REASONING WITH UNCERTAINTY USING NILSSON'S PROBABILISTIC LOGIC AND THE MAXIMUM ENTROPY FORMALISM

Thomas Brett Kane BSc. (Hons)

April 2, 1992

This thesis is submitted in fulfillment of the requirements for the degree of Doctor of Philosophy to the University of Heriot-Watt. The research was conducted at the Department of Computer Science at the University of Heriot-Watt.

Contents

[1]The work reported in this chapter was done in collaboration with A.M. Wallace and P. McAndrew of Heriot-Watt University's Computer Science Vision Group.

List of Figures

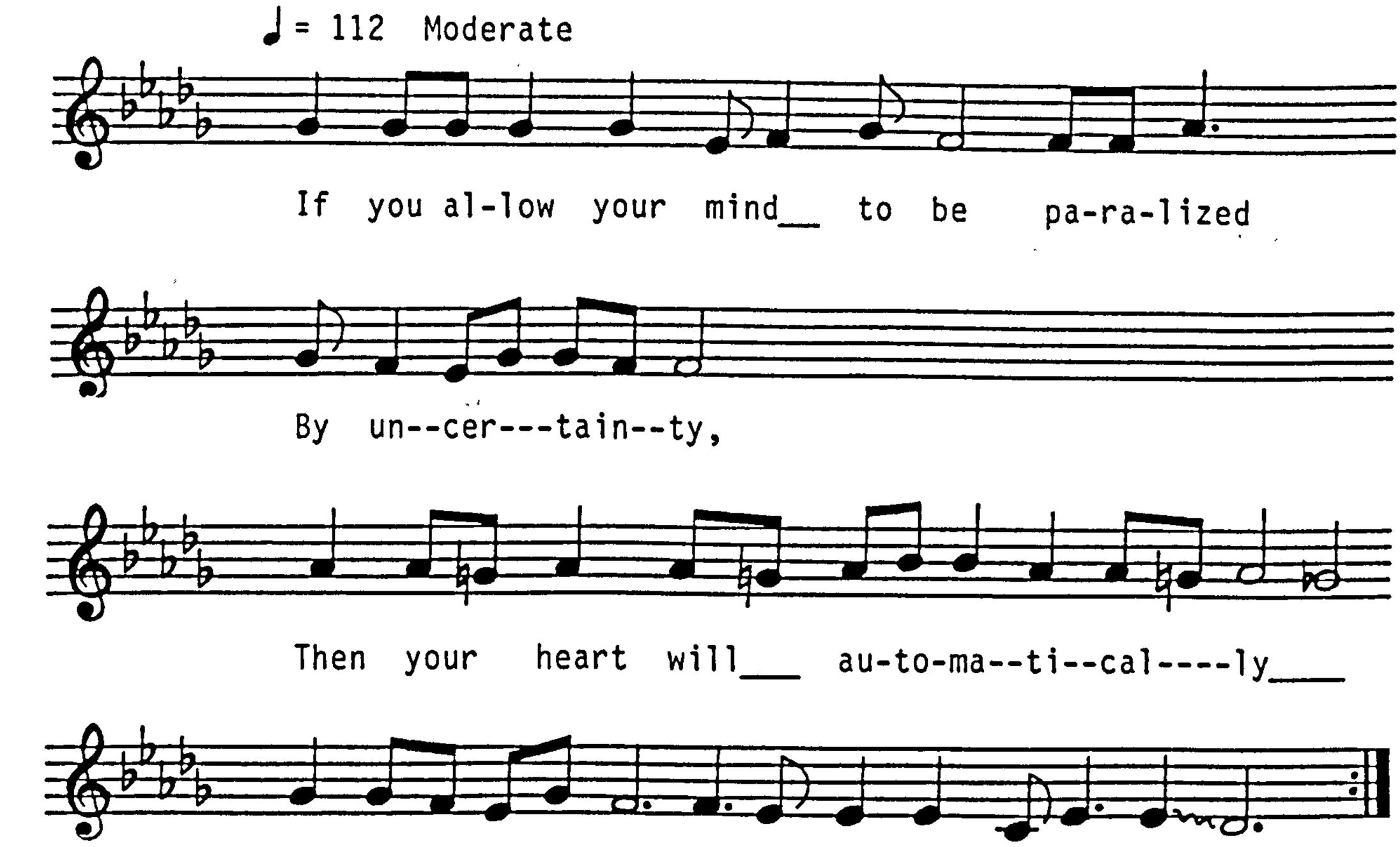
List of Tables

Acknowledgement

I would like to thank many people for helping me one way or another while I have been working on this thesis. Firstly, Alex Gammerman who introduced me to Nilsson's Probabilistic Logic and invited me to come and work with him at Heriot-Watt University. Secondly the rest of the staff at Heriot-Watt Computing Science, especially Greg Michaelson, Kevin Waugh, Patrick McAndrew, Steven Salvini, Andrew Wallace and Hunter Davis who greatly helped over the five years I spent there. Thirdly, all of my friends in the Phd room. Fourthly all of the people external to the department who took an interest in my work, and who helped shape it in one way or another. A special thanks to Alan Bundy, Peter Cheeseman, Max Henrion, Ed Jaynes, Mary McLeish and Ted Shortliffe. Fifthly, I would like to thank the KCM group at Stirling University for allowing me time off from work to finish my write up.

I would also like to thank the members of my family for their tolerance and patience and support (which was not always warranted). Joanna (my wife), Margaret (my mum), Sidell and Patrick (my sister and brother). I would also like to say "here it is!" to my father who was a great inspiration to me, and who got me to promise to finish, and didn't allow me the time to change my mind. Finally, I would like to thank my friends at the Edinburgh Sri Chinmoy Centre, and my Spiritual Teacher Sri Chinmoy for showing me a new world and making me able for the work herein. Thank-you. Thank-you. Thank-you.

If you allow your mind to be
Paralyzed by uncertainty,
Then your heart will automatically
Be paralyzed by unwillingness.



Sri Chinmoy

Abstract

An expert system must reason with certain and uncertain information. This thesis is concerned with the process of Reasoning with Uncertainty. Nilsson's elegant model of "Probabilistic Logic" has been chosen as the framework for this investigation, and the information theoretical aspect of the maximum entropy formalism as the inference engine. These two formalisms, although semantically compelling, offer major complexity problems to the implementor. Probabilistic Logic models the complete uncertainty space, and the maximum entropy formalism finds the least commitment probability distribution within the uncertainty space.

The main finding in this thesis is that Nilsson's Probabilistic Logic can be succesfully developed beyond the structure proposed by Nilsson. Some deficiencies in Nilsson's model have been uncovered in the area of probabilistic representation, making Probabilistic Logic less powerful than Bayesian Inference techniques. These deficiencies are examined and a new model of entailment is presented which overcomes these problems, allowing Probabilistic Logic the full representational power of Bayesian Inferencing. The new model also preserves an important extension which Nilsson's Probabilistic Logic has over Bayesian Inference: the ability to use uncertain evidence.

Traditionally, the probabilistic solution proposed by the maximum entropy formalism is arrived at by solving non-linear simultaneous equations for the aggregate factors of the non-linear terms. In the new model the maximum entropy algorithms are shown to have the highly desirable property of tractability.

Although these problems have been solved for probabilistic entailment the problems of complexity are still prevalent in large databases of expert rules. This thesis also considers the use of heuristics and meta level reasoning in a complex knowledge base. Finally, a description of an expert system using these techniques is given.

Chapter 1

INTRODUCTION

1.1 Introduction

For the purposes of this thesis I will adopt the pragmatic approach and define an expert system as a computer program which works to solve a specialized problem in the same way that a human expert would [88, 112, 45]. To do this, an expert system has to be able to store and manipulate the knowledge of an expert, to be able to reason with the knowledge given to it, and to be able to present results to the end user in an acceptable fashion.

Expert systems have developed within the computer science sub-area of information technology (figure 1.1). With their development, information technology has expanded to include the artificial intelligence paradigms of knowledge processing and expert knowledge processing; and software engineering has been enriched with methods of abstraction which allow program knowledge to be separated from program control [46].

However, expert systems have inherited the standard information technological problems of run-time complexity and algorithmic correctness. This thesis is concerned with the process of deduction that goes on within an expert system, and with the preservation of correctness and the removal of complexity from expert systems.

1.2 The Expert Perspective

In calling someone an expert in a particular field we recognise two things: firstly that the field itself is sufficiently complex as to warrant careful decision making; and secondly that the person to whom we are refering has consistently demonstrated extreme proficiency in

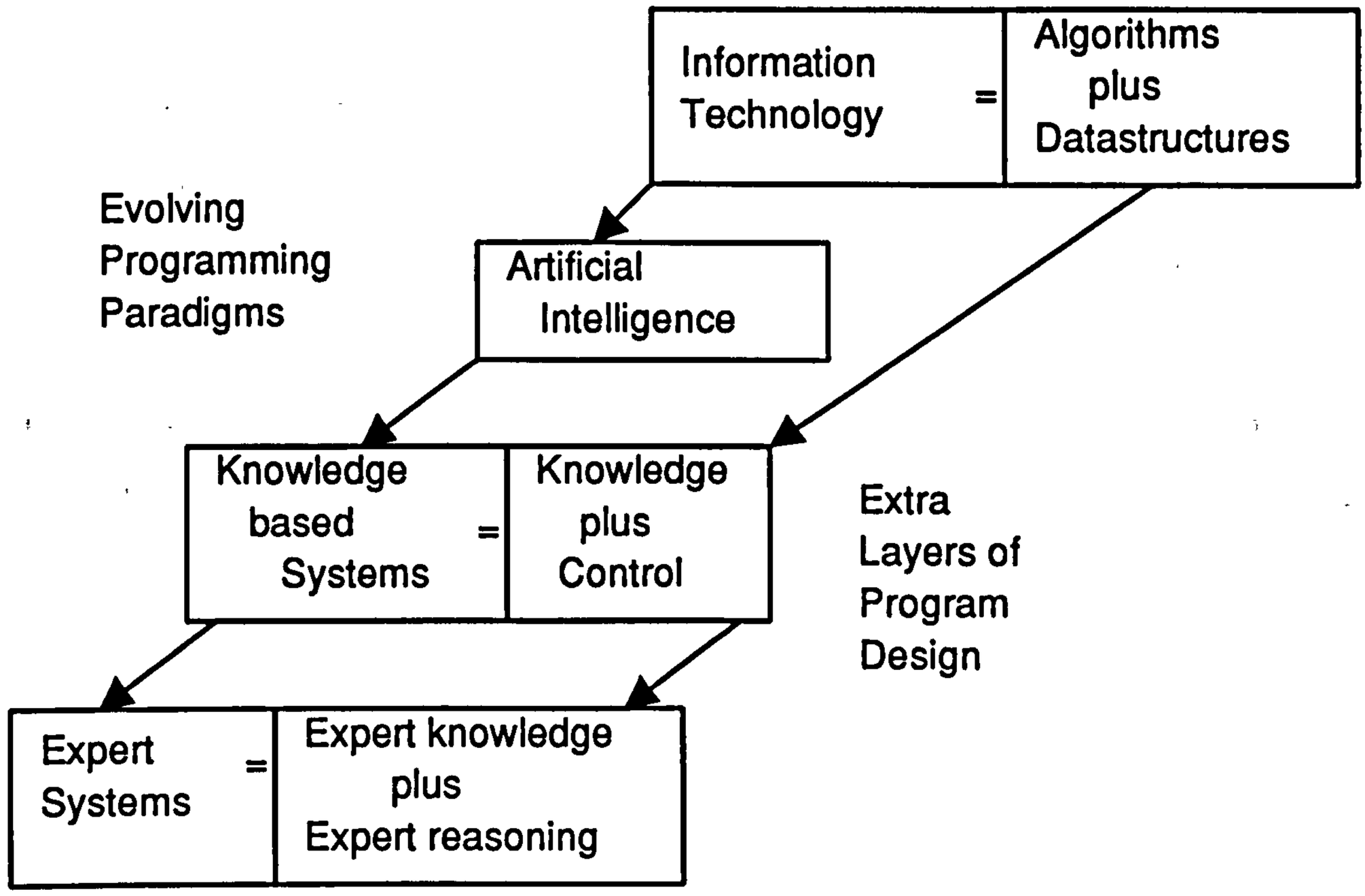


Figure 1.1: Expert System Technology within Information Technology

dealing with everything from simple to very detailed problems in the field. Typical expert fields are: diagnostic medicine, mineral prospecting, programming, process control, financial management, design and planning [133, 37].

But how does an expert differ from another worker in the field who is able, but not as capable as the expert? Patil has shown [92] that in a real life clinical situation, an expert is only ever dealing with a small number of hypotheses (no more than five or six) at one time; whereas, a non-expert is typically entertaining many more. It has become clear that the expert is consciously combining not only information in a logical way so as to manipulate possibility, but also with regard to its importance and certainty [10, 34]. So that the reasoning process is complicated not only with a number of possibilities, but also to the extent that these fluctuate in terms of likelihood as the investigative process proceeds.

1.3 The Software Engineering Perspective

Programmers, trying to write programs which do helpful work in sufficiently complex data fields, turned to expert systems with the hope that it would be possible to distil the subtle

nature of an expert's reasoning process and use this as a knowledge base rather than to try to model the same level of expertise in a computer program by working towards it from first principles. Thus, the feature which distinguishes a true expert system from an ordinary program is its ability to use the knowledge of an expert in coming to a conclusion.

A major motivation in developing expert systems is that interacting with an expert system should be like interacting with the expert who supplied the knowledge. Also, an expert's time will usually cost a lot of money, wheras it costs very little to make an online expert system available to users. Online expert systems offer the added advantage that many people can use the expert's knowledge at the same time. With these points in mind, an expert system must be able to produce plausible, if not certain, solutions to problems (the same way an expert would). It must also be able to explain its reasoning, and demonstrate that its reasoning procedures are valid (the same way an expert might).

Considering an expert system as a piece of software, the algorithms used should conform to a number of rules of good software engineering. Such rules include [31]:

1	Use simple but powerful general solutions
2	Can be easily understood by others
3	Can be easily modified if necessary
4	Are correct for clearly defined situations
5	May be understood on a number of levels
6	Are economical in the use of computational resources
7	Are documented well enough to be used by others
8	Are not dependant on being run on a particular computer
9	Are able to be used as a sub-procedure for other programs
10	Produce pleasing and satisfying solutions

Table 1.1: Desirable Characteristics of Good Algorithms

The state of the art in expert systems technology is such that current expert systems can be shown to exhibit many of these pleasing features of good algorithms, but not all [57, 130]. The problem facing the software engineer in Artificial Intelligence is to provide software which operates within these constraints while at the same time performing the desired function of intelligence.

The Artificial Intelligence Perspective

Computers are symbol manipulators [55, 56]. Early applications for computers have been in storing factual data, and performing quick arithmetical operations on numbers. However, a clear area of development for computer applications was in the psychological field of cognition. The science of Artificial Intelligence (AI) has developed to explore the possibilities of reasoning, learning, perception and language understanding on the digital computer.

So, in the first instance, AI application programs should obey the laws of good algorithms given in section 1.3. Secondly, AI programs must have simple and complete methods for representing and manipulating the basic elements of human knowledge. Thirdly, AI programs must have a way of quantifying uncertainty in knowledge so that the best use of available evidence is made when forming a decision. The approach adopted throughout this thesis is to use knowledge modelling techniques from the mathematics of predicate calculus and modal logics to model uncertain knowledge; and uncertainty management techniques taken from probability theory to quantify the strength of uncertainty. In the event that probability theory cannot be used to quantify uncertainty, (for reasons of complexity or lack of information), heuristic measures are proposed which may be used instead.

1.4 Overview of Thesis

The four disciplines: mathematics, statistics, psychology and computational science are the basic tools for reasoning with uncertainty in expert systems. This thesis examines how far the synthesis of disciplines can be achieved using the particular method of Nils Nilsson's probabilistic logic [89, 83, 52], with particular reference to the maximum entropy formalism [54]. (Throughout the text I shall use the abbreviation of "Probabilistic Logic" to replace "Nils Nilsson's probabilistic logic".) In the course of my investigations various extensions have been made to Probabilistic Logic, and a complex expert system is realised as a fruit of this process.

In the first section of the thesis, there is a brief history of the development of material from mathematics and statistics which has a bearing on this thesis (chapter 2). Also in the

first section is a description of present day systems for reasoning with uncertainty (chapter 3). The second section of the thesis, introduces various extensions to Nilsson's probabilistic logic (chapters 4) which allow an efficient use of the maximum entropy formalism (chapter 5) and relates these new results to Bayesian Inference techniques (chapter 6). The third section (chapters 7, 8 and 9) of the thesis concerns approximation techniques, the use of heuristics in a large system, and finally a full scale application of the major theoretical results of the thesis.

Chapter 2: The Mathematical Tools for Reasoning With Uncertainty

In chapter 2 the mathematical and statistical perspectives on automated reasoning are briefly summarised. This chapter is used to show how vigorously the contributory aspects of reasoning with uncertainty have been debated. The mathematical preliminaries essential to this thesis (automatic deduction techniques, theories of probability, Bayesian Inference) are introduced in this chapter. Also discussed in this chapter are the limits which naturally appear on any system attempting to reason with uncertainty.

Chapter 3: Automated Systems

Chapter 3 shows how the theories of reasoning with uncertainty have been developed in the time of the digital computer. The most difficult problem with providing a complete implementation of Bayesian Inference for instance has always been the problem of computational complexity.

This problem manifests itself in one of two ways: either the computer takes too long to compute and present its results (either because human time considerations render waiting the required time impractical, or simply because the computation will never complete); or because the computer does not have enough space available in which to store its local computations. These two problems are known as time and space complexity, and more than any other single factor they have directed a dominating influence upon the implementation of reasoning with

uncertainty on the digital computer.

Chapter 4: Enhancements to Probabilistic Logic

The main finding in this thesis is that Nilsson's probabilistic logic [89] can be succesfully developed beyond the structure proposed by Nilsson. In this chapter Nilsson's probabilistic entailment model is examined in detail. Various conventions are introduced for dealing with complex entailment problems. Some criticisms are made of Nilsson's proposed model, and ultimately a new interpretation of the semantic tree is presented. This involves the introduction of the "context split", which provides an extensive use for conditional probabilities in Nilsson's probabilistic logic.

Originally, Nilsson intended the task of his probabilistic logic to be that of computing the bounds of probability predictable from an underspecified probability model. A new algorithm for computing the bounds of an entailment procedure is presented in this chapter which will give the absolute bounds of an entailment procedure without having to resort to geometric considerations.

Chapter 5: The Maximum Entropy Formalism

Chapter 5 is concerned with the maximum entropy formalism. The maximum entropy formalism has been succesfully developed as a theory of information measurement [114]. In terms relevant to this thesis it can be viewed as a descriptor of the information content imposed on a probability distribution by the guiding probabilities in an entailment process (that is, the probabilities of the antecedents, and the probability of the rule of inference). In particular the fixed point of maximum entropy identifies a probability distribution over the possible worlds of an entailment problem which makes the least additional assumptions about the nature of the interactions between the sentences involved [79].

The maximum entropy formalism is of special interest to the expert system community in that it has a sound mathematical history from which to draw its credibility. The drawback with this method of choosing a probability distribution to fit the information available, is that

it introduces an extra complexity level into the proceedings. In general, the solution proposed by the maximum entropy formalism is arrived at by iteratively solving simultaneous equations for the aggregate factors of the non-linear terms.

In this chapter the complexity of the traditional solution method is examined. An algorithm for solving the maximum entropy equations for the new semantic tree is given, and its complexity is examined. This algorithm gives an immediate solution for the aggregate factors of the entailment problem. The implications of the result are examined.

Chapter 6:Nilssonian and Bayesian Inference Compared

Chapter 6 is concerned with an examination of Nilsson's probabilistic logic in the light of the Bayesian theory of Inference. This is possible because of the new bounds algorithm which has the ability to use conditional probabilities in Probabilistic Logic, and the ability to quickly determine the factors of the maximum entropy equations. In particular, the meaning of probabilistic entailment is cast in a new light. The information required by the two logics is compared and contrasted. The final analysis in this chapter is a description as to how far the two theories may be unified.

Chapter 7: Incidence Calculus

Bundy's Incidence Calculus [13, 12] is examined in chapter 7. Incidence calculus is similar to Probabilistic Logic in that it uses possible worlds. The theory proposed by Bundy allows the implementor to choose the number of possible worlds, and the system employs a mechanism for fashioning these worlds into possible worlds, and for assigning probabilities in a manner consistent with the rules of probability theory. Incidence calculus puts at our disposal a framework which allows for heuristic mechanisms for probability assignment, which become more attractive as the complexity of the entailment problems increase.

Chapter 8: Entropy and Meta-Level Reasoning

In this chapter the role of the maximum entropy formalism is extended to encompass meta-level reasoning: that is to choose which of a number of rules will provide the most information on being processed. A new measure is introduced which calibrates the certainty in a probabilistic rule of entailment. As a result of this development a new way of viewing the linguistic hedges of Fuzzy Logic is introduced.

Chapter 9: The Use of Heuristics

In an attempt to deal with real world situations in which either not enough information is available to completely specify a probability model, or where a model is too complex to go through the time consuming process of assigning every required conditional probability for a complete probability model, chapter 9 presents various simplification heuristics. Ultimately, these strategies can be used in situations where as little information as that provided in Nilsson's initial model is provided. The problems of knowledge interleaving are examined.

Chapter 10: An Application of Probabilistic Logic in Vision

A system of Probabilistic Logic based on the the results of chapters 4, 5 and 9 has been developed. The new system was succesfully applied to a problem in two dimensional vision, and chapter 10 shows how the system performed.

1.5 Appendices

Appendix A has a list of all the typographical conventions used throughout the thesis to denote logical and probabilistic elements. Also in this appendix is a brief description of the interpretation tables and maximum entropy equations used throughout. It may be found useful to flip back and forth to this section whenever any question of notation arises.

Appendix B is a very brief resumé of the mathematical and statistical prerequisites for reading this thesis. Appendix C is an extension of chapter 5, which shows the proof of the

derivation of aggregates in an entailment rule whose probability is 1. Appendix D is an extension of chapter 9 which compares the effect of two heuristics on a set of rules.

1.6 Scope of the Thesis

The thesis combines two great interests of mine: automated deduction and reasoning with uncertainty. The simple aim of the thesis is to show that the two approaches of probability theory and predicate calculus can be mutually consistent with each other within Nilsson's probabilistic logic. The ultimate use of this will be as a means of reasoning with uncertainty, in a consistent and semantically justifiable way, in expert systems.

1.7 Declaration

The work presented in this dissertation was carried out by myself, except where due acknowledgement is made. This thesis has not been submitted to this, or any other, university. However, material taken from this thesis has been, or will be, published as follows:

- "Reasoning with Maximum Entropy in Nilsson's Probabilistic Logic", which was presented at the International Joint Conference on Artificial Intelligence in 1989 [66].
- "Enhancing the Inference Mechanism of Nilsson's Probabilistic Logic" which appeared in "The International Journal of Intelligent Systems" [65].
- "Reasoning with Maximum Entropy in Expert Systems" which is the text of an invited talk presented to the "10th International Workshop on Maximum Entropy and Bayesian Methods" in Wyoming 1990, published in [67].
- "Model Based Object Recognition using Maximum Entropy and Nilsson's Probabilistic Logic" which was cowritten with Patrick McAndrew and Andrew Wallace of the Vision department of Heriot-Watt Computer Science department and appeared in "The International Journal of Pattern Recognition and Artificial Intelligence" in 1991 [68].

Chapter 2

MATHEMATICAL TOOLS FOR REASONING WITH UNCERTAINTY

2.1 Introduction

Present day methods for reasoning with uncertainty have grown from two major domains. The first is mathematical inference techniques. The second is automatic deduction techniques for the digital computer. In mathematics, the most important step forward for automated reasoning came with the publication of Whitehead and Russell's book "Principia Mathematica", in which rules for reasoning alongside the implication rule, were constructed for organising the knowledge of mathematics in a set theoretic formalism. Another development in twentieth century logical systems is the systemised notion of a "possible world", as opposed to the real world. Theories for dealing with possible worlds are the youngest of all those mentioned thus far, and are reported here insofar as they are relevant to the subject of reasoning with logic and probabilities.

However, semi-decidability in predicate calculus makes it mathematically impossible to implement a calculus for reasoning with logical possible worlds with the property that all of the possible worlds for an uncertain situation can be enumerated. And so, although the work

of the late nineteenth and early twentieth century philosophers laid foundations for reasoning with uncertainty in mathematics, it also discovered that reasoning with uncertainties in mathematics has clearly visible and insurmountable obstacles.

A history is given of the evolution of reasoning with uncertainty from the time of Jakob Bernoulli (1713), through to the ideas of the nineteenth and twentieth century mathematicians and statisticians. These three hundred years have been contentious years for the subject, and have yielded three complementary definitions for the meaning of probability: the range theory of probability, the frequency theory, and the subjectivist theory. Each of these definitions offer different perspectives on the meaning of reasoning with uncertainty. Their similarities and differences are explored in this chapter in a mathematical context. The results of this investigation reveal the probabilistic problems facing the set-theoretic mechanisms.

2.2 Uncertainty and Mathematical Logic

The idea of using a formal system to reason about propositions goes back to Aristotle in the 7th century BC, whose collected works, the "Organon", introduced logic as a tool for sharpening thought. He enumerated fourteen forms which a correct argument could take, and a further five were added by his pupil Ariston. These forms were called "syllogisms" [106]. An example of a syllogism would be:

All Fridays are pay days
Today is Friday

Therefore, today is pay day

Syllogisms can be read from top to bottom. A syllogism is a group of three logical sentences, two antecedents and one conclusion. The line separating the conclusion from the antecedents indicates the process of reasoning.

In the nineteenth century, George Boole developed what is now known as Boolean algebra: a calculus for manipulating variables representing TRUE and FALSE. Frege [44] in his paper "Begriffsschrift" combined the propositional logic formalisms with those of Boolean algebra. He introduced the notion of a language proposition, and the law of "modus ponens" (the method of bridges). For this he needed a definition of sentences which can express facts in the predicate calculus. His definition is used throughout the text and is given below:

Definition 2.1 *A sentence in predicate calculus is a well formed formula, constructed from atomic facts; or facts joined by the logical connectives* and *(&)*, or *(∨)*, *and* not *(∼)*.

The rule of modus ponens, although not strictly necessary since it can be stated in terms of connectivity primitives was also introduced as a symbol of convenience to denote the process of entailment at work in the syllogism. The rule of entailment is also used throughout this text, and its meaning is defined below:

Definition 2.2 *The symbol to denote entailment, (or material implication), is ⇒, and it is used between logical sentences in the form A⇒B, which may be read if sentence A is true, then conclude B.*

In this scenario, A⇒B is an instance of the rule of entailment, the sentence A is the rule's *antecedent* and the sentence B is the rule's *consequent.* Syllogisms can now be written:

$$\frac{\begin{array}{l}A \Rightarrow B \\ A\end{array}}{\text{Therefore, B}}$$

In classical logic the rule of "modus ponens" allows us to use the proposition set (A, $A \Rightarrow$B) to entail proposition B. That is, if A is true, and $A \Rightarrow$B is true, then B must also be true. The truth-table for implication, first given by Russell and Whitehead [107], table 2.1, shows the four possible labellings of truth values to the predicates A, B and $A \Rightarrow$B.

A	B	$A \Rightarrow$B
t	t	t
t	f	f
f	t	t
f	f	t

Table 2.1: Truth Table for Implication

For the purposes of this thesis a familiarity with the concepts of predicate calculus [46] is assumed, so that the problems may be explored in some depth.

Although English sentences can be expressed in terms of formal logic, and although the proof procedure allows the implementation of the entailment rule, there have been misgivings about the semantics involved with the rule of entailment. There have been developed what are now known as "the paradoxes" of material implication [107, 58]:

1. $(p\& \sim p) \Rightarrow q$. That is, from any proposition of the form $p\& \sim p$ any proposition whatsoever can be deduced.

2. $q \Rightarrow (p \vee \sim p)$. For any proposition, any other proposition of the form $(p \vee \sim p)$ can be deduced.

3. $\sim p \Rightarrow (p \Rightarrow q)$. From any false proposition, any proposition whatsoever can be deduced.

4. $q \Rightarrow (p \Rightarrow q)$. That is, every true proposition can be deduced from any proposition whatsoever.

These "paradoxes" derive from the implication table, which equates the meaning of $p \Rightarrow q$ with $\sim p \vee q$ whereby, when an antecedent is false and the rule true, the consequent can be either true or false. This puzzling relation has led logicians to reconsider the meaning of deducability.

The deducability of q from p can be said to only mean that it is logically impossible for p to be true, and q to be false. However, as pointed out in [58], "No one is likely to deny that the logical impossibility of $(p\& \sim q)$ is a "necessary" condition of q's deducability from p, but it has been suggested that it is not a "sufficient" condition on the ground that there should be some connection of 'content' or 'meaning' between p and q."

This semantic problem aside, the logic of propositions and entailment has given rise to a calculus of certain belief management which allows deductions based on the truth of antecedents to be manipulated in a correct and verifiable way. These misgivings about the nature of entailment have relevance only when the antecedents are false. This particular problem reappears [89, 66], when we consider the set-theoretic mechanism for generalising classical logic.

The introduction of a rule of inference over such rules, made computational reasoning a possibility; and heralded the way for knowledge-based and expert systems.

The Set Theoretic Limits of Automatic Deduction

The work of Whitehead and Russell [107] was intended to show that all of mathematics was an elaboration of the laws of logic [27]. They assumed that Set Theory could be used to represent the laws of logic, and set about representing the tools of mathematics in this new notation, using axioms (the basic theorems) and rules for producing new theorems from old ones. Such rules were formulated using logical implication (or entailment). The motivation for this development was the development of a procedure which would discover the truth or falsity of any logical sentence by finding either that it was an axiom (default theorems), or that it could be produced from axioms by repeated application of production rules on earlier theorems [56, 55].

This introduced a strict mathematical concept of provability into the science of automated deduction. For some sentences, such a procedure is guaranteed to find proof either of the sentences' truth or falsity. For these sentences, the question of logical implication is decidable [46]. This property is not true for all sentences however, and there exist sentences for which neither the sentence itself nor its negation can be produced by the above procedure. That is, neither the sentence nor its negation are implied by the axioms and/or repeated application of the rules of inference. For such sentences, the above procedure will never terminate, and for this reason, logical implication is only semi-decidable [56].

Modal Logics

Modal Logic allows for reasoning with uncertainty in mathematics, where uncertainty about the truth or falsity of a proposition means more than one possible scenario has to be considered in the reasoning process [58, 45]. Aristotle introduced the notion of "modal logic"; which extends properties of propositions to include necessity, contingency, possibility and impossibility, as opposed to just true or false. So that, true propositions can be divided into two categories: those that are necessarily true, and those which are true by contingency.

The world in which we live may be called the real world. In this world many things are true, some of which we know about, others we do not know about. We conjecture about

the nature of uncertain things by introducing possible worlds with sentences such as "if some event was actually true in the real world, what would be the consequences of this". Possible worlds fall into several different categories:

1. A logically possible world might be defined as one which conforms to the rules of logic. A world where "Peter is a boy" and "Peter is a girl" is not logically conceivable, and is therefore an "impossible" world.

2. A physically possible world might be defined as a world which has the same physical characteristics as the real world- so that in such a world, the speed of sound in air is restricted to 330 metres per second.

3. A conceivable possible world is a world which could be imagined. For example, a world where everyone had free access to public transport.

4. A temporally possible world is one which we could imagine developing on from the present world over a period of time. For example, we could imagine that in five years time the Russians will have landed on Mars.

All of the above listed examples of possible worlds require a few common characteristics from any calculus which intends to reason over them:

(i): Each possible world has to be dealt with separately from any others.

(ii): In each world there is a strict set of assumptions which, when contravened, would make a world into an impossible world.

(iii): In each world there are rules which allow worlds to be developed, or examined in greater detail.

In possible worlds, we require only that the rules of deduction of ordinary predicate (propositional) logic apply. For example, the proposition "2 + 2 = 4" is true from logical necessity; and the proposition "Britain is not a member of the European Monetary System" is only contingently true. Similarly, false sentences can be split into two groups: those which are false by logical necessity, and those which are contingently false. (For example, "2 + 3 = 4" and "Russia has a democratic system of government").

Resolution Refutation Systems

Robinson [105], in 1965, introduced the resolution principle: a mechanical procedure to perform inference. The resolution principle is applied in resolution refutation systems. In a typical theorem proving problem there is a set S of well-formed-formulas from which we wish to prove a goal formula w. The first step is to negate the goal and add this negation to S. The expanded set is then converted to a set of clauses, and we use the resolution principle in an attempt to derive a contradiction, that is the occurrence of a fact and its negation, which is signified when the empty clause, NIL, is produced. Robinson showed that if the resolution principle is applied to an unsatisfiable set of clauses then NIL can always be produced eventually.

The Development of a Probability Calculus

There were many conflicting views among philosophers, mathematicians and statisticians, as to what may be the precise meaning of probability. The reasons for this may be attributed to the many branches of science and commerce in which the concept of probability has emerged [60, 26, 73]. Commercial insurance against risks which was practiced as early as the fifteenth century; the practice of life insurance; the theory of mathematical games of chance; and the combination of judicial evidence all developed their own concepts of probability.

In statistics the concept of probability developed from studies of games of chance. This subject was developed in the seventeenth and eighteenth centuries into a "geometry of the die" by Pascal, a theory of event combinations by Fermat and finally Jakob Bernoulli's range theory of probability [60].

2.3 Rules of the Calculus of Probability

The four basic rules which any calculus of probability must satisfy were formulated by early twentieth century statisticians [69, 61]. These rules can be regarded as the basic building blocks for reasoning with uncertainty which have been arrived at over much argument and

debate amongst the great statistical mathematicians from the time of Bernoulli.

Definition 2.3 *In all the statistical rules throughout this text the expression p(a|h) is to be read as the probability of a being true given that h is true which is defined to be p(a&h)/p(h) [38].*

We can consider a to be a proposition, and h to be some data or evidence which is relevant to the occurrence or non-occurrence of a. For example, one might be interested to know the probability that it will rain, r, given that the sky has become overcast with dark clouds (sodc), that is, p(r | sodc).

The meaning associated with the word probability has been considerably developed over hundreds of years. Essentially, there are three probability interpretations:

1. the range theory of probability;

2. the frequency theory of probability;

3. the subjective, or, belief theory of probability.

