiddly.

Lemma 5.1.1 (Partitioning). *Given a Cartesian product of finite domains,*

$$D = \prod_{i=0}^{N} D_i \tag{5.1}$$

and functions,

$$f : D_f = \textstyle\prod_{i=1}^{N} D_i \to [0_\chi, 1_\chi] \tag{5.2}$$

and

$$g : D_g = D_0 \times \textstyle\prod_{i=1}^{M} D_{\sigma(i)} \to [0_\chi, 1_\chi] \tag{5.3}$$

where $M \geq 1$ and σ is an injection,

$$\sigma : \{1, 2, \ldots, M\} \to \{1, 2, \ldots, N\} \tag{5.4}$$

such that

a) $f(\overline{x}) \subseteq 1_\chi \ominus \mathop{\sum\!\!\!\!\!\circ}_{\overline{y} \in D_f \,:\, \overline{y} \neq \overline{x}} f(\overline{y})$ *for all* $\overline{x} \in D_f$

b) $g(\overline{x}) \subseteq 1_\chi \ominus \mathop{\sum\!\!\!\!\!\circ}_{\overline{y} \in D_g \,:\, y_0 \neq x_0} g(\overline{y})$ *for all* $\overline{x} \in D_g$

then for all $\overline{x} = (x_0, x_1, \ldots x_N) \in D$

$$\begin{aligned} & f(x_1, x_2, \ldots, x_N) g(x_0, x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(M)}) \\ & \quad \subseteq 1_\chi \ominus \Big(\mathop{\sum\!\!\!\!\!\circ}_{\overline{y} \in D \,:\, \overline{y} \neq \overline{x}} f(y_1, y_2, \ldots, y_N) g(y_0, y_{\sigma(1)}, y_{\sigma(2)}, \ldots, y_{\sigma(M)}) \Big) \end{aligned} \tag{5.5}$$

Proof.

$$
\begin{aligned}
& f(x_1,x_2,\dots,x_N)g(x_0,x_{\sigma(1)},x_{\sigma(2)},\dots,x_{\sigma(M)}) \\
& \subseteq f(x_1,x_2,\dots,x_N)\big(1_\chi \ominus \sum^{\circ}_{y_0\neq x_0} g(y_0,x_{\sigma(1)},x_{\sigma(2)},\dots,x_{\sigma(M)})\big) && (5.6) \\
& \subseteq f(x_1,x_2,\dots,x_N) \ominus f(x_1,x_2,\dots,x_N) \sum^{\circ}_{y_0\neq x_0} g(y_0,x_{\sigma(1)},x_{\sigma(2)},\dots,x_{\sigma(M)})\big) && (5.7) \\
& \subseteq 1_\chi \ominus \sum^{\circ}_{y_i\neq x_i} f(y_1,y_2,\dots,y_N) \ominus f(x_1,x_2,\dots,x_N) \sum^{\circ}_{y_0\neq x_0} g(y_0,x_{\sigma(1)},x_{\sigma(2)},\dots,x_{\sigma(M)})\big) && (5.8) \\
& \subseteq 1_\chi \ominus \sum^{\circ}_{y_i\neq x_i} f(y_1,y_2,\dots,y_N) \sum^{\circ}_{y_0} g(y_0,y_{\sigma(1)},y_{\sigma(2)},\dots,y_{\sigma(M)}) \\
& \qquad\qquad \ominus f(x_1,x_2,\dots,x_N) \sum^{\circ}_{y_0\neq x_0} g(y_0,x_{\sigma(1)},x_{\sigma(2)},\dots,x_{\sigma(M)})\big) && (5.9) \\
& \subseteq 1_\chi \ominus \big(\sum^{\circ}_{\overline{y}\in D\,:\,\overline{y}\neq\overline{x}} f(y_1,y_2,\dots,y_N)g(y_0,y_{\sigma(1)},y_{\sigma(2)},\dots,y_{\sigma(M)})\big) && \square
\end{aligned}
$$

This Lemma shows that the mass functions of a set of conditional and prior discrete linguistic variables (under certain reasonable conditions) can be combined to form a joint distribution. The details of this are spelt out in the following Theorem.

Theorem 5.1.2 (Representation Theorem for linguistic Bayesian networks). *Given a graph, $\mathfrak{G} = (V, Pa)$ and functions, $f_1, f_2, \dots, f_n$ and g with domains $D_1, D_2, \dots, D_n$ and $D^* = D \times \prod_{i=1}^{n} D_i$ respectively, and the common range $[0_\chi, 1_\chi]$, such that for all $i \in \{1,2,\dots,n\}, \overline{x} = (x, x_1, x_2, \dots, x_n) \in D^*$,*

a) $f_i(x_i) \subseteq 1_\chi \ominus \sum^{\circ}_{x_i'\neq x_i} f_i(x_i')$

b) $g(\overline{x}) \subseteq 1_\chi \ominus \sum^{\circ}_{\overline{x}'\neq\overline{x}} g(\overline{x}')$

then there exist random variables $X, X_1, X_2, \dots, X_n$ with respective domains $D, D_1, D_2, \dots, D_n$

such that for all $i \in 1, 2, \ldots, n$

$$\mathrm{lp}_{X_i}(x) = f_i(x) \textit{ for all } x \in D_i \tag{5.10}$$

and for all $\overline{x} \in D^*$

$$\mathrm{lp}_{X, X_1, X_2, \ldots, X_n}(\overline{x}) = g(\overline{x}) \prod_{i=1}^{n} f_i(x_i) \tag{5.11}$$

Proof. The proof follows by induction on the size of the graph, i. Where $i = 0$ the result follows immediately from the preceding Lemma and 4.3.4. Now suppose the result holds for $n = N$, given $f_1, f_2 \ldots f_{N+1}$, the product of the first N f_is satisfies the preconditions for the preceding Lemma, thus so do all $N + 1$. And by a further application of the Lemma,

$$g(\overline{x}) \prod_{i=1}^{N+1} f_i(x_i) \subseteq 1_\chi\big(\ominus \sum_{\overline{y} \neq \overline{x}}^{\circ} g(\overline{y}) \prod_{i=1}^{N+1} f_i(y_i)\big) \tag{5.12}$$

By 4.3.4 the desired result is obtained. □

In ordinary language, this theorem demonstrates that the product of appropriate node functions is of the correct form to represent a linguistic joint probability mass function.

5.2 Completeness

Theorem 2.4.7 demonstrated that every joint probability mass function can be represented as a Bayesian network. This result was an almost trivial consequence of Bayes' Theorem. Given a joint probability mass function for N variables $(X_1, X_2, \ldots, X_N)$ it considered the sort of fully-connected graph displayed in Figure 2.3. Each node was then identified with a random variable and associated with the corresponding conditional probability mass function. By construction, this graphical model exactly represents the original joint probability mass function.

The considerations marshaled in Section 4.2, however show that the constructive proof strategy cannot be used to prove the parallel result for linguistic Bayesian networks. As was demonstrated, even trivial bivariate linguistic joint distributions are not necessarily factorable. In such a case, it is not possible to decompose the joint distribution into a prior and a conditional. Hence no such Bayesian network representation exists.

5.3 Inference in linguistic Bayesian networks

As in the classical case both exact and inexact inference procedures are available. Section 2.4.3 showed that efficient exact inference algorithms for classical Bayesian networks rely on re-arranging and eliminating terms from the sum of products expression for the joint mass function. These optimizations rely on properties of classical Bayesian networks. First, redundancy elimination methods such as Bayes' Ball utilise the fact that screened sections of the network effectively sum to 1. Once redundant terms have been eliminated, the query sum may be most efficiently calculated by rearranging the factors allowing summations to be "pushed in" as far as possible. These rearrangements require both commutativity and distributivity.

Since linguistic probability mass functions share to a great extent these properties, the classical algorithms apply without modification, although there are some differences in interpretation which require further elaboration.

First, since by definition $\mathrm{LP}_A(A) = 1_\chi$, redundancy elimination methods work in exactly the same way. Indeed where redundancy elimination offers only improved efficiency in classical networks, in linguistic Bayesian networks, it also improves the "quality" of the resulting answer. The naive approach to computing a query would involve processing terms of the form

$$\sum^{\circ}_{x \in D_X} \mathrm{lp}_X(x) \tag{5.13}$$

by summing the values of lp_X. Where these are fuzzy the computed sum will also be some fuzzy number subsuming 1_χ. However it is evident that the value must be exactly 1χ. Thus redundancy elimination in linguistic Bayesian networks doubles as a way of removing unnecessary uncertainty.

The rearrangement of terms, via the junction tree algorithm, also carries over to linguistic Bayesian networks, albeit with one small caveat. As stated in Chapter 2, the fuzzy reals whilst commutative are not distributive, but subdistributive. Thus whilst it is legitimate to permute the multiplicands, "pushing in" the summations introduces additional uncertainty. In practice this means that as in manual calculations the output of inference algorithms will not be exactly the quantity required, but rather a value subsuming it.

Since the basic arithmetic operations for fuzzy numbers have a greater algorithmic complexity than their classical equivalents, inference techniques for linguistic Bayesian networks are (weakly) less efficient than their classical counterparts. However, following the observations in Section 2.2.8, depending on the representation used this will be a linear relationship.

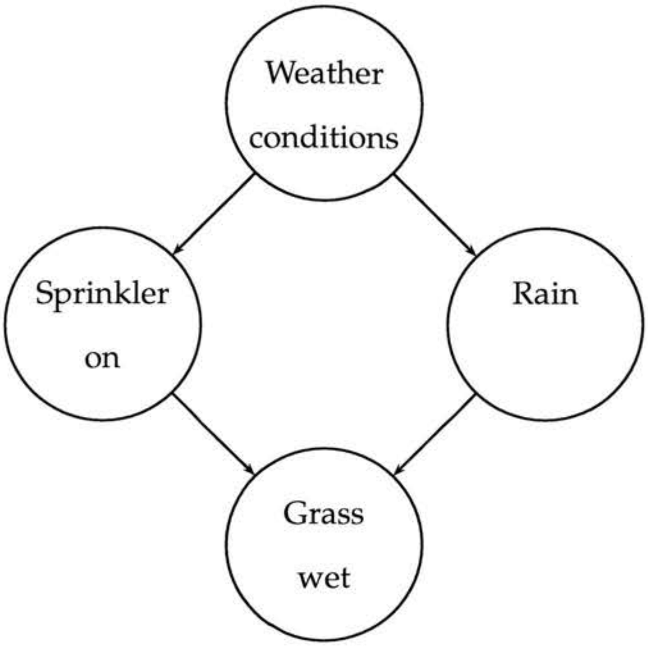


Figure 5.1: The structure of the Water Sprinkler example Bayesian Network

5.4 A simple example

Russel and Norvig (1995) describe a Bayesian Network that has become a standard didactic example. Although very simple, this network can be used to illustrate the key features of the technique such as conditional independence and "explaining away". The network and associated probabilities describe an analysis of the situation found, say, on a golf course in which three causal factors (the general weather conditions, the weather and a sprinkler) collectively determine whether the grass is wet. The network is formally specified by Figure 5.1 and Tables 5.1 to 5.4.

One classic application of Bayesian Networks is *diagnosis*, that is the inference of causes from effects. The sprinkler asserts that there are two (and only two) possible immediate causes for the grass being wet – either it's raining or the

c	$\mathrm{p}(c)$
cloudy	0.5
sunny	0.5

Table 5.1: The prior probability of cloudiness

c	$\mathrm{p}(s = on \mid c)$	$\mathrm{p}(s = off \mid c)$
cloudy	0.1	0.9
sunny	0.5	0.5

Table 5.2: The conditional probability of the sprinkler being on

c	$\mathrm{p}(r = raining \mid c)$	$\mathrm{p}(r = clear \mid c)$
cloudy	0.8	0.2
sunny	0.2	0.8

Table 5.3: The conditional probability of rain

s	r	$\mathrm{p}(g = wet \mid s, r)$	$\mathrm{p}(g = dry \mid s, r)$
off	clear	0.0	1.0
	raining	0.9	0.1
on	clear	0.9	0.1
	raining	0.99	0.01

Table 5.4: The conditional probability of the grass being wet

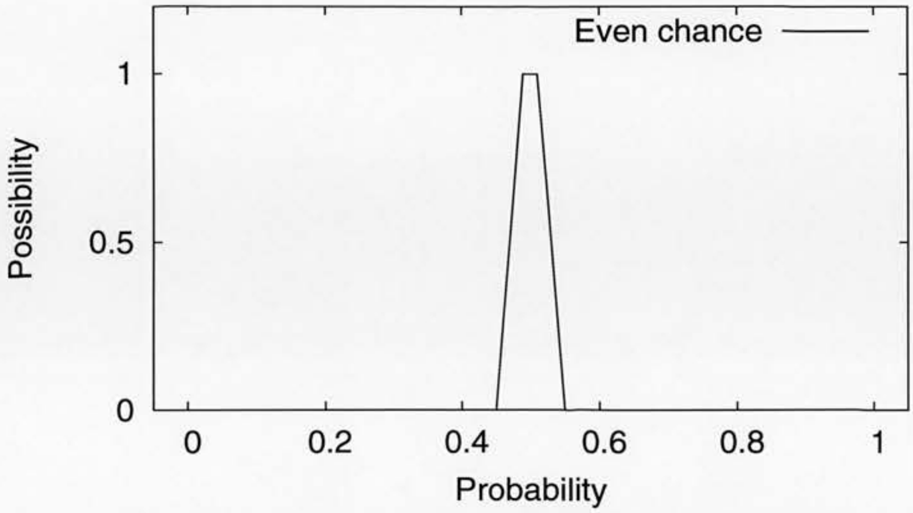


Figure 5.2: The linguistic probability "Even chance"

sprinkler is on. So, given that the grass is wet, which of these causes is most likely may be determined by computing the posterior probabilities as follows,

$$\mathrm{p}(s = on \mid g = wet) = \frac{\mathrm{p}(s = on, g = wet)}{\mathrm{p}(g = wet)} = \frac{0.2781}{0.6471} = 0.430 \tag{5.14}$$

$$\mathrm{p}(r = raining \mid g = wet) = \frac{\mathrm{p}(r = raining, g = wet)}{\mathrm{p}(g = wet)} = \frac{0.4581}{0.6471} = 0.708 \tag{5.15}$$

Thus if the grass is wet, rain is the most likely explanation.