In all of these theories the probability of the occurrence of an event is a number which must satisfy the following four axioms:

$$p(a|h) \geq 0 \tag{2.1}$$

$$p(h|h) = 1 \tag{2.2}$$

$$p(a|h) + p(\sim a|h) = 1 \tag{2.3}$$

$$p(a\&b|h) = p(a|h)p(b|h\&a) \tag{2.4}$$

From equations 2.1, 2.2 and 2.3 it follows that all probability values are within the range 0 to 1. From equations 2.3 and 2.4 comes the addition principle:

$$p(a \vee b|h) = p(a) + p(b) - p(a\&b). \tag{2.5}$$

If a and b are mutually exclusive (that is, they cannot both be true at the same time), then p($a\&b$) is zero, and the addition principle becomes simplified to:

$$p(a \vee b|h) \Rightarrow p(a) + p(b)$$

the special addition principle.

If $p(a|h) = p(a|h\&b)$ then a and b are probabilistically independent. That is, the probability of a given that h and b have happened is in no way different from the expected probability of a given that only h has happened.

2.4 The Range theory of Probability

Jakob Bernoulli can be regarded as the founder of probability theory as a branch of mathematics. His posthumously published "Ars Conjectandi" [7] formed a bridge between the "a priori" methods of combinatory probability, (which by definition required only knowledge of the gaming situation, and no evidence for validation or support), and the early "a posteriori" methods of statistical theory (which deduced results from supportive evidence). This early work was based on a range theory of probability.

In its simplest form, if we consider h to be a hypothesis, then we must break h down into a number N of alternative conditions. That h is fulfilled (or true) means that one of $h_1, h_2, ...h_N$ is fulfilled. Some of these alternatives, say M, entail the occurrence of a; the remaining ones entail the occurrence of $\sim a$. The probability of a given h ($p(a|h)$) is the ratio M/N.

The mutually exclusive alternatives, $(h_1, h_2, ...h_N)$, covered by a proposition are what is called its range. The range theory of probability satisfies the four axioms 2.1 to 2.4, and defines the probability of a given h as the measure of the range of h-and-a divided by the measure of h alone.

The Principle of Insufficient Reason

The main difficulty confronting this range theory of probability concerns measurement of the ranges, and effectively the numbers M and N. Bernoulli stressed that the alternatives into which h is to be analysed ought to be equally possible; this rule is called the principle of insufficient reason [7], or in Keyne's terminology the principle of indifference [69].

Reliance on the principle of insufficient reason for measuring probabilities in a range has certain attractions in games of chance, where there is usually agreement amongst experts as

regards gaming situations where alternatives are equally possible. Bernoulli himself noted that the inventors of games of chance "took pains to set up so that the numbers of cases would be known and- so that all these cases could happen with equal ease" [7].

With these two tools at his disposal, Bernoulli then turned his attention to the relationship between frequency data drawn from many sample runs, and the definition of probability as given above. To this end, he developed the binomial theorem, which may be stated as follows. If an experiment of two possible outcomes has the same probability p, (i.e. M/N), of success from trial to trial, then the probability of seeing m successes in n trials is:

$$p(m|n,p) = \binom{n}{m} p^m (1-p)^{n-m} \tag{2.6}$$

Bernoulli then showed that as the number of trials $n \to \infty$, the observed frequency f=m/n of successes tends to the probability p. With this expression Bernoulli became the first mathematician able to relate the result of a random experiment (m/n) within a mathematical model to the absolutely perfect results which were to be expected (M/N).

Reasoning with the Range Theory: Bayes' Theorem

In the problems considered in probability theory, the population numbers concerning the range of the hypothesis, and the range of the event are known, (i.e. N and M respectively). For this theorem to be of practical use to the theoretical statistician, an inversion of the originally stated goals had to be solved. The inversion problem may be stated thus, given that the sample is known (that is, n the number of trials, and m the succesfull trials), but the population is unknown, how can we predict the population numbers M and N, and hence the true probability of an event, with accuracy.

It is likely that in many trials the observed frequency f will be close to the true probability p. But the question was, how to describe this process in a precise mathematical theorem. That is, the binomial law gives the probability of m, given (M,N,n), so how can we derive from this a formula for the probability of M given (m,N,n)?

Thomas Bayes [5], in 1763, provided the first example of an inversion of the binomial theorem. He developed a theory to answer questions such as: "Suppose a solid or die of

whose number of sides and constitution we know nothing; and that we are to judge of these [the number of sides and the probability of each number showing at any given throw] from experiments made in throwing it." His result states that given the sample data (m,n), he finds that M/N lies in the interval:

$$p < M/N < p + dp \tag{2.7}$$

where:

$$p(dp|m,n) = \frac{(n+1)!}{m!(n-m)!} p^m (1-p)^{n-m} dp \tag{2.8}$$

Pierre Laplace [77, 75], in 1774, developed the theory of inverse probabilities in greater generality. In Laplace's terminology, we let H stand for some observable event which can be analysed into C_1, C_2, ..., C_n mutually exclusive and exhaustive causes. If the causes C_i, as in the principle of insufficient reason, are considered equally likely, then having seen event H to be true, the posterior probabilities of the C_i are:

$$p(C_i|H) = \frac{p(H|C_i)}{\sum_{j=1}^{N} p(H|C_j)} \tag{2.9}$$

It is therefore necessary to know the probabilities $p(H|C_i)$ for each of the causes. If the C_i are not considered equally likely, but have prior probabilities $p(C_i|I)$, where I denotes prior information, then the result becomes:

$$p(C_i|H) = \frac{p(H|C_i)p(C_i|I)}{\sum_{j=1}^{N} p(H|C_j)p(C_j|I)} \tag{2.10}$$

In this case, we also need to know values for $p(C_i|I)$, the conditional probabilities of the causes in the light of whatever other evidence comes to light. This is known as "Bayes' Theorem" [60, 24, 125], and, as can be seen from equations 2.9 and 2.10, it requires a lot of precise information. In fact, if there are x causes, the complete system requires 2^x conditional probabilities $p(H|C_i)$ to represent all possible states of presence and absence of the x causes taken together. The problem is compounded when the conditional probabilities $p(C_i|I)$ have to be supplied for every possible type of prior information having a bearing on causes C_i.

Laplace's best results came from 2.9. His failure to use equation 2.10 to any great effect, and the fact that he did not provide a detailed description of the derivation of the two results 2.9 and 2.10 left the theory of inverse probabilities open to criticisms of not being well-founded.

Statisticians gradually turned towards a definition of probability as something other than a function over ranges. This developmental period led to the frequency theory of probability, and the beginnings of sampling theory.

2.5 The Frequency Theory of Probability

Early proponents of the frequency theory spoke of probability only as a relative frequency "in the long run". This is the view that $p(E|H)$ means the relative frequency with which the event E takes place when condition H is fulfilled. That is, the probability of event E given H is the proportion of H situations which lead to E events. John Venn, was first to develop a mathematical theory for the frequency theory of probability [127]. Venn defined an event's probability as the limiting value which it's relative frequency approaches as the number of occasions of observation are indefinitely increased.

The German mathematician Richard Von Mises [85, 86] further developed the concept by adding a qualification of randomness in the events being measured. He gave an example of a traveller walking along a road on which milestones are placed: large ones at whole miles, small ones at every tenth of a mile (including the whole miles). Von Mises reasoned that a probability could not simply be the limiting value of a relative frequency. "If we walk long enough along this road, calculating the relative frequencies of large stones, the value found in this way will lie around 1/10... the deviations from the value 0.1 will become smaller and smaller as the number of stones passed increases; in other words, the relative frequency tends towards the limiting value of 0.1 This result may induce us to speak of a certain 'probability of encountering a large stone'."

To prevent such a scenario he introduced the idea of a collective: "such sequences of events or observations, which satisfy the requirements of complete lawlessness or 'randomness'". The traveller's scenario is ruled out of being a collective because there is a way of selecting the stones which would cause a fundamental change in the relative frequencies. That is, starting at a whole mile and register every second marker passed, the relative frequencies converge towards 1/5 instead of 1/10. His demand of randomness in the recording of the event he

called "the principle of the impossibility of a gambling system". This extended Venn's work by defining two properties possessed by a collective appropriate for the application of probability theory:

1. The relative frequencies must possess limiting values (as with Venn's definition)

2. These limiting values must remain the same in all partial sequences which may be selected from the original one in an arbitrary way

With this qualification, Von Mises introduced a considerable difficulty into the theory of probability, namely, the problem of how to determine a truly random distribution.

Reasoning with the Frequency Theory

Many problems in statistical inferencing lead to repetitions of experiments having two possible outcomes. For example, "Is the readership of the Sun newspaper equally split between males and females?" In this example, the population which is described by the readership of the Sun newspaper can be tested by asking the simple question: "Are you female?" And recording the answers (yes, or no). Our interest will be in the proportion (p) of responses to the affirmative, divided by the size of the population.

In realistically difficult statistical questions, such as this one, it is not generally possible to collect data from the total population of interest; and so samples from that population are tested instead. And so, reasoning with the frequency theory of probability is called the theory of sampling.

In the example given above, we wish to test the hypothesis that 50% of the readership of the Sun newspaper is female, i.e. $p=1/2$. This is a statistical hypothesis. Let x be the number of female Sun readers in the sample, and n be the number of Sun readers in the sample. If x/n is close to 1/2, the hypothesis gains credence from our sample; whereas, if x/n is continually far from 1/2, we will begin to doubt that $p=1/2$. However, even when $p=1/2$, fluctuations in the random sample could produce a value of x/n far from 1/2. As n becomes larger however, these fluctuations become less and less likely, and eventually, the fluctuations

should be lost as the true value for p is more and more closely approximated by x/n. This behaviour is known as the strong law of large numbers.

In hypothesis testing, the assumption "p=1/2" is known as the null-hypothesis. The number of women, x, in a population of size n, has a Binomial distribution, which may be approximated by a Guassian distribution, whose mean is np, and whose variance is np(1-p), for the null hypothesis.

If there is a 50:50 split female:male in the population then the chance that x will differ from n/2, given by the Binomial theorem, by more than $1.96\sqrt{n/4}$ is 0.05. This means that a number recorded outside this range makes our null hypothesis very probably, (probability ≥ 0.95), wrong. For example, in a sample of size 50, the probability that x is outside the interval (18, 32) is 0.05; and so, any number of successes in the ranges 0-18, 32-50 make the null hypothesis very unlikely.

Problems with Sampling Theory

Choosing a sample to be tested from a population is open to human bias and the results from an investigation lose credence if a sample is improperly chosen. For example, if we were to investigate the null hypothesis on a sample drawn from an all women hospital or in a male dominated work setting. To ensure that the sample is a fair representation of the population, the sample must be chosen with respect to the principle of excluded gambling systems [85].

The problems involved in selecting an unbiased sample reflective of the nature of the population led to the development of stratified sampling, where certain possible samples are eliminated from those possible; cluster sampling, where certain combinations of individuals are grouped into the sample; and multi-stage sampling methods which allow for combinations of cluster and stratified sampling [123].

Coupled with the frequency definition of probability, was the development, in the early twentieth century, of techniques for inference using this notion of probability [98]. This resulted in the production of the Chi-squared test, the principle of maximum likelihood, unbiased and or efficient estimators, confidence intervals, fiducial distributions, conditioning

on ancillary statistics, power functions and sequential methods for hypothesis testing. All of these methods test how well sample data conforms with model predictions.

None of the sampling models take known prior information into account, and the place of such evidence in the reasoning process is replaced by hypothesising the existence of "nuisance" parameters. A nuisance parameter is defined as a parameter which "is included in a probability model for an experiment because it is necessary for the good fit of the model, but that is not of prime interest to the investigator" [73].

Although there are misgivings about the frequency theory, it prevailed strongly throughout the nineteenth and well into the twentieth century. However, Cox [25] points out that pure significance tests, while useful, are of limited importance, particularly because they give no idea of the magnitude of possible departure of the results in the sample from the null hypothesis.

2.6 The Subjective Theory of Probability

Many philosophers and statisticians have spoken of probability as a measure of belief or certainty. Ramsey [104] and De Finetti [40] made the first attempts to systematise the notion of partial belief into a framework of mathematical probability obeying the four laws required of a statistical calculus.

A person's measure of belief in a concept may be measured by proposing a bet as to the truth of the event, and observing the lowest odds that will be accepted. The odds of an event are defined to be the ratio of the probability of the event, and the probability that the event will not happen. A probability can be converted to an odds ratio by using the formula:

$$ODDS = \frac{PROBABILITY}{1 - PROBABILITY} \tag{2.11}$$

For example, if an expert meteorologist is quite sure it is going to rain, we would ask if there was a 4 to 1 chance, or an evens chance, and so on; until the expert finally settles on the odds most appropriate to quantifying the belief. Odds given in this way can be transformed back into probabilities with the following equation:

$$PROBABILITY = \frac{ODDS}{ODDS + 1} \tag{2.12}$$

where a 4 to 1 shot has odds 4, and a 1 to 4 shot has odds of 0.25. Ramsey pointed out that a distribution of partial beliefs contrary to the four laws of the abstract calculus of probability, would be inconsistent in the sense that it would "violate the laws of preference between options". The ideas of Ramsey and De Finetti were further developed by Savage [110], who became the founder of a "subjectivist" or "personalist" school in probability and statistics, and Cox [25].

The Re-emergence of Bayesian Statistics

Jeffreys reintroduced Bayesian Inference to statistics [61] and criticisms of the well foundedness of the range theory of probability (and consequently Bayesian Inference methods) were overcome when Cox [25] answered the question: "is it possible to construct a consistent set of mathematical rules for carrying out plausible, rather than deductive reasoning?" [60]. He found that if degrees of plausibility are represented by real numbers, then the conditions of consistency can be stated in the form of functional equations, whose general solutions can be found. He then defined probability as the scale over which the degrees of plausibility were defined. He deduced the only consistent set of rules of combination to be:

$$p(A\&B|C) = p(A|B\&C)p(B|C) \tag{2.13}$$

$$p(A|B) \; + \; p(\sim A|B) = 1 \tag{2.14}$$

These two equations are precisely those of 2.3 and 2.4; and, with the addition of the equations 2.1 and 2.2, a calculus of probability may be defined within which the theory of inverse probabilities is consistent.

2.7 Probability: Subjective or Objective?

Two schools of thought have emerged as to the conception of probability as a degree of belief, on the part of the subjectivists; and as a relative frequency, or as a theory of ranges on the part of the objectivists. There is however a great overlap between the two methods.

Supporters of the frequency view have found that an adequate analysis of the probability of an event, requires that the event being tested be randomly distributed through the sampling data. Supporters of the range theory of probability have needed the principle of indifference for determining certain situations equipossible in certain fundamental alternatives. The question is asked whether the knowledge of randomness or equipossibility are not forms of knowledge forced onto an inference model. If the answer to this is no, then it is not possible to think of the objectivist models as being able to exist without recourse to subjective knowledge of the sampler, thus bolstering up the case for probability as a subjective entity.

On the other hand, subjective probabilities must be based on knowledge acquired in some knowledge elicitation process. If the sources of this knowledge are to be considered reliable, then the information they provide must objectively describe the nature of events in the domain of interest. In this instance, it is not possible to think of the subjectivist model being supportable without recourse to the objectivist model.

Also, although there seems to be a great conflict of opinion between those championing a frequency theory of probability, and those championing a range theory of probability; in chapter 2 it becomes clear that both lines of development have contributed to the development of set theoretic mechanisms.

2.8 Conclusion

Any attempt to integrate the two branches of statistics and mathematics faces major problems of complexity. The mathematical development of the predicate calculus has bequeathed us modal logic, an efficient means of performing inference (Robinson's resolution theorem) and the problem of semi-decidability in the first-order predicate calculus (an unfortunate setback). The statistical development of probability has (via Cox) produced a sound mathematical basis for Laplace's "Bayes' Theorem" and given us a system which is precise, voracious in its appetite for data and not very flexible.

An understanding of these problems places us in a position to see why simplification methods, such as those discussed in chapter 3, have had to be used in the early expert

systems, and why ultimately, any system of reasoning both with mathematical logic and probability theory must also employ simplification strategies of one sort or another in an attempt to reduce the complexity problem.

Chapter 3

AUTOMATED SYSTEMS FOR REASONING WITH UNCERTAINTY

3.1 Introduction

One of the major requirements of an automatic system for reasoning with uncertainty is that it should demonstrably perform its reasoning task within a reasonable amount of time. Thus, the early methods for reasoning with uncertainty made various simplifications to the reasoning process in order to make it tractable. On the whole, these techniques of simplification have been a good movement within artificial intelligence in that they have opened up the field of expert system technology, and demonstrated a crucial role for subjective probability estimates.

However, the simplification strategies operated by these early systems lead, in the long run, to continued and sustained errors of judgement on the part of the reasoning process. So much so, that criticisms can be made against almost all of the present day methods for reasoning with uncertainty. This chapter will show how the simple methods of reasoning with uncertainty have developed into the complex methods of reasoning with uncertainty which honestly address the problems discussed in the previous chapter.

One interesting dynamic in the development of good methods of reasoning with uncer-

tainty from poorer methods, is the reintroduction of complexity problems. Methods for reasoning with uncertainty using set-theoretic inference mechanisms, have been proposed by Cheeseman [16], Bundy [13], Nilsson [89], and Pearl [95]. Each of the authors were dissatisfied with current methods of reasoning which are non-set theoretic. These dissatisfactions can only be viewed with respect to the major inferencing mechanisms of today. This chapter examines the development of these mechanisms and illustrates their shortcomings.

3.2 Uncertainty in Artificial Intelligence

With the development of the computer in the Second World War, mechanised thinking processes started to seem plausible. Alan Turing, in asking the question "can machines think?" [126] started the search for "artificial" intelligence in a computer system. the quest for developing this "thinking" aspect of computation became topical. The first major breakthroughs in this area were in reasoning using first order calculii ([105]), and in general problem solving using domain independent generate-and-test techniques ([87]).

Robinson provided a computational procedure for performing reasoning with implication rules. Newell, Simon and Shaw provided a system which could solve a family of problems in a general way with recourse only to simple deductive techniques. This system made use of the resolution principle, and was called the General Problem Solver.

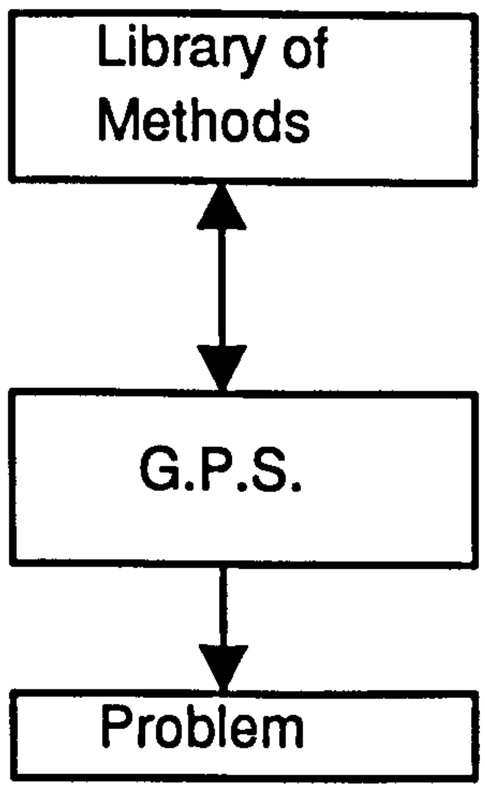


Figure 3.1: General Problem Solver

The authors of GPS believed that a collection of problem solving methods could be

grouped together in a library, and a way to increase the intelligence of the GPS would be to continually add more and more new rules. This turned out not to be the case. In fact, as more and more rules were added, the computation of solutions took longer, and, the reasoning process became so diverse that in some cases, results of relatively easy problems could not be reported in real time. The reasoning process of GPS lacked specific direction.

The introduction of the program DENDRAL [11], altered the way such research was considered. At Stanford University, the Professor of Genetics devised a program to enumerate all possible legal configurations of atoms in a molecule from a chemical formula. The program was then further refined in an attempt to identify molecular compounds from analatical data from a mass spectrograph. The work produced a program to solve a difficult analytical task, using highly domain-dependent knowledge. This was the beginning of the shift from domain independent solution methods, to domain dependent methods [36], and culminated in the development of Knowledge Based Systems, and Expert Systems.

Expert systems are designed to deal with problems whose specifications, and solution methods are complex. Such complex problems usually offer many possible solutions. Expert systems specifically are intended to function as an expert would when presented with a problem in the expert's area of expertise.

The background to this subject is epistemic knowledge [72]. A knowledge based system [45] is a system which is able to manipulate "knowledge" in order to perform a given task. Expert systems, being a sub-area of knowledge based systems, are used in areas where an expert may be found. Expert knowledge is distilled from an expert's experience, and structured symbolically in a computational formalism in such a way as to model relations between data elements in the expert's domain the way the expert would.

In DENDRAL [11], the molecular structure of organic compounds was deduced from the results of a mass spectrograph reading, using expert knowledge on how to interpret the lines. DENDRAL is very much used by chemists, and results from the program have been cited in many academic papers. However, quantified uncertainty as such, was not needed in the reasoning mechanism. The rules were the straightforward if-then rules of propositional and predicate calculus. DENDRAL exhibits causal reasoning procedures whose motivations were

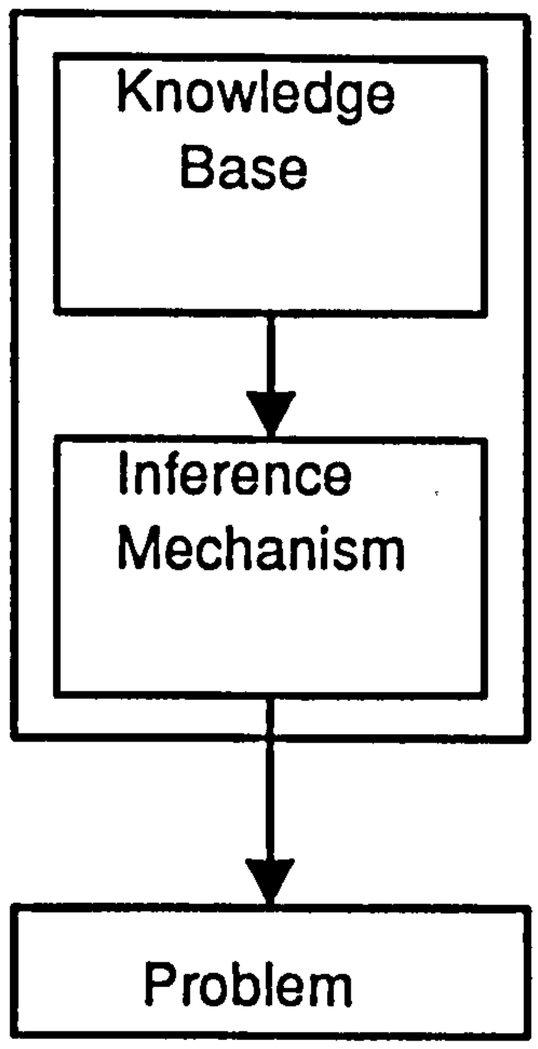


Figure 3.2: Standard Knowledge Based System

provided by the expert encoding the rules. and as such performs the task of an expert very well. However, since there are experts who must, in the course of applying their expertise, deal with uncertainties in reaching towards conclusions, it became inevitable that mechanisms for representing and manipulating uncertainties must also be found. This lead towards methods for "inexact", or "plausible" reasoning.

3.3 The Purely Bayesian Approach

Expert systems were first used most successfully in the field of medical diagnosis where they superceded the use of purely Bayesian methods of reasoning. To understand why this progression took place, consider the very succesfull Bayesian model produced by the designers of the ARF system which was a program to diagnose one of 14 diseases causing acute renal failure [49].

The differentiation among these 14 possibilities was carried out using 31 clinical parameters. Each parameter had approximately three to four possible values, so the sample space of findings was approximately 100. Data tables indicating the prior probabilities of the hypotheses and the conditional probabilities of diseases, given various symptomatic findings, were used by the program to interactively query the user as the reasoning process progressed.

The algorithm is as follows:

1. Construct a vector of prior probabilities for the 14 possible hypotheses.

2. Using Bayes' Theorem, reevaluate the hypotheses based on given information.

3. If any probability reaches a previously defined threshold value, (e.g. 95%), stop the investigative process, and report results.

4. Identify the finding with maximum information content, from entropy considerations. Ask about the finding with maximum expected information content.

5. Go back to step (2).

This procedure produced impressive results when applied in several medical application domains; was economical and directed in its attempt to reason; and was able to arrive at the same clinical diagnoses as experts over 33 cases on which it was tested [50].

The creators of ARF, when looking to further develop the system, turned to techniques of Artificial Intelligence. The reasons for this switch can be listed as follows:

- Bayesian Inference demands a lot of information before it will form an opinion. (750 conditional probability estimates were needed by ARF to discriminate between the 14 causes of acute renal failure.)

- For reasons of implementational simplicity, the list of diseases were considered to be mutually exclusive, and exhaustive — a condition not typical in medicine where a patient may have many correlated illnesses.

- The findings in the program were considered to be conditionally independent, that is, that the probability of a patient having a particular symptom is conditioned only upon the present disease under investigation, and not on the other findings already made. For example, in ARF it is not possible to correlate nausea and vomitting.

- The database of the program is conditioned on the patient population from which it has been derived; and so extending the program with results from a new population, or porting the program to a new work location degrades the program's critical performance.

- The entire repertoire of hypotheses known to the program had to be updated each time a new finding is reported. In internal medicine where there may be thousands of hypotheses, this requirement becomes impractical, and wasteful, given that in an expert situation, an expert clinician is, typically, dealing with only a small number of hypotheses (no more than five or six) [92] whereas, a non-expert is typically entertaining many more.

- In choosing the information to expand into questioning, the program must think ahead, by evaluating probability distributions of expected answers in the light of the entropy changes liable to be induced. This requirement demands heavy usage of computational resources.

3.4 Fuzzy Logic

Uncertainty which is described by people is necessarily couched in a linguistic formalism, corresponding to the language in which the uncertainty is expressed. Probability may be regarded as a language for coping with uncertainty; however natural language itself also provides tools for making subtle distinctions between things.

Winograd's natural language system SHRDLU [132] used a natural language interface to manipulate blocks on a tabletop. As defined, all of the objects were equally distinct; so that a blue-green block was no closer to being a blue block, than a yellow block. Fuzzy logic over Fuzzy set-theory attempts to overcome this kind of inability [135, 34].

Fuzzy set theory [134], allows an element of the language to be a member of any of a number of sets, and assigns a number (ranging from 0 to 1) to the elements of each set which describes the objects level of membership within each set. This number is called a membership value.

For example, the shape of the object shown in figure 3.3 could be fuzzy described as: [(triangular, 0.6) (oval, 0.3) (round, 0.1) (square 0)], where the numbers measure a goodness of fit of the object to each of the possible shapes.

As another example, the colour of an object may be described by one person as 'green',

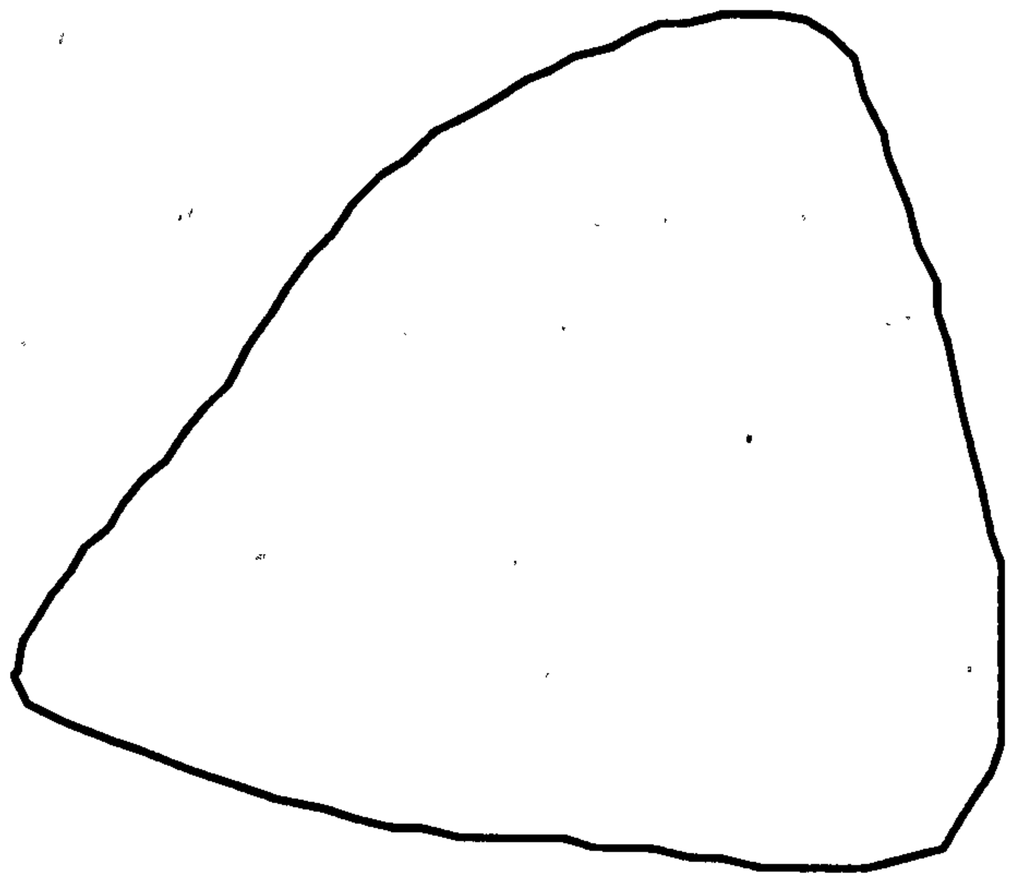


Figure 3.3: A Fuzzy Object

and by another as 'a sort of blue-green'. Internally, the colour of it may then be described: [Colour(Green 0.75) Colour(Blue 0.25)]; or alternatively if a colour is stored in terms of its wavelengths in nanometres, and we know that the wavelengths of green and blue are 445-490 and 500-575 respectively, then the colour 'sort of blue-green' may be represented with a wavelength 495 nanometres. So a linguistic function 'sort-of' can be predefined for each of the colours.

In the above example, 'sort of' is used to demonstrate the concept of a 'linguistic hedge': a linguistic function which causes some amount of shift of an object's membership value within a group, in such a way as to accord with human expectations. Other such hedges are: 'very', 'nearly', 'rather', 'too', 'almost', etcetera.

The truth or falsity of a proposition is not a certain event, but is rather modelled on a standard Gaussian curve [38] which measures possibility. An uncertain event is given a point on the curve and is assigned a possibility number in the region [0,1]. Fuzzy Set theory offers three rules for dealing with conjunctions, disjunctions and negations of uncertain propositions:

$$p(A\&B) = minimum(p(A), p(B)) \tag{3.1}$$

$$p(A \vee B) = maximum(p(A), p(B)) \tag{3.2}$$

$$p(\sim A) = 1 - p(A) \tag{3.3}$$

Problems with Fuzzy Logic

Disadvantages of this formulation are its sensitivity to only the smallest possibility, in the case of the conjunction rule; and to only the largest, in the case of the disjunction rule [119, 130]. This makes the reasoning process optimistic when estimating the strength of conjunction; and pessimistic when estimating the strength of disjunction. For example, for a sentence whose associated possibility is 0.6, Fuzzy logic would give a possibility of $p(A \,\&\, \sim A) = 0.4$; and a possibility of $p(A \vee \sim A) = 0.6$.

3.5 MYCIN: A Method of Certainty Factors

Mycin [10], can be regarded as the first rule-based reasoning mechanism from which reasoning with uncertainty in expert systems has been derived. The primary task of MYCIN is to determine what significant organisms exist within a patient. Aspects of diagnosis are broken down into triples of context-attribute-value groups.

For example, if our current patient is called John Knox, then this information is coded in the following way. The context is 'the patient', the attribute is 'name' and the value is 'John Knox'. In such a way, the address of John Knox could be stored as context 'the patient', attribute 'address', value 'High Street, Edinburgh'. Or a context might be a particular organism, an attribute of the organism might be it's shape, and the value would then be the organism's actual shape.

Rules in MYCIN are then of the type:

$$IF < antecedents > THEN < action\ or\ conclusion > \tag{3.4}$$

And an example of a rule (taken from [10]) is:

IF: 1) The stain of the organism is gram positive and

2) The morphology of the organism is coccus and

3) The growth confirmation of the organism is chains

THEN: There is suggestive evidence (0.7) that the identity of

the organism is streptococcus

Attached to each context-attribute-value (c-a-v) group, there is a certainty factor, which is an indicator of how certain the fact is. So that, we may break the above rule down to:

IF (ORGANISM-STAIN-GRAMPOS C1)

(ORGANISM-MORPHOLOGY-COCCUS C2)

(ORGANISM-GROWTH-CHAINS C3)

THEN:(ORGANISM-IDENTITY-STREPTOC Cf)

where C1, C2, C3 represent the present certainty factors of the respective antecedents, and Cf is the certainty factor associated with the rule. (In this case 0.7.) The factors C1 to C3 are calculated by MYCIN, and it chooses the minimum of these (using the Fuzzy and-rule [6, 108] to be the certainty factor of their all being true together. Call this certainty factor af.

If the rule creates the first value for the certainty of c-a-v, then the new certainty factor is:

$$af * Cf \tag{3.5}$$

If a certainty factor already exists for c-a-v, then let CC = af * Cf and let Co be the old factor for c-a-v. The new certainty factor for c-a-v is:

$$Cn = CC + Co - (CC * Co) \tag{3.6}$$

If $Co > 0$ and $CC > 0$.

$$Cn = CC + Co + (CC * Co) \tag{3.7}$$

If $Co < 0$ and $CC < 0$.

$$Cn = 1. \tag{3.8}$$

If CC = 1 and Co = -1, or, CC = -1 and Co = 1.

$$\frac{Cn = CC + Co}{1 - min(|CC|, |Co|)} \tag{3.9}$$

In all other cases.

Problems with MYCIN

This method for handling uncertainty is very computationally efficient, and was developed through a heuristic process of trial and error. This has led to many criticisms about its *ad-hoc* nature [15, 130]. In particular, the system has been criticised for the way its results deviate from expected probabilistic results as the reasoning process becomes deeper [1].

3.6 The Prospector Model for Handling Uncertainty

The Prospector model combines both the standard Bayesian techniques, and techniques of Fuzzy Logic, on a database of rules. A rule in the Prospector model is of the form:

$$if\ E\ then\ (to\ degree\ LS,\ LN)\ H \tag{3.10}$$

where LS is known as the sufficiency factor, and LN is known as the necessity factor. Attached to each proposition is its currently estimated odds of being true, where the odds of an uncertain proposition are defined:

$$o(H) = \frac{p(H)}{1 - p(H)} \tag{3.11}$$

The odds of an event are as defined in the subjective theory of uncertainty. As the probability of an event approaches 1, the odds of the event approach infinity; and as the probability of an event approaches 0, so do the odds of the event. The probability of an event can be recovered from its odds by the transformation:

$$p(H) = \frac{o(H)}{1 + o(H)} \tag{3.12}$$

The sufficiency factor of a rule is usually a number very much greater than one, and the necessity factor is usually a number very much less than 1. So that, when the odds of an event are multiplied by a sufficiency factor, we expect its probability to be increased; and when multiplied by a necessity factor, we expect its probability to decrease.

3.7 Prospector's Inference Mechanism

As in Bayesian Inference, when there is no uncertainty about the evidence there is no uncertainty about the hypothesis in the light of that evidence. That is, when evidence E is definitely present, the odds of hypothesis H in the light of E are:

$$o(H|E) = LS * o(H) \tag{3.13}$$

and when E is absent:

$$o(H|E) = LN * o(H) \tag{3.14}$$

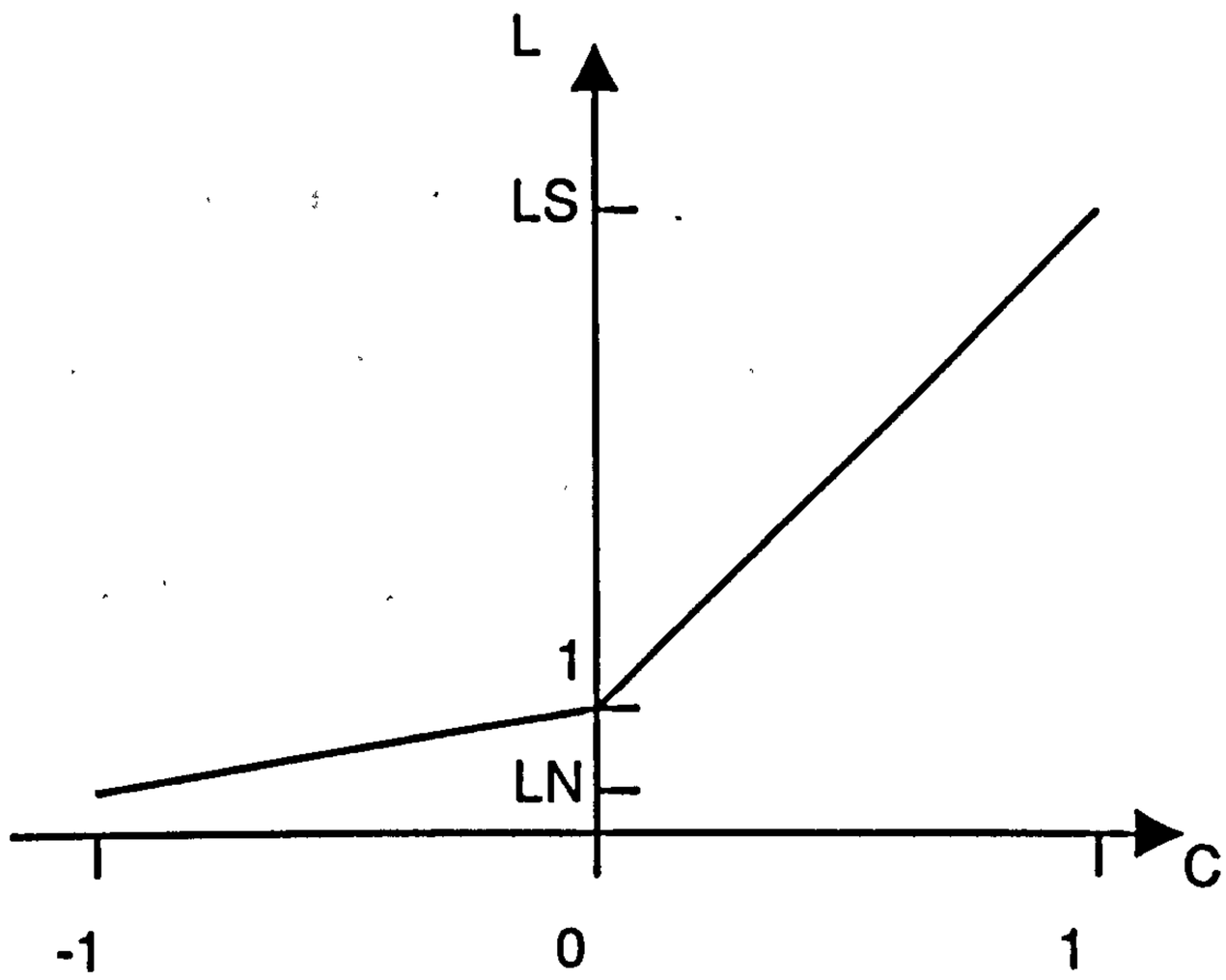


Figure 3.4: PROSPECTOR's Interpolation Schema

Uncertainty about E is expressed on a scale of belief from -1 to 1. A belief of -1 means E is false, +1 means E is true, 0 means there is no knowledge about E one way or the other. Numbers intermediate to these extremities are to quantify uncertainty in the truth or falsity of the proposition E (see figure 3.4). If the belief number is positive, then the new odds are interpolated between o(H) and o($H|E$) depending on the value of belief and equation 3.13; and if the number is negative, then the odds are interpolated between o(H) and o($\sim H|E$), depending on the belief value and equation 3.14.

When two or more pieces of evidence affect the same hypothesis, Prospector allows the

evidences to be considered conditionally independent given the hypothesis. So that:

$$p(E_1 \& E_2 | H) = p(E_1|H)p(E_2|H) \tag{3.15}$$

and

$$p(E_1 \& E_2 | \sim H) = p(E_1| \sim H)p(E_2| \sim H) \tag{3.16}$$

It is now no longer possible to evaluate the true odds of hypothesis H, so the o function is replaced by an o^* function, an approximation of the odds, such that for one hypothesis and n pieces of evidence:

$$o^*(H|E) = \Pi_{i=1}^{n} L_i O(H)$$

Problems with Prospector

The assumption that all evidences are conditionally independent at the level of the hypothesis whose probability is to be updated is unwarranted. The situation becomes clearly wrong when we consider the example of finding one piece of evidence capable of proving the hypothesis definitely true. In this situation, all the other evidences being independent from this piece of evidence, are also independent of the hypothesis [10].

In local computations Prospector's approximation method deviates only slightly from probability theory; but White has shown [130], that as the process of reasoning becomes deeper the results become extremely unreliable.

3.8 The Use of Entropy in Reasoning with Uncertainty

The entropy of a probability distribution [54, 16, 2], is a function which operates over a whole probability space and is a measure of the extent to which the probability is concentrated on a few points or dispersed over many. Probability distributions with a low entropy have probability concentrated on certain elements of the probability space; distributions with a high entropy have the probability spread more throughout the space. More formally, the entropy of the probability mass function $p_X(x)$ may be regarded as a descriptive quantity,

just as the median, mode, variance and coefficient of skewness may be regarded as descriptive parameters. It is an indicator of the degree of disorder in a probability space.

Definition

If p_i is the probability that the discrete random variable X takes the value x_i and $p_i \geq 0$, i=1,2,...,n, and $\sum_{i=1}^{n} p_i = 1$, then the entropy of X is H(X) where

$$H(X) = H(p_1, p_2, ..., p_n) = -\sum_{i=1}^{n} p_i log p_i. \quad (3.17)$$

In the examples here we will use 2 as the logbase; although any base can actually be used [54].

Example 1

We are provided with four coins and told that one of the coins is counterfeit. The situation is shown pictorially in figure 3.5, where the four probability distributions are labelled D1 to D4. In each of these distributions, probabilities p_1 to p_4 are associated with the four coins, where p_n describes the likelihood of coin n being the counterfeit. The entropy of each of the distributions is show in column H.

Distribution	p(n) = Prob associated with coin n				Entropy (H)
	p1	p2	p3	p4	
D1	1/4	1/4	1/4	1/4	2
D2	1/2	1/6	1/6	1/6	1.8
D3	3/4	1/12	1/12	1/12	1.2
D4	1	0	0	0	0

Figure 3.5: Entropy as a Changing Property of Probability Distributions

The distribution with maximum entropy is D1. More precisely, D1 is the maximum entropy distribution for a sample space of four points when there are no probability constraints on the points other than that they must sum to one. More generally, for any n points, the distribution which has the maximum entropy is that which assigns 1/n to each. The reduction in entropy from D1 to D4 demonstrates the effect of having more information about the change

in probabilistic likelihood of one of the coins over the others. D4, (entropy 0), represents the case where there is no uncertainty as to which coin is counterfeit.

Information is embodied in each of the distributions, and we can see that the distribution which says least about the identity of the counterfeit coin is D1. This equation of information with entropy leads to the maximum entropy principle: *Of all probability distributions which satisfy the constraints imposed by the known aggregate probabilities, choose that distribution which has the maximum entropy or, equivalently, contains the least information.*

When there are more probability constraints on the points in the possibility space, the absolute maximum entropy distribution (D1 above), cannot normally be applied. The solution method for dealing with this problem is given in example 2.

Example 2

This example is adapted from one provided by Bard [3]. Jack and Jill work in an office with several co-workers. The probabilities are π_1 that there is somebody in the office, π_2 that Jack is in the office, and π_3 that Jill is in the office. What is the probability that both Jack and Jill are in the office?

Sentence	a	b	c	d	e	Probability
τ	1	1	1	1	1	1
SOMEBODY	1	1	1	1	0	π_1
JACK	0	1	0	1	0	π_2
JILL	0	0	1	1	0	π_3

Table 3.1: Interpretation Table for Office Example

The possible worlds are labelled with small letters a, b, c, d and e. The possibilities are: (a) somebody in the office, but it is neither Jack nor Jill, (b) Jack but not Jill is in the office, (c) Jill but not Jack is in the office, (d) Jack and Jill are both in the office, (e) nobody is in the office. Table 3.1 is an interpretation table for this scenario, with the possibilities (a) to (e) explicitly represented by worlds a to e respectively. The tautology τ is true in all possible worlds and is included to ensure that all the probabilities sum to 1.

Solution

The probability of world e is known. It is $1 - \pi_1$. The probability constraints are formed thus: if a sentence is true in any of the possible scenarios (a-e), then the probability of that scenario helps to make up the probability of the sentence. From the table, the equations to be solved are:

$$\begin{aligned} a + b + c + d &= \pi_1 \\ b + d &= \pi_2 \\ c + d &= \pi_3 \end{aligned} \tag{3.18}$$

The procedure for obtaining a solution is based on standard Lagrangian methods and can be expressed as follows: associate an unknown variable with each unknown aggregate, one for each row of the semantic tree. We will assign a_1, a_2, a_3, and a_4 to the rows for τ, SOMEBODY, JACK, and JILL respectively. Each possible world can now be rewritten in terms of the multiplication of aggregates where the aggregate is included in the multiplication list only if the world has the value one in the corresponding row of the semantic tree.

$$\begin{aligned} a &= a_1 a_2 \\ b &= a_1 a_2 a_3 \\ c &= a_1 a_2 a_4 \\ d &= a_1 a_2 a_3 a_4 \\ e &= a_1 \end{aligned} \tag{3.19}$$

Substituting expressions 3.19 into equations 3.18 we obtain:

$$\begin{aligned} a_1 a_2 (1 + a_3)(1 + a_4) &= \pi_1 \\ a_1 a_2 (1 + a_4) &= \pi_2 \\ a_1 a_3 (1 + a_3) &= \pi_3 \end{aligned} \tag{3.20}$$

So that: $(d) = a_1a_2a_3a_4 = \frac{\pi_2\pi_3}{\pi_1}$; which is the required probability from the maximum entropy distribution.

Problems With Entropy Solutions

The two main problems with using maximum entropy directly in an expert system are:

- The enumeration of all of the semantic possibilities becomes a non-trivial task as the number of sentences increases, and without these semantic enumerations any entropy solution is rendered invalid.

- In general, the solution of the non-linear entropy equations can only be acheived using an iterative approximation procedure which has been shown to be NP-Complete [91]. As n (the number of random variables) increases, such procedures eventually saturate and produce no result [56].

3.9 The Dempster-Shafer Theory of Evidence

This method was first proposed by Arthur Dempster (1968) and later extended by Glen Shafer (1976). The set of all possibilities is called the "frame of discernment", denoted Θ. These possibilities are assumed to be mutually exclusive, and exhaustive. Dempster-Shafer theory uses a real number in the range [0,1] to indicate the degree to which a piece of evidence supports a hypothesis. The impact of each piece of evidence on the subsets of Θ is represented by a function called a "basic probability assignment" (bpa). The bpa assigns a measure of belief to subsets of Θ, and is a specially developed generalisation of the statistical probability density function (pdf) [38]. The power set of Θ (2^Θ) is assigned a bpa by the special function "m" which assigns a number to each of the subsets of 2^Θ such that the numbers all sum to 1. In a pdf, a number is assigned to each singleton of the hypothesis set such that the numbers sum to 1.

The quantity "m" (A) is a measure of the belief assigned to proposition A, where A is some subset of 2^Θ, and the total belief sums to 1. This belief assignment may not be subdivided

amongst the subsets of A. The remaining belief (1 - 0.7) is then assigned to Θ. That is, we cannot further choose between any of the subsets of Θ with the remainder of the belief.

A function is introduced which gives the total amount of belief in hypothesis A, not only belief committed exactly to A, but also belief committed to all subsets of A. This function is called a *belief function*, denoted Bel.

If Bel_1 and Bel_2 are two belief functions whose bpa's are m_1 and m_2 respectively, then Dempster's rule allows us to compute a new bpa, denoted by $m_1 \otimes m_2$, which represents the combined effect of m_1 and m_2 on the frame of discernment.

The Dempster-Shafer theory of evidence defines that the belief associated with the empty set (*empty*) must always be 0. A heuristic is employed on the orthogonality principle to achieve this. Dempster deals with the problem by normalising the computed values so that $m_1 \otimes m_2(empty)=0$, and all other values of the new bpa lie between 0 and 1. This is achieved by defining κ as the sum of all non-zero values attached to *empty* in a given case. He then assigns 0 to $m_1 \otimes m_2(empty)$, and divides all other values of $m_1 \otimes m_2$ by $1 - \kappa$.

Bel(A) gives the total amount of belief committed to subset A. The complement of A can be denoted A^C. And so, the information contained in Bel(A^C) is the amount of belief attributed to A^C. Therefore, the quantity 1 - Bel(A^C) expresses the plausibility of A, i.e. the extent to which the evidence fails to doubt A. Therefore, the information contained in Bel concerning a given subset A may be conveniently expressed by the interval: [Bel(A), 1 - Bel(A^C)].

Critiscisms of the Dempster-Shafer theory have been made by Zadeh [108] and Pearl [97] about the nature of deduction and the admissibility of the orthogonality principle. These are that, although this technique has a method of broadening out the scope of its answers, there is no sound justification for this means of assigning beliefs, nor of reassigning belief which is initially assigned to a null hypothesis. Pearl, in comparing the Dempster-Shafer theory of evidence with Bayesian Inference has said "in the Bayesian approach a proposition is believable when it is provably probable; in the D-S approach, when it is probably provable. Thus, the former uses probability as the object language and logic as a meta-language; the latter reverses these roles."

3.10 From Extensional to Intensional Methods

The artificially intelligent reasoning methods considered so far have been classified as *extensional* [99, 97], which is typified by rule-based systems or production systems [88]. In such systems uncertainty values are directly attached to sentences, and the uncertainty of any formula is computed as some function of the uncertainty of the respective sub-formulae. Furthermore, all of these systems, necessarily, offer the user the ability to see into the system how conclusions and their uncertainty values have been built up: that is, a "window" into the system.

The advantages of such systems are that no semantic information between propositions need be modelled, and that the speed of producing a resultant certainty factor from given information is very quick. However the computational advantages have been acquired by a loss of reliable semantics in the inferring mechanism. White [130] has said that models already developed in statistics should be the tools for reasoning with uncertainty, and criticised artificially intelligent reasoning mechanisms with the phrase "we are better off without windows if they are obtained at the cost of distorting what is seen through them".

Bundy, in introducing Incidence Calculus [13] called these extensional methods "purely numeric" mechanisms and, in examining them in detail, clarified some fundamental limitations in their ability to function in a correct way over a probability space. These limitations can be summarised in the following manner [13]; where τ represents a universally true sentence, and f represents a universally false sentence.

$$p(\tau) = 1 \tag{3.21}$$

$$p(f) = 0 \tag{3.22}$$

Sentences representing propositions whose truth value is uncertain may be probabilistically quantified with a number between 0 and 1. The following equations assign arithmetic functions to the propositional connectives:

$$p(\sim A) = 1 - p(A) \tag{3.23}$$

$$p(A \vee B) = p(A) + p(B) - p(A\&B) \tag{3.24}$$

and, provided A and B are statistically independent:

$$p(A\&B) = p(A).p(B) \tag{3.25}$$

The caveat attached to equation 3.25 asserts that the relationship of independence holds between the propositions A and B. This assumption of independence between random variables makes the calculation of probabilities of compound events computationally feasible. However, the truth or falsity of either of the propositions must have no effect on the other.

The positive aspects of the independence assumption may be seen in the example of throwing a seven-sided dice once, and tossing a coin. If we wanted to know the probability of simultaneously obtaining a head on the coin, and a three on the dice, we can use the independence assumption to calculate the probability as: 1/7 * 1/2 = 1/14. Such an example illustrates the nature of the independence assumption: that perturbations in the sample space caused by one event being true have no effect on the result of the other event being measured.

One might expect that the independence assumption has such a small effect that its convenience warrants its inclusion into the assumption axioms for the reasoning process. However, rules of inference which are based on a number of antecedents e.g.$A_1\&A_2\&...A_n \Rightarrow B$, suggest that there is some dependence, or at least some relationship between the antecedents $A_1, A_2, ..., A_n$. i.e. since their truth together implies the truth of proposition B, there is obviously some relationship between them.

Bundy uses a heavily weighted example to show how dangerous, both probabilistically and logically, the unconstrained assumption of independence can become. If we assume that the probability of proposition A being true is 0.75, then, using equation 3.25 and ignoring the caveat, p(A&~A) = 0.75 * 0.25 = 0.1875; also, from equation 3.25, p(A∨~A) = 0.75 + 0.25 - 0.1875 = 0.8125. Logically speaking, the real relationship between propositions A and ~A has been muddied over by the independence assumption, and therefore the probabilities, (which should be 0 and 1), are unreliable.

What the independence assumption asserts is that there is no correlation between the two propositions. In probability theory, whenever we are aware that two propositions A and B

are correlated, equation 3.25 may be replaced by:

$$p(A\&B) = p(A).p(B) + c(A,B).\sqrt{p(A).p(\sim A).p(B).p(\sim B)} \qquad (3.26)$$

where the correlation, c(A,B) is a number between -1 and 1, such that 0 represents no correlation between the two, (i.e. the independence assumption), 1 represents the case where B is present whenever A is present, and -1 represents the case where B is absent whenever A is present.

Although this equation can be used to replace equation 3.25, Bundy goes on to prove that a probability calculus which used this replacement would not have truth functional connectives. Truth functional connectives have the ability to generate compound sentences from simple ones in such a way that the truth values of the compound sentences are determined solely by the truth values of the simpler sentences [13].

With these problems in mind, some development of systems for reasoning with uncertainty has taken place on the "intensional" approach [99]. In this approach uncertainty is attached to sets of "possible worlds", and is manipulated in accordance with the rules of set theory. For this reason Bundy has called such mechanisms "set-theoretic". The three methods in this category which have emerged since 1985 are: Incidence Calculus [13], Probabilistic Logic [89], and Stochastic Simulation [95]. In all these methods the semantics are clear and mathematically justifiable, but the inference mechanisms have shown themselves to be computationally expensive. Because these problems are in the areas of data complexity the problems are those faced by the Bayesian Inferencing community, with the added limitations of mathematical deduction.

These mechanisms, on shifting the focus of system development away from the difficulty of creating an extensional calculus back to the consistent handling of uncertainty, have been less well integrated into the expert system community as a whole because of their inability to produce quick results. However, the developers of these mechanisms have chosen the simplest computational strategies which preserve consistency in the data sets, while still maintaining a way of examining the working of the system: a consistent "window" into the system. The development of the set-theoretic mechanisms is a movement towards the creation

of mathematically and statistically sound methods of reasoning with uncertainty which still maintain ease of overall system visualisation.

3.11 Network Models

Inferno [103] is a "cautious approach to uncertain inference", which uses the idea of belief and plausibility, as per Dempster-Shafer, except over a network model of probability assignments. The method is probabilistic, but is particularly cautious, (within a few steps the bounds of uncertainty approach [0,1]). There are problems in the uncertainty handling mechanism. Spiegelhalter [119] has said "although this seems to be a suitably "cautious" approach to probability propagation, Inferno appears to confuse *conflicting* evidence with *inconsistent* evidence."

Inferno does not use the set-theoretic notion of possible worlds. Pearl has classified Inferno as a "local approximation to Nilsson's probabilistic logic" [97], and as such, it is not treated distinctly in this thesis.

3.12 Nilsson's Probabilistic Logic

Probabilistic Logic has been anticipated both in Bernoulli's range theory of probability [7], and by De Finnetti's pioneering work on subjective probability [41]. The computer is the new ingredient which brings a freshness to this subject. In Probabilistic Logic the sample space over which probabilities are defined is taken to be the total number of logically allowable possible worlds given by the state of uncertainty. Theorem proving software is used to collect all of the possible worlds (termination problems notwithstanding).

If we are interested in only one sentence, S say, we could imagine two sets of possible worlds, W1 containing the worlds in which S is true, and, W2 containing the worlds where S is false. Nilsson's probabilistic logic allows probabilities to be assigned to logical sentences. In this example, we assign a probability π_s to sentence S. If we have enumerated all the logically possible worlds in the set W1 *union* W2, then the actual world must be one of these. We model our uncertainty by imagining S to be in W1 with probability π_s, and in W2 with

probability 1-π_s. Since, all of the distinct possible worlds have been enumerated, the sum of the probabilities of the worlds equals one, and the worlds are mutually exclusive.

$$\begin{pmatrix} P \\ P \Rightarrow Q \\ Q \end{pmatrix} \begin{pmatrix} true & true & false & false \\ true & false & true & true \\ true & false & false & true \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

Figure 3.6: Nilsson's Possible Worlds Notation

These possible worlds are derived by an exhaustive theorem prover [14] to completely produce all of the possible worlds. This is hampered by semi-decidability problems, but, for the moment, we will consider this production of all possible worlds to be non-problematical. An example used by Nilsson is of the set P, $P \Rightarrow Q$, Q, where the consistent possible worlds are modelled in figure 3.6. In the first column are the names for each sentence, in the second to the fifth column are the possible worlds, and in columns six to nine is an abbreviated shorthand notation for this information. For the purposes of this example consider that the four possible worlds are individually labelled (from left to right) a, b, c and d.

When the proposition P, and the rule $P \Rightarrow Q$, are given probabilities (say π_1 and π_2 respectively), then Probabilistic Logic provides a method for assigning these probabilities amongst these four worlds consistently. If a consistent probability distribution can be assigned to the random variable represented by the possible worlds, we will have a probability for worlds 'a' and 'd' (the only two worlds in which proposition Q is true). The sum of these two probabilities then is the *entailed* probability of proposition Q.

The set of consistent possible worlds for a set of uncertain sentences can be found by constructing a binary tree of all possible assignments of true and false to each of the sentences and then testing each possible assignment using a theorem prover. Those assignments which generate an inconsistency are removed from the set. The remaining assignments are all of the logically consistent assignments of true and false to the uncertain sentences which is, throughout this thesis, referred to as the "semantic tree" of the logical sentences. This semantic tree is represented in a matrix notation, and Nilsson calls the resultant matrix the V-matrix. i.e.

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

From this he constructs the V'-matrix, which collects together all of the sentences for which there are probabilities. The difference between the V-matrix and the V'-matrix is that although they both have the same number of rows, the V'-matrix does not contain the last row of the V-matrix; but it does contain a new row (the first row) which is all 1's. This row is to represent the tautology sentence, which is true in all possible worlds. So the V'-matrix is:

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \end{pmatrix}$$

He introduces the matrix II which holds the probabilities for all the included sentences, which has a corresponding matrix II' to represent the probabilities of the sentences we know. By default the probability of the tautology is 1, and from information provided we have $p(P)=\pi_1$, $p(P \Rightarrow Q)=\pi_2$. So, the II' matrix is:

$$\begin{pmatrix} 1 \\ \pi_1 \\ \pi_2 \end{pmatrix}$$

The final matrix used by Nilsson is called P, which holds the probabilities of each of the possible worlds. The solution to a problem of probabilistic entailment is to solve the equation:

$$\Pi' = V'P \tag{3.27}$$

to find the probabilities of possible worlds, and hence the probabilities of the entailed sentence. So that, in this example, we would be looking to solve the matrix expression shown in figure 3.7

$$\begin{pmatrix} 1 \\ \pi_1 \\ \pi_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} p(a) \\ p(b) \\ p(c) \\ p(d) \end{pmatrix} \tag{3.28}$$

Figure 3.7: Method of Multiplying Matrices

$$\begin{pmatrix} p(a) \\ p(b) \\ p(c) \\ p(d) \end{pmatrix} \tag{3.29}$$

for p(a), p(b), p(c) and p(d); which are the probabilities of the possible worlds a, b, c and d respectively. The probability of Q is then the sum of p(a) and p(d), where a and d are the two possible worlds in which Q is true.

3.13 The Process of Probabilistic Entailment

Usually there is more than one way of assigning probability to the possible worlds; and because of this, all we can say with certainty is that the probability of a probabilistically entailed conclusion, (Q in the example above), is bounded. This is best demonstrated geometrically.

Consider the possible worlds a, b, c and d shown in the V-matrix (figure 3.6). These worlds, when considered as points in 3-space, can be used to outline a three-dimensional object (figure 3.8). The four axioms of probability theory (2.1 to 2.4) only require that probabilities attached to the propositions P, P⇒Q, and Q must lie in the convex hull of these points [46].

Therefore, when probabilities π_1 and π_2 are assigned to P and P ⇒ Q respectively, the probability of Q is found by moving to point (π_1,π_2) on the P(P)-P(P⇒Q) plane, and projecting a line parallel to the P(Q) axis through this point. The line only touches the object at each of the two points (1,1) and (1,0) (that is, (1,1,1) and (1,0,0) respectively). At the other points, either the projected line misses the object (in which case the assigned probabilities are inconsistent), or it passes through the object. Where the line passes into the object is the lower bound of the probability of the entailed sentence Q, and where it passes out of the object is the upper bound of the probability.

Another way to perform probabilistic entailment would be to use Gaussian elimination methods. The row which was removed from the V-matrix, [1,0,0,1], represents the worlds in which Q is true (that is, worlds a and d). The method is to manipulate the rows of the V'-matrix, using arithmetic operations of addition and subtraction, to produce the row [1,0,1,0] and to use the same transformations on the Π' matrix to get the corresponding probability

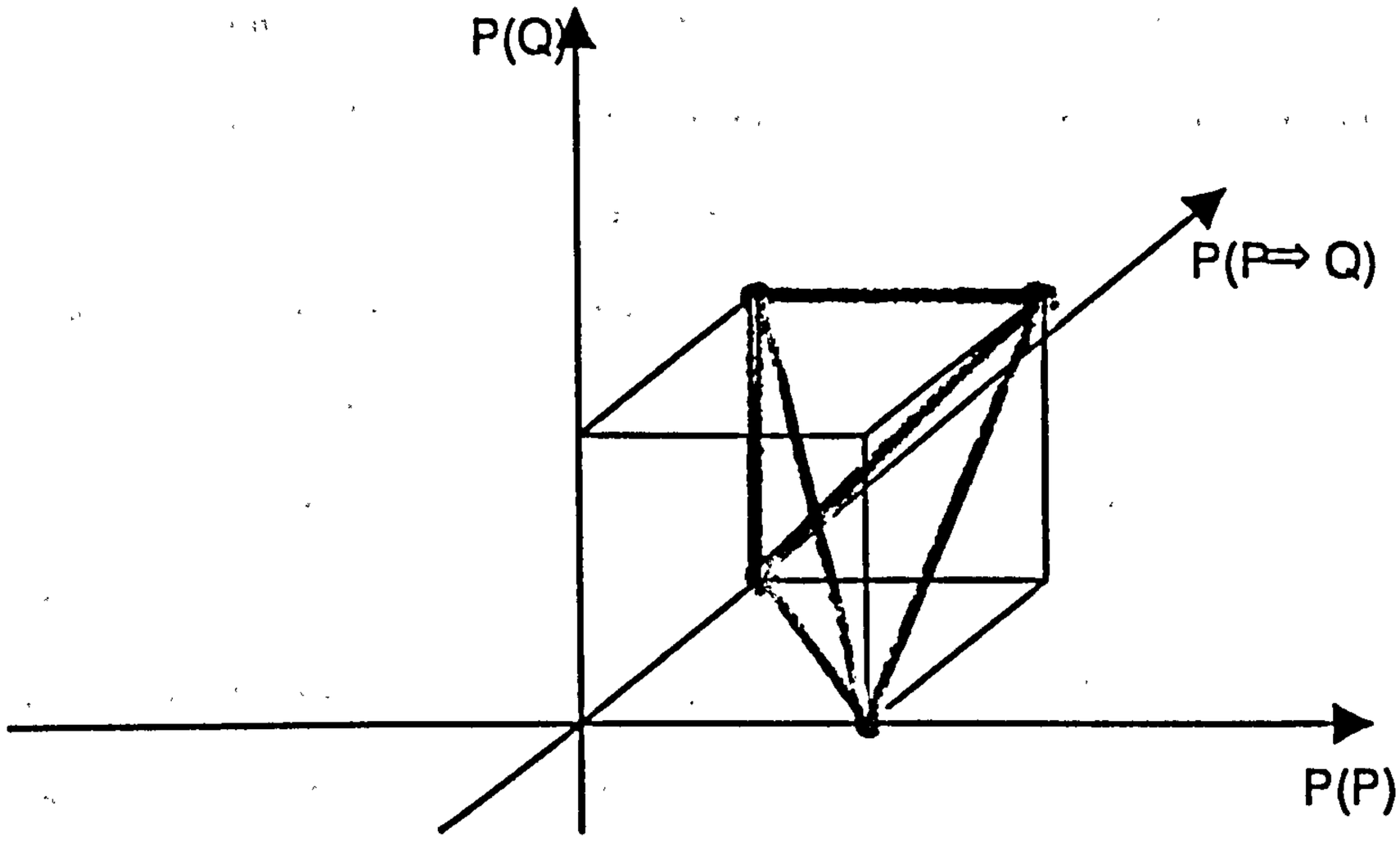


Figure 3.8: Geometric Considerations

P(Q).

In the event that this method does not work, Nilsson proposed a number of approximation methods which reduced the complexity of the entailment process by pruning the semantic tree in various ways. In particular he suggested a use for Jaynes's maximum entropy formalism [60] for producing the least commitment probability distribution to the possible worlds. This method is examined in detail in chapter 5. Extensions to this basic model of Probabilistic Logic are presented in chapter 4.

3.14 Incidence Calculus

Incidence Calculus is a set theoretic mechanism for reasoning with uncertainty which overcomes problems which have been perceived in the purely numeric methods of uncertainty management [13, 45]. Bundy expands the idea of the probability of a logical sentence from being represented by a number to being represented by a set of points, each of which has a probability attached. The set of all points is the sample space and is denoted by 'w'. Bundy says "The sample space is to be an exhaustive and disjoint set of points".

Bundy's statement "Each point can be regarded as a situation, Tarskian interpretation, or possible world in which a sentence will be either true or false" indicates that 'w' is supposed to encompass all of the logically possible worlds. The fact that the production of 'w' does

not require a theorem prover, but could perhaps be guided and shaped by a human expert, or some other, mechanical, process of design makes Incidence Calculus a good complement to Probabilistic Logic, but necessarily an approximation to it.

If A is a proposition sentence, then i(A) is defined to be the *incidence* of A, and is the subset of points in 'w' in which sentence A is true. The dependence between two proposition sentences A and B is coded in the amount of intersection between their two incidences i(A) and i(B). Once incidences are assigned to each sentence proposition such that the probability associated with each proposition is honoured; the mathematical processes of set-union and set-intersection provide the incidences for the application of the logical 'or' rule (V) and 'and' rule (&) respectively.

For each incidence assignment, a pair of lower bound and upper bound assignments are kept. These are directly analogous to the belief and plausibility estimates provided in Dempster Shafer theory, and represent the amount of producible evidence which can support a proposition; and the extent to which it is not possible to prove the proposition wrong [113].

On acquiring an incidence assignment for a proposition, Bundy employs an "Inconsistency Detector" to check that the new assignment does not violate the bounds of the known assignments. If the assignment is inconsistent with incidences which have already been assigned, Bundy uses a "Legal Assignment Finder" to attempt to find a new incidence assignment pattern for the conflicting predicates which will code all of the known probabilistic information sufficiently. If no such assignment can be found, the program terminates with failure. Bundy proves his Legal Assignment Finder to be sound and complete for first order predicate Incidence Calculus.

Problems with Incidence Calculus

Bundy states [13] "Incidence Calculus can be implemented reasonably efficiently by representing the incidences of sentences as bit strings and manipulating them with logical operations. Each incidence can be represented by a bit string of a fixed length, say 100 bits, each bit corresponding to an element of 'w'. The longer the string, the greater the accuracy, but the

greater the cost in terms of space and time."

The most pressing problem with Incidence Calculus is the way it assigns the number of points to the sample space. When too few points are given to the set, uncalled for relations will be forced between sentences. This is one of the points Bundy criticised about the numeric mechanisms. On the other hand, when the set is overly large, there is redundancy in the system. This added redundancy must be removed as much as possible from the already computationally expensive inference mechanism. The problem of incidence assignment itself is still an open question, and a complete mechanism to deal with conditional probabilities is also lacking.

3.15 Pearl's Stochastic Simulation

Stochastic simulation is a method of computing probabilities using the frequency theory of probability. That is, by counting out how many times an event occurs over a number of samples, and dividing by the total number of sample data events [95]. A causal model of a domain is used to generate random samples of hypothetical scenarios (possible worlds) that are likely to develop in the domain. The probability of any event or combination of events can then be computed by counting the percentage of samples in which the event is true.