Although this risks targetting a straw man, it seems clear that the probabilities utilised in even this very simple example would be difficult to obtain experimentally. Instead they are the product of (perhaps informed) guesswork. It is interesting therefore to examine how sensitive the conclusion drawn is to these premises.

Using the probabilities shown in Figures 5.2 to 5.12, Tables 5.5, 5.7, 5.6, 5.8 provide a complete specification of a linguistic Bayesian network. Since these

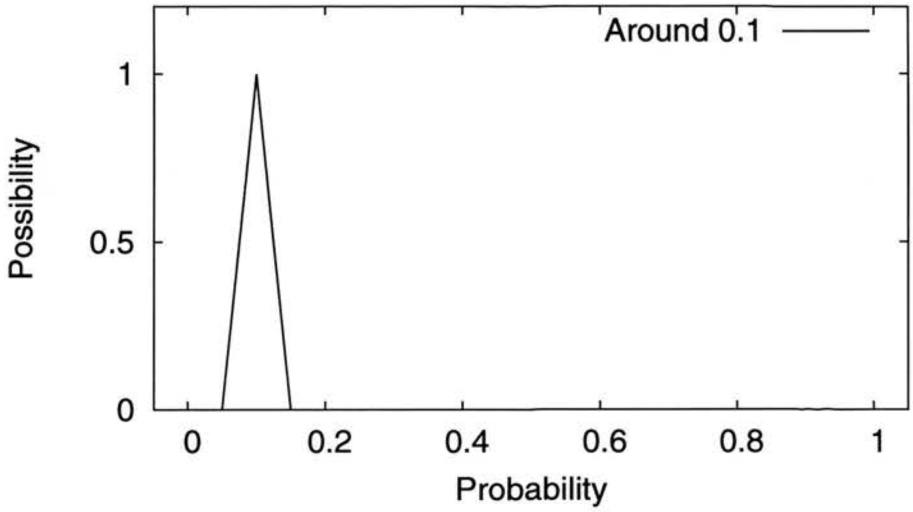


Figure 5.3: The linguistic probability "Around 0.1"

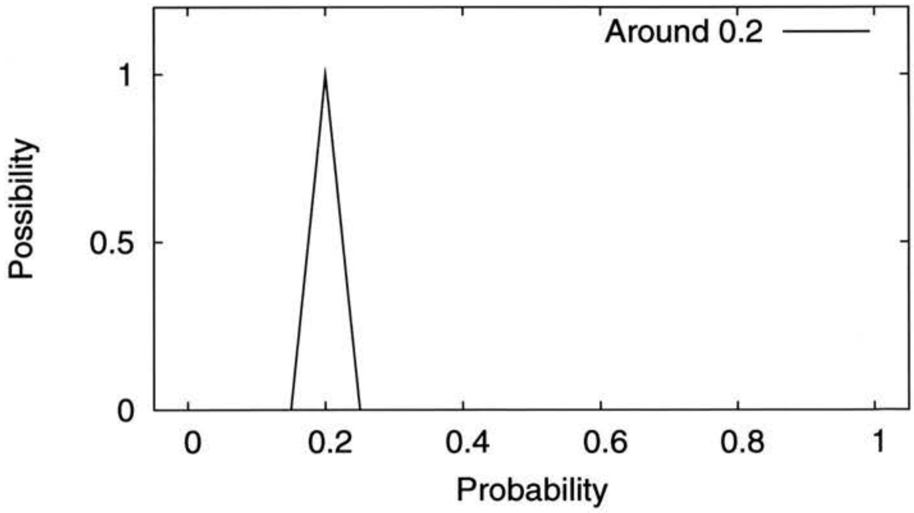


Figure 5.4: The linguistic probability "Around 0.2"

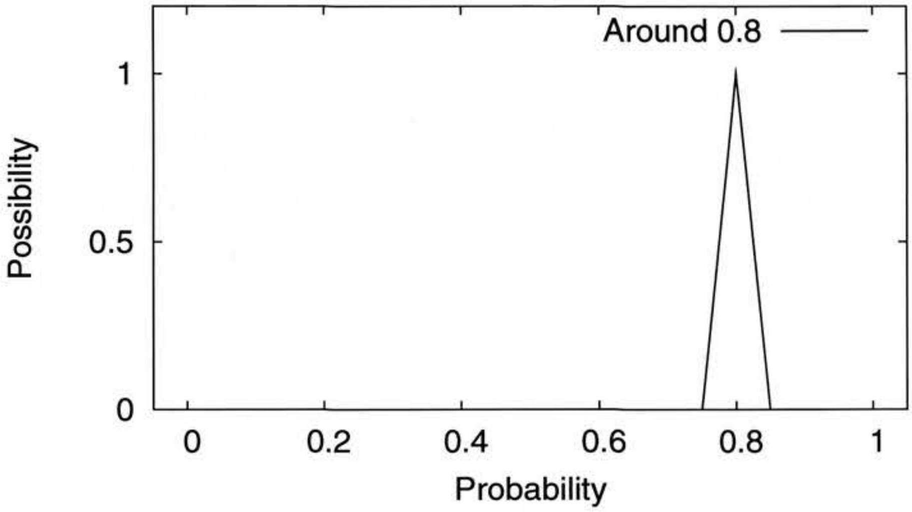


Figure 5.5: The linguistic probability "Around 0.8"

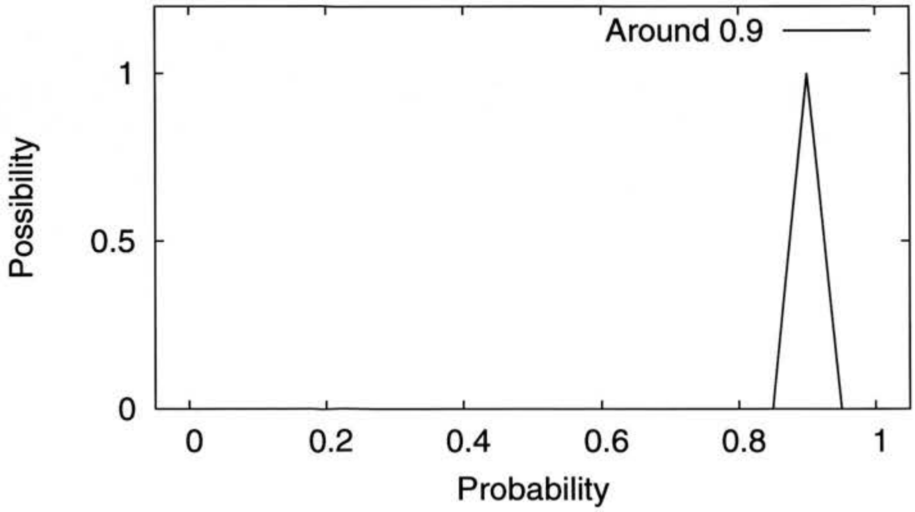


Figure 5.6: The linguistic probability "Around 0.9"

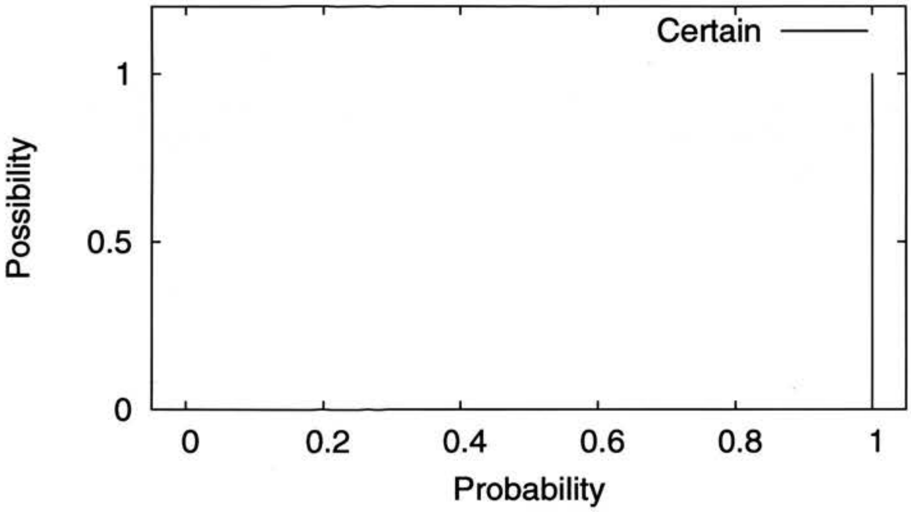


Figure 5.7: The linguistic probability "Certain"

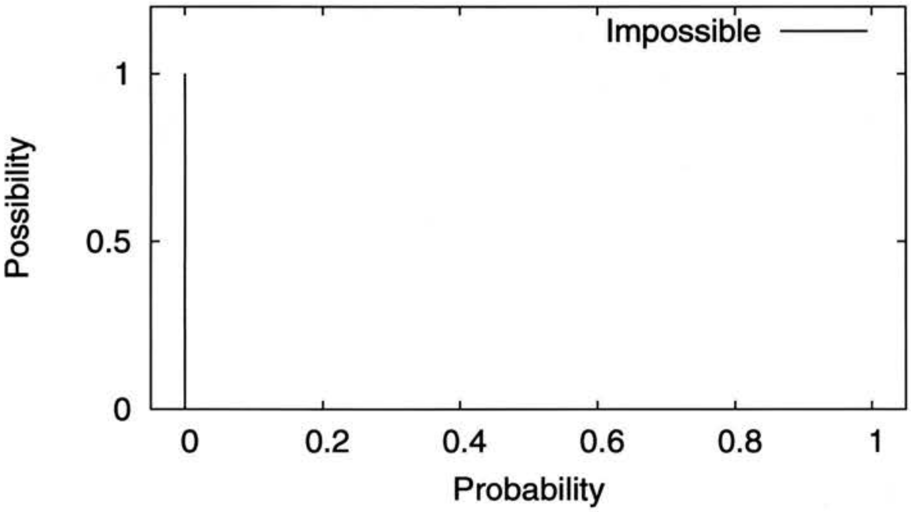


Figure 5.8: The linguistic probability "Impossible"

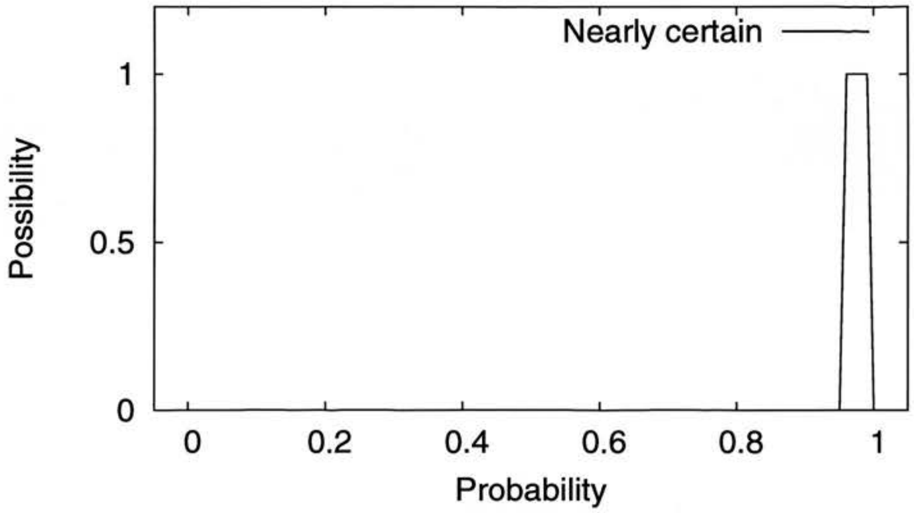


Figure 5.9: The linguistic probability "Nearly certain"

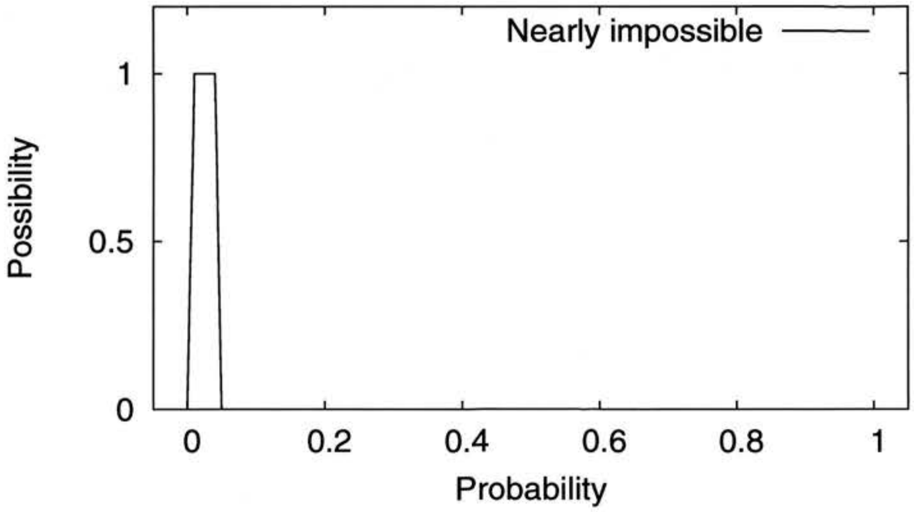


Figure 5.10: The linguistic probability "Nearly impossible"

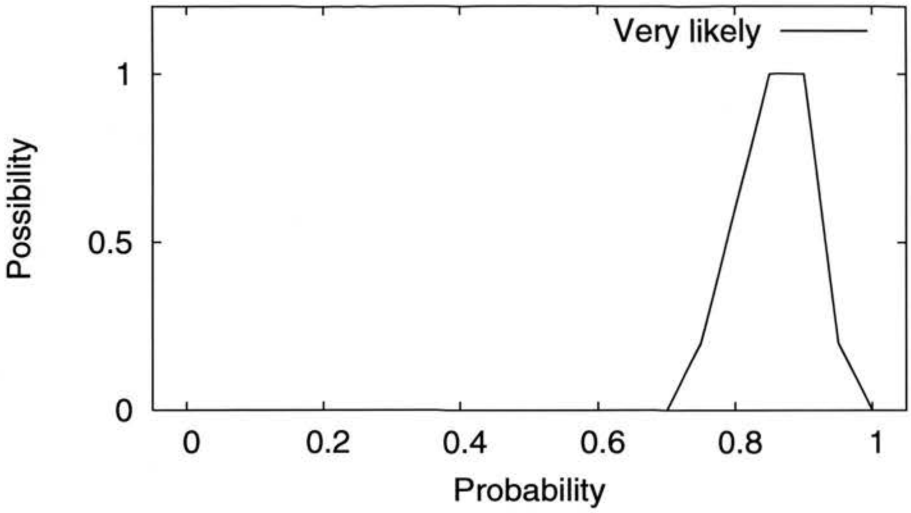


Figure 5.11: The linguistic probability "Very likely"