Pearl uses the following example, first proposed by Cooper [20]: *Metastatic cancer (A) is a possible cause of a brain tumor (C) and is also an explanation for increased total serum calcium (B). In turn, either of these could explain a patient falling into a coma (D). Severe headache (E) is also possibly associated with a brain tumor.* The Bayesian network associated with this information is shown in figure 3.9, and the corresponding conditional probability information, which Pearl calls the *link matrix* is:

$$
\begin{array}{rr}
P(a) = 0.2 & \\
P(b|a)=0.8 & P(b|\sim a)=0.2 \\
P(c|a)=0.2 & P(c|\sim a)=0.05 \\
P(d|b,c)=0.8 & P(d|\sim b,c)=0.8 \\
P(d|b,\sim c)=0.8 & P(d|\sim b,\sim c)=0.05 \\
P(e|c)=0.8 & P(e|\sim c)=0.6
\end{array}
$$

Pearl [97] uses uppercase letters to represent propositional variables in the Bayesian network. A Bayesian network is a directed acyclic graph in which nodes represent proposition

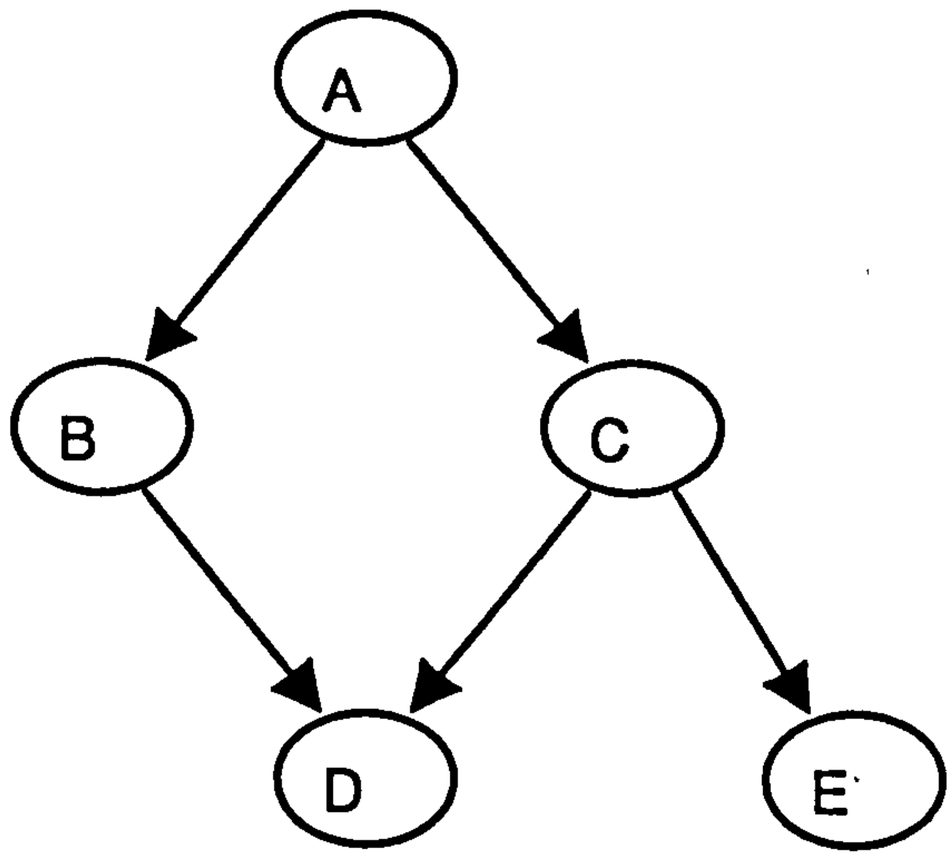


Figure 3.9: Diagram of Causal Connections

sentences and the arcs represent the existence of direct causal influences between linked propositions. The strength of these influences are quantified by conditional probabilities. When assigning truth values, true or false, to the variables he uses the lower case equivalent letter. For example, in figure 3.9 A can be true or false, A=1 or A=0 respectively; or alternatively, a or $\sim a$, respectively. Given the information in the link matrix, the goal is to compute the posterior probability of every proposition in the system, given that a patient is suffering from severe headaches (e), but has not fallen into a coma ($\sim$d); that is, E=1, and D=0.

The first step is to initialise all of the unobserved variables (A, B, C) to some arbitrary initial state (for example A=B=C=1), and then let each variable in turn choose another state in accordance with the variable's conditional probability given the current state of the other variables. Pearl denotes w_A to be the state of all variables except A. So that, in the initialisation stage, w_A is {B=1,C=1,D=0,E=1}; "and the next value of A will be chosen by tossing a coin which favours 1 over 0 by a ratio of $P(a|w_A)$ to $P(\sim a|w_A)$. The expression $P(x|w_X)$ is called the "transition" probability of variable X. Pearl then provides a way to derive the conditional probability of any variable X conditioned on the values w_X of all other variables in the system.

For this he needs to inspect only neighbouring variables of X, that is the Markov blanket of X. He calculates the Markov blankets of all the nodes. The Markov blanket of a node X is the direct parents of X, the direct successors of X, and all direct parents of X's direct

successors. Denoting B_X as the Markov blanket of X, then from figure 3.9:

$$B_A = [B, C]\,; \quad B_B = [A, C, D]\,;$$

$$B_C = [A, B, D, E]\,; \qquad B_D = [B, C]\,; \qquad B_E = [C] \tag{3.30}$$

From this information, and the knowledge that D and E are kept constant (0 and 1 respectively), we can compute the transition probabilities from:

$$P(a|w_A) \;=\; P(a|b, c, d, e) \;=\; \alpha P(a)P(b|a)P(c|a) \tag{3.31}$$

$$P(b|w_B) \;=\; P(a|b, c, d, e) \;=\; \alpha P(b|a)P(d|b, c) \tag{3.32}$$

$$P(c|w_C) \;=\; P(a|b, c, d, e) \;=\; \alpha P(c|a)P(d|b, c)P(e|c) \tag{3.33}$$

where the α's are normalising constants that make the respective probabilities sum to 1.

The transition cycle then repeats itself in the order A, B, C until a query, for example, "what is the posterior distribution of A?" is to be addressed. Pearl allows the answer to such a query to be the percentage of times A registers the value TRUE; or, in a more complex manner, the answer is the average of the conditional probabilities $\mathrm{P}(A = 1|w_A)$ computed in transition.

Problems with Stochastic Simulation

The major problem with Stochastic Simulation is the amount of time required before a result can be found; but a secondary problem has to do with the random world generation procedure. If the system ever gets into an undesirable position, it is possible never to generate some possible worlds. In this situation, the probability of these possible worlds is kept at zero for all time. Another problem is that, running the system twice in succession with the same data is not liable to give the same result, again because of the random fluctuations in the selection of the next world.

3.16 Bayesian Networks and Influence Diagrams

Related to Stochastic Simulation are "Influence Networks" [78] and "Bayesian Networks" [97]. The starting point for this work is to "view a Bayesian network not merely as a passive parsimonious code for storing factual knowledge but also as a computational architecture for reasoning about that knowledge" [93].

This work is strongly linked with developing issues in probability theory and seeks to provide efficient methods of structuring knowledge and fusing results both downwards through the network and back up to the top — hence allowing both top-down and bottom-up reasoning. However, the architectures themselves do not involve the representation of possible worlds specifically, and do not make specific use of predicate calculus. As such, although it is a very interesting area, it is outside the scope of this thesis and is not dealt with.

3.17 Conclusion

The previous two chapters outline the strands which led towards the development of the set theoretic mechanisms. The most important factor is a desire for clarity in the uncertainty management process. This conceptual clarity is an overhead on the computation process, which, up until now, has been the most persuasive argument for making do with approximation schemes of inferencing such as those treated in this chapter. The simple numeric mechanisms of PROSPECTOR or MYCIN are easily implemented on a computer. However, problems of complexity, become evident even when using the Dempster-Shafer mechanism, and are a major issue in the possible worlds methods. In general, as the clarity of the inferencing mechanism becomes increased, the complexity of the mechanism also increases.

This thesis develops an inference mechanism, from within Nilsson's probabilistic logic, which adheres strictly to the theories of uncertainty in mathematical logic and statistics. The maximum entropy formalism is used extensively throughout. The most important question addressed is "Can reasoning with uncertainty using the set theoretic mechanisms be computationally efficient?". I intend to show that it can. All of these points will be covered in the discussion of the extensions proposed for Nilsson's probabilistic logic, which also have

ramifications for Incidence Calculus and Stochastic Simulation.

In summary, the complexity problems of the three set theoretic mechanisms are in the areas of possible worlds generation, and probability assignments to these generated worlds. For Probabilistic Logic, the complexity problems are in the areas of semantic tree generation, and entailment solutions. The former problem leads to computational complexity problems in space; the latter in time. This problems in space and time are equivalent to those shared by Incidence Calculus and Stochastic Simulation. These problems of Nilsson's probabilistic logic are considered in more detail in chapter 4, where solutions are presented for the representational problems.

Chapter 4

ENHANCEMENTS TO NILSSON'S PROBABILISTIC LOGIC

4.1 Introduction

Nilsson's probabilistic logic is chosen as a paradigm for set-theoretic reasoning with uncertainty because it requires knowledge of all of the possible worlds for an uncertain situation before it attempts to employ a statistical calculus in an attempt to reason with the uncertainty. Nilsson's probabilistic logic has the added feature that it is a combination of ideas from first-order logic and probability theory and thereby lies its usefulness in the field of rule based systems in which either data or rules of inference may be uncertain.

Since we intend to investigate the nature of Nilsson's probabilistic logic when dealing with a large number of sentences, and therefore possible worlds, we introduce naming conventions to standardise this process. We also introduce a shorthand notation for viewing the sentences, the possible worlds of Nilsson's V′-matrix, and the probabilities in Nilsson's Π′-matrix in one diagram.

Pearl, in his book, "Probabilistic Reasoning in Intelligent Systems" [97], summarises the difference between Bayesian Theory, Dempster-Shafer Theory, and Nilsson's probabilistic logic

in the following way. "While Bayesian theory requires the specification of a complete probabilistic model and the Dempster-Shafer sidesteps the missing specifications by compromising its inferences, Probabilistic Logic considers the space of all models consistent with the specifications that are available and computes bounds instead of point values for the probabilities required." Pearl points out the major strength of Probabilistic Logic: the ability to produce the upper and lower bounds of probability for an uncertain sentence. But implicitly, he also points out a failing which is that in the model for Probabilistic Logic proposed by Nilsson there is no way to specify a complete probabilistic model so that point probabilities may be produced.

The final aspect of this chapter is an interpretation of Probabilistic Logic slightly altered from Nilsson's proposed model. This new interpretation allows the inclusion of conditional probabilities, an extended role for the maximum entropy formalism, and a new proof of the absolute bounds of an entailment problem.

The extension introduced allows Probabilistic Logic to use conditional probabilities in such a way that it is now possible to specify a complete probabilistic model for Probabilistic Logic, as for Bayesian Theory, and so to get point probability results. A proof is given as to how to deduce the bounds of an entailment without resorting to tracing the path of a convex hull in multi-dimensions [89]. A presentation of Probabilistic Logic is made which can incorporate heuristic information and rule integration into the reasoning process. These results are the first steps in opening up Nilsson's probabilistic logic to powerful aspects of both Bayesian Inferencing and Heuristic Reasoning.

4.2 Probabilistic Entailment and the Interpretation Table

Nilsson defines probabilistic entailment as an analogue of logical entailment such that, when we wish to infer from (A_1, $A_1 \Rightarrow$B) to deduce B probabilistically, and there is uncertainty about whether or not A_1 or $A_1 \Rightarrow$B is true, the real world, which has the true values of the sentences A_1, $A_1 \Rightarrow$B, and B, becomes a random variable, and can be one of a number of logical possibilities. These logical possibilities are produced exhaustively, typically by a

semantic tree theorem prover, and form the parameters of the uncertainty equations. In conventional set theoretic terms, this set of all possibilities is the *universal set.* In statistical terms, this set is called the *sample space* or *possibility space* [38]. In mathematical terms, each of these possibilities is an *extension* [13, 97], of the original uncertain sentences; or alternatively *possible worlds* [46, 58].

To introduce his probabilistic logic, Nilsson uses the sentences (P, P⇒Q) to estimate the probability of logically entailed sentence Q. However, in order to standardise the use of many antecedent sentences in an entailment rule, we introduce a number of conventions. When there are n antecedent sentences, we label the antecedents A_1 to A_n; and the entailed sentence is labelled B. In this notation, Nilsson's example becomes one of using (A_1, A_1⇒B) to estimate the probability of logically entailed sentence B. (The set of all antecedent sentences and the rule of entailment sentence is known as the "base set" for a probabilistic entailment [89].)

In order to examine the elements of probabilistic entailment clearly, we introduce the "interpretation table". This is a shorthand notation for describing the entailment procedure of section 3.12. The interpretation table is a means of gathering together the denoted sentences, the possible worlds, and the probabilities attached to sentences into one diagram. A complete interpretation table for the worlds which form the base set for the inference is:

Sentence	a	b	c	d	Probability
τ	1	1	1	1	1
A_1	1	1	0	0	π_1
$A_1 \Rightarrow$B	1	0	1	1	π_R

Table 4.1: Interpretation Table Reduced from (A_1,A_1⇒B,B)

The possible worlds of the V′-matrix are headed with small letters (a, b, c and d) and are collected in the middle column. The sentences which make this V′-matrix are represented in the leftmost column, and the probabilities associated with these sentences (the elements of the Π′-matrix) are in the rightmost column. These probabilities are labelled π_1, π_2, ... , π_n for the antecedent sentences A_1, A_2, ... , A_n; and π_R for the rule sentence. As in section 3.12 for Q, sentence B is true in worlds a and d. That is, it could be represented by the row matrix [1,0,0,1] in the in the middle column of interpretation table 4.1.

The entailment problem becomes one of assigning a probability to each of the possible worlds, such that, if the probability of a sentence S is π_S, and S is true in worlds a and b, then p(a) + p(b) = π_S. The tautology τ is true in all possible worlds and is included in the set to ensure that all the probabilities sum to 1.

From the above example we get the following constraints:

$$\begin{aligned} a+b+c+d &= 1 \\ a+b &= \pi_1 \\ a+c+d &= \pi_R \\ \Rightarrow c+d &= 1-\pi_1 \\ b &= 1-\pi_R \\ a &= \pi_1+\pi_R-1 \end{aligned} \tag{4.1}$$

which are the equations which must be solved from Nilsson's model of:

$$\Pi' = V'P$$

$$\begin{pmatrix} 1 \\ \pi_1 \\ \pi_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} p(a) \\ p(b) \\ p(c) \\ p(d) \end{pmatrix} \tag{4.2}$$

A solution to these equations for p(a), p(b), p(c) and p(d) should provide enough information to determine the entailed probability of sentence B. That is, $p(B) = p(a) + p(d)$.

4.3 Defeciencies of Nilsson's Entailment Model

Structurally, worlds c and d in equations 4.1 and correspondingly, the last two columns of table 4.1 above, have the same representation. This is because in worlds a and b, B can only assume one logical value, (true and false respectively). When A_1 is false however, and $A_1 \Rightarrow B$ is true, B can logically assume either of the values true or false. Hence, in the above example, c, (c.f. column 3),represents the world where B is false, and d, (c.f. column 4), represents the world where B is true.

An argument against the inclusion of both c and d in the interpretation table [66], is that if we were to pick a world at random from the interpretation table of table 4.1 and it so happened to be one of c or d, we would not be able to tell which of the two it was. All that we could say with certainty is that the chosen world represents the situation where A_1 is false, the rule $A_1 \Rightarrow$B is true, and the sentence B can be either true or false.

Pearl [97], comes to "the obvious conclusion that the material implication P $\Rightarrow$ Q is the wrong interpretation of the conditional sentence 'if P then Q' ". This problem is a fundamental criticism of the nature of entailment and has already surfaced in consideration of the paradoxes of material implication presented by Russell and Whitehead [107] discussed in section 2.2.

Considering the solution of equations 4.1 the best estimate of the probabilities of worlds c and d is:

$$c + d = 1 - \pi_1. \tag{4.3}$$

Nilsson [89] solved this equation:

$$c = d = \frac{1 - \pi_1}{2} \tag{4.4}$$

thus imposing an unnecessary condition on the relationship between the possible worlds ($\sim A_1, A_1 \Rightarrow$B, B) and ($\sim A_1, A_1 \Rightarrow$B, $\sim$B), namely that they have the same probabilistic likelihood. In this way, we are forced to resort to a method of probability estimation which incorporates information into our reasoning process which is not necessarily true, and which in the long run threatens the integrity of the entailed probability.

The Inadequacy of Nilsson's Equal Split

Consider table 4.2, which shows the effect of using Nilsson's method of making a half-split on such possible worlds for the entailed probability for the two cases of: p(A_1) = 0.9999, p(A_2) = 0.9999, p($A_1 \& A_2 \Rightarrow B$) = 0.5; and p(A_1) = 0.0001, p(A_2) = 0.0001, p($A_1 \& A_2 \Rightarrow B$) = 1.

Both of these example sets give the same answer because of the assumption of equal probabilities for each of the possible worlds whose stalks are the same. However, a human observer, in the light of no other information would put more credence on the usefulness of

$p(A_1)$	$p(A_2)$	$p(A_1 \& A_2 \Rightarrow B)$	Entailed $p(B)$
0.9999	0.9999	0.5	0.5
0.0001	0.0001	1.0	0.5

Table 4.2: The Inadequacy of Uncontrolled 50:50 Splits

the first result, simply because the antecedent probabilities are very strong, and the entailed probability is (almost) the total strength of the rule. In the second example, the world in which both antecedents are false, and the rule is true is by far the most probable world (0.99); and it is this world which contributes almost 0.5 to the strength of the entailment.

This anomaly in the entailment process is very prevalent in the results from this method of entailment. The problem is compounded in situations where the level of uncertainty is produced through delicate fluctuations of antecedent probabilities over a wide range of variables. It is clear that if the entailment procedure is to be employed with confidence in a problem involving uncertain variables that these offending possible worlds must be satisfactorily dealt with.

Other Problems with Probabilistic Logic

If Probabilistic Logic is to be used in earnest by the expert systems community, there are a number of problems which have to be solved.

1. When there are a reasonable number of sentences in the base set, generation of the set of all possible worlds becomes time consuming and complex.

2. Since in general the probabilities given will underdetermine the probability distribution to worlds, we are forced to make do with a range of possible probabilities for worlds. Choosing the best distribution among these worlds becomes increasingly problematical as more sentences are involved. Statistically speaking, this may be achieved by employing the method of maximum entropy, which is a notoriously time consuming process.

3. The probability bounds need to be discovered by geometrically modelling the shape of the search space, and then finding upper and lower bounds for the probability of

the entailed sentence. This means having some means of modelling shapes in a multi-dimensional coordinate system, and tracing out to the upper and lower bounds of an entailed sentence.

4. All of the sentences are represented on independent axes, thus allowing no possibility of structuring correlation information between sentences. That is, there is no mechanism for including conditional probability information.

These problems would have to be solved by any calculus of reasoning with uncertainty which attempted to reason with probabilities ranging over all logically possible worlds. It is with these problems in mind that the new interpretation table of section 4.4 is proposed.

4.4 The New Interpretation Table

A new model for probabilistic entailment is proposed, which allows the option of specifying a full probabilistic model if one is available. Another option is the use of a partial model to narrow the probability bounds produced in the probabilistic entailment. The basic premise is that every world which is generated twice in the base set is only represented once in the interpretation table. Consider interpretation table 4.3. In this new layout the question arises

Sentence	a	b	c	Probability
τ	1	1	1	1
A_1	1	1	0	π_1
$A_1 \Rightarrow B$	1	0	1	π_R

Table 4.3: Interpretation Table for (A_1, $A_1 \Rightarrow$B).

of what to do with worlds (e.g. world c), where the entailed sentence can be either true or false.

An attractive aspect of the distribution shown in equations 4.1 is that the variables c and d can be treated separately, and therefore assigned independently. In the case where the last equation to be solved is $c + d = 1 - \pi_R$, computationally speaking, the most obvious way to resolve this problem is to assign half of this value to each variable, as Nilsson proposed. This leads to the unacceptable consequences discussed above.

4.4.1 The Context Split

The context split is introduced as a vehicle for assigning certainty to worlds in which the conclusion can be either true or false. Each uncertain world in which a split can be applied provides the contextual information necessary to estimate the certainty of the conclusion in this world, hence it is named the "context" under which the split is applied.

In the new model the probability of B (of table 4.3) is all of the value of a, plus some proportion x of the value of c, where x indicates how likely it is that B will be true in the context of A_1 being false, and $A_1 \Rightarrow$B being true.

There are three ways in which a *context split* can be assigned:

- the conditional probability of the conclusion in the light of the context can be used. This use of conditional probability is different to the way it is used in Bayesian Inference, (that is p($B|context$) rather than p($context|B$)). The relationship between Bayesian Inference and Nilsson's probabilistic logic is further discussed in chapter 6.

- a subjective probability (section 2.6) estimate from the expert may be used. It may be that the expert feels that when A_1 is false, there is little chance of B being true, and may therefore wish x to be small.

- a heuristic measure may be used. One way of estimating x would be to use the ambient prior probability of B. Another way would be to base the calculation of x on how many conditions are being met in the context in which x is being applied. Another way would be just to assume that x is 0.5, and force equal probabilities between worlds ($\sim A_1$, $A_1 \Rightarrow$B, B) and ($\sim A_1$, $A_1 \Rightarrow$B, $\sim$B) as in table 4.1.

These assignment techniques are further discussed in chapter 8 and examples of the use of heuristic measures are shown in chapter 10.

4.5 Interpretation Tables for Larger Semantic Trees

Consider table 4.4, which shows a rule with two antecedents, and whose equations are:

Sentence	a	b	c	d	e	Probability
τ	1	1	1	1	1	1
A_1	1	1	0	0	1	π_1
A_2	1	1	0	1	0	π_2
$A_1 \& A_2 \Rightarrow B$	1	0	1	1	1	π_R

Table 4.4: Interpretation Table for (A_1, A_2, $A_1 \& A_2 \Rightarrow$ B)

$$\begin{aligned} a+b+c+d+e &= 1 \\ a+b+e &= \pi_1+\pi_R-1 \\ a+b+d &= \pi_2+\pi_R-1 \\ a+c+d+e &= \pi_R \end{aligned} \tag{4.5}$$

This example demonstrates a further convention employed throughout the text. When drawing a protracted interpretation table for an entailment, the worlds are ordered in a particular way. The world with all of the antecedents true, and the rule true is drawn in the leftmost column, (and is consequently always labelled a); the world with all of the antecedents true and the rule false is always next to this world, (and is therefore always labelled b); all the other worlds are to the right of this world.

This convention makes it possible to visualise the two most important worlds with respect to the rule itself (that is, the world in which all antecedents are true and the rule is true; and the world where all antecedents are true and the rule is false). World b can always be assigned immediately because it is the only allowable extension of the uncertain sentences in which the rule can demonstrably be proven to be false (section 4.6). That is, for any uncertain rule (whose probability is greater than zero, and less than one),

$$b = 1-\pi_R$$

so the sentences become:

$$a+c+d+e = \pi_R \tag{4.6}$$

$$a + e = \pi_1 + \pi_R - 1 \tag{4.7}$$

$$a + d = \pi_2 + \pi_R - 1 \tag{4.8}$$

that is: n+1 equations with 2^n possible worlds to solve for. When the strength of the rule is 1 the number of unknown variables remains the same because world b is the only world in this interpretation table which becomes an impossible world.

4.6 Semantic Tree Case Analysis and Probabilistic Entailment

In this section I show that $2^n + 1$ possible worlds are created for an entailment of the form $A_1 \& A_2 \& ... A_n \Rightarrow$B, where n is the number of propositions in the antecedent list of the rule. This produces n+1 equations and 2^n possible worlds to solve for.

4.6.1 Propositional Calculus

Firstly consider entailment in the propositional calculus of the set:

$$\{A_1, A_2, ..., A_n,\ A_1 \& A_2 \& ... A_n => B\}, \{B\}. \tag{4.9}$$

. The three cases are:

1. All $A_1, A_2, ..., A_n$ true, and rule true. With the rule clause having all the $A_1, A_2, ...A_n$ negated, the rule is continuously resolved away by each literal, eventually releasing the literal B. Only the inclusion of $\sim B$ in the set, would produce a contradiction.

2. All $A_1, A_2, ..., A_n$ true, and rule false. The negation of the rule means that n+1 clauses replace the rule, where all of the $A_1, A_2, ...A_n$ will be true, and the consequent is false. The inclusion of B in the set, would produce a contradiction.

3. At least one of $A_1, A_2, ..., A_n$ false, and rule true. The literals in the rule cannot all be resolved away from the premises, and so no statement can be made about B from the rule. Consequently, either B or $\sim B$ will be consistent with the set. The number of worlds produced is $2^n - 1$. (i.e. only removing the all true world from the list of possibilities.)

There is no analogous case for case 3 where the rule is false. This is because, the rule will split into n+1 clauses, with $A_1, A_2, ...A_n$ all true; and B false. Consequently, if any of the $A_1, A_2, ...A_n$ premises are false, a contradiction is immediately produced.

4.6.2 Predicate Calculus

Now consider the predicate calculus case:

$$\{\exists A_1(x_1), ..., \exists A_n(x_n), \forall(x_1, ..., x_n).A_1(x_1)\&...\&A_n(x_n) => \ B(x_1, ..., x_n)\}, \{\exists B(a_1, ..., a_n)\}. \tag{4.10}$$

When the conclusion is negated, the clause produced is: $\sim B(v1, v2, ..., vn)$, where $v_1, ..., v_n$ are variables. When it is just simplified the clause produced is: $B(g1, g2, ...gn)$ where $g_1, ..., g_n$ are constants. The cases are:

1. All $\exists A_1(x_1), ..., \exists A_n(x_n)$ true, and rule true.

   When all of the antecedent propositions are true this produces the clause list $A_1(c_1), ..., A_n(c_n)$ where the $c_1, ..., c_n$ are constants. The rule being true gives $\sim A_1(x_1), .. \sim A_n(x_n)$ where the $x_1, ..., x_n$ are variables. From these we can resolve away to produce from the rule clause: $B(c1, c2, ..., cn)$ and the inclusion of $\sim B(v1, v2, ..., vn)$ produces nil.

2. All $\exists A_1(x_1), ..., \exists A_n(x_n)$ true, and rule false. The rule converts to n+1 clauses: the $A_n(c_n)$, for i=1 to n, where c_n is a unique constant for each predicate functor A_n; plus a final clause which is $\sim B(c_1, ..., c_n)$. The inclusion of $B(g_1, ..., g_n)$ would not produce nil, and neither would the inclusion of $B(v_1, ..., v_n)$.

3. At least one of $\exists A_1(x_1), ...\exists A_n(x_n)$ false, and rule true. A directly analogous case to case 3 above. The rule does not free any information about the consequent, and so the second set can be true or false. Again we get $2^n - 1$ possible worlds.

4.6.3 Results of Case Analyses

The nature of these worlds is listed below, and for completeness, the effects of the worlds on the conclusion B is shown in brackets.

1. All $A_1, A_2, ...A_n$ true, Rule true. (B true.)
2. All $A_1, A_2, ...A_n$ true, Rule false. (B false.)
3. At least one of $A_1, A_2, ...A_n$ false, Rule true. (B can be true or false.)

In a predicate calculus rule of the form:

$$\forall(x1, x2, ..., xn).A_1(x_1), ...A_n(x_n) \Rightarrow B(x_1, ..., x_n) \tag{4.11}$$

with antecedent predicates existentially quantified, the same number of worlds is produced in direct analogy with the above cases, except that in case 2 the conclusion can be either true or false. So, in the predicate calculus the expert is given the opportunity of providing another context split. Throughout the thesis points about Probabilistic Logic will be drawn from the propositional rather than the predicate calculus, on the understanding that such examples are easily generalised to first order predicate calculus, and that the system can equally well cope with the generalised predicate calculus rule, given the extra context split for the world where the antecedents are all true; and the rule is false. This convention is adopted purely for typographical convenience.

Consistency in the semantic tree.

The implication rule imposes a consistency relation on the probabilities of the premises $A_1, A_2, ...A_n$. Namely, that in the world where the rule is false, (case 2), all of the premises are true. This logical necessity imposes the following probabilistic constraint:

For any probabilistic rule of the form $p(A_1 \& A_2 \& ...A_n \Rightarrow B) = \pi_R$, the probabilities of the premises are consistent if and only if:

$$\forall i.(i = 1...n)\ p(A_i) > 1 - \pi_R \tag{4.12}$$

Simply because the probability of the world where the rule is false is $(1 - \pi_R)$, and this is a possible world shared by all the premises. Consequently, the probability assigned to each premise must be at least this large, with some probability left over to be dispersed amongst its other possible worlds.

4.7 New Method of Entailment

The use of a context split on the reduced semantic tree suggests the following new method of entailment to infer probabilistically from the set $A_1, A_2, ..., A_n, A_1 \& A_2 \& ... A_n \Rightarrow B$ to the set B , where $\pi_1, \pi_2, ... \pi_n$ are the probabilities of the premises $A_1, A_2, ..., A_n$, and π_R is the probability of the rule:

1. Make the semantic tree for the base set to find the number of possible worlds.

2. Assign the probabilities $\pi_1, \pi_2, ..., \pi_n, \pi_R$ amongst the worlds consistently.

3. Each of these worlds provides a context in which to test the consistency of B. Find in which of these worlds B can be true, false, or either. Allow the expert to provide context splits for the worlds where B can consistently be either true or false.

The entailed probability of B is then the sum of the probabilities of all the worlds in which it can only be true, plus the respective context split proportions of the worlds where it can be either true or false.

It is important to note that this model is peculiar to the rule of entailment, and as such bears some differences with regard to the standard Bayesian probabilistic model. In most basic terms, the entailment procedure does not need knowledge of the prior probabilities for each of the antecedent sentences (c.f. the problems of Bernoulli and Laplace which led ultimately to the development of the frequency theory of probability). It also does not need to work back from the singularity of all antecedents known with certainty to be in one state or another to approximate an uncertain condition in the state of the antecedents, and consequently to approximate the uncertain state of the inference (c.f. the PROSPECTOR approximation scheme [33]).

4.8 A New Algorithm to Produce The Absolute Bounds of an Entailment Problem

Grosof [51] has examined the consequences of Probabilistic Logic as a reasoning tool which performs the job that Dempster-Shafer theory purports to perform: namely the provision and

manipulation of bounded probabilities.

However, a strong point in favour of the Dempster-Shafer theory of beliefs and of the corresponding orthogonality principle [113, 30], is its ease of implementation, and the tractability of the orthogonality function. However, it has been noted that the semantic inconsistencies which can be produced, in particular when conflicting evidences are combined, make the theory unreliable [97, 108].

I present a new algorithm for producing the bounds of a probability entailment in Probabilistic Logic, which does not require the tracing of the vertices of a multi-dimensional convex hull [89].

The formula can be expressed as follows. For any probabilistic rule of entailment of the form: $p(A_1 \& A_2 \& ... \& A_n \Rightarrow B) < 1$. The probabilities are labelled $1, \pi_1, \pi_2, ..., \pi_n, \pi_R$ such that 1 is the probability of the tautology; $\pi_1, ..., \pi_n$ are the probabilities or propositions $A_1, ..., A_n$; and π_R is the probability of the rule. The bounds of the entailment are given by the following expression:

$$max(0, \pi_R - \sum_{i=1}^{n}(1 - \pi_i)) \leq p(B) \leq \pi_R, \tag{4.13}$$

4.9 Inductive Proof of Bounds Algorithm

In the semanic tree for a set of sentences (c.f. tables 4.3 and 4.4) the maximum which can possibly be assigned to the worlds with all sentences true is the value π_R. This assignment is made when all of the antecedent sentences are assigned a probability of 1, producing two worlds in the semantic tree — one with all antecedents true and the rule true (and consequently the conclusion true), and the other with all antecedents true and the rule false (and consequently the conclusion false). The former world is assigned a probability π_R, the latter is assigned a probability of $1 - \pi_R$. We are interested in the former world. The proof proceeds by showing how this upper limit can be maximally reduced in this world (with all of the antecedents true and the rule true) so that it holds the minimum probability possible for that world. This value is then the minimum probability which can be assigned to the conclusion (as it would be if all of the context splits were 0). The maximum probability

of the conclusion is always π_R because the context splits for each of the worlds could be 1, (totally in favour of the conclusion), making the probability assigned to it π_R.

Base

The base case, where n=1, is as shown in table 4.3. From this table we get the equations:

$$b = 1 - \pi_R \tag{4.14}$$

$$c = 1 - \pi_1 \tag{4.15}$$

$$a = \pi_R - (1 - \pi_1) = \pi_R - (1 - \pi_1) \tag{4.16}$$

These are the only solutions to the equations, and they satisfy the algorithm with n = 1.

Step

The algorithm is true for n antecedents, now to prove it true for n+1 antecedents.

The old tree for n antecedents can be separated into:

1. one world with the rule false and all other antecedents true — the probability of this world is $1 - \pi_R$ (c.f. equation 4.9 case 2 and equation 4.12); and

2. 2^n worlds with the rule true and all possibilities of truth values attached to the antecedents (c.f. equation 4.9 cases 1 and 3). The sum of the probabilities of these worlds is therefore π_R.

To add in the new antecedent, take the world with the rule false, (1 above), and add the new antecedent after the nth in this list with a value of true (c.f. equation 4.12). This is the only world with the rule false. Since A_{n+1} is true in this world the probability which is available to assign among the other worlds (with the rule true) is reduced by the amount $(1-\pi_R)$. This means that only $\pi_{n-1} - (1-\pi_R)$ ($= \pi_{n+1} + \pi_R - 1$) is available to be assigned amongst these other possible worlds.

Next take the old tree with the rule true, (2 above), and make two copies of this. In the first add the new premise A_{n+1} = true after premise A_n in the premise list. In the second add the new premise A_{n+1} = false after premise A_n in the premise list. The difference between

the tree for n+1 propositions and n propositions is that in row n+1 there are now 2^n 1's and 2^n 0's, and the rule is pushed down to position n+2. The new tree is made up of two identical copies of the old tree, one of which has a 1 in row n+1, the other of which has a zero in row n+1. As described in section 4.5 the number of worlds for $n+1$ antecedents is therefore:

$$2 * 2^n + 1 = 2^{n+1} + 1$$

There are three cases to be considered:

1. Even after adding the new antecedent A_{n+1} the new expression $\pi_R - \sum_{i=1}^{n+1}(1-\pi_i)$ is still greater than 0.

2. Before adding the new antecedent A_{n+1} the new expression $\pi_R - \sum_{i=1}^{n}(1-\pi_i)$ is less than or equal to 0.

3. Before adding the new antecedent A_{n+1} the expression $\pi_R - \sum_{i=1}^{n}(1-\pi_i)$ was greater than 0 and after adding A_{n+1} the expression $\pi_R - \sum_{i=1}^{n+1}(1-\pi_i)$ is less than or equal to 0.

Case 1: $\pi_R - \sum_{i=1}^{n+1}(1-\pi_i) > 0$

Add π_{n+1}, A_{n+1}.

1. set prob := $\pi_{n+1} + \pi_R - 1$ (this is the probability to be assigned to the possible world set).

2. take the old tree, and arrange its probabilities in increasing size.