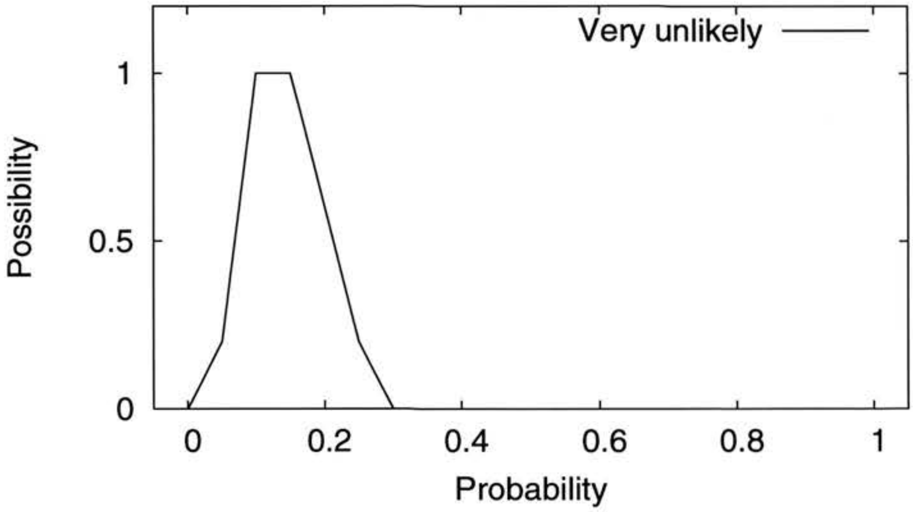


Figure 5.12: The linguistic probability "Very unlikely"

c	$\mathrm{p}(c)$
cloudy	even chance
sunny	even chance

Table 5.5: The prior probability of cloudiness

c	$\mathrm{p}(s = on \mid c)$	$\mathrm{p}(s = off \mid c)$
cloudy	around 0.1	around 0.9
sunny	even chance	even chance

Table 5.6: The conditional probability of the sprinkler being on

c	$\mathrm{p}(r = raining \mid c)$	$\mathrm{p}(r = clear \mid c)$
cloudy	around 0.8	around 0.2
sunny	around 0.2	around 0.8

Table 5.7: The conditional probability of rain

s	r	$\mathrm{p}(g = wet \mid s, r)$	$\mathrm{p}(g = dry \mid s, r)$
off	clear	impossible	certain
	raining	very likely	very unlikely
on	clear	very likely	very unlikely
	raining	nearly certain	nearly impossible

Table 5.8: The conditional probability of the grass being wet

probabilities subsume the (crisp embeddings of) the classical probabilities in the original example, the output of that example may be used as a sanity check on any computations performed using this extended model.

5.5 ARBOR: a linguistic Bayesian network tool

The various fuzzy numbers and probability tables needed to fully specify a linguistic Bayesian network are laborious to represent on paper and even more cumbersome to work with. To simplify this process and reduce the sources of error, a graphical tool, Arbor, was created.

Arbor is an extensible tool for creating and editing graph structures and performing operations upon them that was written to provide a comfortable environment for experimenting with linguistic Bayesian networks.

Graph types (their topological constraints, node annotations and associated actions) are defined using a plugin architecture. It utilises a plugin architecture for defining graph types. The use of the high-level, object-oriented programming language Python allows functionality to be abstracted at various levels of granularity across the system and in a way that reflects the structure of the underlying mathematics.

So, for example, directed acyclic graphs are simply a sub-class of a directed graph type, adding an acyclicness constraint check. Similarly, the Linguistic Bayesian network graph class directly inherits the (suitably abstracted) inference algorithm from a Bayesian network superclass – with comments the linguistic Bayesian network graph type amounts to just 60 lines of code. Re-using functionality in this way helps to ensure robustness of the system as well as significantly reducing implementation time.

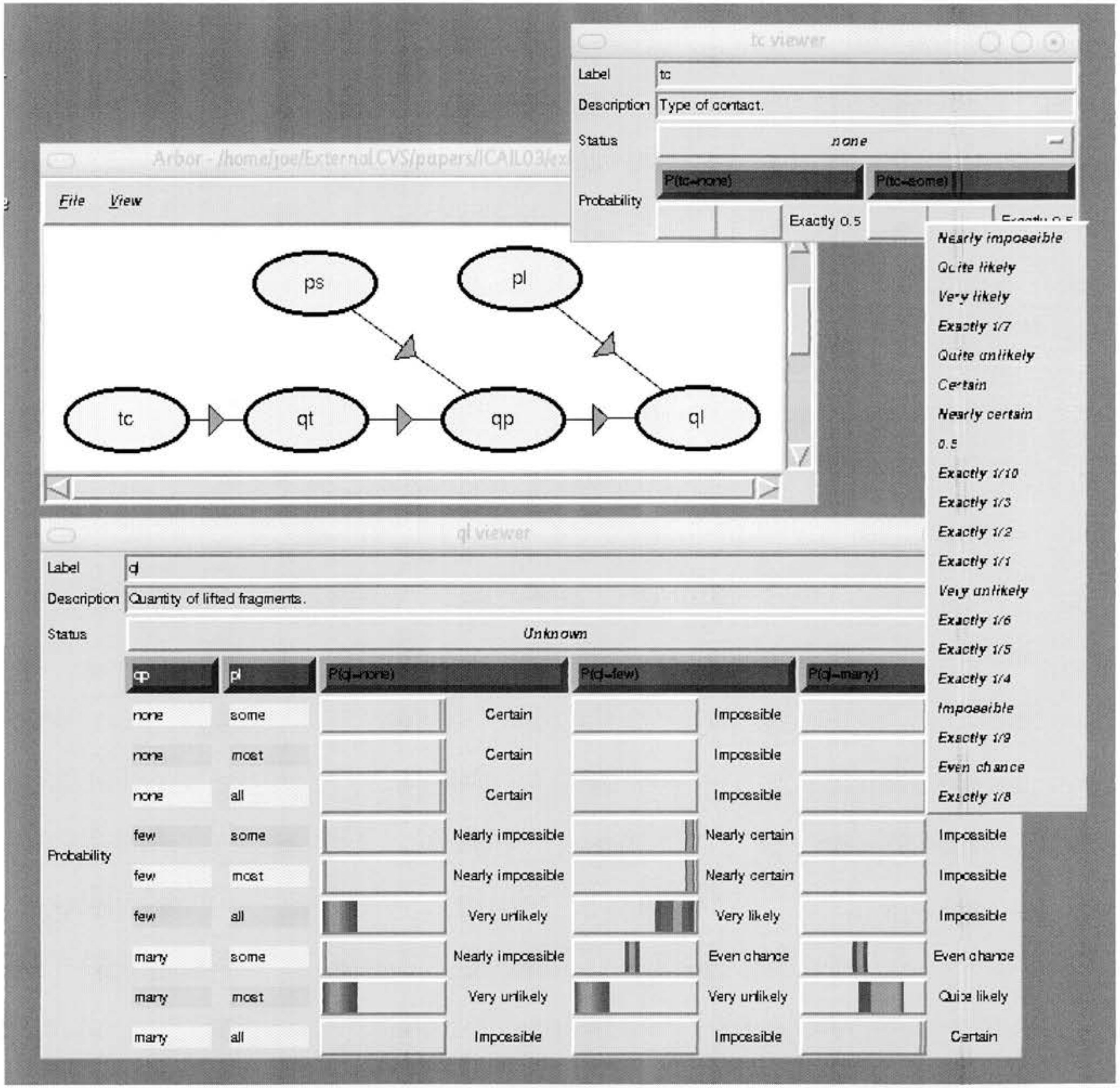
The system also makes use of other advanced language features offered by Python. Functional programming styles are exensively employed in the inference algorithms to take advantage of lazy evaluation/implicit state and closures. The native serialization provided by the pickle module is also used to provide save and restore functionality across hetrogenous graph types with no additional code.

Finally the system enforces a clean separation of core logic and presentation through the MVC (model-view-controller) paradigm. This is a useful engineering strategy as it simplifies interactions between components, but also directly enables batch-mode operation. This enhances the reusability of core algorithms and data-structures – so, for example, it would be relatively trivial to embed a linguistic Bayesian network model in a "headless" server application.

5.6 Example computations

With the data fed into Arbor, it is a point and click operation to set the "grass" variable to "wet" and determine the posterior distributions for the sprinkler and weather. These are displayed in Figures 5.14 and 5.15 respectively.

As in the classical case (and as expected) if the grass is wet then rain is the most likely explanation.



Label tc

Description Type of contact.

Status none

Probability	P(tc=none)	P(tc=some)
	Exactly 0.5	Exactly 0.5

Label ql

Description Quantity of lifted fragments.

Status Unknown

Probability

qp	pl	P(ql=none)	P(ql=few)	P(ql=many)
none	some	Certain	Impossible	
none	most	Certain	Impossible	
none	all	Certain	Impossible	
few	some	Nearly impossible	Nearly certain	Impossible
few	most	Nearly impossible	Nearly certain	Impossible
few	all	Very unlikely	Very likely	Impossible
many	some	Nearly impossible	Even chance	Even chance
many	most	Very unlikely	Very unlikely	Quite likely
many	all	Impossible	Impossible	Certain

Figure 5.13: A screenshot of the ARBORextensible Bayesian network editor showing the network used in the forensics case study.

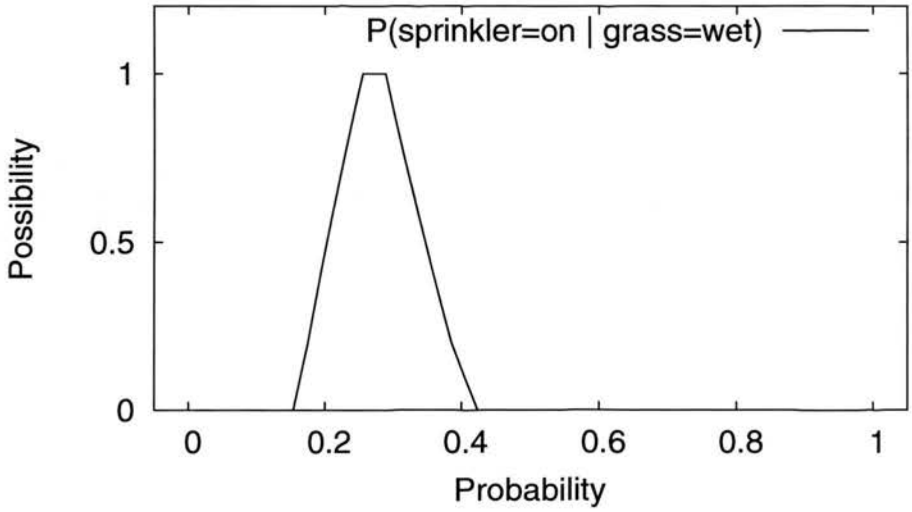


Figure 5.14: The computed linguistic probability of the sprinkler being on given that the grass is wet.

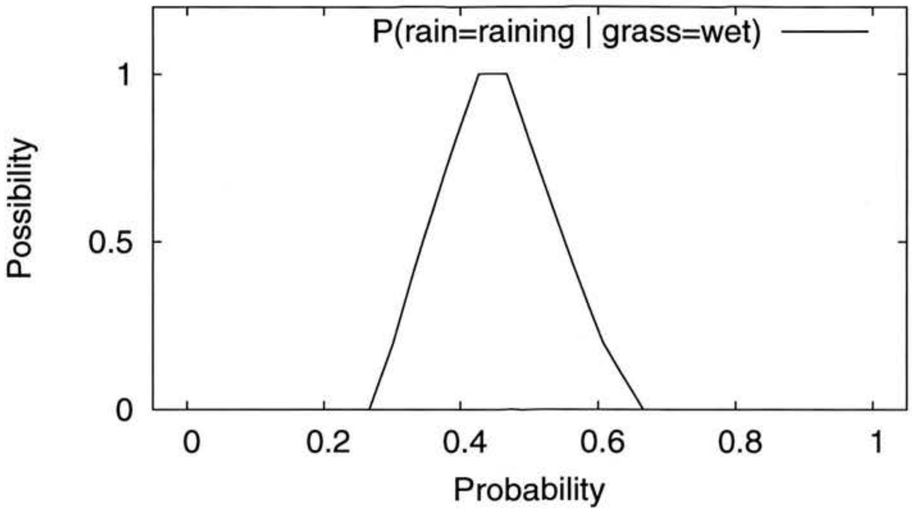


Figure 5.15: The computed linguistic probability of its raining given that the grass is wet.

5.7 Summary

This chapter applied the theory of linguistic probabilities developed in Chapter 4 to the construction of graphical probabilistic models analogous to Pearl's Bayesian networks. It was shown that a graphical model under a certain set of constraints could be used to represent linguistic joint probability distributions but that it is not clear whether all linguistic joint probability distributions are representable in this way.

It was then shown that the underlying algebraic properties of the model allow the use of standard, efficient belief propagation and relevance determination algorithms. Finally, a standard didactic example of a Bayesian network was extended to a sample linguistic Bayesian network. Sample computations performed with the aid of the Arbor software yielded results consist with, but extending the classical case.

The following chapter will utilise this theory in connection with a contemporary application of Bayesian networks in forensic science.

Chapter 6

Case study: forensic statistics

Forensic statistics is a discipline that is mainly concerned with the experimental design of forensic examinations and the analysis of the obtained results. The issues it studies include hypothesis formulation, deciding on minimal sample sizes when studying populations of similar units of evidence and determining the statistical significance of the outcome of tests. Recently, the discipline has been branching out to the study of the statistical implications of forensic examinations on defence and prosecution positions during crime investigation and criminal court proceedings.

This chapter explains one application of classical Bayesian networks to forensic statistics, namely the evaluation of the strength of support collected evidence provides for a given crime scenario. An example is provided of such a network and the calculations dictated by the protocol are performed. These results are then contrasted with those obtained by using a similar, but more expressive model utilizing linguistic probabilities.

The material presented here was developed in conjunction with Jeroen Keppens (then a researcher at the Joseph Bell Centre for Legal reasoning) and presented to a an audience of specialists (Halliwell et al., 2003) in 2003.

6.1 Bayesian networks in forensic statistics

The most high-profile forensic application of Bayesian reasoning was in the 1996 case of Crown versus Denis John Adams. Here a statistical expert witness explained Bayes' theorem to a jury and invited them to use it as a basis for assessing a moderately complex set of evidence. Although the jury delivered a guilty verdict, the case was appealed on the grounds that no alternative method was provided for jurors who chose not to accept the mathematical approach. Ultimately, the Court of Appeal upheld the conviction and gave their opinion that "To introduce Bayes' Theorem, or any similar method, into a criminal trial plunges the Jury into inappropriate and unnecessary realms of theory and complexity, deflecting them from their proper task." Further appeal was precluded, but the form of the ruling has ensured that the use of explicit Bayesian assessment in the courtroom remains controversial.

Nevertheless, Bayesian reasoning has firmly established itself within the crown prosecution service. In Cook et al. (1998b), a method is proposed to assess the impact of a certain piece of forensic evidence on a given case. This method is the result of a significant research effort by the Forensic Science Service (FSS), the largest provider of forensic science services in England and Wales. It involves 1) formalising the respective claims of the prosecution and the defence Cook et al. (1998a); Evett et al. (2000), 2) computing the probability that the evidence is found given that the claim of the prosecution is true and the probability that the evidence is found given that the claim of the defence is true, and 3) dividing the former probability by the latter to determine the likelihood ratio Balding and Donnelly (1995):

$$LR = \frac{\mathrm{P}(E \mid C_p)}{\mathrm{P}(E \mid C_d)}$$

where E, C_p, C_d respectively represent the evidence, the prosecution claim and the defence claim, and $\mathrm{P}(E \mid C)$ is the probability that evidence E is found if claim C is true.

The likelihood ratio is a numerical evaluation of the extent to which the evidence supports the prosecution claim over the defence claim. It has two important applications. Firstly, the potential benefit associated with performing forensic procedures (which are often expensive and resource intensive) may be assessed in advance by examining the effect of their possible outcomes on the likelihood ratio. Increasingly, police forces must purchase forensic services. Likelihood ratio based calculations can support this difficult decision making process. Secondly, the likelihood ratio can be used to justify the testimonies of forensic experts during the court proceedings. To this end, a verbal scale to help forensic experts interpret the *LR* is suggested by the FSS Evett et al. (2000). This is reproduced in Table 6.1 for reference.

6.1.1 Bayesian networks and likelihood ratio

The likelihood ratio method is, of course, crucially dependent upon a means to compute the probabilities $P(E \mid C_p)$ and $P(E \mid C_d)$. Bayesian networks have

LR	Support of evidence to prosecution claim over defence claim
1 to 10	limited
10 to 100	moderate
100 to 1,000	moderately strong
1,000 to 10,000	strong
$>$ 10,000	very strong

Table 6.1: Interpretation of the likelihood ratio.

	Event	Domain
q_t	quantity of transferred fragments	{none,few,many}
q_p	quantity of persisted fragments	{none,few,many}
q_l	quantity of lifted fragments	{none,few,many}
t_c	type of contact	{none,some}
p_s	proportion of fragments shed	{none,small,large}
p_l	proportion of fragments lifted	{some,most,all}

Table 6.2: Variables in the one-way transfer case.

emerged as a helpful technique in this context Aitken et al. (2003); Cook et al. (1999); Dawid et al. (2002).

A Bayesian network is a directed graph whose nodes represent events. Causal relationships between these events are, in turn, represented by arcs. Since the network is patterned after these real world relations, the decision to use a Bayesian network can guide the knowledge acquisition process.

An example may best illustrate this application of Bayesian networks. Con-

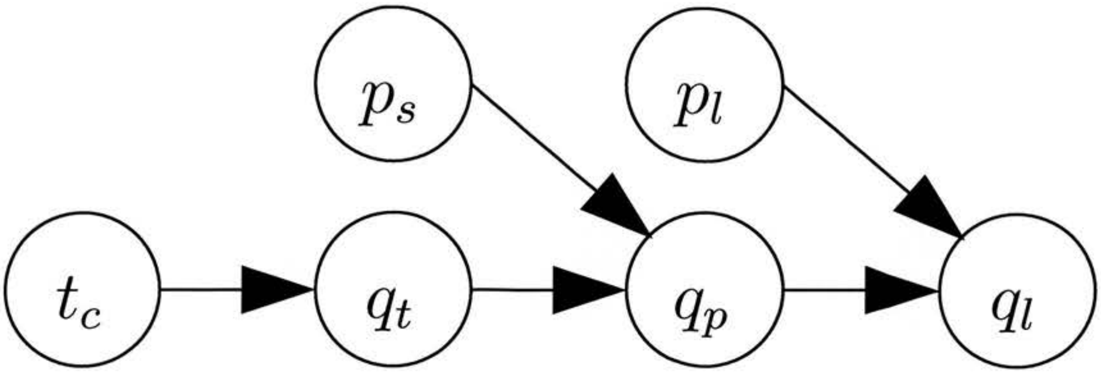


Figure 6.1: Bayesian network of a one-way transfer case.

sider the following scenario:

A burglar smashes the window of a shop, steals some money from the cash registry and flees the scene of the crime. A bystander witnessed this event and reports a description of the perpetrator to the police who arrest a man, matching the description of the witness half an hour after the event. The suspect, Mr. Blue, denies having been near the shop. However, q_l glass fragments, matching the type of glass of the shop's window, are retrieved from Mr. Blue's clothes.

Figure 6.1 shows a Bayesian network that models the probabilistic relationship between the retrieval of q_l glass fragments from the garment of Mr. Blue in the forensic laboratory and the type of contact, t_c, between Mr. Blue and the shop's window. The number of glass fragments q_l that are retrieved from Mr. Blue's clothes depends on the number of glass fragments that have persisted in the clothes q_p and on the effectiveness of the retrieval technique p_l, where p_l represents the proportion of glass fragments lifted from the garments under examination. The number of glass fragments q_p that have persisted in the clothes until the time of the examination, in turn, is dependent upon the number of glass fragments q_t that were transferred in the first place and the proportion of fragments p_s shed between the time of transfer and the time of the examination. Finally, the number of transferred fragments q_t depends on the type of contact t_c. The domains of these variables are reproduced in Table 6.2.

By Bayes' Theorem the joint distribution of variables involved in a Bayesian network is given by the product of the (conditional or prior) distributions at each node. Hence, $P(q_l \cap t_c)$ may be found by marginalizing over the other variables:

$$\begin{aligned} \mathrm{P}(q_l \cap t_c) = \sum_{p_l, q_p, q_t, p_s} & \mathrm{P}(q_l \mid q_p, p_l)\,\mathrm{P}(p_l) \\ & \mathrm{P}(q_p \mid q_t, p_s)\,\mathrm{P}(p_s)\,\mathrm{P}(q_t \mid t_c)\,\mathrm{P}(t_c) \end{aligned}$$

which, by rearranging terms to minimise the number of arithmetic operations required, equals

$$\begin{aligned} & \mathrm{P}(t_c) \sum_{p_l} \mathrm{P}(p_l) \sum_{q_p} \mathrm{P}(q_l \mid q_p, p_l) \\ & \sum_{q_t} \mathrm{P}(q_t \mid t_c) \sum_{p_s} \mathrm{P}(q_p \mid q_t, p_s)\,\mathrm{P}(p_s) \end{aligned}$$

Suppose that the prosecution case is that the defendant has had *some* contact with the window in question and that a given forensic procedure has yielded *many* matching fragments. The probabilities required to evaluate the likelihood ratio are provided in Tables 6.3, 6.4, 6.5 and 6.6. The relevant calculation is,

$$\begin{aligned} LR &= \frac{\mathrm{P}(q_l = many \mid t_c = some)}{\mathrm{P}(q_l = many \mid t_c = none)} \\ &= \frac{0.428586}{0.038813} = 11.042\ldots \end{aligned}$$

p_s	$\mathrm{p}(p_s)$
none	0.03
small	0.3
large	0.67

p_l	$\mathrm{p}(p_l)$
none	0.06
few	0.29
many	0.65

Table 6.3: Classical prior probabilities $\mathrm{p}(p_s)$ and $\mathrm{p}(p_l)$.

t_c	$\mathrm{p}(q_t = none \mid t_c)$	$\mathrm{p}(q_t = few \mid t_c)$	$\mathrm{p}(q_t = many \mid t_c)$
none	0.9	0.05	0.05
some	0.1	0.25	0.65

Table 6.4: Classical conditional probabilities $\mathrm{p}(q_t \mid t_c)$.

q_t	p_s	$\mathrm{p}(q_p = none \mid q_t, p_s)$	$\mathrm{p}(q_p = few \mid q_t, p_s)$	$\mathrm{p}(q_p = many \mid q_t, p_s)$
none	none	1	0	0
	small	1	0	0
	large	1	0	0
few	none	0	1	0
	small	0.1	0.9	0
	large	0.3	0.7	0
many	none	0	0	1
	small	0.05	0.1	0.85
	large	0.07	0.48	0.45

Table 6.5: Classical conditional probabilities $P(q_p \mid q_t, p_s)$.

q_p	p_l	$\mathrm{LP}(q_l = none \mid q_p, p_l)$	$\mathrm{LP})q_l = few \mid q_p, p_l)$	$\mathrm{LP}(q_l = many \mid q_p, p_l)$
none	some	1	0	0
	most	1	0	0
	all	1	0	0
few	some	0.05	0.95	0
	most	0.05	0.95	0
	all	0.02	0.6	0.38
many	some	0.08	0.46	0.46
	most	0.2	0.2	0.6
	all	0	0	1

Table 6.6: Classical conditional probabilities $\mathrm{p}(q_l \mid q_p, p_l)$.

Thus, according to Table 1, this item of forensic evidence provides *moderate* support to the prosecution case.

6.2 Linguistic Bayesian networks for forensic statistics

This section details a proof-of-concept case study created in conjunction with Jeroen Keppens (at the time a researcher at the Joseph Bell Centre for Legal reasoning). This study together with a sketch of the underlying theory was presented to an audience of specialists in the area of legal applications of artificial intelligence techniques at ICAIL03.

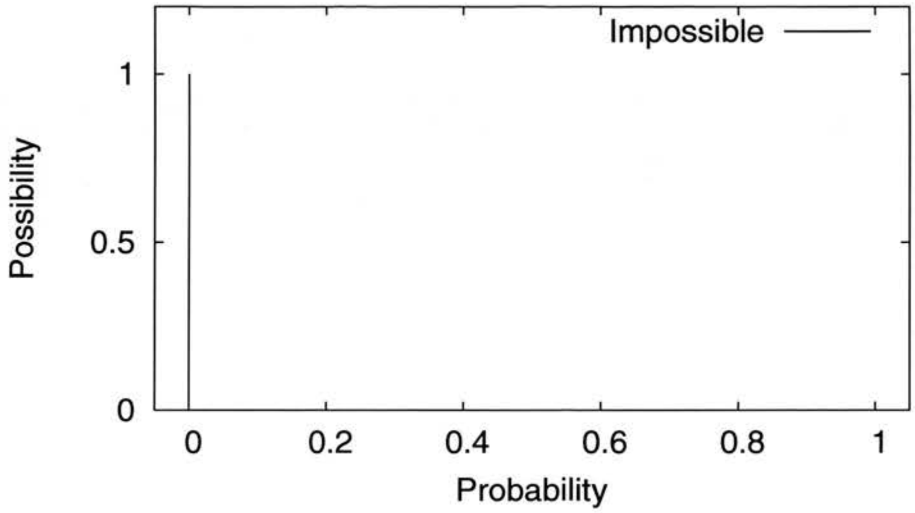


Figure 6.2: The linguistic probability labelled "Impossible"

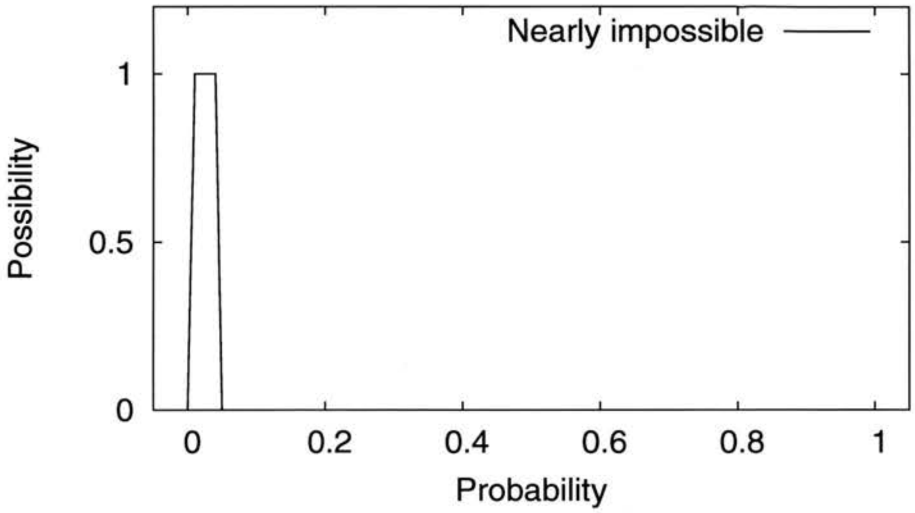


Figure 6.3: The linguistic probability labelled "Nearly impossible"

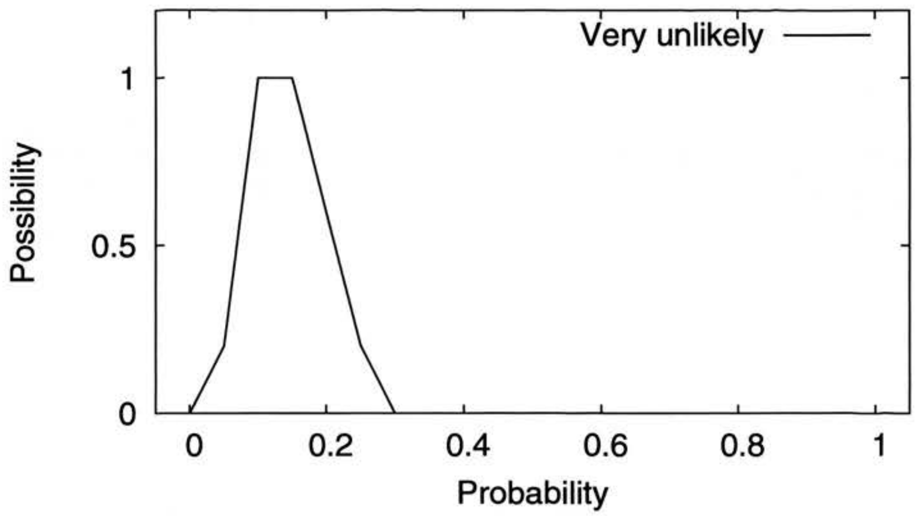


Figure 6.4: The linguistic probability labelled "Very unlikely"