3. take the world with all antecedents true and the rule true, make the antecedent A_{n+1} false in this world and assign it the probability $1 - \pi_{n+1}$. Make the world with all of the old antecedents true and A_{n+1} true and assign this the residual of:

$$\pi_R - \sum_{i=1}^{n}(1-\pi_i) - (1-\pi_{n+1}) = \pi_R - \sum_{i=1}^{n+1}(1-\pi_i)$$

4. For each of the other worlds make A_{n+1} true in the world, and assign it the value it had before.

5. The total probability assigned to A_{n+1} is then:

$$\sum_{i=1}^{n}(1-\pi_i) + (\pi_R - \sum_{i=1}^{n}(1-\pi_i)) - (1-\pi_{n+1}) = \pi_R + \pi_{n+1} - 1$$

which is the required probability when $1-\pi_R$ is subtracted from the probability π_{n+1}.

Since the tree for n antecedents has been correctly filled, this new tree, which maintains all of the probability assignments of the old also has a correct assignment for the tree extended by predicate A_{n+1}.

Case 2: $\pi_R - \sum_{i=1}^{n}(1-\pi_i) \leq 0$

In this case it is necessary that there is a way of assigning the probability for antecedent A_{n+1} without assigning probability to the world where all of the $n+1$ antecedents are true and the rule is true.

1. set pset := $\pi_{n+1} + \pi_R - 1$ (this is the probability to be assigned to the possible world set).

2. take the old tree, and arrange its probabilities in increasing size.

3. choose the world from the world set which has the smallest non-zero probability. Call this world pw, whose probability is w_x.

4. if (pset - w_x) > 0 then pset := pset - w_x; make two new worlds from the world pw with a new space for predicate A_{n+1} such that p(A_{n+1} true, pw true) = w_x and p(A_{n+1} false, pw true) = 0. Return to step 3.

5. if (pset - w_x) < 0 then prest := w_x - pset; resid := w_x - prest; make two new worlds from the world pw with a new space for predicate A_{n+1} such that p(A_{n+1} true, pw true) = prest and p(A_{n+1} false, pw true) = resid.

Make two copies of each of the other worlds. To one of the copies add the sentence A_{n+1} false, and assign this world the probability w_x, where w_x was the probability of the original. To the second add the sentence A_{n+1} true, and assign this the probability 0. Stop.

Since the tree for n antecedents was built without the use of the world where all antecedents are true, this new tree has been constructed without use of this world, and so, consequently it has a probability of 0. The sentence A_{n+1} has been added to the previous set with probability π_{n+1} in a consistent manner.

Case 3: $\pi_R - \sum_{i=1}^{n}(1-\pi_i) > 0$; $\pi_R - \sum_{i=1}^{n+1}(1-\pi_i) \leq 0$

It is necessary to show that in this case, the bounds can be reduced to 0 to π_R and that the probability of A_{n+1} can be assigned consistently. Add π_{n+1}, A_{n+1}.

1. set pset := $\pi_{n+1} + \pi_R - 1$ (this is the probability to be assigned to the possible world set). Also, set pnot := $1 - \pi_{n+1}$ (the probability of A_{n+1} being false).

2. take the old tree, and create the assignment which maximally reduces the alltrue world (the world with all antecedents true and the rule true). The probability of the alltrue world is therefore:

$$\pi_R - \sum_{i=1}^{n}(1-\pi_i)$$

   Call this prall.

3. Let plast := prall, and prem := $(1 - \pi_{n+1}) - plast$.

4. Take the previous alltrue world, and make A_{n+1} false in this world with probability plast. We still need to use up probability prem where A_{n+1} is false, as well as assign A_{n+1} true the probability $\pi_{n+1} + \pi_R - 1$.

5. set prob := prem

6. take the old tree, and arrange its probabilities in increasing size.

7. choose the world from the world set which is the smallest non-zero probability. Call this world pw, whose probability is w_x.

8. if (prem - w_x) > 0 then prem := prem - w_x; make two new worlds from the world pw with a new space for predicate A_{n+1} such that p(A_{n+1} false, pw true) = w_x and p(A_{n+1} true, pw true) = 0. Return to step 7.

9. if (prem - w_x) < 0 then prest := w_x - prem; resid := w_x - prest; make two new worlds from the world pw with a new space for predicate A_{n+1} such that p(A_{n+1} false, pw true) = prest and p(A_{n+1} true, pw true) = resid.

10. For all other worlds pw, p(A_{n+1} true, pw true) = w_x (the probability of pw in the assignment for n antecedents) and p(A_{n+1} false, pw true) = 0. Stop.

We have found a way of assigning worlds in the old tree where A_{n+1} is false the value of $1 - \pi_{n+1}$, and have assigned the worlds where A_{n+1} is true a value of $\pi_R - (1 - \pi_{n+1})$, i.e. $\pi_{n+1} + \pi_R - 1$ without using the all true world.

4.10 Consequences of the Boundary Algorithm

The result makes it possible to provide the absolute bounds of an entailment without resorting to the multi-dimensional projection method provided by Nilsson in his derivation of probabilistic entailment. However, as Nilsson reports, the bounds of an entailment rapidly widen as the probabilities of propositions drop away from 1. Very soon, the bounds of a proposition are the simple lower limit of 0 and upper limit of π_R.

However, using boundary information in conjunction with conditional probabilities provided by an expert we can narrow these bounds. It may not be practical for an expert to provide all of the relevant conditional probabilities; but the expert may be able to provide some of the important ones, in which case, the bounds of the deduced probability may be reduced. If the expert can provide the conditionals: p($B|C1$), p($B|C2$), ... , p($B|Cn$), where C1, C2, ... ,Cn are contexts in which at least one of the antecedents are false, and the rule is always true, then, from equation 4.13, the bounds may be succesively altered:

$$L_n = L + p(B|C1) \tag{4.17}$$

$$U_n = U + p(B|C1) - 1 \tag{4.18}$$

And so with each successive application of the context splits, the absolute bounds will be narrowed. This procedure simply makes use of the fact that p($B|C1$) + p($\sim B|C1$) = 1. The

use of this information in conjunction with a reduction in entropy (uncertainty) is considered in chapter 8.

4.11 Conclusion

In this chapter several criticisms have been made against the structure proposed by Nilsson. These criticisms are: the inability to use a complete probability model, and the conservative estimate of probability using the half split. To answer these criticisms a new model for entailment has been proposed which allows the inclusion of conditional probabilities, in the form of *context splits*. The context split notion is introduced as a vehicle for supporting conditional probabilities in Probabilistic Logic, but it can also be used to support subjective heuristic estimates of how often a conclusion may be inferred in a particular possible world.

Making the logic of conditional probabilities available to the entailment structure allows Bayesian Inference within Probabilistic Logic, a topic which is further explored in chapter 6. Furthermore, if the context split is a heuristic function, we can use partial information to help the reasoning process to proceed. The better the heuristic function, the more accurate the results. This topic is further discussed in chapter 8.

We have also introduced a new algorithm for estimating the probability bounds of an entailment process. We are now left facing the problem of underdetermination of the actual state of the antecedents. With this in mind, we must model a probability distribution through our uncertainty space. The most acceptable such distribution is provided by the maximum entropy formalism, which is examined in chapter 5.

Chapter 5

THE MAXIMUM ENTROPY FORMALISM IN NILSSON'S PROBABILISTIC LOGIC

5.1 Introduction

In the previous chapter the structural aspects of Probabilistic Logic were developed to include conditional probabilities. In this chapter, the maximum entropy formalism is investigated with relation to choosing the most probable probability distribution for a set of uncertain sentences. In this sense, we are concerned with the information theoretic aspects of the maximum entropy formalism.

Maximum entropy, as applied in information theory, is concerned with the semantic content of a message passed between a transmitter and a receiver [114]. Typically, if the receiver receives the message as it was broadcast by the transmitter, then the receiver is said to have received perfect information from the process. In the theory of communications, a signal is altered from some human readable form at the transmitter's end, into some transmittable form, then transmitted, and then reconverted into human readable form at the receiver's end. The most likely points for message degradation are at the conversion stages, and the transmission stage. This process has a direct analogy in expert systems.

Where expert systems are concerned, antecedent information is transmitted to the rule of inference which then transmits information about the rule's consequent with some certainty value. That is, when all of the antecedents attached to the rule are correct, then the conclusion may be drawn with the predefined certainty attached to the rule. In this case perfect information has been transferred.

However, degradation arises in this process when any, or all, of the antecedents is, or are, uncertain. In this situation, information transfered to the rule has suffered degradation. The maximum entropy principle has been shown to give the least commitment [115] probability distribution subject to the probability constraints of the antecedents and the rule, to the entailed conclusion. The summation function which has been used to measure the entropy of a probability distribution has three simple properties which make it well suited to be a measure of the information content of the distribution [60, 54, 3]. For illustration purposes, consider an example of a set of n mutually exclusive possible worlds, and consider all the possible ways probability can be assigned to these possible worlds so that it sums to 1. The function used to measure the entropy of the distribution is $H = \sum_{i=1}^{n} p_i log p_i$.

The first property is that it does not matter in which order the summation of the individual terms ($p_i log p_i$) is applied; the function always gives the same result. The second is that the function is at a maximum when each of the possible worlds p_1 to p_n are assigned a probability of $1/n$, the point of maximum dispersion of probability throughout the sample space. The third is that the function is continuous and reduces monotonically from this maximum, towards a value of 0 at the point where one of the possibilities p_1 to p_n is assigned a value of 1; and all of the others are assigned a probability of 0. For probability distributions whose entropy value is between these two extremes, the fact that the function is necessarily monotonically decreasing means that, as probabilities are applied to the possibilities the information content in each of the intermediate states can be compared.

For these reasons the entropy function has been used throughout this thesis as the measure of the information content within a probability distribution, with a view to reasoning with uncertainty in a mathematically justifiable way.

In this chapter the Lagrangian derivation of the maximum entropy distribution is given,

and we examine the way the maximum entropy principle may be given a deeper role in Nilsson's probabilistic logic, which ultimately allows the user to specify a complete probabilistic model for the entailment process if one is available. A new solution to the maximum entropy equations for probabilistic entailment is derived which provides the correct values for the non-linear factors of the probability equations in a time small enough so that the mechanism of maximum entropy can easily sit within an expert system inference engine without causing intractable complexity problems.

5.2 Derivation of The Maximum Entropy Solution

The entropy equations will be derived within the matrix framework of Nilsson's probabilistic logic, as described in chapters 3 and 4. The Entropy Equation is approximated by a new function H' which is written:

$$H' = -\sum_i p_i log p_i + l_1(\Pi_1 - R_1 P) + l_2(\Pi_2 - R_2 P) + ...l_n(\Pi_n - R_n P) \qquad (5.1)$$

The variables l_1 to l_n are Lagrange multipliers. The variables R_1 to R_n are the rows of the V′ matrix. The variables Π_1 to Π_n are the rows of the probability matrix Π. P is the matrix with the probabilities of the possible worlds. H' is exactly the entropy function when there is a probability assignment to each of the possible worlds which meets the marginal probabilities provided by the expert.

We concern ourselves with each world individually (that is, each of the columns in V′, and its corresponding probability in P). The new entropy expression is differentiated with respect to each world, setting the result to zero to derive the distribution with maximum entropy, giving for each world i:-

$$-(log p_i + 1) - l_1 v_{1i} - .... - l_n v_{l_n} = 0 \qquad (5.2)$$

which can be written:

$$log p_i = -1 - l_1 v_{1i} - .... - l_n v_{l_n} \qquad (5.3)$$

$$\Rightarrow p_i = e^{-1} e^{-l_1 v_{i1}} ... e^{-l_n v_{l_n}} \qquad (5.4)$$

The aggregate factors of this multiplicative list can be simplified, in the sense that the column elements of V′ for possible world i (v_i1 to v_in) may take values 1 or 0. If the value of a row x of world i is zero (i.e. $v_{xi} = 0$), we note that the expression for p_i loses the term $e^{-l_x v_{ix}}$, which simply evaluates to 1 in the expression. If the value $v_{xi} = 1$, then the exponential for dealing with row x of world p_i can be simplified to e^{-l_x}. With these observations in mind, the following definitions can be used to simplify the expression:

$$a_1 = e^{-1}e^{-l_1} \tag{5.5}$$

$$a_j = e^{-l_j}, \ (for\ j = 2\ to\ n); \tag{5.6}$$

and, consequently each possible world p_i can be written as a product of some of the aggregate factors a_1 to a_n, where the factor a_x is included in the multiplication list only if there is a 1 at position v_{xi} in the V′ matrix.

Now, the maximum entropy equations only require a solution for the factors a_1 to a_n; and this being done, the maximum entropy probability for each possible world can easily be reconstructed. The problem is that although for our n sentences we have n unknowns, the equations are nonlinear multiplications of these unknowns, and typically require solution by iteration.

5.3 Entropy Equations

As n increases, the difference between these numbers will increase rapidly, introducing $(2^n - n)$ extra degrees of freedom. One way to remove the additional degrees of freedom is to maximise the entropy of the system [60, 16, 2].

In this approach each possible world is rewritten in terms of a multiplication of aggregate factors [3, 16, 89]. The notation used to associate these factors with the corresponding sentences for a rule with n antecedents is as follows: a_τ represents the factor for the tautology; the factors a_j are associated with propositions A_j, for j equal 1 to n, and factor a_R is associated with the entailment rule. An aggregate factor is included in the multiplication list for a

possible world only if the factor's associated sentence is true in that world. So, for table 4.3, we have:

$$a = a_\tau a_1 a_R; \quad b = a_\tau a_1; \quad c = a_\tau a_R \tag{5.7}$$

and the equations are rewritten as shown in table 5.1.

1.	$a_\tau a_1 a_R + a_\tau a_1 + a_\tau a_R =$	1
2.	$a_\tau a_1 a_R + a_\tau a_1 =$	π_1
3.	$a_\tau a_1 a_R + a_\tau a_R =$	π_2

Table 5.1: M.E. Probabilistic Equations for Table 4.3

The equations which need to be solved from table 4.1 are shown in table 5.3.

1.	$a_\tau a_1 a_R + a_\tau a_1 + 2a_\tau a_R =$	1
2.	$a_\tau a_1 a_R + a_\tau a_1 =$	π_1
3.	$a_\tau a_1 a_R + 2_\tau 1 a_R =$	π_2

Table 5.2: M.E. Probabilistic Equations for Table 4.1

Once the entropy equations for table 4.3 are solved to give the required aggregate factors, the probability of B is simply $a_\tau a_1 a_R + a_1 a_R$. In the case of table 5.1, the probability of B is $a_\tau a_1 a_R$ plus the context split proportion of the value $a1a_R$. The provision of this context split will never lead to any difficulties, because in the worst possible case, the system can automatically assign the size of the split, using prearranged rules agreed with the knowledge engineer.

5.4 The Iterative Method of Solution

In Probabilistic Logic and Bayesian Inference when there are n uncertain sentences and hence 2^n different possible worlds, this is the number of *degrees of freedom* [73] of the system. Using the maximum entropy method the probabilistic equations relating to n uncertain sentences can always be rewritten in terms of n aggregate factors, and hence the degrees of freedom are reduced to only n.

However, when we use the maximum entropy formalism to remove the additional degrees of freedom of an entailment problem, it is still necessary to solve the non-linear entropy

equations. Such equations are solved iteratively if no generic pattern can be found within them. The update method of solution [16], is a particular case of the general one-point method of solution which covers all iterative methods of the form:

$$a_n = F(a_c) \tag{5.8}$$

where the next iterative approximation of a value (a_n) is found by applying some function (F) to the current value of a variable (a_c). The solution is found when none of the variables in the equations move by more than a predefined amount (ϵ) from their current approximation. This is the method Nilsson suggested for solving the entropy equations for entailments involving small numbers of antecedents. The method can be expressed as follows.

1. Number the equations 1...n.
2. For each a_i solve equation i in terms of the other variables.
3. Assume initial values for each of the a_i.
4. Choose the next a_i, and recalculate it in terms of the others.
5. If the change in any of the a_i's is more than ϵ then continue from 4.
6. Otherwise stop with success.

Where ϵ represents the tolerance in the approximation. This method will always converge to a solution, but the time taken is dependent on two things:

1. the number of a_i's to be solved for, and

2. the initial starting values for the a_i's.

Table 5.4 shows times taken for this algorithm. As with all of the reported times, the hardware is a High Level Hardware 'Orion' with 8 megabytes of memory, and all programs have been written in C-Prolog [19]. The first column shows how many antecedents are involved in the rule, and the second shows how many cpu seconds the program took to provide a stable result with ϵ equal to 0.1%.

The accuracy achieved may be improved by making ϵ smaller, but this vastly increases the time taken for solution. Paris et al [91], show that the problem of computing these factors to a reasonable accuracy is NP-hard, and consequently such methods are probably unfeasible. In fact, Pearl [97] dismisses Nilsson's use of the maximum entropy formalism within the Probabilistic Logic because "computational techniques for finding a maximum-entropy distribution are usually intractable". An algorithm is presented in section 5.5 which

Antecedents	Time (cpu secs)
1.	2.883
2.	12.283
3.	45.2
4.	150.52
5.	475.33

Table 5.3: Times for Iterative Solution

will discover the factors for the extended system of Probabilistic Logic introduced in chapter 4.

5.5 The New Algorithm for Solving the Maximum Entropy Equations

The algorithm can be expressed as follows. For any probabilistic rule of entailment of the form: p(A_1&A_2&...&A_n $\Rightarrow$ B) < 1, the corresponding aggregate factors are a_τ, a_1, $a2$,... a_n, a_R such that a_τ is for the tautology, a_1... a_n are for the propositions A_1... A_n; and a_R is for the rule of entailment. The probabilities are labelled $1, \pi_1, \pi_2, ..., \pi_n, \pi_R$ such that 1 is the probability of the tautology; $\pi_1, ..., \pi_n$ are the probabilities of propositions $A_1, ..., A_n$; and π_R is the probability of the rule. The solution is as follows:

$$a_{j(j=1,...,n)} = \frac{\pi_j + \pi_R - 1}{1 - \pi_j} \tag{5.9}$$

$$a_\tau = \frac{1 - \pi_R}{\prod_{j=1}^{n} a_j} \tag{5.10}$$

$$a_R = \frac{\pi_R}{a_\tau \prod_{j=1}^{n} (1 + a_j)}; \tag{5.11}$$

5.5.1 Inductive Proof of Entropy Algorithm

The proof proceeds in four stages. First, to derive the expression for the world where the rule is false. Second, to show that for each of the terms a_j ($j = 1$ *to* n) there is a direct match of terms on the numerator and denominator of the expression: $(\pi_j + \pi_R - 1)/(1 - \pi_j)$, i.e. all the unknown worlds where sentence A_j is true divided by all the worlds where A_j is false. The third stage is related to the first and allows us to solve for a_τ. The fourth stage

Sentence	Possible Worlds			Equations	
τ	1	1	1	$a_\tau a_1 a_R + a_\tau a_1 + a_\tau a_R =$	1
A_1	1	0	1	$a_\tau a_1 a_R + a_\tau a_1 =$	π_1
$A_1 \Rightarrow B$	0	1	1	$a_\tau a_1 a_R + a_\tau a_R =$	π_R

Table 5.4: Sentences, Worlds, and Equations

is for the final factor a_R and is based on the worlds in which the rule is true, and a recursive expression for describing the contribution of each of the possible worlds to this probability: $a_\tau a_R \prod_{j=1}^{n} (1 + a_j) = \pi_R$.

Base Case (n=1)

From the equations of table 5.4:

$$a_\tau a_1 = 1 - \pi_R \tag{5.12}$$

$$a_1 = \frac{a_\tau a_1 a_R}{a_\tau a_R} = \frac{\pi_1 - (1 - \pi_R)}{(1 - \pi_1)} = \frac{\pi_1 + \pi_R - 1}{1 - \pi_1} \tag{5.13}$$

$$a_\tau = \frac{a_\tau a_1}{a_1} = \frac{1 - \pi_R}{a_1} \tag{5.14}$$

$$a_R = \frac{a_\tau a_1 a_R + a_\tau a_R}{a_\tau a_1 + a_\tau} = \frac{a_\tau a_R (1 + a_1)}{a_\tau (1 + a_1)} = \frac{\pi_R}{a_\tau (1 + a_1)} \tag{5.15}$$

And the above equations satisfy the algorithm with n = 1.

Step

The algorithm is true for n antecedents, now to prove it true for n+1 antecedents.

The new premise A_{n+1} is added to the antecedent arm of the rule, and placed after premise A_n in the premise list. We now have aggregate factors $a_\tau, a_1, ..., a_{n+1}, a_R$.

1. $a_\tau a_1 ... a_{n+1} = 1 - \pi_R$

2. In each row there are now 2^{n+1} possible worlds, where there used to be 2^n. The difference between the tree for n+1 propositions and n propositions, being that in row n+1 there are now 2^n 1's and 2^n 0's, and the rule is pushed down to position n+2.

For the half of the tree with 0's in row n+1 we proved that there is a direct match to give each of the previous a_i's. For the other half, we use the same enumeration, and find that the factor for proposition n+1 cancels out on top and bottom. Furthermore the numerator still only holds the worlds where sentence A_j is true, and the denominator the worlds where A_j is false. Therefore the equation still holds.

Is the formula true for new row n+1?

The new tree was made up of two identical copies of the old tree, one of which has a 1 in row n+1, the other of which has a zero in row n+1. Consequently, again it is possible to cancel the terms of the true worlds divided by the false worlds so that there is only a factor of a_{n+1} left.

3. a_τ is trivially $\frac{1-\pi_R}{a_1 a_2 \ldots a_{n+1}}$

4. The expression for all the worlds where the rule is true is: $a_\tau a_R \prod_{j=1}^{n} (1 + a_j)$

   When we include the new row, we have a new multiplicative factor: We have two copies: one with an a_{n+1} in row n+1, and one with a 1. So the new expression for all the worlds is:

$$a_{n+1} a_\tau a_R \prod_{j=1}^{n} (1 + a_j) + a_\tau a_R \prod_{j=1}^{n} (1 + a_j) \;=\; (1 + a_{n+1}) a_\tau a_R \prod_{j=1}^{n} (1 + a_j) \qquad (5.16)$$

$$= a_\tau a_R \prod_{j=1}^{n} (1 + a_j) \qquad (5.17)$$

In the event that the probability of the rule is 1, the world where the rule is false becomes an impossible world, and consequently the rule is subsumed into the tautology. In this case there are only n+1 factors, where a_j $(j = 1\ to\ n) = \frac{\pi_j}{1-\pi_j}$, and a_τ is $\frac{1}{\prod_{j=1}^{n}(1+a_j)}$. The proof is a generalisation from steps 2 and 4 above, and is detailed in appendix C in the interests of brevity.

Consequences of the New Algorithm

Once these aggregate factors are found for any consistent probability problem, the possible worlds can be rebuilt from the appropriate multiplication of factors. Not only will we have

the probability of a conclusion but also a detailed breakdown of the probabilities of the contributing possible worlds. Each of these possible worlds are contexts for the conclusion which the user of the system may require to see before committing him/herself to a decision. Thus, the underlying nature of the probability distribution is available, and this allows the user of the system to examine the probabilities of each of the possible worlds also.

For the case where the probability of the rule is one, the rule is subsumed into the tautology, and so we lose one of the aggregate factors. We use the same reasoning on a simplified version of the semantic tree, which gives the results:

$$a_{j(j=1,\ldots,n)} = \frac{\pi_j}{1-\pi_j} \tag{5.18}$$

$$a_R = \frac{1}{\prod_{j=1}^{n}(1+a_j)}; \tag{5.19}$$

These aggregate factors are now the building blocks for the probabilities of each of the possible worlds. Their values can range between 0 and infinity (non inclusive). (infinity cannot be attained, because this would mean that a proposition antecedent was definitely true, and therefore we adopt the collapsing procedure detailed above; and 0 cannot be attained because probabilities are constrained to be above the probability of (1-π_R)). Because the factors are multiplied together to give the probability of a possible world, those factors greater than 1 will have a promotional effect on the probability of a possible world, while those factors less than 1 will have an demotional effect on the probability of a possible world.

Table 5.5 shows times for this algorithm executing on the same problem data and hardware as that introduced in section 5.4. The first column shows how many antecedents are involved in the rule, and the second shows how many cpu seconds the program took to calculate the values of all of the aggregate factors. The results can be compared with those reported in table 5.4.

Observing the formula for deriving the factor a_i associated with proposition A_i, we can see that there are three regions which the probability of a proposition can take an entailment process, whose values are derived:

$$p(A_i) < 1 - \pi_R \tag{5.20}$$

$$1 - \pi_R < p(A_i) < \frac{1-\pi_R}{2} \tag{5.21}$$

Antecedents	Time (cpu secs)
1.	0.017
2.	0.017
3.	0.017
4.	0.033
5.	0.033

Table 5.5: Times for Quick Solution

$$\frac{1-\pi_R}{2} < p(A_i) < 1 \tag{5.22}$$

In the region of equation 5.20, the probability of the antecedent proposition renders the application of the entailment rule logically inconsistent. In the region of equation 5.21, the probability of the antecedent has a debiliting effect on that of the conclusion ($a_i < 1$), and in that of equation 5.22, the probability of the antecedent has a positive effect on that of the conclusion.

The factors can also be sorted in decreasing order, such that those at the front have the most positive effect on the probability of a possible world. This list can then be split into those factors which increase the probability of a world, and those which would reduce it. Thereby providing a mechanism for reporting any number of the most probable contributory possible worlds in an entailment process.

5.6 Entropy Equations in Probabilistic Logic

The simple expressions used in this formulation have as their basis the algorithm for the quick determination of the aggregate factors of an entailment result. We wish now to generalise the procedure to make the probabilities of the possible worlds easier to calculate. We know that:

$$a_\tau a_1 ... a_n \;=\; (1-\pi_R)$$

This is the world where the rule of inference is false, and all the antecedents are true. We also know that the factor $a_\tau a_R$ is present in all of the other possible worlds (that is, in all of the other worlds the tautology is true, and the rule of inference is also true). From equations

5.10,5.11 and 5.9:

$$a_\tau a_R = \frac{(1-\pi_1)(1-\pi_2)...(1-\pi_n)}{\pi_R{}^{n-1}} \tag{5.23}$$

we also know from equation 5.9:

$$a_i = \frac{\pi_i + \pi_R - 1}{1-\pi_i} \tag{5.24}$$

If, for each possible world, we collect the sentences which are true into the set W, (and use the variables f and g to denote aggregate factors), the probability for any possible world can be directly calculated from:

$$p(world) = a_\tau a_R \Pi_{f \in W} \frac{\pi_f + \pi_R - 1}{1-\pi_f} \tag{5.25}$$

$$= \frac{\Pi_{f \in W}(\pi_f + \pi_R - 1)\Pi_{g \notin W}(1-\pi_g)}{\pi_R^{n-1}} \tag{5.26}$$

Where the probability of the rule is 1, this expression simplifies to:

$$= \Pi_{f \in W} \pi_f \Pi_{g \notin W}(1-\pi_g) \tag{5.27}$$

These expressions make it possible to evaluate the probability of each of the possible worlds by multiplying n expressions, and when the strength of the rule is less than one, dividing by one variable. This makes the algorithm deterministic [56] but when all of the worlds are to be assigned a probability, the complete process is still, necessarily, non-deterministic. With this new method of finding the probabilities of the possible worlds we are now in a position to explore the possibility of using the maximum entropy result as a tool in Meta-Level Reasoning within a reasoning task. The complexity of this algorithm is examined in the next section. With this new method of finding the probabilities of the possible worlds we are now in a position to explore the possibility of using the maximum entropy result as a tool in Meta-Level Reasoning within a reasoning task; and this topic is explored in chapter 8.

5.7 Complexity of the Algorithms

Throughout this section n is used to refer to the number of antecedents involved in a rule of entailment.

In section 5.5 the composition of each aggregate factor is n multiplications, plus one division. The number of operations in total is n * (n + 1), which is of the order n^2. And so, the discovery of the aggregate factors may be achieved with a deterministic algorithm.

In section 5.6 each possible world is assigned a probability in n multiplications plus 1 division. To assign a value to all of the possible worlds in an entailment problem will require $2^n * (n+1)$ operations, which is exponential. However, this is a very low cost exponential equation for a process which not only works out the maximum entropy factors, but assigns each world a probability. The attractiveness of this function is best considered by remembering that any probability assignment to all of the possible worlds is necessarily exponential of the order 2^n, and the algorithm of section 5.6 is linear in that function.

It is clear that for entailments involving anything up to 30 uncertain antecedents, it is plausible to work out the probability of each individual possible world. However, the complexity problem eventually becomes insurmountable. The exponential increase of the algorithm in section 5.6, is, although slow, intractable. It is at this point that we must look for simplification strategies if we wish to have rules with more than 30 or so uncertain antecedents attached.

5.8 Conclusion

Maximum entropy solutions to probability problems have been known for some time. They generally require iterative solution by constant updating of the terms of non-linear variable equations. In this way, they are provide a formidably difficult problem to computational solution methods. The conception of merging the theory of the maximum entropy principle, and the representation schema of Nilsson's probabilistic logic looks from the outset to be an impossible task- when one wants to model all possible worlds; and the other wants to assign each of these worlds a least commitment probability value commensurate with the probability constraints given by the problem setup.

However, in this chapter, it has been shown that when Nilsson's probabilistic logic is extended to allow the inclusion of conditional probabilities, there is a polynomial time algo-

rithm for solving for the terms of the non-linear equations. These terms can be used in the reasoning process in many ways. For instance, if the probabilities of all the possible worlds is required then, the algorithm to build all of the worlds from these terms is necessarily non-deterministic. In this case, it is only feasible to build the worlds for rules of a small number (up to about 30) antecedents. However, beyond this point, (in fact more likely considerably below it), an expert will not be able to descriminate the probability fluctuations in a useful manner. In ths situation, key worlds can be evaluated, and probability bounds reduced to within an acceptable tolerance level. This topic is further explored in chapters 8 and 10.

Chapter 6

NILSSON'S PROBABILISTIC LOGIC AND BAYESIAN INFERENCE

6.1 Introduction

It has been noted by Grosof [51], that the mechanism proposed by Nilsson can use conditional probabilities insofar as that once probabilities have been assigned to possible worlds, it becomes possible to measure the intersections of groups of possible worlds and thereby produce values of conditional probabilities. The conditional probabilities are not used to form the probability distribution and are only used to describe it. This he refers to as conditional probabilities in terms of "post-construction". That is, conditional probabilities as a consequence of the probability assignment, rather than conditional probabilities used to shape the probability assignment.

He then suggests that there should be a method of representing conditional probabilities explicitly in Probabilistic Logics. With the extensions developed in chapter 4 We are now in a position to do this. When the new model is combined with the ability to quickly derive the maximum entropy probability distribution of an entailment problem, developed in chapter 5, Probabilistic Logic is in a position to tackle probability problems for which a complete

probability model is present. It is now able to solve problems that the Bayesian model solves, that is, to provide a point probability from a completely specified probability model. One advantage which Probabilistic Logic holds over Bayesian Inference is in its use of Jaynes's maximum entropy formalism to generate the most-probable probability distribution through our uncertain antecedent-rule data. This makes it possible for Probabilistic Logic to quote most-probable point probabilities for conclusions without recourse to interpolation schemas.

Both mechanisms require the same amount of pieces of information to derive a point probability [52] and the two reasoning processes are compared in this chapter. The findings are that a valid statistical meaning for probabilistic entailment is more in line with the definition of conditional probability than with a generalisation of the rule of modus ponens and Bernoulli's rule of indifference. However, the two models use different conditional information. Bayesian inference is derived from a knowledge of the hypothesis whereas Nilssonian inference is derived from a knowledge of the evidence. In this regard Nilssonian inference is more amenable than Bayesian inference to expert systems situations where the evidence may be uncertain or incomplete.

6.2 Information Cross Comparison

There are two separate issues in the comparison:

1. what is needed for a complete problem setup; and

2. what information is used to make the final deduction?

To answer the questions, consider the information shown in table 6.1 which shows the information required to specify the complete probability model for estimating the probability of event B conditioned on information about event $A1$.

Information Needed for a Problem Setup

That is, prior to the deduction, to apply Bayesian inference to find the probability of B, we need a prior probability for B (p'(B)). Bayesian inference is a method of updating probability and so it is necessary at all times to have a present probability for the deduced event. We also

Bayesian	Original Nilssonian	Extended Nilssonian
p'(B) p(A1 | B) p(A1 | ~ B)	p(A1 ⇒ B)	p(A1 ⇒ B) p(B | ~ A1)

Table 6.1: Information required by B.I. and P.L.

need two conditional probabilities (2^n for $n > 1$). So, it requires $2^n + 1$ pieces of information in setup.

Nilsson's original mechanism only asks for one piece of information — the probability of the rule of inference — but it does not produce a complete probability model. The Extended version of Probabilistic Logic requires the probability of the rule of inference and $2^n - 1$ context splits, that is, 2^n pieces of information.

Information Needed for the Deduction

To answer the second question from this example, Bayesian inference requires information about A1 — that is, it needs to know either that A1 is true, or that A1 is false. When there is uncertainty about whether or not the conditioning event has actually happened, we need an approximation schema, as, for example, PROSPECTOR's interpolation method [33], for estimating the effect of the probability of the evidence on the hypothesis.

The original model for Probabilistic Logic is shown in column 2. In this model, no conditional probability information can be explicitly encorporated. However, uncertainty in the conditioning event A1 is handled from within the model, probabilistic bounds can be produced. The original model is not able to give any result when A1 is false, which is something the Bayesian model does easily.

The extended model for Probabilistic Logic, shown in column 3, requires the strength of a probabilistic entailment plus the conditional probability of the hypothesis given that the evidence is false. Not only can the extended model give bounds when there is uncertainty, but with the use of context splits, a point probability can be produced. The meaning of probabilistic entailment and context splits are discussed in section 6.3.

The General Case

For the general case where there are n antecedents (A_1 to A_n) which have a bearing on a conclusion (or, in Bayesian terminology, a hypothesis) B, the information required by each formalism is as follows:

Bayesian Inference the prior probability of the conclusion, the conditional probability of each of the possible world in the light of the hypothesis being true

Probabilistic Logic the probability of the rule, the conditional probability of the conclusion in the light of each possible world.

It is one of the contentions of this thesis that the best form of context split (introduced in section 4.4.1) is a conditional probability (section 6.3). Since, in many cases it is not possible to give a conditional probability for a problem (see chapter 10) and in many others it is not possible to give enough conditional probabilities (see chapter 8) the more general vehicle of the context split has been introduced so that subjective probabilities and heuristics can be used. However, any subjective probability estimate given by an expert or heuristic method developed should aim to approximate the conditional probability.