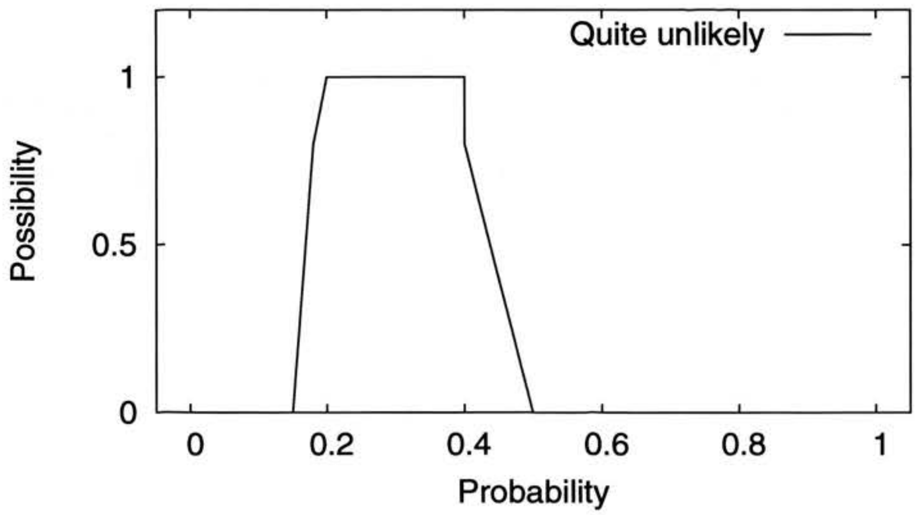


Figure 6.5: The linguistic probability labelled "Quite unlikely"

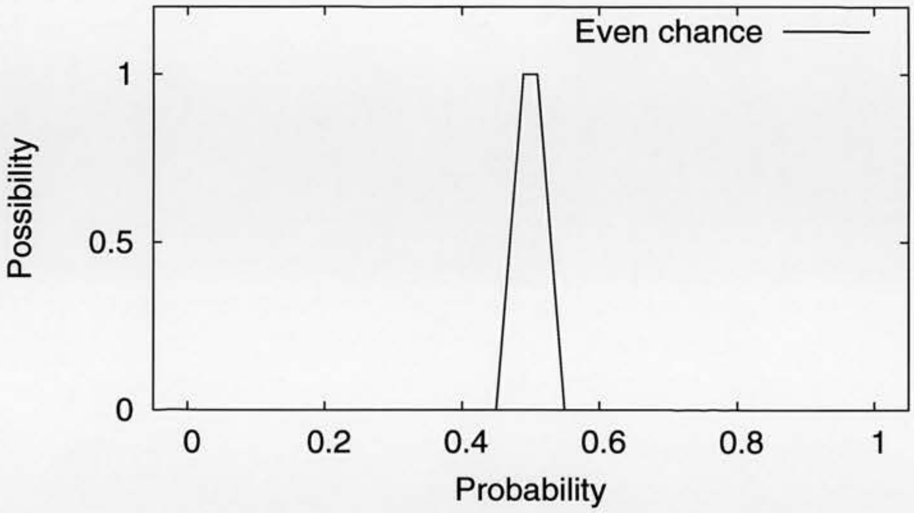


Figure 6.6: The linguistic probability labelled "Even chance"

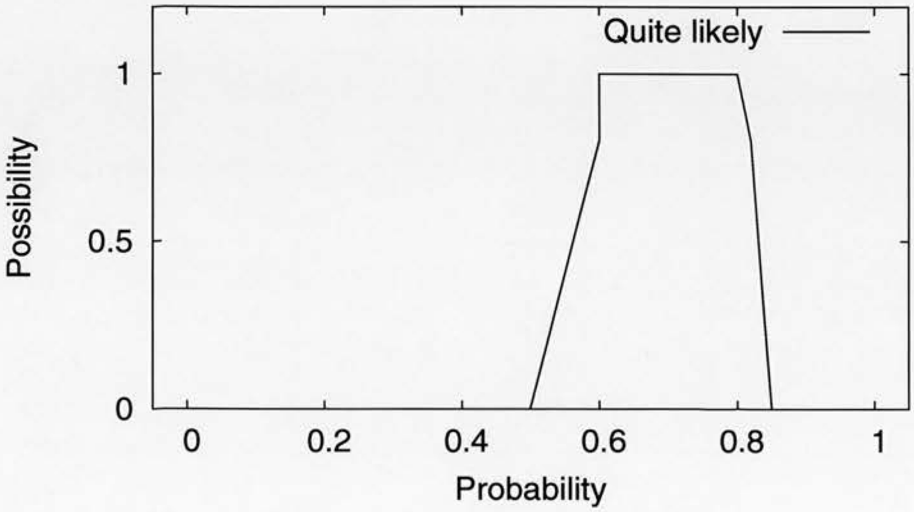


Figure 6.7: The linguistic probability labelled "Quite likely"

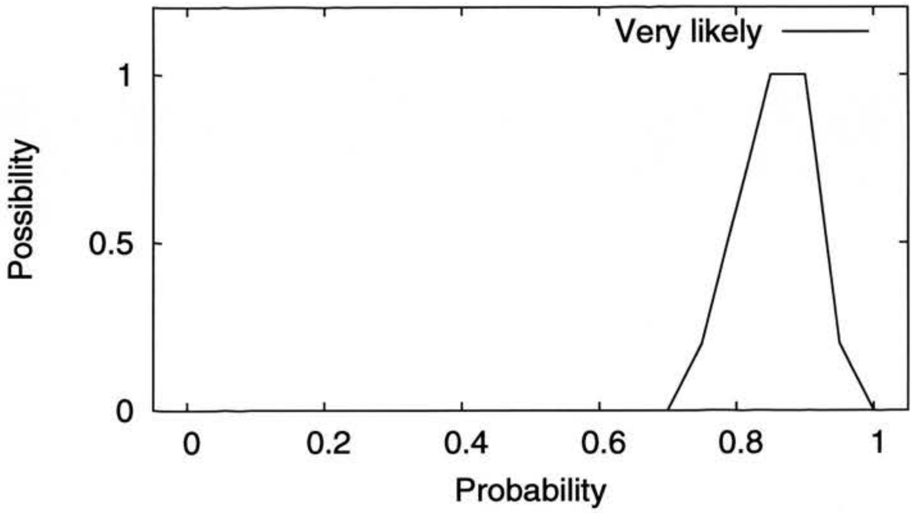


Figure 6.8: The linguistic probability labelled "Very likely"

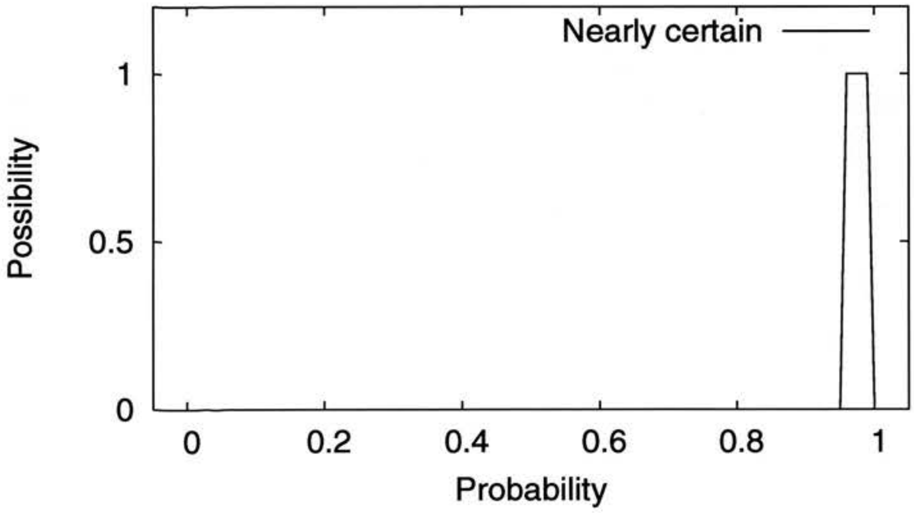


Figure 6.9: The linguistic probability labelled "Nearly certain"

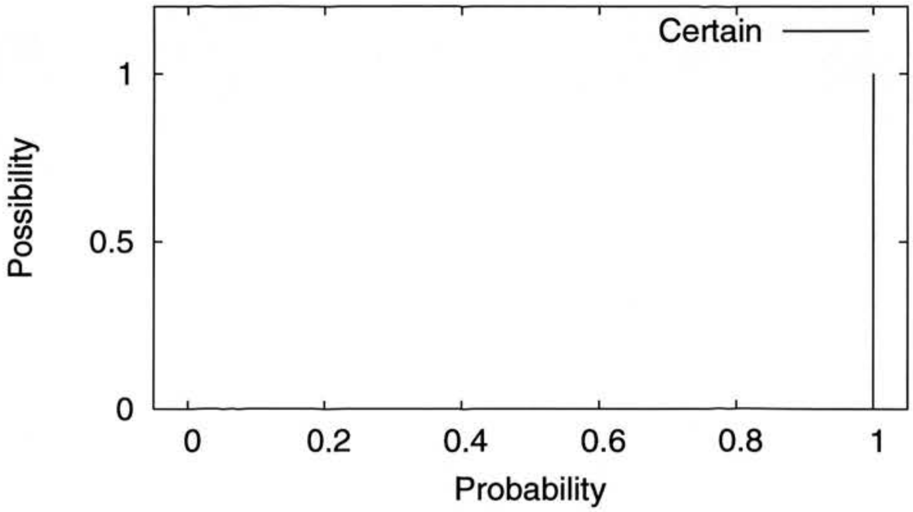


Figure 6.10: The linguistic probability labelled "Certain"

Table 6.7: Linguistic prior probabilities $\mathrm{lp}(p_s)$ and $\mathrm{lp}(p_l)$

p_s	$\mathrm{lp}(p_s)$	p_l	$\mathrm{lp}(p_l)$
none	nearly impossible	none	nearly impossible
small	quite unlikely	few	quite unlikely
large	quite likely	many	quite likely

Table 6.8: Linguistic conditional probabilities $\mathrm{lp}(q_t \mid t_c)$.

t_c	$\mathrm{lp}(q_t = none \mid t_c)$	$\mathrm{lp}(q_t = few \mid t_c)$	$\mathrm{lp}(q_t = many \mid t_c)$
none	nearly certain	nearly impossible	nearly impossible
some	impossible	quite unlikely	quite likely

Table 6.9: Linguistic conditional probabilities $\mathrm{lp}(q_p \mid q_t, p_s)$.

q_t	p_s	$\mathrm{lp}(q_p = none \mid q_t, p_s)$	$\mathrm{lp}(q_p = few \mid q_t, p_s)$	$\mathrm{lp}(q_p = many \mid q_t, p_s)$
none	none	certain	impossible	impossible
	small	certain	impossible	impossible
	large	certain	impossible	impossible
few	none	impossible	certain	impossible
	small	very unlikely	very likely	impossible
	large	quite unlikely	quite likely	impossible
many	none	impossible	impossible	certain
	small	nearly impossible	very unlikely	very likely
	large	nearly impossible	even chance	even chance

6.2.1 Extended example: Glass transfer

Tables 6.7, 6.8, 6.9 and 6.10 present a linguistically specified version of the network discussed in Section 2. The qualitative probability terms themselves are in turn graphed in Figures 6.2 to 6.10. Computations are performed in exactly the same sequence as in the classical case, but with fuzzy arithmetic operators and numbers. This yields fuzzy values for $\mathrm{lp}(q_l = many \mid t_c = some)$, $\mathrm{lp}(q_l =$

Table 6.10: Linguistic conditional probabilities $\mathrm{lp}(q_l \mid q_p, p_l)$.

q_p	p_l	$\mathrm{lp}(q_l = \mathit{none} \mid q_p, p_l)$	$\mathrm{lp}(q_l = \mathit{few} \mid q_p, p_l)$	$\mathrm{lp}(q_l = \mathit{many} \mid q_p, p_l)$
none	some	certain	impossible	impossible
	most	certain	impossible	impossible
	all	certain	impossible	impossible
few	some	nearly impossible	nearly certain	impossible
	most	nearly impossible	nearly certain	impossible
	all	nearly impossible	very likely	very unlikely
many	some	nearly impossible	even chance	even chance
	most	very unlikely	very unlikely	quite likely
	all	impossible	impossible	certain

$\mathit{many} \mid t_c = \mathit{none})$ and the likelihood ratio. These are presented in Figures 6.11 and 6.13 respectively. Note that the membership functions, as expected, subsume their classical counterparts as calculated in Section 2.

The value calculated for $\mathrm{lp}(q_l = \mathit{many} \mid t_c = \mathit{none})$ is particularly interesting. Practical forensic applications typically use conservative (high) estimates for $\mathrm{P}(E|C_d)$ (i.e. the denominator in the likelihood calculation) thereby biasing the case in favour of the defence (Cook et al., 1999). Additionally, the probabilities typically associated with the subsets of events modelling the case where evidence originates not with the crime, but with some other source, are vanishingly small. Moreover, these probabilities are typically the most difficult to obtain experimentally. The use of linguistic probabilities to represent such probabilities allows the uncertainty that prompts this conservatism to be explicitly included in the model.

The fuzzy value calculated for the likelihood ratio has an extremely broad plateau (α-cut at 1), dramatically exhibiting the sensitivity of this statistic to small perturbations in the subjective probabilities on which it is based. That the set's membership function is greater than zero in each of the Forensic Science Service's recommended interpretation classes that are reproduced in Table 1 is, of course, partly a result of the rather "low-resolution" term set used for convenience of presentation here. Nevertheless, to re-iterate the central argument of this paper, the effects of propagating uncertainties should not be brushed aside. It is clear from the graph that the support provided by the evidence is roughly speaking moderate to strong. Note that, given a fuzzification of the likelihood ratio quantity space, it would be possible to automatically generate this description. However, the newly acknowledged uncertainties in the subjective probability estimates are certainly consistent with much more limited support.