6.3 Assigning Probabilistic Meaning to Entailment

Generalising the classical logic implication rule raises a number of questions, which we shall consider in this section with the aid of a dice throwing example. Suppose we have a fifteen sided dice, and proposition sentences $A1$ and B which denote the propositions 'the number on the topmost face of the dice is odd' and 'the number on the topmost face of the dice is divisible by three'. And suppose we join the two propositions with the rule $A1 \Rightarrow$B. The questions we will address with this example are:

1. What is the probability of $A1$?
2. What is the probability of $A1 \Rightarrow$B?
3. What is the probability of B?

To answer these questions, consider the sample space for the dice shown in table 6.1.

We shall consider this to be a 'fair' dice in so far as each of the possibilities is equally likely, and will assign a probability of 1/15 to each. $A1$ is true in worlds (1,3,5,7,9,11,13,15) and so

1 A1, ~B	2 ~ A1, ~B	3 A1, B	4 ~ A1, ~B	5 A1, ~B
6 ~ A1, B	7 A1, ~B	8 ~ A1, ~B	9 A1, B	10 ~ A1, ~B
11 A1, ~B	12 ~ A1, B	13 A1, ~B~	14 A1, ~B	15 A1, B

Figure 6.1: Example of a 15-sided Dice

has probability 8/15. The rule of entailment ($\sim A1 \vee B$) is true in worlds (2,3,4,6,8,9,10,12,14,15) and so has probability 10/15. To answer the final question, we will solve the equations associated with the probabilistic entailment rule with reference to the representations shown in table 6.2 and table 4.3 respectively.

Sentence	a	b	c	d	Probability
τ	1	1	1	1	1
$A1$	1	1	0	0	π_1
$A1 \Rightarrow B$	1	0	1	1	π_2

Table 6.2: Interpretation Table for Dice Example

In terms of the representation of table 6.2 (Nilsson's original model) the equations are:

$$\begin{aligned} a+b+c+d &= 1 \\ a+b &= 8/15 \ (\pi_1) \\ a+c+d &= 10/15 \ (\pi_R). \end{aligned} \tag{6.1}$$

Which solve to give:

$$\begin{aligned} a &= 3/15(\pi_1+\pi_R-1), \\ b &= 5/15(1-\pi_R), \\ c &= d = 7/30((1-\pi_1)/2). \end{aligned} \tag{6.2}$$

and so a probability for B of a+c=13/30.

In the extended model of Probabilistic Logic, shown in table 6.3 the equations are:

Sentence	a	b	c	Probability
τ	1	1	1	1
$A1$	1	1	0	π_1
$A1 \Rightarrow B$	1	0	1	π_2

Table 6.3: Interpretation Table for Dice Example

$$\begin{aligned} a+b+c &= 1 \\ a+b &= 8/15 \\ a+c &= 10/15. \end{aligned} \tag{6.3}$$

Which solve to give a = 3/15, b = 5/15, (both as before), and c = 7/15 $(1-\pi_1)$; and so a probability for B of a + x.c. Where x is the expert's assessment of the conditional probability p($B|\sim A1$). If the expert knows the problem domain and gives the correct conditional probability of 2/7 for x, the probability of B is 1/3 from this method.

Bayesian inference gives the following solution when proposition $A1$ becomes true:

$$\begin{aligned} p(B|A1) &= \frac{p(A1|B).p'(B)}{p(A1|B).p'(B)+p(A1|\sim B).p'(\sim B)} \\ &= \frac{\frac{3}{15}}{\frac{3}{15}+\frac{5}{15}} \\ &= \frac{3}{8} \end{aligned} \tag{6.4}$$

where p'(X) represents the prior probability of predicate X, and $p(X|Y)$ is the probability of X conditional on Y. Neither of the methods of Probabilistic Logic obtain this value.

A Probabilistic Meaning for Probabilistic Entailment

The discrepancy arises because of the way probability has been assigned to the rule of entailment. This was simply done using the logical equivalence of $\sim A1 \vee$B with $A1 \Rightarrow$B, and summing the probabilities of the worlds in which this formula holds true. In fact, as the probability of $A1$ approaches 1, the probability of the conclusion necessarily approaches 10/15 (the strength of the rule). This result is too high, and forces us to consider what *probabilistic* meaning to assign to the rule.

Logically speaking, we have not taken into account the fact that our sentence $A1$ has also to be consistent with these worlds. (This consistency relation is discussed in section 7). If this is done, we reduce our consistent possibilities for the rule to those worlds (3,9,15), of the worlds (1,3,5,7,9,11,13,15); and so the probability of the rule is 3/8. So in conclusion, one way to remove the discrepancy, is to equate the probabilistic meaning of p($A1 \Rightarrow B$) with p($B|A1$). More generally, the question we could ask the expert in order to get him to estimate the strength of a rule of the form $A1\&...\&An \Rightarrow$B is: 'if the conditions $A1, ...An$ were all true, how often would you expect the rule of entailment to be true?'

If we use the value of 3/8 as the strength of our entailment rule, both equations 6.3 and equations 6.4 will produce equations which give a probability of 3/8 to B in the event of $A1$ being true. However, we also know that for probabilities of $A1$ greater than 5/8, Probabilistic Logic can derive a consistent probability for B in a manner similar to PROSPECTOR. Equations 6.3 can give us a probability value for proposition B which ranges from 0 to 3/8. Equations 6.1 will produce an optimistic result in the range 3/16 to 3/8.

This interpretation of the probabilistic meaning of entailment ensures that the result produced at the extreme case of all the antecedents being true is in accordance with that obtained using Bayesian Inference. So that this result is a necessary upper bound, and that for the case where the antecedents are uncertain, the framework will bound the region of correct possibilities. Using the maximum entropy method will then give us the best estimate of what the uncertain probability is in this region.

6.4 Probabilistic Logic plus Conditional Probabilities

The addition of conditional probabilities to Nilsson's Probabilistic Logic, discussed in chapters 4 and 5, means that to completely describe the probability model for a rule of entailment with n antecedents, $2^n - 1$ conditional probabilities are required, plus the probabilistic strength of the rule. This is exactly the amount of conditional probabilities required in the Bayesian model.

They both use the same amount of information, but the nature of the required information

is different [67]. Nilsson's is a MYCIN-like formalism with an extension to allow for the complete specification of a probabilistic model. A further aspect of the Nilssonian inference with the above model is that it addresses the problem of entailment using implication. With the above model we are able to code the exact amount of attachment between the antecedents and the conclusion, that is to describe the details of the connection between the antecedents and the conclusion across the rule of entailment [58].

The most outstanding problem with Probabilistic Logic now is how to make use of the semantic clarity, while controlling the complexity problem which Bayesian Inference has been trying to deal with for over a hundred years.

Inference Conditioned on Hypotheses vs. Inference Conditioned on Evidence

Observing the type of inference performed by the Bayesian schema, we see that it is conditioned on knowledge about the hypothesis; whereas inference performed under probabilistic entailment is based on knowledge about the evidence. It may be that for some situations, the latter information may be more readily available from an expert than the former. For example, where an expert medical consultant is creating an expert system to be used for diagnosis of a newly emerging, and constantly evolving, disease. In this case, the prior probability of the presence of a disease in a patient may not be known, and the expert may feel more at ease estimating the probabilities of the hypothesis based on evidence rather than the probabilities of the evidence based on the hypothesis.

Nilsson's probabilistic logic opens up a new way of reasoning with uncertainty to the expert system community, and is particularly suited to experts who may think of their reasoning processes more in terms of entailment (and MYCIN), and less in terms of Bayesian Inference techniques (and PROSPECTOR).

6.5 Consistency in Probabilistic Logic

If all of the contexts splits are conditional probabilities, and the rule of entailment is a conditional probability then it is clear that Probabilistic Logic can be described as Bayesian Inference acting in another way. Exactly the same amounts of conditional information are used by both. This has an interesting consequence for consistency problems in Probabilistic Logics. In section 4.6.3 a formula was produced for the calculation of consistency in an entailment problem:

$$p(A_i) \geq (1 - \pi_R)$$

for any rule of entailment involving n antecedents (A_i for i = 1 to n), in a rule of entailment $A_1 \& ... \& A_n \Rightarrow B$ which has probability π_R.

Consider the example where we use the sets: $\{A_1,\ A_1 \Rightarrow B\}$ shown in table 6.4 which has a probability model of: $\{p(A_1) = 0.2,\ p(A_1 \Rightarrow B) = 0.7,\ p(B|\sim A_1) = 0.4\}$. From

Conditionals	Probabilities	Entailment 1	Entailment 2
$p(B\|A_1)$	0.7	$p(A_1 \Rightarrow B)$	$p(B\|A_1)$
$p(B\|\sim A_1)$	0.4	$p(B\|\sim A_1)$	$p(\sim A_1 \Rightarrow B)$

Table 6.4: Swapping the Entailment Rule

the consistency constraint, this rule cannot be fired. However, since all of the information is conditional probability information, the same information can be re-expressed: $\{p(\sim A_1) = 0.8,\ p(\sim A_1 \Rightarrow B) = 0.4,\ p(B|A_1) = 0.2\}$. and the mechanism of Probabilistic Logic can use this information to give the probability of B as: $0.2 + 0.7 * 0.2 = 0.34$ (see section 4.4).

The point is that the complete specification of conditional probabilities makes it possible to change the rule of entailment. This may be done by choosing one of the other context splits, making this the rule of inference, and forming the base set of antecedents as the corresponding truth values specified in the context split. If an inconsistency has been found for a particular ordering of the information, a new one might be found which renders the entailment consistent. An entailment is only truly inconsistent if there is no conditional probability p for which the probabilities of the antecedents in the context, as formed in the rule all are equal or greater to the value $1 - p$. This is an expansion of Nilsson's probabilistic logic into the area of Bayesian

inference and, by the way, it provides a simple method of ensuring consistency in a Bayesian setup.

6.6 Conclusion

In this chapter the effects of the extensions to Probabilistic Logic have been demonstrated and the probabilistic entailment has been compared with the Bayesian Inference mechanism. It has been shown that the logical nature of probabilistic entailment needs to be given a statistical context if consistent results are to be derived from within the Probabilistic Logic. This context is an identification of the probabilistic rule of inference with the conditional probability statement. Once this connection has been made, Probabilistic Logic can give the results with the same accuracy as Bayesian Inferencing with one major extension.

When any or all of the antecedents attached to the rule of inference are uncertain, the apparatus of Probabilistic Logic plus the inference engine of the maximum entropy formalism make it possible to extract a probability distribution from the uncertainty that takes into account the nature of the uncertainty and which is the least commitment probability distribution based on that information. This process comes naturally to Probabilistic Logic and is a major extension to the Bayesian concept of reasoning with uncertainty.

In the next three chapters the logical aspects of Probabilistic Logic are examined in their relationship to Incidence Calculus and heuristic reasoning and control.

Chapter 7

APPROXIMATION TECHNIQUES: INCIDENCE CALCULUS AS A PROBABILISTIC LOGIC

7.1 Introduction

The essence of Bundy's *incidence calculus* is that a sample space of points is chosen independently of the level of uncertainty of the sentences. In fact, although Incidence Calculus uses a set theoretic foundation, it is possible for two implementations to be perfectly correct, and to give different inference values. This is in contrast to the situation in Probabilistic Logic, where the number of points in the sample space is only known when the semantic tree is produced. Three algorithms are presented which allow a complete implementation of Incidence Calculus within the framework of Probabilistic Logic. Ultimately, because Incidence Calculus is a weaker set-theoretic probabilistic logic, it can be useful in situations where a complete implementation of Nilsson's probabilistic logic is impossible or impractical. It can therefore be used as an approximation to Nilsson's probabilistic logic.

7.2 Incidence Calculus

Assignment in Incidence Calculus is not done with reference to a Semantic Tree, as it is in Probabilistic Logic. Corlett and Todd took the problem of assignment, and suggested the use of Monte Carlo techniques to find legal incidence assignments for Incidence Calculus [22]. This gives rise to two problems. Firstly, the same information run on two different days can give different results. And, more importantly, the size of the incidence sets is arbitrarily chosen before the process begins. So that, Incidence Calculus can be wasteful of space when the set size is unnecessarily large, and it can force relationships between propositions when the set is too small. The latter of these two problems is one of the things Bundy wanted to remove when he proposed Incidence Calculus.

The version of Incidence Calculus proposed here has neither of these problems. Two algorithms are proposed for incidence assignment, both of which will terminate. They are called into operation after the Semantic Tree has been generated, and now it is only the generation of the tree which is problematic. There is no problem in the Propositional Calculus, as the rule of resolution is sound and complete here [105]. However the first order Predicate Calculus is only semi-decidable [55]. A number of approximations are proposed to alleviate this situation.

7.3 The Method of Reasoning

The manner of reasoning is as follows. If a condition C is entailed from a set of data, whose truth is represented by predicate functions P1 .. Pn, where n is the number of conditions involved in the entailment process, and each of these probabilities can be estimated; then the set is input as shown in table 7.1 where the bottom sentence in the table is the one whose probability value is to be estimated.

Unless the expert connects predicates in one of the two manners listed below:

1. By some entailed value being used in another reasoning process.

2. By some predicates being dependent.

$$\exists(x, P1(x)).$$
$$\cdot$$
$$\cdot$$
$$\cdot$$
$$\exists(x, Pn(x)).$$
$$\exists(y, P1(y)\&...\&Pn(y)).$$
$$\forall(z, P1(z)\&...\&Pn(z) \Rightarrow C(z)).$$
$$\exists(u, C(u)).$$

Table 7.1: Form of Probabilistic Entailments in Predicate Calculus

then the reasoning processes are dealt with as independent. In fact, a system using these ideas would probably run best as a number of communicating sequential processes. This work suggests that not only is it crucial to keep the semantic tree as the basis for the startup procedures, but that, based on what the expert says, and the conditions which become positively identified as true, (or false), as the reasoning process progresses, the semantic tree should be pruned and manipulated. The ultimate goal being to reduce the tree to the only possible worlds which are feasible for information given, and to distribute the given probabilities to each of these worlds in a controlled and mathematically sound manner.

In the case of a predicate whose value is entailed, and which is used by another reasoning set; the user must wait until the producer has provided the answer. Communication between processes might hurry up, or slow down this process, depending on probability levels. How these processes run to alter the texture of 'w' is covered below.

Incidence Assignment and The Semantic Tree

Before Incidence Assignment is attempted an interpretation table for the uncertain sentences is produced and the probability constraints on the possible worlds are recorded as in section 4.5. Any solution to these equations, which assigns a non-negative probability to each possible world, will constitute a valid incidence assignment.

The process for determination of incidences is then applied, and this consists of:

1. Whenever a probability for a possible world appears in it's own probability constraint, remove it from all the other constraints (subtracting the probability from the constraint's probability) and declare it as a deduction.

2. If there are possible worlds on their own, call on the simplification algorithm to find intersections between the worlds.

7.4 The Simplification Algorithm

The basis for all of the following assignment algorithms is a simplification algorithm which has access to a procedure for finding minimal subsets. The motivation for this simplification routine is taken from the following premises. If $[a_1, a_2, ..., a_n]$ are to be assigned probability P_x then:

1. the maximum probability for any of the worlds a_1 to a_n is P_x.

2. the minimum probability for any of the worlds a_1 to a_n is 0.

Therefore when all of the minimal subsets and associated probabilities are generated, as long as the assignments do not contradict these simple premises, the assignment will be consistent. Care must be taken to ensure that probabilities are only assigned from the smallest probabilities to the largest. The algorithm that ensures this is given below.

1. Give each world a unique label.

2. For each proposition P_i, take the associated probability π_i, and make two constraints thus. Sum the worlds associated with P_i to π_i. Sum the rest of the worlds to $(1 - \pi_i)$.

3. Sort this set of constraints into ascending probabilistic order. So that the first constraint in the list is the one with the smallest probability, and the last has the largest probability.

4. Remove all subsets from supersets thus: If the worlds associated with proposition A are W_1, and the worlds associated with proposition C are W_2, then, if $W_1 \in W_2$, then make a new set of worlds W_3 to represent a new proposition D such that,

$$p(D) = p(C) - p(A); \quad W_3 = W_2 \backslash W_1$$

5. Stop if there is an inconsistency detected. ie. a set of worlds getting two different probabilities attached. Or if a set of worlds is assigned a probability value less than zero. This is failure.

6. Stop if all worlds are assigned a positive probability value less than or equal to one. This is success.

7. Sort them into increasing probability assignments again. These are our minimal relationships. Stop when no more removals can be performed. Otherwise go back to step 3.

7.5 Discussion of Simplification Algorithm

This will generate more probabilistic constraints. However, it will always terminate, in one of two ways. Either, it will not be able to remove a subset from a superset to provide a new proposition- in which case it terminates succesfully. Or, secondly, it will encounter an inconsistency. There are two forms of inconsistency. Either a set of worlds is produced twice, with different probability numbers. Or, to use the above example of getting proposition D from propositions C and A,

$$p(A) > p(C)$$

.

In both of these cases, the offending results can be noted, and an explanation as to why the inconsistency has arisen may be offered. This form of inconsistency is easily fixed, and in the case of logical entailment using the rule of modus ponens, section 4.6.3 has presented a simple method of ensuring that such inconsistencies never arise.

7.6 Assignment Algorithm 1

The simplest approximation algorithm is the following:

1. Run the Simplification algorithm. If there are no worlds yet to be assigned a probability, then a consistent assignment has been succesfully applied within the uncertainty space.

2. Otherwise, take the first of the constraints C_1, such that, $p(C_1) = \pi_n$ Where length(C_1) = L, and assign each of the worlds in C_1 a probability of π_n/L.

and again, this algorithm will not introduce inconsistencies when there is not a certain assignment.

7.7 Assignment Algorithm 2

Without loss of generality assume that for random propositional sentences X and Y that the uncertainty of sentence X is represented by the possible worlds B, and the uncertainty of sentence Y is represented by the possible worlds A. Assume also p(X) is Xv and p(Y) is Yv. An estimate of A ∩ B (the intersection of the sets A and B) is Xv * Yv.

7.8 Justification for the Assignment Algorithms

The simplification algorithm produces groups of worlds which share common elements. For each member in the groups, the maximum probability value it may have is the minimum of the probabilities of the groups it appears in. A restriction on any assignment method then, is that a world may only be assigned a probability value less than or equal to the smallest probability value of the minimal groups it appears in.

In the case of algorithm 1, this condition will always apply when the group of worlds with the smallest probability is given an equal share of the probability of the group.

Consider the case of a single shared world being assigned a probability value by algorithm 2. Since, the combination will be between worlds whose probability must always be less than 1, if one of the groups being combined has the maximum possible value for the shared world, then an estimate for the probability of the shared worlds which is less than the world's maximum will be made. So, removing this world, and assigning the reduced probability value, does not cause inconsistency, since there is always a residual probability to be redistributed among the remaining worlds.

If it is not possible, at some level, to use the compound event which shows the maximum possible value for that world, to remove it, then it will be possible at a later level. This is because, in this case, the world shares at least one common world in all it's compound events. And, these will be discovered when greater numbers of worlds are collected in the

form of intersections. So, the process starts again, looking for bigger intersections between compound events, and, when these intersections are found, new restrictions are imposed on the maximums. Inconsistency will only be seen when the sum of the maximums for a compound event is less than the actual probability assigned.

One last point about inconsistency is that it is possible for a world to assume it's maximum possible value, (this happens when all the other worlds in the group which shows it's maximum possible value are false). So, in the case where the probability assigned to a world is less than the maximum allowed, the rest of the group should be able to redistribute the residual probability (maximum possible value - assigned value), between it's elements, since, consistently, the world assigned in question can range between probability zero, and it's maximum.

The process will eventually terminate when there is an assignment for each world; or when there are two, mutually exclusive groups left, each assigned a positive value. In the latter case, these values can be equally divided between the worlds; since a good estimate on them is that they are equally likely.

This assignment, when the sentence set is consistent, is a reasonable estimate. The expert now can be employed here to redistribute assignments within the relationships above, and this is one way in which he/she can model the reasoning process to his/her knowledge.

7.9 Semi-Decidability and the Semantic Tree

A major drawback to the use of exhaustive theorem prover's to provide all of the possible worlds before the assignment process is applied, is the semi-decidable nature of predicate calculus. For example, a set of sentences for which at least one possible world will not stop generating new possibilities to test with the resolution principle is:

$$P_1(c_1)$$
$$(\forall x).(P_1(x) \Rightarrow P_1(f_1(x)))$$
$$\sim (P_1(f_1(f_1(c_1))))$$

1. Since a vast proportion of the worlds can be dispensed with, because they will be shown to be inconsistent, perhaps worlds should be added from the most restrictive predicate equation in the initial set. This would mean that if we run out of space to represent the worlds, we would have done so in any reasonable mechanism for incidence assignment. Secondly, there may be a more efficient use of time by doing this.

2. This would mean that the sentences would have to be pre-ordered before the tree generation should be attempted, and that the semantic tree would be built up from small cliques.

3. If we can order the sentences such that we can remove the ones which mutually recurse, we could make a partial semantic tree for the rest; which we could call the base. Maybe we could then have some higher level rules saying how we could add the recursive sentences in. We might involve the person with knowledge of the proof domain in telling us the semantics of the region. There is not have enough relevant information to deal adequately with the problem.

The offending world in the above set of sentences is:

$$P_1(c_1)$$
$$(\forall.x)(P_1(x) \Rightarrow P_1(f_1(x)))$$
... consistent clause set ...

which produces: $P_1(c_1), P_1(f_1(c_1)), P_1(f_1(f_1(c_1))), \ldots$

and so on without terminating.

7.10 Functionality and the Expert

We clearly need to know about the function f_1. Asking the expert who provided the knowledge for such like information might be one way of dealing with the problem. Automatically flagging such functions could be achieved in the following ways:

- Surrogate functions could be useful in such cases. For example for any function f, we can find another function g such that g $\sqsubseteq$ f (g approximates f) under the following

circumstances:

$$(\forall i.\ \ 1 \leq i \leq k) \quad g^i(x) \ = \ f^i(x) \tag{7.1}$$

where k is a preset constant defining the maximum depth of compatibility between the functions f and g. And also:

$$(\forall i.i > k) \quad g^i(x) \ = \ \perp \tag{7.2}$$

where $\perp$ represents failure of the function. Any possible world in which it is produced can be flagged and referred to the expert. Such approximations make the problem of generating the set of all possible worlds quite feasable, given that recursion is in fact limited on all computers to a maximum depth anyway.

- Another option might be to have distributed processes run for each of the possibilities, and say that a world is inconsistent until proven consistent. Bearing in mind that this theorem proving mechanism must be run before subjective probability estimates are applied.

7.11 Conclusion

Nilsson, in his initial paper suggested that it might be worthwhile to look for approximation techniques to solve the basic problem of inference in the Probabilistic Logic. He saw the biggest problem being the estimation of probabilities for entailed sentences when the dataset of sentences becomes too detailed. In his approximations he considered only matrix solutions to problems of inference, thereby restricting himself to points in the allowed range (collapsing Probabilistic Logic to a version of Incidence Calculus).

This idea is pursued in this chapter and culminated in a complete implementation of Bundy's incidence calculus in Probabilistic Logic. The Semantic Tree has various properties when it is used to deal with rules. Proofs are provided on how to deal with inconsistency and tautology from within a rule. However, the process of semantic tree generation becomes a task in which various restrictions have to be made to the theorem proving process to force termination for all possible worlds; and the expert can be called in to adjudicate on possible

worlds which terminated without direct proof of true or false.

Three algorithms which allow incidence assignment in incidence calculus have been introduced. The basis of these algorithms is a simplification strategy based on matrix manipulation techniques, and so the implementation will always discover if the subset we are concerned with can be generated from simple matrix manipulation methods on the rows and columns. Thus the reasoning process of Incidence Calculus can be standardised.

Chapter 8

ENTROPY AND META LEVEL REASONING

8.1 Introduction

A real problem in expert system reasoning is "what to do when it is possible to expand more than one rule?". Is there an intelligent, and computationally practicable method for choosing the next rule to expand? In this chapter the maximum entropy formalism in Probabilistic Logic is examined to determine when it can be used as a tool for meta-level reasoning. In this regard, entropy diagrams and an uncertainty measurement are introduced. A way of implementing the linguistic hedges of Fuzzy Reasoning in the entropy mechanism is also introduced.

8.2 The Complexity of the Large Database

When a large number of rules are to be coordinated in a database various procedural problems arise which are in the domains of administration and control. Rule saturation is a key cause of complexity in a knowledge base. Alternative proposals for dealing with the problem are:

1. Imposing an ordering on the rules. The situation arises that we need a strategy for choosing what to do next from what has been done before. Easily implemented approaches are breadth-first, or depth-first searches through the knowledge base. Both of

these approaches are static techniques, whose merits are discussed in [88]. A problem with these approaches is the firing of many irrelevant rules, which cause confusion in the diagnostic process, and make the selection of further rules for firing more and more difficult.

2. Employing Meta-Level Reasoning. The problems faced are that the consultation process is usually creative on the part of the consultant, and, that speed is an important factor in the process. Meta-Level Reasoning is a process which purports to offer a way of judging the information content of the present diagnostic scenario with enough accuracy as to allow the next stage in the diagnostic process to be arrived at from rational calculations.

A human expert dealing with uncertainty in a diagnostic session is able to assimilate the information at hand to limit the uncertainty in order to best proceed with the diagnostic procedure. This requirement of a human expert is the ability to summarise the current situation and make a rational decision as to how to proceed. Meta-Level Reasoning is used to perform this operation in an automatic system.

Entropy is already widely used in the field of Information Theory [114] as an indicator of the information content in a message. In the next sections the relationship between the probabilities of the antecedents, and the entropy of a rule waiting to be fired is explored. The intention is to find the best way to minimise the uncertainty in the knowledge base, and thereby to efficiently discover the strongest conclusion attainable from the evidence.

8.3 Entropy as a Tool to Aid Meta-Level Reasoning

The quick assignment algorithms introduced in chapter 5, put us in a position where it is possible to do some pre-processing in order to see which of a number of rules it would be most appropriate to fire next in a diagnostic session. Ultimately, we would like to have a system which implemented Meta-Level Reasoning on the current state of the reasoning task. That is, at any moment when there is a lull in the reasoning procedure, we should like to take a snapshot of the current position, and decide from this what would be the most fruitful course of action to pursue.

In section 3.8 it is shown that as the quality of information pertaining to an uncertain situation increases, so the entropy of the resultant probability distribution decreases. That is, as the uncertainty decreases the uncertainty space develops structure, (or equivalently, some of the possible worlds begin to look more likely than others). In chapter 5 expressions are derived for the three interesting probability regions of a logical sentence which have an affect on the entropy of the distribution. They are:

$$p(A_i) < \quad 1 - \pi_R \tag{8.1}$$

$$1 - \pi_R < \quad p(A_i) \quad < 1 - \frac{\pi_R}{2} \tag{8.2}$$

$$1 - \frac{\pi_R}{2} < \quad p(A_i) \quad < 1 \tag{8.3}$$

where in region 8.1, the probability of the antecedent proposition renders the application of the entailment rule logically inconsistent. In region 8.2, the probability of the antecedent has a debiliting effect on that of the conclusion ($a_i < 1$), and in region 8.3, the probability of the antecedent has a positive effect on that of the conclusion. We shall focus on the sections shown in 8.2 and 8.3.

We expect that for any assignment of probability to antecedents and rule the entropy of the maximum entropy distribution will lie between clearly distinguishable bounds. This can be tested since it is possible to evaluate the entropy of any probability distribution using the expression:

$$Entropy = -\sum_{i=1}^{n} p_i log p_i. \tag{8.4}$$

We shall use a structured assignment of probability to the antecedents, and plot the entropy of the maximum entropy distribution for each assignment on a graph of entropy versus probability assignment. In the entropy calculations 2 is used as the logbase.

8.4 Entropy Diagrams

To examine the relationship between entropy and the probability of antecedents and rules, consider the example of a rule with four antecedents, $(A_1 \& A_2 \& A_3 \& A_4 \Rightarrow B)$, whose proba-

bility is 0.72. The change of entropy as the antecedent probabilities are reduced is plotted in figure 8.1 using the following procedure:

1. Assign all of the antecedents a probability of one.

2. Let x = 0.

3. Assign y the value of the entropy value of the resultant maximum entropy distribution. Plot the point (x,y) on the entropy diagram.

4. Make antecedent 1 the current antecedent.

5. If the probability of the current antecedent can be reduced by a preset amount, (in this case 0.025), and still render the antecedent in region 2 or region 3 then, reduce the probability by this amount. Increase the x-coordinate by 1. Assign y the value of the entropy value of the resultant maximum entropy distribution. Plot the point (x,y) on the entropy diagram.

6. When another subtraction of the set amount would create an inconsistent entailment process the current antecedent is kept at its minimum.

7. If antecedent i is the current antecedent then update i such that:

$$if\ i = n, i := 1; \quad if\ i < n, i := i + 1$$

and make antecedent i the next current antecedent.

8. If the current antecedent is at its minimum, then stop; otherwise return to step 5.

In this ordered way the process plots out the change of entropy of the maximum entropy distribution for all the possible probability values of the antecedents.

In the entropy diagram of figure 8.1 the probabilities of antecedents are decreased in sequence as explained above. That is, each probability is decreased, the entropy of the distribution is plotted, and the list of antecedents is rotated. This diagram shows the two points of minimum entropy (when all probabilities are 1, or $1 - \pi_R$ respectively- both these fixed points corresponding to minimum entropy of 0.59). The maximum entropy point is when all probabilities have a probability of $1 - \pi_R/2$, giving maximum entropy of 2.59.

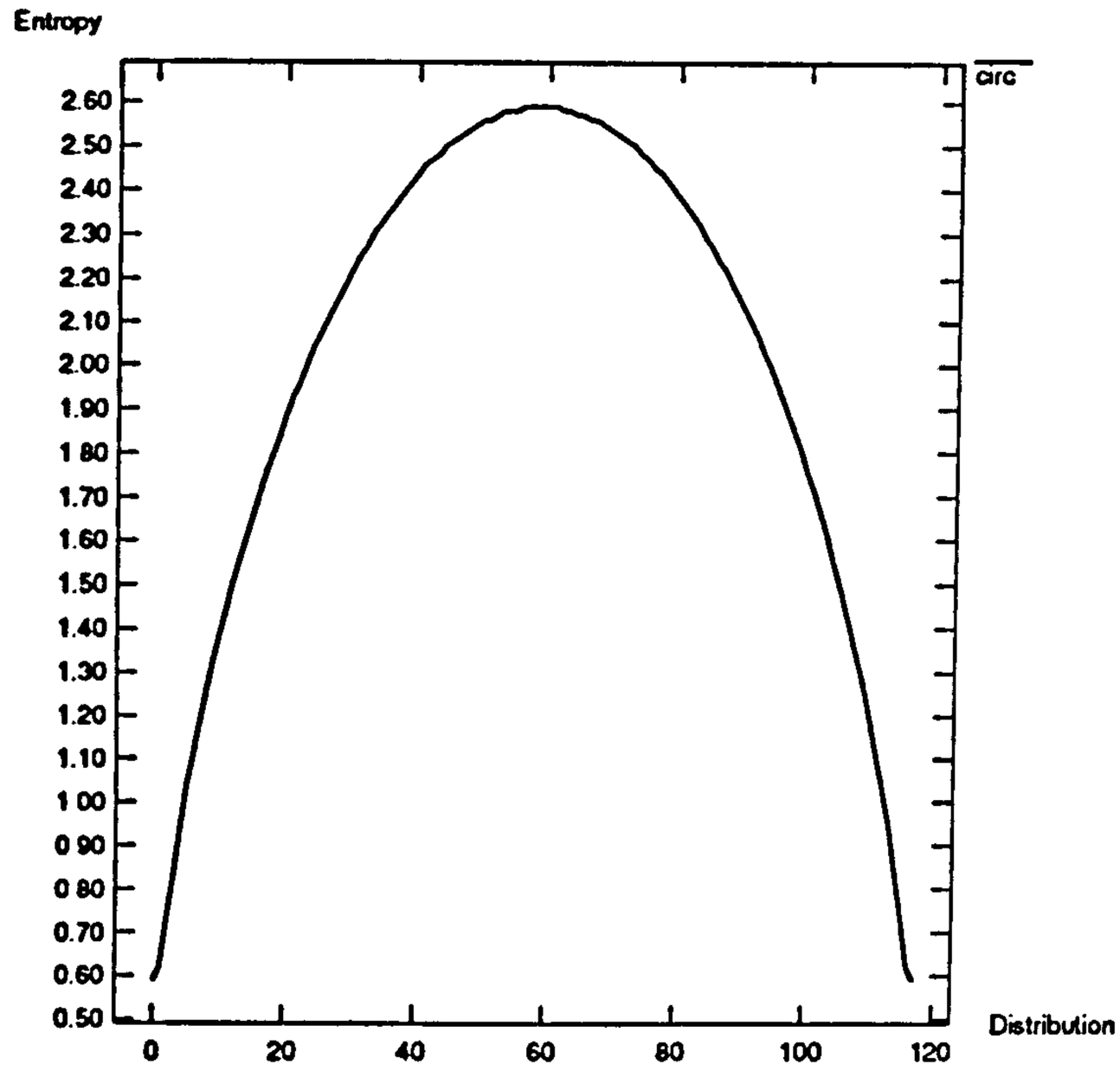


Figure 8.1: Four Antecedent Changes In Entropy

8.5 Explaining The Shape of The Entropy Diagram

8.5.1 Maximum Possible Entropy

In chapter 3 (section 3.8) it was shown that the maximum entropy possible for a set of uncertain sentences is when each of the associated possible worlds are assigned the same probability value. So for n possible worlds which span the probability space each world is assigned a probability of $1/n$.

For a rule of inference whose probability is π_R applied over n antecedents there are $2^n + 1$ possible worlds (section 4.6). Of these, one world represents the case where all antecedents are true and the rule false. This world is immediately assigned a value $1 - \pi_R$, leaving 2^n possible worlds which sum to the probability π_R. The maximum possible dispersion of probability is to give each of these worlds the probability $\pi_r/2^n$. Since each antecedent is true in exactly half of these worlds the probability to be assigned to each antecedent is $\pi_R/2$. Including the probability of the other world (with the rule false) in which all antecedents are true, the probability of each antecedent is therefore:

$$\frac{\pi_R}{2} + (1 - \pi_R) \;=\; 1 - \frac{\pi_R}{2} \tag{8.5}$$

which is the point as shown in equation 8.2, and the turning point in figure 8.1.

In summary, the calculation of the maximum possible entropy for a rule of inference π_R is therefore:

$$\begin{aligned}(1 - \pi_R).log(1 - \pi_R) + 2^n.\frac{\pi_R}{2^n}.log\frac{\pi_R}{2^n} \\ = \; (1 - \pi_R).log(1 - \pi_R) + \pi_R.log\frac{\pi_R}{2^n}\end{aligned} \tag{8.6}$$

The terms are: the probability calculation related to the rules uncertainty; and the 2^n probability calculations for each possible world with the rule of inference true. This is a straightforward calculation which can be evaluated at any time: all that is required is the strength of the rule and the number of antecedents which will be involved.