6.3 Summary

This chapter illustrated how linguistic Bayesian networks might be used in a real-world setting. Classical Bayesian networks are already utilised by the UK's Forensic Science Service to calculate the evidential support for a given criminal scenario. However, some of the prior and conditional probabilities involved in these models are obtained "subjectively" through expert consultation. This is then an ideal application for linguistic probabilities.

The increased expressivity of linguistic probabilities allows second order uncertainty to be represented intuitively as is surely appropriate where probabil-

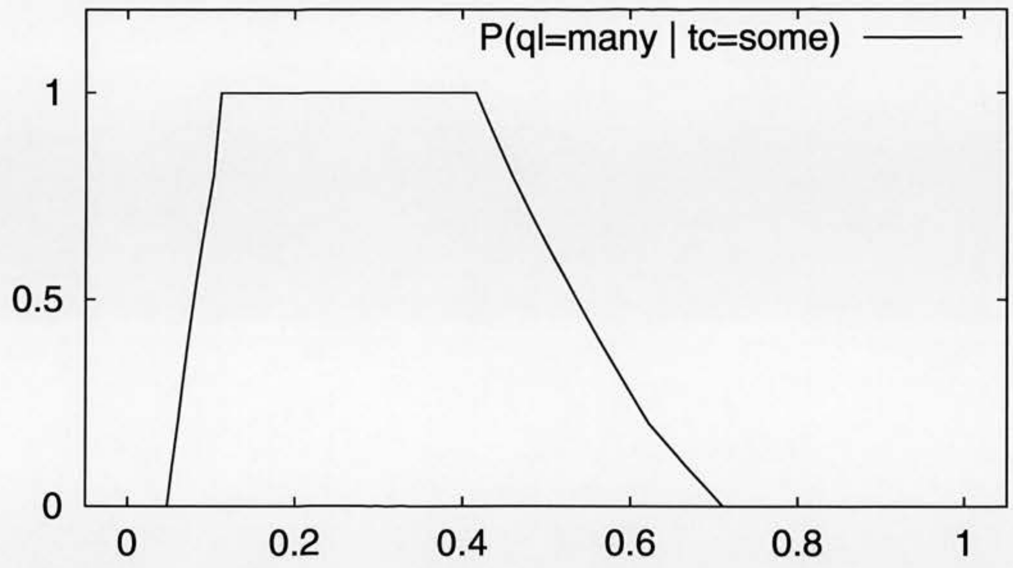


Figure 6.11: The computed linguistic probability of $\mathrm{lp}(q_l = \mathit{many} \mid t_c = \mathit{some})$

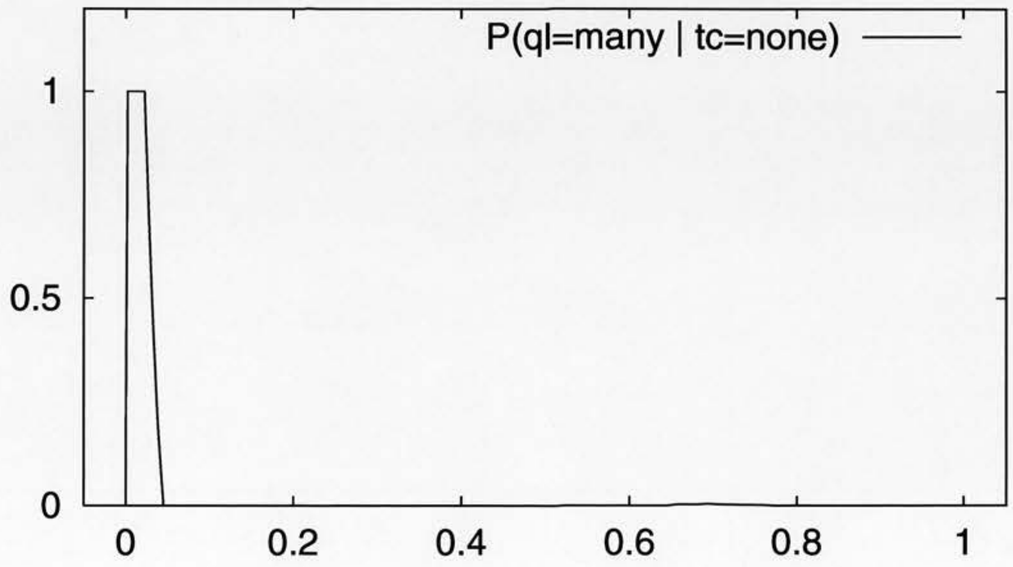


Figure 6.12: The computed lingustic probability of $\mathrm{lp}(q_l = \mathit{many} \mid t_c = \mathit{none})$.

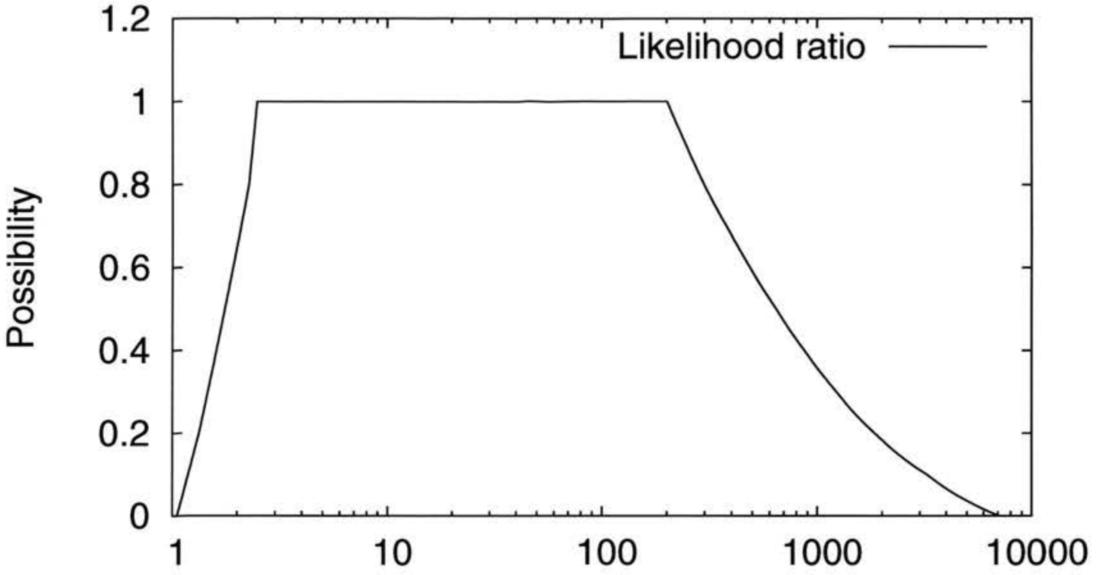


Figure 6.13: The computed fuzzy likelihood ratio plotted on a logarithmic scale.

ities cannot be determined experimentally for practical or economic reasons. That the likelihood ratio calculated above spans the entire range considered by FSS protocols suggests that the greater expressivity provided by linguistic probabilities would have important consequences for decision making within criminal prosecutions. Although the probabilistic relationship between factors in the scenario considered is not sufficiently determinate to allow a definitive view on its likelihood to be formed, it would be misleading to suppress this uncertainty.

Chapter 7

Conclusion

This chapter begins with a recapitulation of the major results presented in the foregoing text. This is followed by a discussion of some of the limitations of the present work together with some suggestions as to how these might be addressed in future.

7.1 Summary

Chapter 1 was primarily concerned with establishing the context and motivation for this work. A number of studies were cited each suggesting that fuzzy numbers might be a more suitable representation for subjective probability estimates than classical point probabilities. Taking this as read, the central research question was stated: what form should a theory of fuzzy probabilities take?

Chapter 2 sought to establish the philosophical and technical background underlying the work as a whole. Although most of the material there is well established, some of the observations about fuzzy numbers and arithmetic have not been widely recognised. Furthermore some notational conventions, especially those concerned with alphacuts, fuzzy numbers and fuzzy intervals which have not yet found wide acceptance were introduced for the sake of clarity.

Building on this foundation Chapter 3 presented a brisk summary of previous work on hybridizing probability theory and fuzzy logic. There have been a reasonable number of assays in this general direction most of which have been rather tentative or positional in character. In the context of the present work Chapter 3's principle contribution was to introduce and distinguish between the two most widely cited attempts to provide a theory of "fuzzy probabilities". It was shown that, though intuitively appealing both Zadeh and Jain and Agogino's theories exhibit serious technical problems that render them incapable of expressing non-trivial fuzzy probability measures.

In response to these criticisms, Chapter 4 developed the core contribution of this thesis – the theory of linguistic probabilities. Unlike the somewhat ad hoc approach taken by earlier researchers, linguistic probability theory was explicitly patterned after the standard measure-theoretic axioms of contemporary probability theory. It was shown that linguistic probability measures, like their classical counterparts, are monotonic and continuous. Analogues for the classical concepts of conditional probability, independence and discrete and continuous random variables were introduced and discussed.

Chapter 5 developed the theory into a computational application by showing how linguistic probabilities might be used in conjunction with a Bayesian-network-like graphical knowledge representation. It was shown that the representational scheme is consistent in the sense that networks with an identified set of properties do indeed represent joint linguistic probability mass functions. The issue of representational completeness was also examined and it was shown that the classical constructive proof strategy is invalid in this context.

In order to illustrate how such linguistic Bayesian networks might be used, Chapter 6 introduced the application area of forensic statistics and presented a simple case-study contrasting the fuzzy and the classical approaches. Whilst tentative, the results of this comparison indicate the potential significance of adopting the proposed representational scheme as small second-order uncertainties in prior and conditional probability estimates can have an extremely significant effect on established decision making protocols.

7.2 Discussion

The core theory of linguistic probabilities developed in Chapter 4 provides a much more robust foundation for further investigation of "fuzzy probabilities". The relative ease with which sophisticated concepts such as random variables and expectation were developed illustrates the advantages of adopting a principled, measure-theoretic approach over the earlier statistical (Zadeh, 1984) and mass-function-based (Jain and Agogino, 1990) analyses.

Furthermore, the demonstration that linguistic probabilities may be used to construct graphical knowledge representations shows that this is not merely a dry, mathematical theory, but a practical calculus capable of efficient implementation.

Finally, the utility of the approach as whole has not been demonstrated here (although studies such as Budescu and Wallsten (1985) lend strong support) however the case-study in forensic statistics was well-received by an expert audience.

7.3 Future work

Nevertheless, the present work has only scratched the surface of each of these three aspects. The following sections examine some of these limitations and sketches a programme of work that might lead to a fuller picture.

7.3.1 Interpretation

Hitherto the interpretation of linguistic probabilities – what they *mean* – has been treated as unproblematic. The intent has been to suggest a subjectivist approach consistent with the predominant understanding of classical probability theory. However it is possible to see the beginnings of connections with alternative semantics.

One extremely natural approach is to align linguistic probabilities with second-order classical probabilities. For the purposes of easy exposition, suppose that A and B are disjoint events (so $A \cap B = \varnothing$) such as a die roll's resulting in a 1

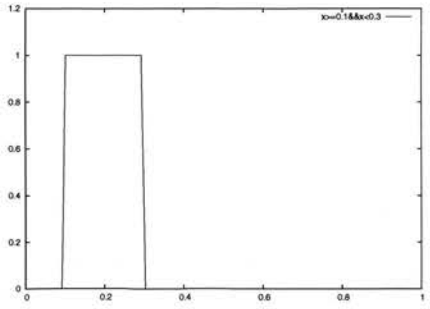


Figure 7.1: A graph of the density distribution of p_A.

or 6 respectively. Suppose further that since the probabilities are not known precisely but are instead given by random variables, p_A and p_B.

This model is a very simple second-order probability distribution i.e. we have probability density functions, f_{p_A} and f_{p_B}. Ontologically this is usually thought of as corresponding with a probability distribution over all possible probability distributions with respect to some (fixed) sigma algebra. Consider the uniform case where p_A and p_B are evenly distributed between a_1 and a_2, and b_1 and b_2 respectively. For notational convenience define,

$$a = a_2 - a_1 \quad \text{and} \quad b = b_2 - b_1 \tag{7.1}$$

Then we have the second-order density functions,

$$f_{p_A}(x) = \begin{cases} \frac{1}{a} & \text{if } x \in [a_1, a_2] \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad f_{p_B}(x) = \begin{cases} \frac{1}{b} & \text{if } x \in [b_1, b_2] \\ 0 & \text{otherwise} \end{cases}$$

In order to picture of what the various functions involved in the following look like it is necessary to plug in a set of values. Although what follows is perfectly general the visualisations will use $a_1 = 0.1, a_2 = 0.3, b_1 = 0.1$ and $b_2 = 0.5$. This density function is graphed in Figure 7.1.

Now consider the confidence intervals this generates. First the expected value of p_A may be calculated as

$$m_A = E(p_A) = \int_0^1 x f_{p_A}(x) dx = \int_{a_1}^{a_2} \frac{x}{a} dx = \frac{a_1 + a_2}{2} \tag{7.2}$$

Next, consider regions either side of this mean. Given $0 \leq \delta \leq \frac{a2-a1}{2}$,

$$\begin{aligned} P(p_A \in [m_A - \delta, m_A + \delta]) = & \int_{m_A-\delta}^{m_A+\delta} f_{p_A}(x) dx \\ = & \frac{m_A+\delta-m_A+\delta}{a} \\ = & \frac{2\delta}{a} \end{aligned} \tag{7.3}$$

This equation may be solved to determine the function d_A such that

$$P(p_A \in [m_A - \delta_A(\alpha), m_A + \delta_A(\alpha)]) = 1 - \alpha \tag{7.4}$$

This may be calculated as

$$\delta_A(\alpha) = \frac{a(1-\alpha)}{2} \tag{7.5}$$

A similar calculation can be made for $C = A \cup B$. Since A and B are exclusive events we know that the probability of C is just the sum of their probabilities i.e. $p_c = p_a + p_b$.

Now assuming that the second-order priors are independent, that $a_2 + b_2 \leq 1$, and without loss of generality that that $a \leq b$ (if not then we can "swap" A and B. This final condition introduces an element of asymmetry which will persist through the following calculations that would otherwise be surprising. This immediately implies that $a_2+b_1 \leq a_1+b_2$.

It is then possible to derive the following expression for the convolution,

$$\begin{aligned} f_C(x) &= \int_0^x f_{p_A,p_B}(t,x-t)dt \\ &= \begin{cases} 0 & x < a_1+b_1 \\ \frac{x-a_1-b_1}{ab} & a_1+b_1 \leq x < a_2+b_1 \\ \frac{a_2-a_1}{ab} & a_2+b_1 \leq x < a_1+b_2 \\ \frac{a_2+b_2-x}{ab} & a_1+b_2 \leq x < a_2+b_2 \\ 0 & a_2+b_2 \leq x \end{cases} \end{aligned} \tag{7.6}$$

As a sanity check one can verify that this is indeed a density function (i.e. $\int_0^1 f_C(x)dx = 1$). It is graphed in Figure 7.2. Now it is possible to calculate the expected value and confidence intervals. By symmetry (or calculation) $m_C = E(p_C) = m_A + m_B$ as expected.

Now the confidence interval equation for p_C can be computed as,

$$P(p_C \in [m_C-\delta, m_C+\delta]) = \begin{cases} \frac{2\delta}{b} & \text{if } 0 \leq \delta \leq \frac{b-a}{2} \\ \frac{\delta(a+b-\delta)-\frac{1}{4}(b-a)^2}{ab} & \text{if } \frac{b-a}{2} < \delta \leq \frac{a+b}{2} \\ 1 & \text{otherwise} \end{cases} \tag{7.7}$$

It then remains to analyse the relationship between δ and α i.e. find the function δ_C such that

$$P(P_C \in [m_C-\delta_C(\alpha), m_C+\delta_C(\alpha)]) = 1-\alpha$$

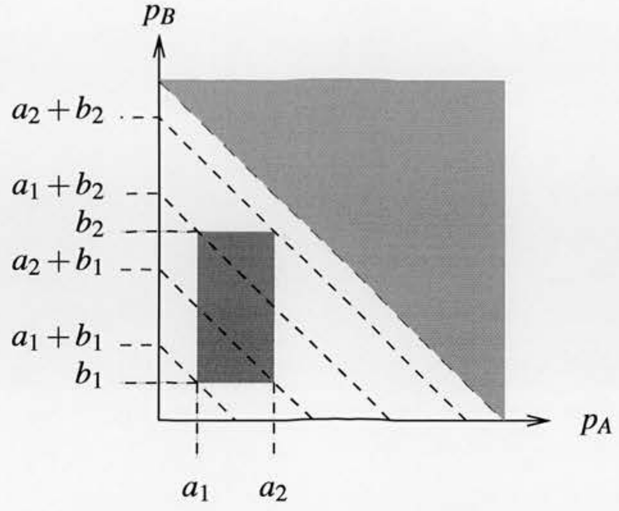


Figure 7.2: Diagram of the joint distribution f_{p_A,p_B} showing the various regions used in the case analysis of its density function and confidence intervals.

This can be calculated as,

$$\delta_C(\alpha) = \begin{cases} \frac{b(1-\alpha)}{2} & \text{if } \alpha > \frac{a}{b} \\ \frac{a+b}{2} - \sqrt{\alpha ab} & \text{otherwise} \end{cases} \tag{7.8}$$

The delta function "predicted" by linguistic probability theory is $\frac{(1-\alpha)(a+b)}{2}$. It is relatively straightforward to show that this dominates δ_C. The (slightly) difficult case is where $\alpha \leq \frac{a}{b}$. Note that,

$$\begin{aligned} \alpha ab &= \alpha\tfrac{a}{b}\tfrac{(b+b)^2}{4} \\ &\geq \alpha^2\tfrac{(b+b)^2}{4} && \text{since } \alpha < \tfrac{a}{b} \\ &\geq \alpha^2\tfrac{(a+b)^2}{4} && \text{since } a < b \end{aligned}$$

Hence,

$$\frac{\alpha(a+b)}{2} \leq \sqrt{\alpha ab}$$

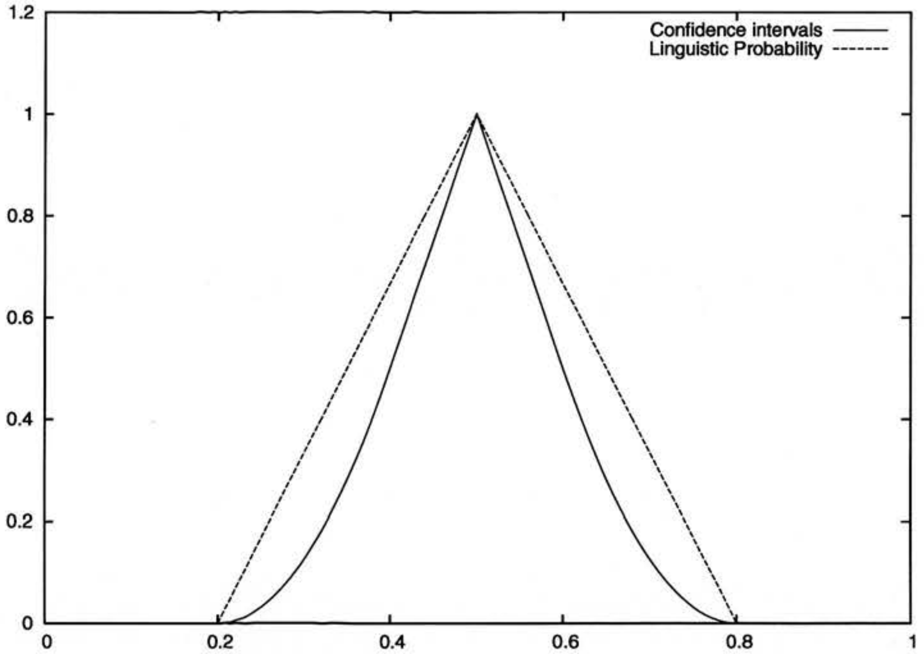


Figure 7.3: The computed confidence intervals for the second-order probability p_C and the corresponding linguistic probability.

and so,

$$\frac{(1-\alpha)(a+b)}{2} \geq \frac{a+b}{2} - \sqrt{\alpha ab} = \delta_C(\alpha)$$

as required.

So for this rather simple example, linguistic probabilities seem at least consistent with an interpretation based on second-order probabilities and their confidence intervals. The general case has, however, proven less amenable to this style of analysis. But that a proof has not yet been found, does not entail no proof exists. The general idea of linguistic probabilities as an an independence agnostic approach to using second-order models remains an attractive goal and interesting prospect for future research.

Another line of investigation that was inconclusively pursued as a possible basis for the interpretation of linguistic probabilities was their connection with rough-sets. Rough set theory is an extension of classical set theory that takes as its "atoms" pairs of classical sets, representing, so to speak, a most conservative and most liberal estimate of a concept's extension given a quantised reference space.

The intuition behind this connection is that event spaces may be thought of as just such a quantisation. Now given a probability measure $(\Omega, \mathcal{E}, \mathrm{P})$ rough set theory gives us a very natural means of representing the probability of a set, A of outcomes that is not in the set of events. So the lower approximation would be

$$\sup\{\mathrm{P}(E) \,:\, E \in \mathcal{E} \wedge E \subseteq A\} \tag{7.9}$$

and the upper

$$\inf\{\mathrm{P}(E) \,:\, E \in \mathcal{E} \wedge A \subseteq E\} \tag{7.10}$$

So rough-sets naturally generate interval-valued probabilities, but how might this link with fuzzy-valued probabilities? One natural possibility would be to examine a mapping from alpha values to sub-algebras of a given event space. In principle such a mapping seems, formally at least, consistent with the calculus of linguistic probabilities. However, fuzzy truth values are not obviously aligned with differing resolutions, so this approach might raise more questions than it promises to answer.

7.3.2 Evaluation as a method for knowledge engineering

Although the idea of using linguistic probabilities to capture expert knowledge has been strongly endorsed by the studies cited in Chapter 1, it would be helpful to examine to what extent linguistic Bayesian networks are a useful technique in this area.

Ideally one would collect two groups of experts in a particular domain and invite them to create networks modelling a particular scenario. One group would utilise conventional Bayesian networks and the other a linguistic system. In order to reduce the scope of the experiment and facilitate interpretation of the results, the topology might be fixed in advance.

It would then be possible to examine quantitative properties of the knowledge elicitation process, such as how long it took. This is often used to evaluate such processes (Menzies and van Harmelen, 1999).

In evaluating the resulting models themselves, several metrics might be considered. First, since they generalise the classical case any use of the additional expressivity allowed by linguistic Bayesian networks could be construed as immediate confirmation that this expressivity is welcome.

Next, it would be interesting to contrast the output of the models with historical decisions. This would require a corpus of internal FSS records discussing cases that fit the pre-agreed modelling scenario. However, the finding an appropriate scenario and adequately addressing the ethical and legal issues surrounding the creation and utilisation of such a corpus would require substantial effort.

Obviously the use of different design tools would further complicate the evaluation of these results. ARBOR's ability to edit both classical and linguistic Bayesian networks would however substantially mitigate this.

Nevertheless, it remains difficult to quantitatively evaluate knowledge engineering techniques. Human factors, priming effects, and the absence of a "correct" target output are amongst the many problems in creating a rigorous empirical study. For these reasons a qualitative approach would likely prove more illuminating.

So, participants in the proposed study would be asked to fill out a questionnaire enquiring about their experience of formalising their expert knowledge using the two representational schemes. The questions would focus as much as possible on their "comfort" with the resulting model, although the precise framing would benefit from collaboration with an expert in human-computer interaction.

Although much work remains to be done before linguistic Bayesian networks can be broadly accepted as a knowledge engineering technique, the promise of vastly enhanced expressivity at an affordable computational cost should not be ignored.

Appendix A

List of the author's publications

Halliwell, J. and Q. Shen (2001). From fuzzy probabilities to linguistic probability theory. In Q. Shen (Ed.), *Proceedings of the 2001 UK Workshop on Computational Intelligence*, pp. 129-135.

Halliwell, J. and Q. Shen (2002). Towards temporal linguistic variables. In *Proceedings of the 11th International Conference on Fuzzy Systems*, pp. 596-601.

Halliwell, J., J. Keppens and Q. Shen (2003). Linguistic bayesian networks of reasoning with subjective probabilities in forensic statistics. In *Proceedings of the 9th International Conference of Artificial Intelligence and Law*, pp. 42-50.

Halliwell, J. and Q. Shen (2007). Linguistic Probability Theory. To appear in *Soft Computing*.

Bibliography

Agogino, A. M. and A. Rege (1987). IDES: Influence Diagram Based Expert Systems. *Mathematical Modelling 8*, 227–233.

Aitken, C., F. Taroni, and P. Garbolino (2003). A graphical model for the evaluation of cross-transfer evidence in DNA profiles. *Theoretical Population Biology 63*(3).

Aristotle. *Metaphysics IV*. Harvard University Press.

Balding, D. J. and P. Donnelly (1995). Inference in forensic identification. *Journal of the Royal Statistical Society Series A 158*, 21–53.

Bertoluzza, C. and A. Bodini (1998). A new proof of Nguyen's compatibility theorem in a more general context. *Fuzzy Sets and Systems*, 99–102.

Brass, P. (2002). On the nonexistence of Hausdorff-like distance measures for fuzzy sets. *Pattern Recognition Letters 23*, 39–43.

Budescu, D. V. and T. S. Wallsten (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behaviour and Human Decision Processes 36*, 391–405.

Charniak, E. (1991). Bayesian networks without tears. *Artificial Intelligence Magazine 12*(4), 50–63.

Cook, R., I. Evett, G. Jackson, P. Jones, and J. Lambert (1998a). A hierarchy of propositions: deciding which level to address in casework. *Science and Justice 38*, 231–239.

Cook, R., I. Evett, G. Jackson, P. Jones, and J. Lambert (1998b). A model for case assessment and interpretation. *Science and Justice 38*, 151–156.

Cook, R., I. Evett, G. Jackson, P. Jones, and J. Lambert (1999). Case pre-assessment and review in a two-way transfer case. *Science and Justice 39*, 103–111.

Cooper, G. (1990). Probabilistic inference using belief networks in np-hard. *Artificial Intelligence 42*, 393–405.

Corral, N., M. A. Gil, M. T. López, A. Salas, and C. Bertoluzza (1998). Statistical models. See Ruspini et al. (1998).

Cox, R. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics 14*, 1–13.

Czogala, E. C. and J. Drewniak (1984). Associative monotonic operations in fuzzy sets theory. *Fuzzy sets and systems 12*, 249–270.

Dagum, P. and M. Luby (1993). Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence 60*, 141–153.

Dawid, A. P., J. Mortera, V. L. Pascali, and D. W. van Boxel (2002). Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics 29*, 577–595.

de Cooman, G. (2003a). A behavioural model for vague probability assessments. Personal communication expected to appear for journal publication soon.

de Cooman, G. (2003b). A behavioural model for vague probability assessments. *Fuzzy Sets and Systems*.

Dong, W. M. and F. S. Wong (1987). Fuzzy weighted averages and implementation of the extension principle. *Fuzzy Sets and Systems 21*, 183–199.

Druzdzel, M. J. and L. C. van der Gaag (1995). Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*, pp. 141–148.

Dubois, D. and P. H (1989). Fuzzy sets, probability and measurement. *European Journal of Operational Research 40*, 135–154.

Dubois, D. and H. Prade (1979). Fuzzy real algebra: some results. *Fuzzy sets and systems 2*, 327–348.

Dubois, D. and H. Prade (1980). *Fuzzy sets and Systems: Theory and Applications*. London: Academic.

Dubois, D. and H. Prade (1986). Fuzzy sets and statistical data. *European Journal of Operational Research 25*, 345–356.

Evett, I., G. Jackson, and J. Lambert (2000). More on the hierarchy of propositions: exploring the distinction between explanations and propositions. *Science and Justice 40*, 3–10.

Evett, I., G. Jackson, J. Lambert, and S. McCrossan (2000). The impact of the principles of evidence interpretation on the structure and content of statements. *Science and Justice 40*, 233–239.

Fagin, R. and J. Y. Halpern (1994). Reasoning about knowledge and probability. *Journal of the ACM 41*(2), 340–367.

Feron, R. (1976). Ensembles alétoires flous. *C. R. Acad. Sci. Paris Ser. A 282*, 903–906.