8.5.2 Minimum Possible Entropy

when all of the antecedents are true there are only two possible worlds: one with all antecedents true and the rule true; the other with all antecedents true and the rule false. This corresponds to the uncertainty in the rule (probability π_R). The entropy of this distribution is then: $-(\pi_R.log(\pi_R) + (1 - \pi_R).log(1 - \pi_R))$

8.6 Maximum Entropy and Fuzzy Logic

These diagrams will be of particular interest to the proponents of the linguistic hedges of Fuzzy Logic [133, 108, 102]. Consider the example of a rule with four antecedents. If there are varying degrees of uncertainty on the four antecedents, the Fuzzy-and rule is to assign the value of least certainty to the conjunction of the uncertainties (section 3.4).

Alternatively, we could project the level of positive or negative influence, guided by the use of linguistic hedges, onto the correct part of the entropy curve, and read off the appropriate entropy probability value, which could then be reworked into a linguistic hedge. The entropy diagram of figure 8.1 shows how the uncertainty in a probability distribution can be related to entropy.

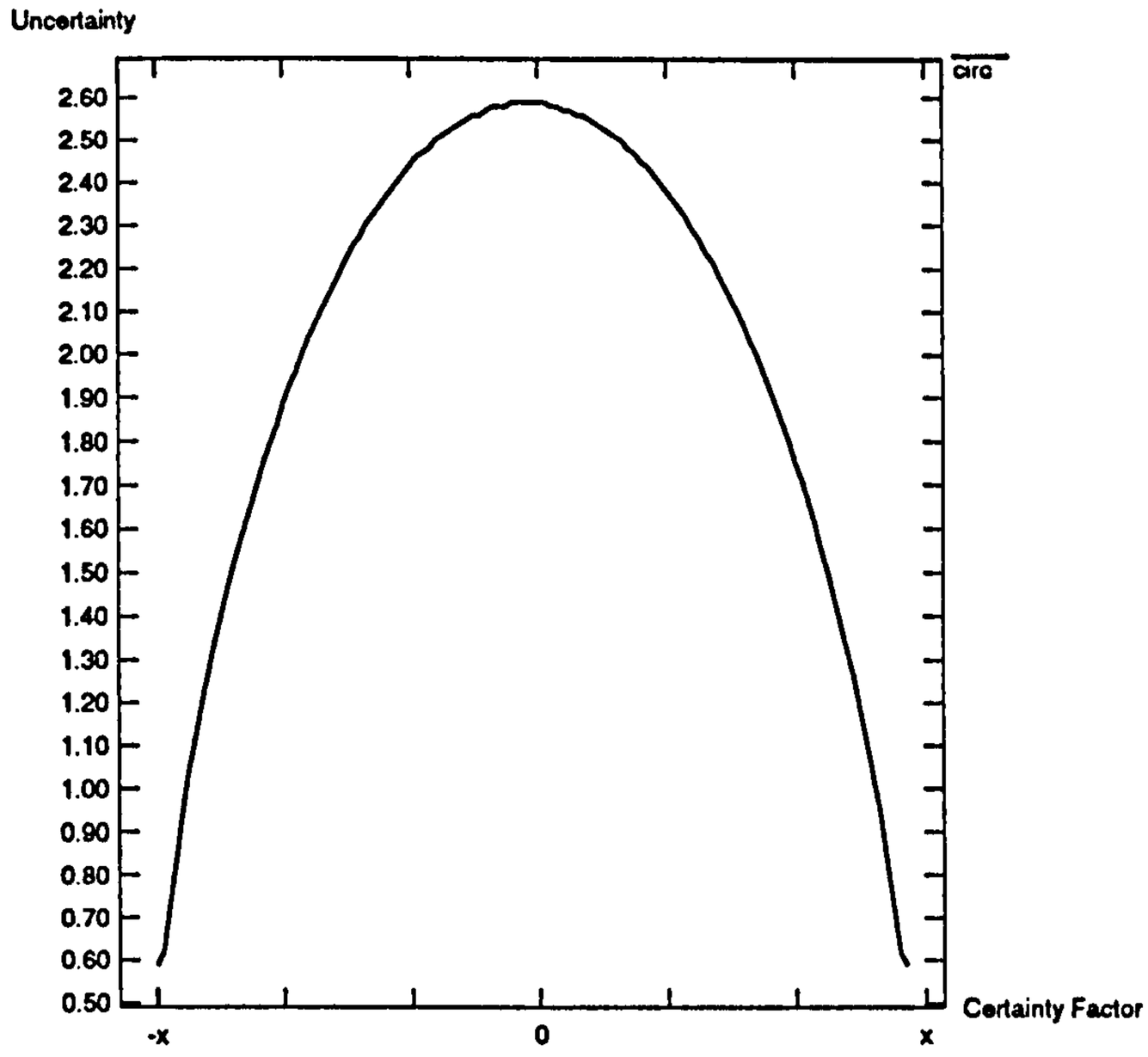


Figure 8.2: Uncertainty as a Linguistic Hedge

Instead of using a numerical formalism for the x-coordinate of this diagram, figure 8.2 shows how a linguistic hedge could be used directly to model the uncertainty. A system of Probabilistic Logic could use "certainty factors" (c.f. section 3.5) which projected a number in the region $[-x,x]$ onto the linguistic hedge "uncertainty" of figure 8.2 where 0 (the point of maximum entropy), $-x$ is false, $+x$ is true (the points of minimum entropy). For values between 0 and x (or $-x$) an interpolation can be made on the x-axis, the probability distribution chosen from the y-axis and a resultant uncertainty produced.

Probability itself could be modelled as a linguistic hedge. Consider figure 8.3 where the curves shown model the linguistic hedges {definitely not (A), very unlikely (B), unlikely (C), improbable (D), possible (E), likely (F), very likely (G), definitely (H)} on a scale of probability (0—1). On answering a question a user may be offered each of these linguistic hedges. If the user wants to further quantify their uncertainty, a certainty factor could be used with any of these hedges as explained above, and the required probability read off the x-axis directly.

The probability (linguistic hedge) of the conclusion can be calculated by using conditional probabilities attached to possible worlds or alternatively, heuristic measures [66, 65]. The

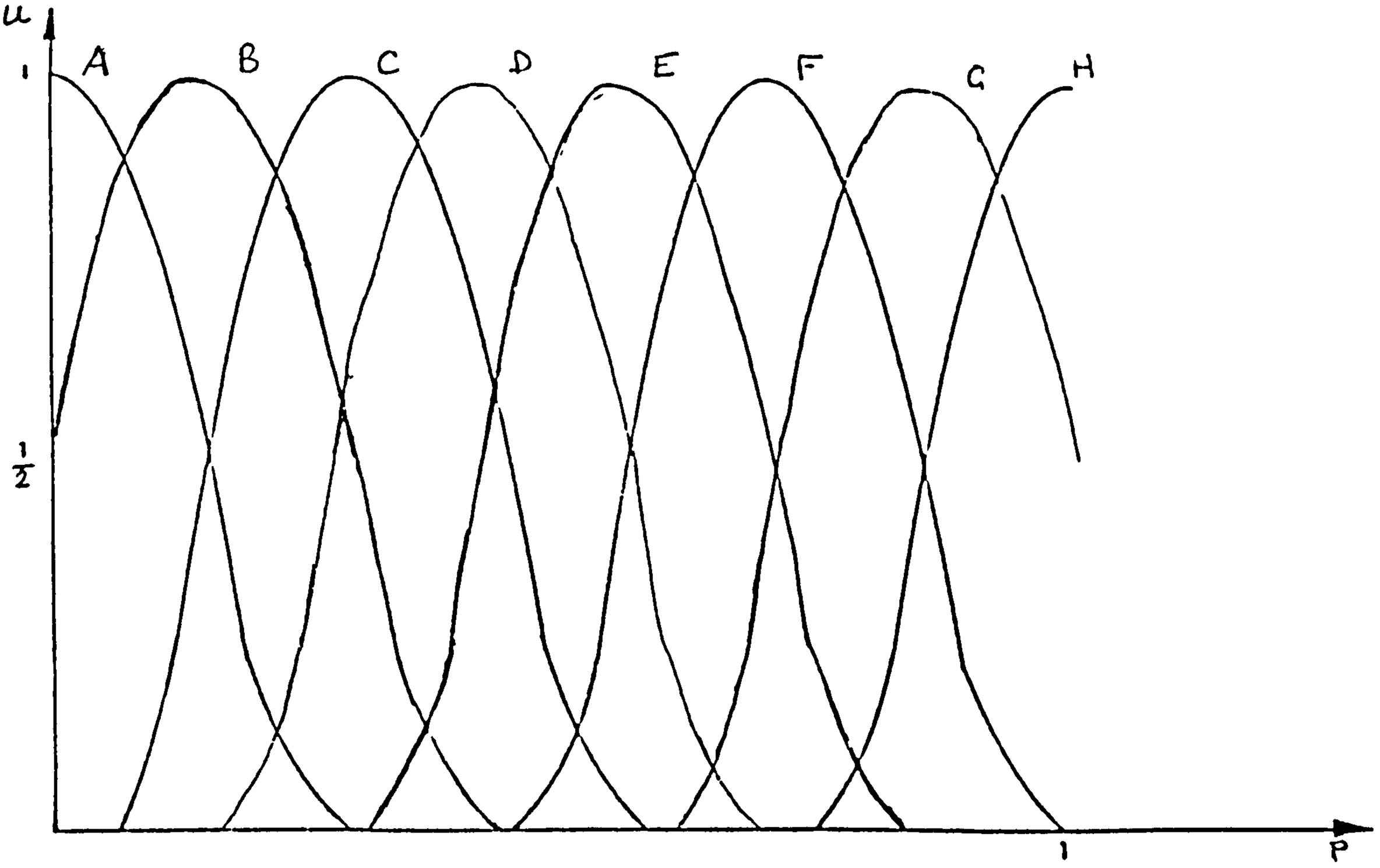


Figure 8.3: Probability as a Linguistic Variable

result of combining these uncertainties with the uncertainty of the rule of inference could produce a maximum entropy probability distribution from which can be recovered:

1. The most probable world

2. The certainty of the distribution

3. The probability (linguistic hedge) attached to the conclusion

8.7 Meta Level Inferencing

This information can be used in Meta-Level reasoning if, when a rule is ready to be fired, (that is, the antecedent probabilities are all available, and all consistent with the strength of the rule), the entropy of the situation is recorded. Since, the higher the entropy of a rule situation, the more uncertain we are about the state of the real world from this rule, the rules to be expanded should be chosen on the basis of increasing entropy. In this way, the most positive statements, (either in terms of truth or falsity), are uncovered first.

8.7.1 A Function Describing Specificity of Probabilistic Rules

Using this information, and the fact that minimum entropy means maximum specificity in the probability distribution, a function returning the usefulness of a particular rule in reducing the uncertainty of a reasoning situation can be found.

From consideration of the cases for minimum possible entropy and maximum possible entropy both equations use the probability of the rule of inference: π_R. The probability of the rule is an important factor in the entailment. It is the means of producing an uncertain deduction in Probabilistic Logic, and when coupled with the maximum entropy formalism the strength of the antecedent probabilities only have a relative meaning with respect to the strength of the rule.

8.8 A Certainty Function

A simple certainty function which places the specificity of the probability distribution in a scale between zero and one is:

$$specificity(Rule) = \frac{Ent - Minpossent}{Maxpossent - Minpossent} \tag{8.7}$$

where Ent is the entropy of the maximum entropy distribution with the probability constraints proposed in the problem, Minpossent and Maxpossent are the minimum possible entropy and the maximum possible entropy of the rule with the number of antecedents it has.

Using this function, the specificity of the first five rules are shown in table 8.1. The results suggest a specificity ordering of: R2, R4, R1, R3, R5. The greater the uncertainty in the

R1	R2	R3	R4	R5
0.750	0.592	0.760	0.601	0.864

Table 8.1: Specificity of Probability Distributions

entailment procedure the closer the specificity is to 1. So a way to choose the next rule for expansion is to evaluate the specificity function for each entailment rule it is possible to expand and then choose the entailment rule which has the smallest specificity number.

This behaviour is consistent with the use of entropy as a measure of information [114], and in this form it can be used to quantify the uncertainty which is inherent in a set of antecedents and a rule when each has a probability attached. What comes out immediately from this investigation is the dominant influence which the strength of probability attached to the rule has over the shape of the entropy diagrams.

8.9 An Example of Meta Level Reasoning

A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
0.83	0.57	0.91	0.78	0.45	0.95	0.9	0.8	0.6	0.9

Table 8.2: Ten Predicates and their Uncertainties

These diagrams provide us with a way of using the experts knowledge without necessarily resorting to numerical probabilities. Consider the following example, where there are ten predicates, labelled A_1 to A_{10} with a level of uncertainty attached as show in table 8.2. Consider also that the rules which can be fired from this information are as given in table 8.3. In this table the probability of the rule is shown in column one, the rules are shown in column 2 and columns 3 and 4 show the minimum possible entropy distribution for the rule and the maximum possible entropy distribution respectively (see section 8.5).

Since there are probabilities for the antecedents of the first five rules these are in a state to be fired. The last three rules cannot yet be fired: they all depend on outcomes from R1 to R5. When the maximum entropy distribution is calculated for R1 to R5 the entropy values

Prob		Rule	Minent	Maxent
0.8	R1:	$A_1 \& A_2 \& A_3 \Rightarrow B_1$	0.5	2.18
0.6	R2:	$A_4 \& A_5 \& A_6 \Rightarrow B_2$	0.67	1.82
0.7	R3:	$A_7 \& A_8 \& A_9 \& A_{10} \Rightarrow B_3$	0.61	2.44
0.9	R4:	$A_1 \& A_{10} \Rightarrow B_4$	0.33	1.57
0.9	R5:	$A_5 \& A_8 \Rightarrow B_5$	0.33	1.57
0.8	R6:	$B_1 \& A_7 \Rightarrow B_6$	0.5	1.61
0.7	R7:	$B_4 \& A_4 \Rightarrow B_7$	0.61	1.58
0.8	R8:	$B_3 \& B_4 \Rightarrow B_8$	0.5	1.61

Table 8.3: Rules over the Ten Predicates

returned are: 1.75, 1.41, 2.08, 1.08 and 1.4 respectively. However, a more interesting question would be: how specific is this answer? That is, how far is it from the simple case where all of the possible worlds are assigned the same probability?

8.9.1 An Examination of Specificity

To answer these questions we need to see the dispersion of probability amongst the possible worlds which has been assigned by the maximum entropy formalism. The probability diagrams (figures 8.4 to 8.8) are for the first five rules of inference shown in table 8.3. In these diagrams the probability attached to each possible world is plotted. The probabilities (see appendix D for the actual maximum entropy probabilities used) are arranged in decreasing size so that distributions can be compared.

In each diagram the line denoting the maximum possible entropy distribution is also drawn. This shows the distribution where each of the possible worlds could be assigned the same probability ($\pi_R/2^n$), and is included to show how the probability constraints imposed by the probabilities of the antecedents have caused the maximum entropy distribution to specify most likely possible worlds.

It can be seen from these diagrams what is meant by specificity. The position of no information is shown by the maximum possible entropy line and so the distribution whose probability peaks sharply for a small number of possible worlds and drops away sharply for the all the others is the one we wish to choose as the next rule to apply in the knowledge base.

Of special interest is the closeness of certain worlds to the point of maximum ignorancein the distribution. The most obviously skew distributions are in figures 8.5 and 8.7 and it is interesting to note the preference of rule 2 for expansion over rule 4. This is because in figure 8.7 the second most probable world was not forced to move very far from its point of maximum ignorance; wheras in figure 8.5 all of the possible worlds have been forced to move away from this point of maximum ignorance.

The composite picture of all these distributions overlaid is shown in figure 8.9.

8.10 Conclusion

This chapter addresses the problem of how to deal with large knowledge bases of many facts and rules. Ultimately, the complexity problem will saturate any implementation of Probabilistic Logic which does not employ some simplification strategies.

With this problem in mind, I have proposed an extended role for the maximum entropy formalism: namely, as a tool to aid meta-level reasoning. This new role for the maximum entropy formalism allows rules to be chosen for expansion on the basis of the information content in the probability of the rule and the probabilities of it's associated antecedent probabilities.

The entropy diagrams which have been introduced in this chapter also show how Fuzzy Logic's linguistic hedges can succesfully be embedded in Probabilistic Logic.

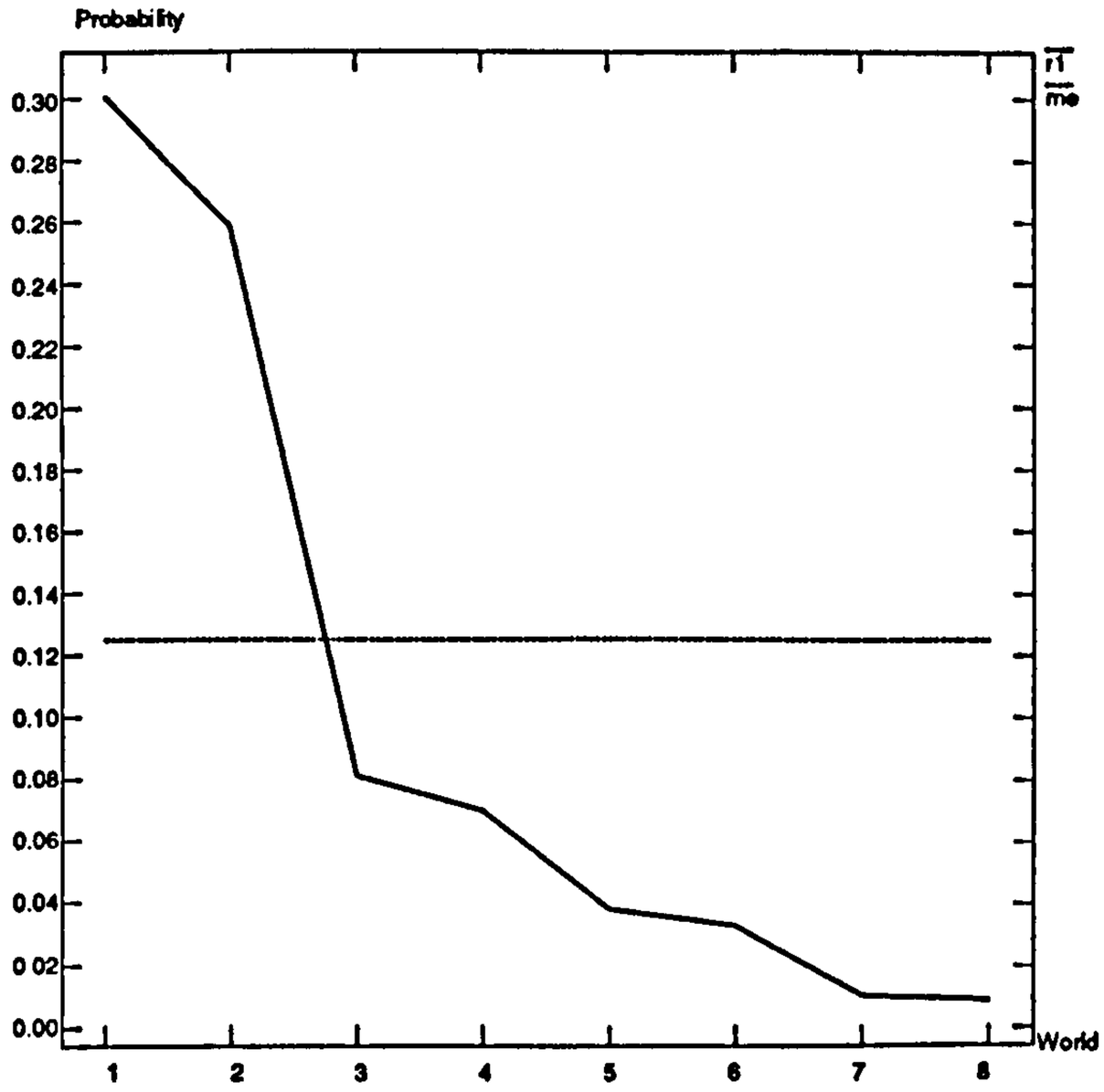


Figure 8.4: Probability Dispersion: R1

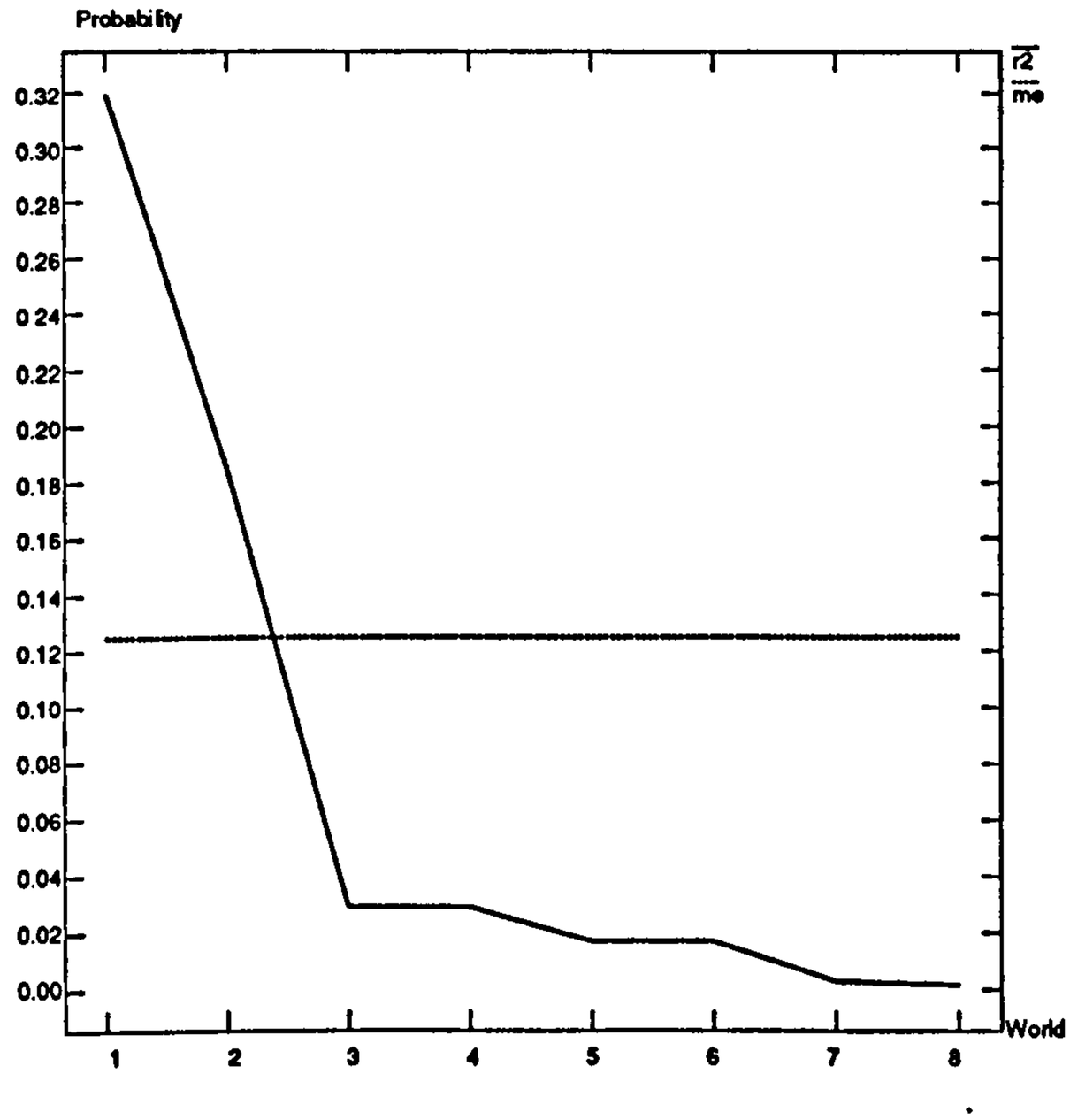


Figure 8.5: Probability Dispersion: R2

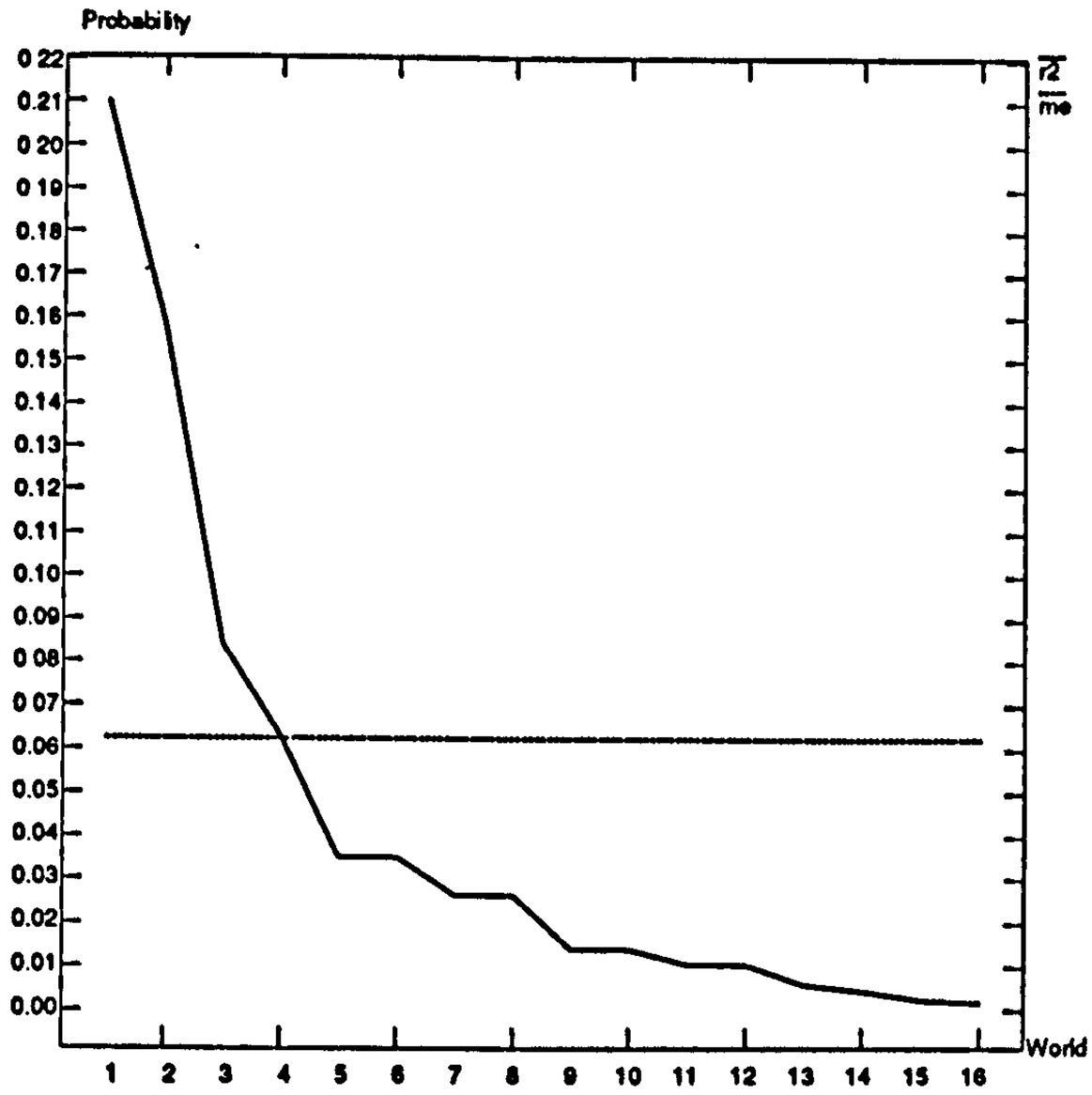


Figure 8.6: Probability Dispersion: R3

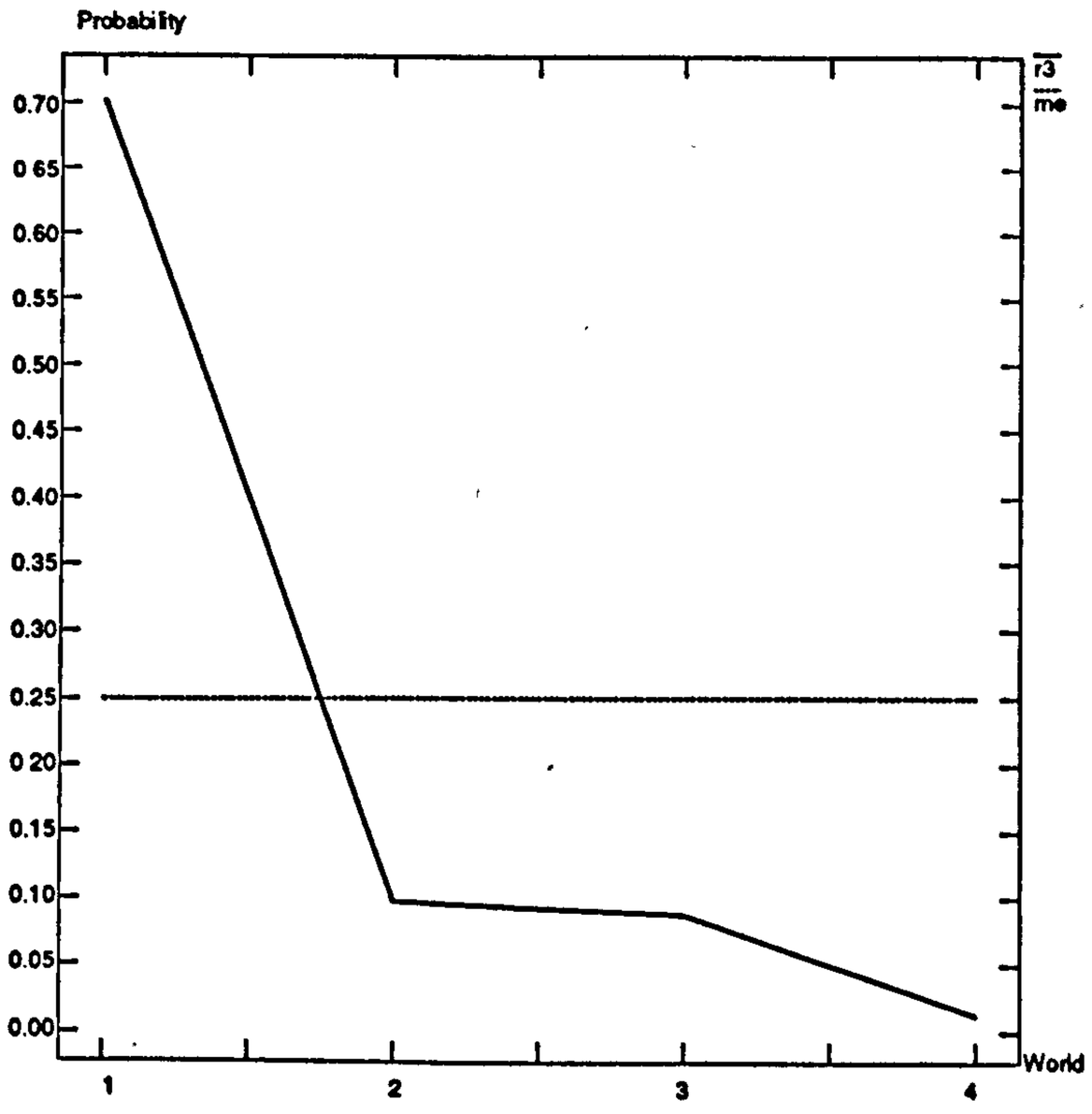


Figure 8.7: Probability Dispersion: R4

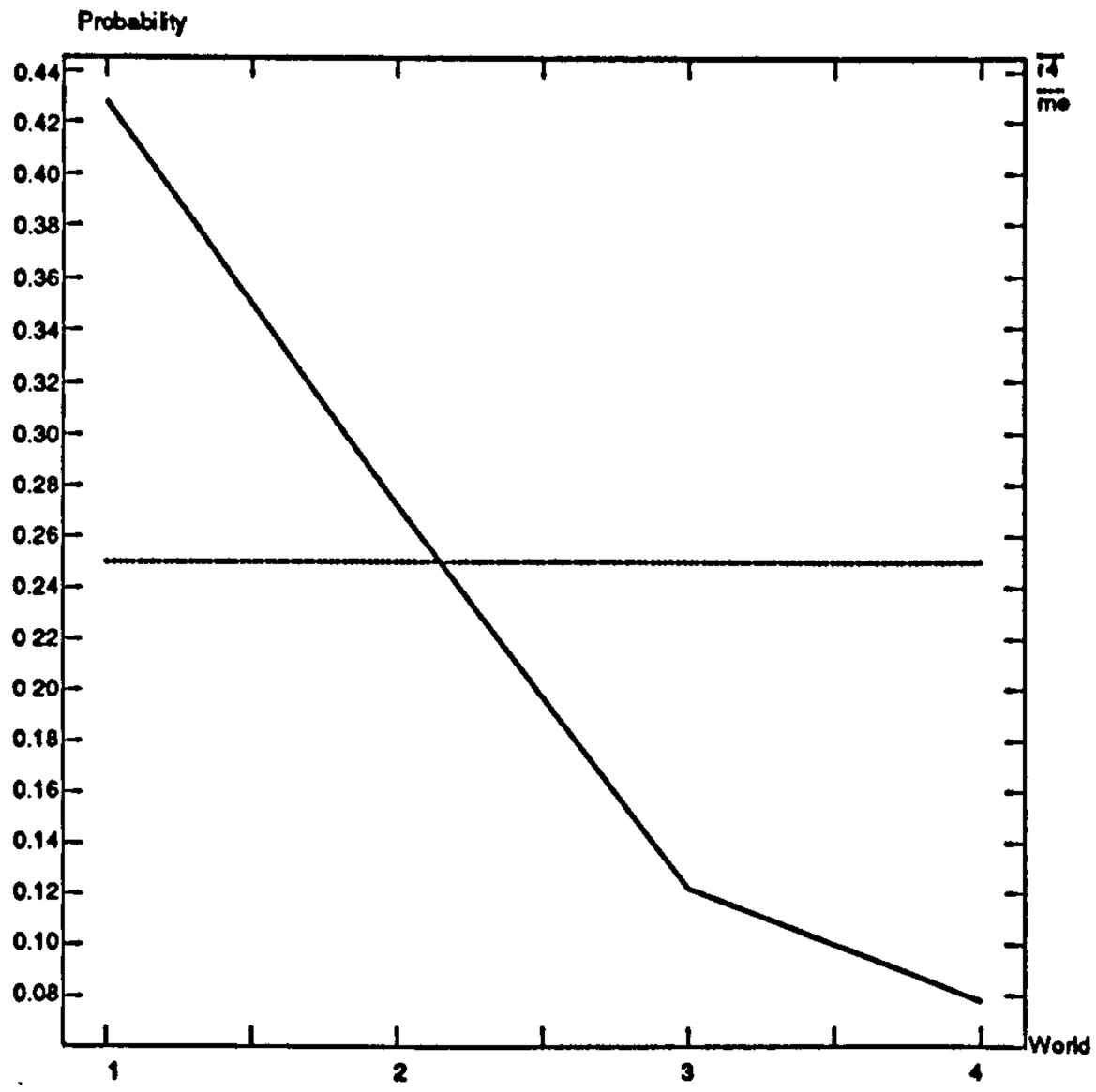


Figure 8.8: Probability Dispersion: R5

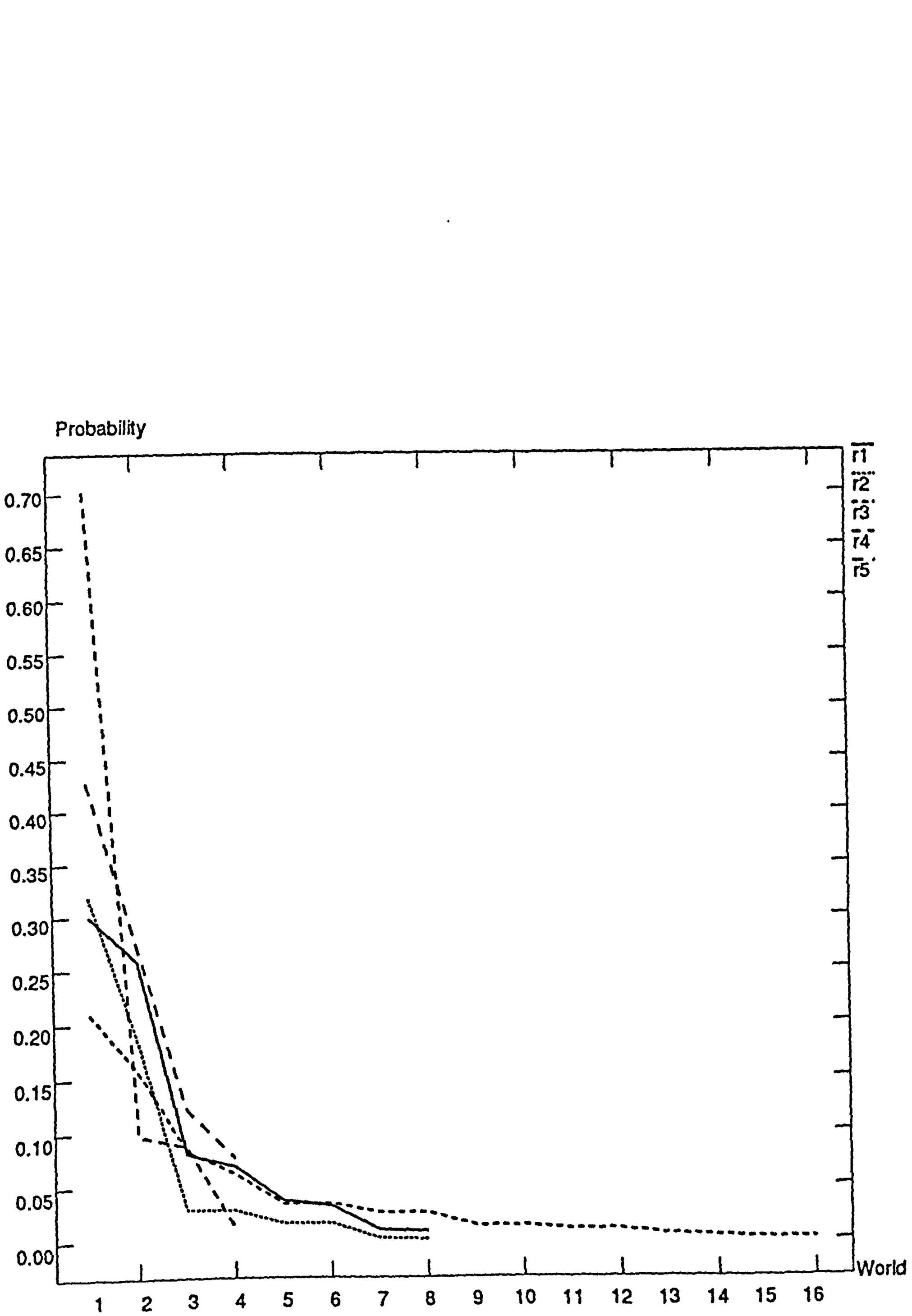


Figure 8.9: Composite Picture: R1 — R5

Chapter 9

HEURISTICS IN PROBABILISTIC LOGIC

9.1 Introduction

A way of dealing with uncertainty, when not enough precise information is available, is to use heuristic measures. It is said that Archimedes was the inspiration for the word "heuristic" when in his bathtub he solved a problem pertaining to the propensity of some objects to float while others sink. It suddenly occurred to him that if the weight of the body in the water was heavier than the weight of water it displaced, that it would sink, and he exclaimed "Heureka!".