Fertig, K. W. and J. S. Breese (1990). Interval influence diagrams. See Schachter et al. (1990).

Freeling, A. N. S. (1980). Fuzzy sets and decision analysis. *SMC 10*, 341–354.

Fullér, R. and T. Keresztfalvi (1990). Generalization of Nguyen's theorem. *Fuzzy Sets and Systems 41*, 371–374.

Gärdenfors, P. (1988). *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. Cambridge, Massachusetts: MIT Press.

Gardner-Medwin, T. (2005). What probability should a jury address? *Significance 2*(1).

Garey, M. R. and D. S. Johnson (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman.

Giachetti, R. E. and R. E. Young (1998). A parametric representation of fuzzy numbers and their arithmetic operators. *Fuzzy Sets and Systems: Special Issue on Fuzzy Arithmetic*.

Gil, M. A. and M. R. Casals (1988). An operative extension of the likelihood ratio test from fuzzy data. *Stat. Papers 29*, 191–203.

Gil, M. A., N. Corral, and M. R. Casals (1989). The likelihood ratio test for goodness of fit with fuzzy experimental observations. *IMSC 19*, 771–779.

Gil, M. A., N. Corral, and P. Gil (1988). The minimum accuracy estimates in χ^2 tests for goodness of fit with fuzzy observations. *Journal of Statistical Planning and Inference 19*, 95–115.

Gmytrasiewics, P. J. and E. H. Durfee (1995). A rigorous, operational formalization of recursive modelling. In *Proceedings of the First International Conference on Multi-Agent Systems*, pp. 125–132.

Gödel, K. (1931). ber formal unentscheidbare stze der principia mathematica und verwandter systeme. *I. Monatsh. Math. Phys. 38*, 173–198.

Goldszmidt, M. and J. Pearl (1992). Rank-based systems: A simple approach to belief revision, belief update and reasoning about evidence and actions. In *KR-92, Principles of Knowledge Representation and Reasoning: Proceeding of the Third International Conference*, San Mateo, CA, pp. 661–672. Morgan Kaufmann Publishers, Inc.

Gómez Marin-Blazquez, J., Q. Shen, and A. Gómez-Skarmeta (2000). From approximative to descriptive models. In *Proceedings of the 9th International Conference on Fuzzy Syste ms*.

Grimmet, G. and D. Welsh (1986). *Probability: an introduction*. Oxford: Oxford University Press.

Halliwell, J., J. Keppens, and Q. Shen (2003). Linguistic bayesian networks for reasoning with subjective probabilities in forensic statistics. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law*, pp. 42–50.

Halliwell, J. and Q. Shen (2001). From fuzzy probabilities to linguistic probability theory. In Q. Shen (Ed.), *Proceedings of the 2001 UK Workshop on Computational Intelligence*, pp. 129–135.

Halliwell, J. and Q. Shen (2002). Towards temporal linguistic variables. In *Proceedings of the 11th International Conference on Fuzzy Systems*, pp. 596–601.

Halpern, J. Y. (1999). Cox's theorem revisited. *Journal of Artificial Intelligence Research 11*, 429–435.

Huang, C. and A. Darwiche (1994). Inference in belief networks: A procedural guide. *International Journal of Approximate Reasoning 5*(3), 225–263.

Jain, P. and A. Agogino (1988). Calibration of fuzzy linguistic variables for expert systems. In *Computers in Engineering*, pp. 313–318. New York: American Society of Mechanical Engineering.

Jain, P. and A. M. Agogino (1990). Stochastic sensitivity analysis using fuzzy influence diagrams. See Schachter et al. (1990).

Jensen, F. (1989). Updating in recursive graphical models by local computations. Technical Report R-89-15, Department of Mathematics and Computer Science, University of Aalborg.

Kahneman, D., P. Slovic, and A. Tversy (1985). *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press.

Kanal, L. N. and J. F. Lemmer (Eds.) (1986). *Uncertainty in Artificial Intelligence*. Amsterdam: North-Holland.

Kerre, E. E. and A. van Schooten (1988). A deeper look on fuzzy numbers from a theoretical as well as from a practical point of view. In M. Gupta and T. Yamakawa (Eds.), *Fuzzy Logic in Knowledge-Based Systems, Decision and Control*, pp. 173–196. Elsevier Science Publishers.

Klir, G. and B. Yuan (1995). *Fuzzy Sets and Fuzzy Logic: Theory and Applications.* Upper Saddle River, NJ: Prentice-Hall.

Kruse, R. and J. Gebhardt (1989). On a dialog system for modelling and statistical analysis of data. In J. Beadete (Ed.), *Proc. 3rd Int. Fuzzy Systems Assoc. Conf. on The Coming of Age of Fuzzy Logic, IFSA '89*, pp. 157–160.

Kruse, R. and K. D. Meyer (1987). *Statistics with Vague Data.* Dordrecht: Reidel.

Kwakernaak, H. (1978a). Fuzzy random variables. Part I: definitions and theorems. *IS 15*, 1–29.

Kwakernaak, H. (1978b). Fuzzy random variables. Part II: algorithms and examples for the discrete case. *IS 17*, 253–278.

Lauritzen, S. and D. Speigelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society 50*, 157–224.

Lemmer, J. F. and L. N. Kanal (Eds.) (1988). *Uncertainity in Artificial Intelligence 2*. North-Holland: Elsevier Science Publishers B. V.

Lichtenstein, S. and J. R. Newman (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychometric Science 9*, 563–564.

Mendel, J., H. Hagras, and R. John (2007). Standard background material about interval type-2 fuzzy logic systems that can be used by all authors. Technical report, IEEE CIS Standards Committee.

Mendel, J. and R. John (2002, April). Type-2 fuzzy sets made simple. *IEEE Transactions on Fuzzy Systems 10*.

Menzies, T. and F. van Harmelen (1999). Evaluating knowledge engineering techniques. *Int. J. Hum.-Comput. Stud. 51*(4), 715–727.

Milch, B. and D. Koller (2000). Probabilistic models for agents' belief and decisions. In *Proceedings of the Sixteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-00)*, pp. 389–396.

National Research Council Governing Board Commitee on the Assessment of Risk (1981). *The handling of risk assessments in NRC reports*. Washington D.C: National Research Council.

Nebel, B. (1986). How hard is it to revise a belief base? Technical report, Albert-Ludwigs-Universiät Frieburg, Institut für Informatik.

Negoita, C. V. and D. A. Ralescu (1987). *Simulation, Knowledge-Based Computing and Fuzzy Statistics*. New York: Van Nostrand Reinhold Comp.

Nguyen, H. T. (1978). A note on the extension principle for fuzzy sets. *Journal of Mathematical Analysis and Applications 64*, 369–380.

Nguyen, H. T. (1979). Some mathematical tools for linguistic probabilities. *Fuzzy Sets and Systems 2*, 53–65.

Okuda, T., H. Tanaka, and K. Asai (1978). A formulation of fuzzy decision problems with fuzzy information, using probability measures of fuzzy events. *Information Control 38*, 135–147.

Okuda, T., H. Tanaka, and K. Asai (1991). Maximum likelihood estimation from fuzzy observation data. In R. Lowen and M. Roubens (Eds.), *Proceed-*

ings of the 4th Int. Fuzzy Systems Assoc. Conf. On Computer, Management and Systems Science IFSA'91 (Brussels), pp. 185–188.

Paris, J. B. (1994). *The Uncertain Reasoner's Companion*. Cambridge, UK: Cambridge University Press.

Pearl, J. (1986). Fusion, propagation and structuring in belief networks. *Artificial Intelligence 29*(3), 231–288.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Pepper, S. and L. S. Prytulak (1974). Sometimes frequently means seldom: Context effects in the interpretation of quantitive expressions. *Journal of Research in Personality 8*, 95–101.

Puri, M. L. and D. A. Ralescu (1986). Fuzzy random variables. *Journal of Mathematical Analysis and Applications 114*, 409–422.

Ralescu, D. A. (1995a). Fuzzy probabilities and their application to statistical inference. In B. Bouchon, R. Yager, and L. Zadeh (Eds.), *Proc. Int. Conference on Information Processing and the Management of Uncertainty in Knowledge-based Systems, IPMU'94*, Number 945 in Lecture Notes in Computer Science, Berlin, pp. 217–222. Springer.

Ralescu, D. A. (1995b). Fuzzy random variable revisited. In *Proc. 4th IEEE Int. Conf. on Fuzzy Systems/2nd Int. Fuzzy Engineering Symposium, FUZZ-IEEE/IFES'95 (Yokohama)*, Volume 2, New York, pp. 993–1000. IEEE.

Ralescu, D. A. (1996). Statistical decision-making without numbers. In *Proc.*

26th Iranian Mathematical Conf. (Kerman), pp. 403–417. Iranian Mathematical Society.

Rappaport, A., T. S. Wallsten, and J. A. Cox (1987). Direct and indirect scaling of membership functions of probability phrases. *Mathematical Modelling 9*, 397–418.

Ruspini, E. H. (1970). Numerical methods for fuzzy clustering. *IS 2*, 319–350.

Ruspini, E. H., P. P. Bonissone, and W. Pedrycz (Eds.) (1998). *Handbook of Fuzzy Computation*. Bristol, UK: Institute of Physics Publishing.

Ruspini, E. H. and E. H. Mamdami (1998a). Approximate reasoning. See Ruspini et al. (1998).

Ruspini, E. H. and E. H. Mamdami (1998b). Probability, imprecision and vagueness. See Ruspini et al. (1998), pp. A2.3:1–A2.3:7.

Russel, S. and P. Norvig (1995). *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, New Jersey: Prentice Hall.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.

Schachter, R. D., T. S. Levitt, L. N. Kanal, and J. F. Lemmer (Eds.) (1990). *Uncertainty in Artificial Intelligence 4*. Amsterdam: North-Holland.

Scholz, R. W. (Ed.) (1983). *Decision making under uncertainty*. Amsterdam: North-Holland.

Schwiezer, B. and A. Sklar (1963). Associative functions and abstract semigroups. *Publicationes Mathematicae, Debrecen 10*, 69–81.

Shafer, G. (1986). Savage revisited (with discussion). *Statistical Science 1*, 463–501.

Song, Q., A. Kandel, and M. Schneider (2003). Parameterized fuzzy operators in fuzzy decision making. *Internation Journal of Intelligent Systems 18*(9), 971–987.

Steyaert, H., F. Van Parys, R. Baekland, and E. E. Kerre (1995). Implementation fo piecewise linear fuzzy quanitities. *International Journal of Intelligent Systems 10*, 1049–1059.

Tanaka, H., T. Okuda, and K. Asai (1979). Fuzzy information and decision in statistical model. In M. Gupta, P. Rage, and R. Yager (Eds.), *Advances in Fuzzy Sets Theory and Applications*, pp. 303–320. Amsterdam: North Holland.

Thomas, S. F. (1995). *Fuzziness and Probability*. Wichitea, KS: ACG.

Utkin, L. V. (1993). Uncertainty importance of system components by fuzzy and interval probability. *Microelectronics and reliability 33*, 1357–1364.

Vidal, J. and E. Durfee (1996). Building agent models in economic societies of agents. In *Working Notes of the AAAI-96 Workshop on Agent Modeling*.

Vidal, J. M. and E. H. Durfee (1998). Learning nested agent models in an information economy. *Journal of Experimental and Theoretical Artificial Intelligence (special issue on learning in distributed artificial intelligence systems)*, 291–308.

Voorbraak, F. (1996). Probabilistic belief expansion and conditioning. Technical report, Department of Mathematics, Computer Science, Physics and Astronomy, University of Amsterdam. (a revised and condensed version of this report was presented at UAI99).

Walley, P. (1982). The elicitation and aggregation of beliefs. Technical report, University of Warwick, Coventry, UK.

Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. London: Chapman and Hall.

Walley, P. and G. de Cooman (1998). A behavioural model for linguistic uncertainty. In P. Wang (Ed.), *Computing with words*. New York: John Wiley & Sons.

Wallsten, T. S., D. V. Budescu, A. Rapoport, R. Zwick, and B. Forsyth (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General 115*(4), 348–365.

Wallsten, T. S., S. Fillenbaum, and J. A. Cox (1986). Base rate effects on the interpretation of probability and frequency expressions. *Journal of Memory and Language 25*, 571–587.

Watson, S. R., J. J. Weiss, and M. L. Donnell (1979). Fuzzy decision analysis. *SMC 9*, 1–9.

Wellman, M. P. (1988). *Qualitative probabilistic networks for planning under uncertainty*. In Lemmer and Kanal Lemmer and Kanal (1988).

Wood, K. L., K. N. Otto, and W. K. Antonsson (1992). Engineering design calculations with fuzzy parameters. *Fuzzy Sets and Systems 52*, 1–20.

Zadeh, L. (1965). Fuzzy sets. *Information Control 8*(338).

Zadeh, L. (1968). Probability measures of fuzzy events. *Journal of Mathematical Analysis and Applications 23*, 421–427.

Zadeh, L. (1975a). The concept of a linguistic variable and its application to approximate reasoning. Part 1. *IS 8*, 199–249.

Zadeh, L. (1975b). The concept of a linguistic variable and its application to approximate reasoning. Part 2. *IS 8*, 301–353.

Zadeh, L. (1975c). The concept of a linguistic variable and its application to approximate reasoning. Part 3. *IS 9*, 43–80.

Zadeh, L. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computational Mathematics 9*, 149–184.

Zadeh, L. (1984). Fuzzy probabilities. *Information Processing Management 20*, 363–372.

Zadeh, L. (1985). Syllogistic reasoning in fuzzy logic and its application to usuality and reasoning with dispositions. *SMC 15*, 754–763.

Zadeh, L. (1996). Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems 2*, 103–111.

Zhong, C. and G. Zhou (1987). The equivalence of two definitions of fuzzy random variables. In *Proceedings of the 2nd International Fuzzy Systems Association Conference (Tokyo)*, pp. 59–62.

Zimmer, A. C. (1983). Verbal vs. numerical processing of subjective probabilities. See Scholz (1983).

Zimmer, A. C. (1984). A model for the interpretation of verbal predictions. *International Journal of Man-Machine Studies 20*, 121–134.

Zimmer, A. C. (1986). What uncertainty judgements can tell about the underlying subjective probabilities. See Kanal and Lemmer (1986).

Zimmerman, H. (1991). *Fuzzy Sets Theory and its Applications*. Deventer: Kluwer.