Heuristics can be coded in rules in a knowledge base, or can be coded into the ways uncertainty is stored and manipulated by the reasoning mechanism. These measures replace strict probability measures, trying to approximate them. As knowledge in an area increases these measures are replaced by probability measures; but heuristic measures are central to expert system reasoning where information is generally only undergoing development. In this chapter, a group of heuristics is presented which can be made available to the knowledge engineer when the structure of the probabilistic knowledge available is under-constrained.

A way of evaluating selected possible worlds is presented which allows the uncertainty bounds of a problem to be effectively narrowed without the evaluation of many, essentially

non-informative, possible worlds. Finally, the notion of rule interleaving is examined, this involves combining together results from different rules which share antecedents and/or conclusions.

9.2 Using Context Splits

If the expert provides all of the context splits, Probabilistic Logic is now able to produce point probabilities from the probabilistic entailment. If the expert wishes to specify all of these for a rule of the form $A_1\&...\&A_n \Rightarrow B$, the system requires:

1. the probabilistic strength of the rule $p(A_1\&...\&A_n \Rightarrow B)$
2. $2^n - 1$ context splits.

As the number n increases, providing reliable context splits will become impracticable. This section suggests two heuristic mechanisms for dealing with this problem.

9.2.1 Heuristic 1: The Equal Split

If it is assumed that the probability of B and $\sim B$ is the same in each split world, then the probability of the conclusion is given by:

$$\begin{aligned} p(B) &= p(A_1,...,A_n, A_1\&...\&A_n \Rightarrow B) + \frac{1}{2}(\pi_R - p(A_1,...,A_n, A_1\&...\&A_n \Rightarrow B)) \\ &= 1/2(\pi_R + p(A_1,...,A_n, A_1\&...\&A_n \Rightarrow B)) \\ &= 1/2(\pi_R + a_r a_R \prod_{i=1}^{n} a_i) \end{aligned} \tag{9.1}$$

Table 9.1 gives the times for this algorithm. Again, as in section 5.4, the left hand column gives the number of antecedents, and the right hand column reports the amount of cpu seconds needed to reach the final result.

These times compare very favourably with those shown in table 5.4.

9.2.2 Heuristic 2: Contextual Weights

Another way of dealing with the problem would be to associate a contextual weight w_i with each of the antecedent propositions A_i in the rule $A_1\&...\&A_n \Rightarrow$B, such that

$$\sum_{i=1}^{n} w_i = 1 \tag{9.2}$$

Antecedents	Time (cpu secs)
1.	0.017
5.	0.017
10.	0.033
15.	0.044
20.	0.062

Table 9.1: Times for Rapid Calculation of Factors

and the weight w_i, given to proposition A_i, is a descriptor of how important the truth of proposition A_i is to the entailment of the conclusion. So that, in a possible world where the rule is true, the size of the context split is determined by adding together the weights of all the propositions which are true in that world. This reduces the amount of information expected from the expert to the strength of the rule, and n contextual weights.

Assigning Contextual Weights

One way of assigning contextual weights would be to commit a weight of 1/n to each of the antecedent sentences in the rule. e.g. for the case of n = 3, the contextual weight to each sentence is 1/3, and for the tree:

Sentence	a	b	c	d	e	f	g	h	i
A_1	1	1	0	0	0	0	1	1	1
A_2	1	1	0	0	1	1	0	0	1
A_3	1	1	0	1	0	1	0	1	0
$A_1 \& A_2 \& A_3 \Rightarrow B$	1	0	1	1	1	1	1	1	1
B	X	X	X	X	X	X	X	X	X
context split:	1	0	0	1/3	1/3	2/3	1/3	2/3	2/3

Table 9.2: Entailment Using Contextual Weights

$$p(B) = a + \frac{1}{3}(d + e + g) + \frac{2}{3}(f + h + i)$$

For completeness, a final table of cpu times is given for this method of assigning weights.

Another method is to get the expert to assign contextual weights to each of the propositions A_1 to A_n. So, for example, in the rule given, the expert may assign weights: ($A_1 = 0.5$,

Antecedents	Time (cpu secs)
1.	0.017
5.	0.100
10.	0.201
15.	9.9
20.	16.17

Table 9.3: Speed of Results Using Weights

$A_2 = 0.3$, $A_3 = 0.2$), and with reference to table 9.2

$$p(B) = a + 0.2d + 0.3e + 0.5f + 0.5g + 0.7h + 0.8i \tag{9.3}$$

This system of Probabilistic Logic gives the expert all the necessary tools to fully design a subjective probability distribution which fully describes their level of expertise. The above methods were used succesfully in the expert system described in [68] and chapter 10.

9.3 A Comparison of Results

The two automatic methods for assigning the context split: the equal split, and the contextual weight of 1/n can be compared. Consider the example of the ten antecedents and five rules of section 8.9. The maximum entropy distribution is calculated for each of the problems, and then the heuristic splits are applied to each world.

Consider the first rule $A_1 \& A_2 \& A_3 \Rightarrow B_1$ whose antecedents have probabilities 0.83, 0.57 and 0.91 respectively. The probability of the rule is 0.8. The maximum entropy formalism assigns a probability distribution to the constraints as shown in column 3 of table 9.4. This table also shows how the half-split and n-split are made from the distribution.

From these results the resultant probability of the conclusion (in this case B_1) is 0.57 for the nsplit, and 0.5293 for the half split. Of particular interest as regards a comparison of the two are the following points:

- N-split gives no probability to the conclusion in the world where all of the antecedents are false. Half split gives a half split.

- As the probability of the rule approaches 1, the half split gives a probability close to

World	Column	ME Probability	Half Split	N Split
1:	[0,0,0]	0.01028	0.00514	0
2:	[0,0,1]	0.081095	0.0405	0.027032
3:	[0,1,0]	0.0088453	0.0044	0.0029484
4:	[0,1,1]	0.06978	0.03489	0.04652
5:	[1,0,0]	0.038095	0.01905	0.012698
6:	[1,0,1]	0.30053	0.150265	0.20035
7:	[1,1,0]	0.03278	0.01639	0.021853
8:	[1,1,1]	0.25859	0.25859	0.25859

Table 9.4: An Example: Possible Worlds and Heuristic Splits

0.5 (as discussed in section 4.3).

For completeness, the heuristic splits for all of the rules of section 8.9 is given in appendix D.

9.4 Entropy as a Tool to Narrow the Bounds of Entailment Results

With this information we are in a position to deal with situations where perhaps there are too many possible worlds to evaluate all of them; or perhaps only a small number of possible worlds need be evaluated before the probability bounds of an entailment process fall within a narrow band of uncertainty.

As shown in chapter 5, the aggregate factors for an entailment problem can be calculated immediately. This calculation is only possible if the probabilities of each of the antecedent probabilities are above a certain threshold limit which is imposed by the strength of the probability of the rules of inference. We could use knowledge of the other two regions, that is, where the factors are consistent, but are either supportive or inhibitive to the probability of a possible world. An inhibitive factor is one which is less than 1, and will therefore reduce the probability of any possible world in which it is applied; and a supportive factor will increase the standing of the probability of a possible world.

Once the aggregate factors are calculated, they can be ordered in increasing order, since now they represent the presence of a world as well as give an indicator as to how the sentence

being true in the world alters its probability. Those factors greater than 1 can be skimmed off from the rest. In this way we can separate the good influences from the very bad, and thus narrow the bounds by working out the probabilities of each of these possible worlds.

Since for each of the resultant worlds we have a context split (cs) either provided manually by the expert, or heuristically by the system, it is possible to provide accurate probability bounds quickly, using the algorithm in chapter 4, and choosing the information from the aggregate factors which will make the biggest impact on the bounds. For example consider the sentence set $\{ A1,\ A2,\ A3,\ A4,\ A5,\ \ A1\&A2\&A3\&A4\&A5 \Rightarrow B \}$ whose probabilities are: { 0.99, 0.06, 0.98, 0.96, 0.92, 0.95 }.

From equation 4.13 the bounds are shown to be 0 and 0.95, thus giving a 95% uncertainty in the result. The calculation of two possible worlds, that is, where all the antecedents are true and rule true, and all where all the antecedents true except A2 and the rule are true, gives probabilities of 0.01 and 0.8 respectively. Using this information, and the context split information on the latter world, the bounds become narrowed to:

$$BOUNDS = 0.01 + cs * 0.8 + [0, 0.14] \tag{9.4}$$

That is, the uncertainty is now reduced to 14% with two calculations from a possible 32; and we have the added advantage of using conditional probabilities in the calculation. The interest of this is that if $A2$ is important to the entailment, the context split (cs) will be very low; whereas, if it is not so important, the split on this world may well be high. It is interesting to note, that for this example, Nilsson would give a result of 0.47925, which is almost exactly halfway between the points 0 and 0.95. This point is discussed further in section 10.5.

9.5 Information Interleaving in the Knowledge Base

I have presented a completely sound method of providing the maximum entropy result from a probabilistic rule of inference, within the constraints of consistency. All of these rules may exist in a database independently of each other and be called on only when needed. One final problem that arises is how to combine the results of two reasoning processes, both of which it is consistent to fire, and both of which entail the same conclusion? One solution would

Form	Simplification	Probability
$A1\&A2 \Rightarrow Z$	$\sim A1 \vee \sim A2 \vee B$	π_1
$A3\&A4 \Rightarrow Z$	$\sim A3 \vee \sim A4 \vee B$	π_2

Table 9.5: A Method of Combining Rules

be to join the two rules together logically, and join the two probabilities using the maximum entropy principle. We join them using the logical or operator, since the conclusion can be entailed from either of the rules.

If the only shared variable in a rule is the conclusion, then the logical connection using the or-rule means collecting together all the antecedents on the joined with the and-rule. And using the independence assumption which is built into maximum entropy to combine the probabilities.

Example: Combining Two Rules, Four Antecedents

The rules are shown in table 9.5.

Logical connection of the two rules using the or-operator gives:

$$\sim A1 \vee \sim A2 \vee \sim A3 \vee \sim A4 \vee B = A1\&A2\&A3\&A4 \Rightarrow B. \qquad (9.5)$$

And the probability of the new rule is $\pi_1 + \pi_2 - \pi_1 * \pi_2$.

A further problem occurs when an antecedent proposition becomes true, i.e. the strength of the probability becomes 1. In this case the rule can be collapsed (using the resolution principle [105, 14]) to an entailment involving each of the antecedents except the tautological one. The antecedent proposition is then subsumed into the tautology, and instead of the tautology having no weight in calculating the probability of the conclusion, it's weight is now increased by the weight of the tautological antecedent. This process allows the reduction of the rule, and the preservation of the solution methods described above.

9.6 Conclusion

This chapter has examined the problem of knowledge management which faces all probabilistic logics, when they must deal with large knowledge bases of many facts and rules. Ultimately,

the complexity problem will saturate any implementation of Probabilistic Logic which does not employ some simplification strategies. This problem is inherent in the nature of uncertainty management where each possible world is modelled, and assigned a probability value from consideration of an appropriate probability model, (e.g. Bayesian Inference). I have examined the nature of the complexity problems and find that the new system of Probabilistic Logic proposed in this thesis offers significant advantages over that originally proposed by Nilsson.

I have considered the measured use of heuristics in Probabilistic Logic. As with all heuristics, these are employed when there is a serious deficiency of reliable knowledge, or a major complexity problem which must be simplified. And, as with all heuristics, these mechanisms are designed upon semantic considerations of the problem domain area.

The next chapter, describes how these ideas have been realised in an expert system designed to solve a formidable recognition problem in the area of two dimensional vision.

Chapter 10

AN APPLICATION OF PROBABILISTIC LOGIC IN TWO DIMENSIONAL VISION [1]

10.1 Introduction

This chapter is devoted to a description of a vision expert system which was implemented using the enhanced Probabilistic Logic developed in chapters 4 and 5; and the heuristic methods introduced in chapter 8.

The problem tackled was to match (imperfect) scene reconstructions against a known catalogue of possible objects and to report the certainty of matches. Very simple heuristics were used for the strength of the rule, and for the context splits attached to each possible world. The results of this work compare very favourably with alternative approaches examined by Wallace and McAndrew [81, 82, 128].

[1]The work reported in this chapter was done in collaboration with A.M. Wallace and P. McAndrew of Heriot-Watt University's Computer Science Vision Group.

10.2 Segmentation and Feature Recognition

The original scenes are captured as 256x256 pixel images with 256 grey levels using a video digitiser. These are then processed using a conventional edge detector which converts each pixel to give a local edge strength and direction. These sets of edge points are then used in a Hough transformation [8, 81] to detect significant alignments which indicate the presence of the particular feature being sought. The types of features sought can be complex shapes or more basic structures such as circular arcs or straight lines. For simplicity the case where all the features are straight lines is discussed here, though other features can be handled in an almost identical manner.

Following the segmentation process each straight line in the scene is described in terms of its start and end positions, length and orientation. When the object is being viewed it is assumed that the scale of the object is unchanged but that any position and orientation is possible (this corresponds to the case of components being viewed from a fixed camera above a conveyor belt). In isolation the parameters of each segmented feature will not necessarily correspond to the stored parameters in the model. However it is possible to construct relations by considering the features in pairs which will be fairly stable under changes caused by different viewing positions. By forming such relations for both the models in the database and for the scene it is possible to come up with sets which can be compared meaningfully for agreements.

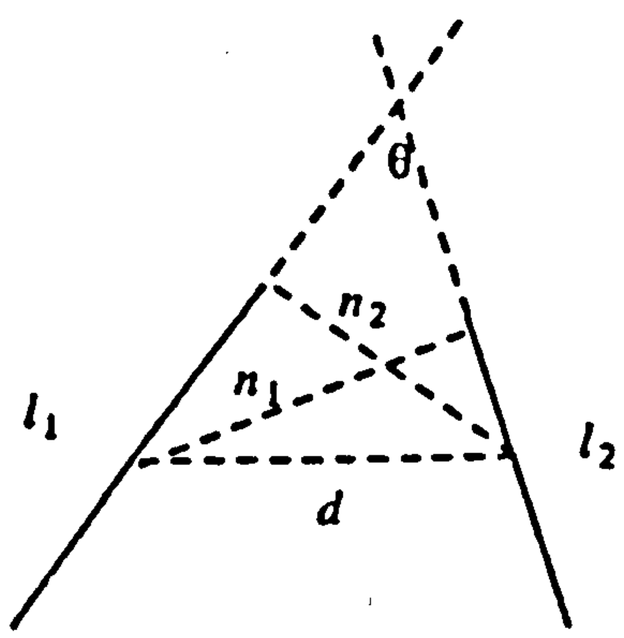


Figure 10.1: Parameters in the Pairwise Relations

The translation and rotation invariant parameters which are constructed are the angle formed at the intersection of the continued lines, θ, the distance between the mid-points of the lines, d, and the normal distances from those points to the opposite line, n_1 and n_2. These parameters are illustrated in figure 10.1. This was discussed fully in [81], with the exception that in this case both normal distances are used. This choice of parameters is not the only one possible but does carry certain advantages particularly in the context of extension to other basic features. The angle between the lines is a particularly stable factor in the automatic segmentation of straight lines; this fact is used in the matching strategy of the expert system. Both the normal distance and the distance between points can also be found in the case where it is not possible to associate a direction with one of the features, that is where one of the features is a point alone.

These coordinates of the form (d, n_1, n_2, θ) will be considered as points in *dntheta* space. This space is the set of all possible (d, n_1, n_2, θ) relationships which exist between any two lines in a picture. A match between a scene and a model will be made by comparing the *dntheta* points of the segmented scene with those of the stored models in an attempt to locate higher level relationships between the points which indicate a particular object being present. The difference in overall match between relations is then used within the reasoning process to decide the quality of a possible match.

10.3 Rule Heuristic

The heuristic is simply that when we have a rule of the form: $A_1\&...\&A_n \Rightarrow B$, and a world in which the conclusion can be either true or false, then the context split for that world is: t/n, where t is the number of antecedents which are true in the world. The final expression for the probability of B is then:

$$P(B) = \frac{a_r a_R}{n}\left(\sum_{j=1}^{n} a_j + 2\sum_{i=1}^{n}\left(\sum_{j=i+1}^{n} a_i a_j\right) + ... + n\prod_{j=1}^{n} a_j\right) \qquad (10.1)$$

as explained in section 9.2.2.

This approach avoids the need to specify a large number of weightings explicitly but still requires the summation of the factors from each of the possible worlds. If a simpler

assumption is made, that in each uncertain world the proposition is equally likely to be true or false a single expression for the probability of B can be derived. Under this assumption the probability of the proposition is given by:

$$P(B) = \frac{1}{2}(\pi_R + a_\tau a_R \prod_{j=1}^{n} a_j) \tag{10.2}$$

as explained in section 9.2.1. In this case the situation is equivalent to that described by Nilsson where the equal split is imposed on each uncertain possible world. The approach used allows calculation of this figure in a simple manner but also permits additional flexibility to specify the heuristic context split (t/n) when this assumption is no longer realistic.

10.4 Applying Probabilistic Reasoning in a Visual Context

Considering the visual context, we may express

1. The set of antecedents, $A_1, ..., A_n$ representing the propositions that a degree of similarity has been discovered between scene and model features. Normally this is based on a pairwise *dntheta* relation in the scene and a pairwise relation in a model. These propositions may be assigned a probabilistic value based on the similarity of the *dntheta* parameters.

2. B is the proposition that the model is present in the scene. Each rule, for example $A_1 \Rightarrow B$ or $A_1 \& ... \& A_n \Rightarrow B$, has a probabilistic strength attached which is calculated in proportion to the importance of the constituent pairwise (or single) features.

The maximum entropy method combines these to produce the probabilistic strength of B. If the approach is subject to no search constraints, the matching problem is combinatorially explosive. Therefore it is necessary to introduce additional constraints to obtain a workable system.

In order to illustrate the technique, consider the limited number of object models shown in figure 10.2, and the corresponding idealised scene of figure 10.3, which has been used [128]. The relations detected in figure 10.3 are the pieces of evidence and the goal is to match the models of figure 10.2 against the scene data.

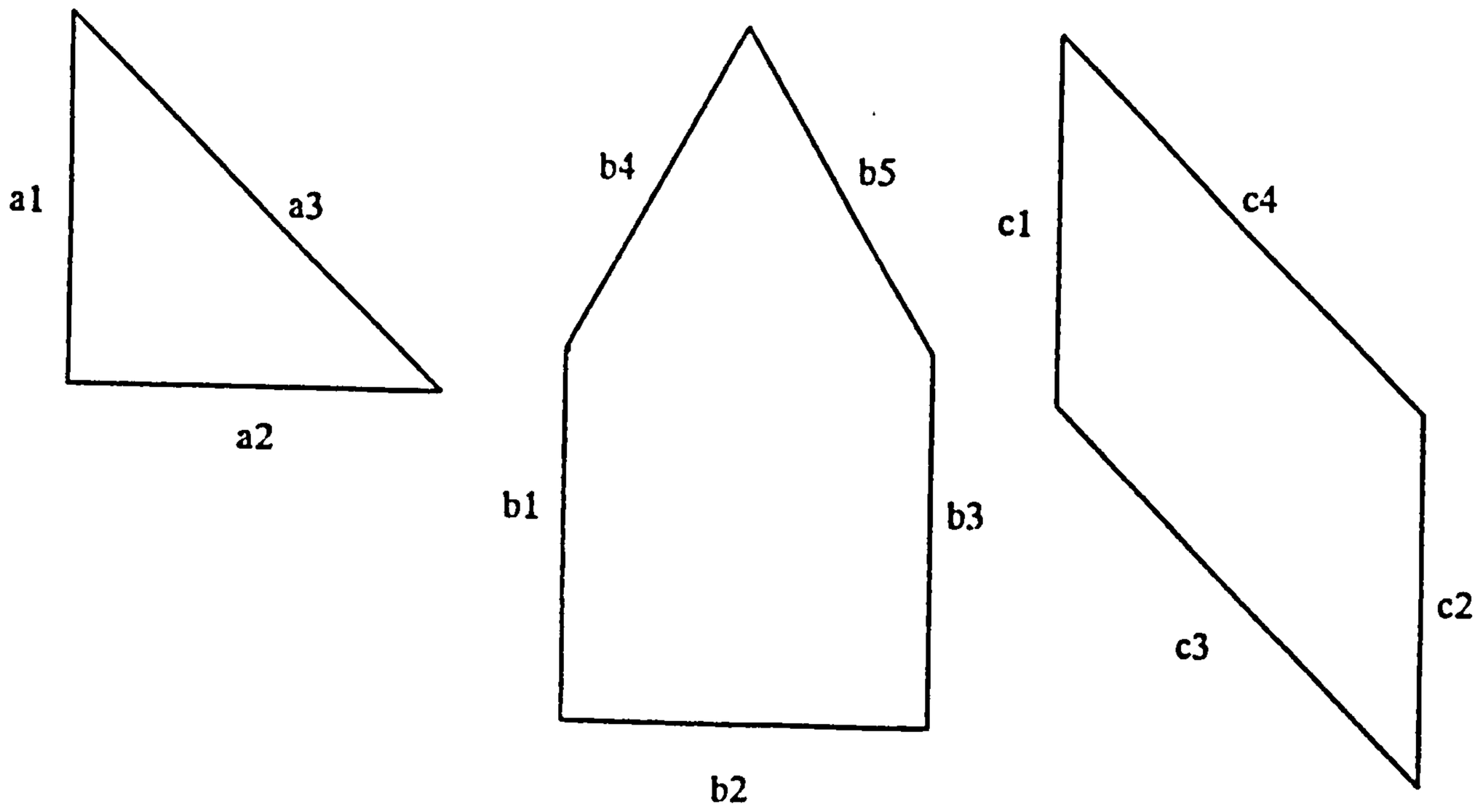


Figure 10.2: Models — Triangle, Pentagon and Quadrilateral

To create the model database, all the pairwise *dntheta* relations for each of the possible models are derived and stored [128]. This can be rapidly accessed during the subsequent scene interpretation phase. In this case, there are three relations for the triangle, ten for the pentagon and six for the quadrilateral. To analyse the scene, pairwise *dntheta* relations for the scene are derived. When features in the scene are matched to features in the model, propositions are formed stating that the same pairwise relation holds between two features in the scene as between the two corresponding features in the model.

In practice, there are many ways to match the scene features to the features of the various models. For each possible match of scene features to model features an interpretation table can be constructed containing a number of antecedent proposition $A_1, ..., A_n$ and a rule of inference. This information, together with the subjective probabilities, forms the base set for the entailment of B (see sections 3.12 and 4.2), the proposition that the features chosen from the scene do bear the same relationship to each other as the features chosen from the model.

The method of entailment is as follows:

1. Chose features to match from the scene with those in a model.

2. Assign probabilistic values to $A_1, ..., A_n$ dependent on the match between *dntheta* parameters.

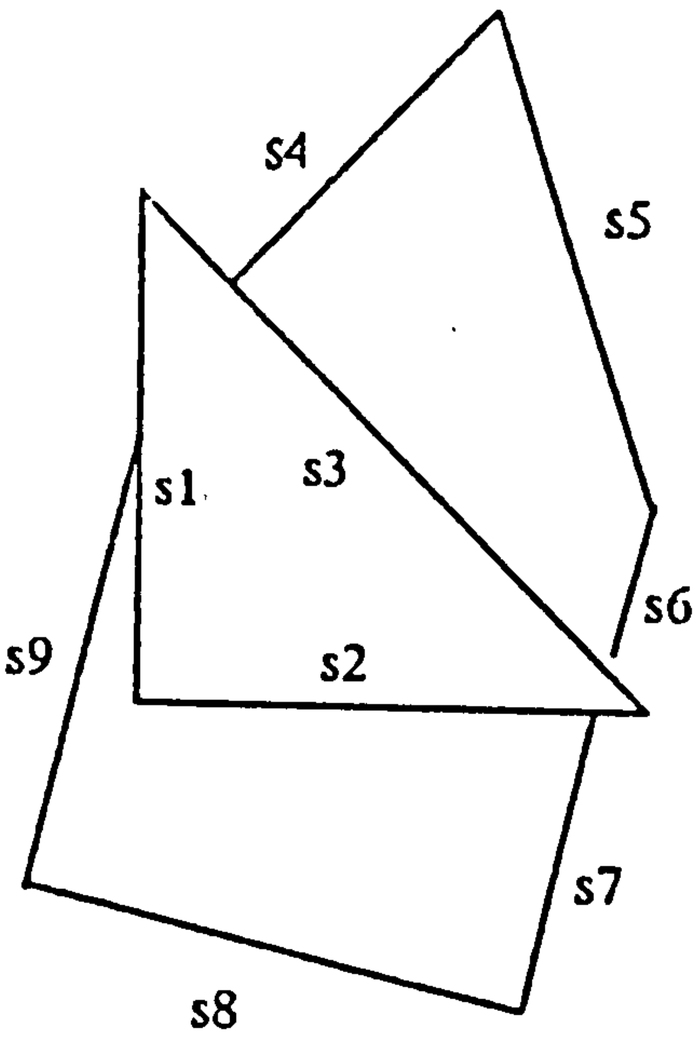


Figure 10.3: Test Scene

3. Assign a probabilistic value to the rule of entailment dependent on the strength of the inference.

4. Derive probability assignments for the possible worlds consistent with the sentence probabilities.

5. Each of these worlds provides a context in which to test the consistency of B. Deduce the entailed probability of B as the sum of the probabilities of all the worlds in which it is true.

Once a selection of m feature matches have been made, there will be $n = m(m-1)/2$ relations which can be constructed. The match between the relations formed in the scene with those in the model produce the n antecedent propositions $A_1, ..., A_n$, to which is added the rule $A_1\&...\&A_n \Rightarrow B$ resulting in an interpretation table with $n+1$ rows for which probabilities must be assigned.

The necessary values for the probabilities of the antecedents and rules may be determined heuristically. This is in contrast to the use of Bayesian techniques as applied in [128] where a predefined database is used together with assumptions about the probability of the occurrence of false and missing features to find values for conditional probabilities such as $P(B|A_1)$. For the antecedent propositions, $A_1, ..., A_n$, a linear weighting function based on the similarity of

the *dntheta* parameters is used such that a weighting of 1.0 corresponds to an exact match, and a weighting of 0.0 corresponds to a deviation greater than 20 degrees in angle, and of 40% in the 3 length parameters. These heuristic weighting functions were provided by the experts at Heriot-Watt University's Computer Science Vision Group.

$$P(A_i) = 0.25a_\theta + 0.25a_d + 0.25a_{n_1} + 0.25a_{n_2} \tag{10.3}$$

where

$$a_\theta = max\left(0, \frac{(20 - |\theta_{model} - \theta_{scene}|)}{20}\right) \tag{10.4}$$

$$a_d = max\left(0, \frac{(0.4 - |\frac{d_{scene} - d_{model}}{max(d_{scene}, d_{model})}|)}{0.4}\right) \tag{10.5}$$

With a_{n_1} and a_{n_2} defined in the same manner as a_d.

The strength of the rule, $A_1, ..., A_n \Rightarrow B$, is based on the number of consistent feature relationships between the scene and the model which have been established as a function of the number of feature relationships which are possible for a perfect match. In the simplest case, the strength of the rule may be expressed as

$$P(A_1 \& A_2 \& ... \& A_n \Rightarrow B) = \frac{N}{m} \tag{10.6}$$

where N is the number of matched features and m the number of model features.

Whenever a set of feature matches is hypothesised the above heuristics can be applied and then the evidence combined using either equation 10.1 or equation 10.2 to determine the scene-model match probability. Allowing arbitrary selection of feature matches will not generally constrain the problem sufficiently, since there are many ways in which the features can be matched each leading to a different interpretation table.

We considered two strategies for the selection of features from the scene and model — *random selection of features* and a heuristic search technique. These strategies were applied to the example of figures 10.2 and 10.3, and also to the more realistic example of figure 10.4. This latter illustration shows a scene consisting of three metal brackets, selected from a possible set of eleven metallic and plastic objects. A model of one of the brackets is also shown.

Model	Equal Prob	Weights	Best Set
Triangle	1.0000	1.0000	$\{s_1 = a_1, s_2 = a_2, s_3 = a_3\}$
Pentagon	0.5197	0.5004	$\{s_9 = b_1, s_8 = b_2, s_7 = b_3, s_4 = b_4, s_5 = b_5\}$
Quadrilateral	0.3948	0.5013	$\{s_4 = c_1, s_7 = c_2, s_6 = c_3, s_5 = c_4\}$

Table 10.1: Best Matches for Scene

10.5 Random Selection of Features

Random selection of features can be a valid approach to model matching [42], either where the subset of evidence supporting a model is significant, so that a random selection is likely to be correct, or if small data sets can be used to infer a more complete match. Table 10.1 illustrates the final derived probabilities of the presence of objects in the scene shown in figure 10.3 based on the alternative formulations of equations 10.1 and 10.2 and considering the *dntheta* relations as antecedents.

For the scene of figure 10.3 the simplest strategy of complete random selection proved capable of locating the perfect match between the model of the triangle and the scene applying either equation 10.1, where the context split is based on the number of true antecedents in each possible world, or equation 10.2 where an equal split is assumed. For the more complex models the weakness in the use of an equal split is apparent. As the term $\frac{1}{2}\pi_R$ (=0.5) dominates, the predicted probabilities occupy a small range making it difficult to determine when a match is satisfactory. In addition, the incorrect match to the quadrilateral is preferred over the correct match to the pentagon.

This is avoided by using equation 10.1 which shows the correct behaviour in ranking the match of the pentagon above that for the quadrilateral. For these models the best set can be located using random selection over a large number of trials as the time to calculate the required probability is small for each hypothesis. Nevertheless, even in this case, the number of possible interpretation tables rapidly becomes large, for the triangle $504(9*8*7)$ different interpratation tables could be constructed; while for the pentagon there could be 15120. In more complex cases, such as the real scene shown in figure 10.4, this approach is unsatisfactory and some restriction on the combinatorial search is required.

Set Size N	Partial Probability	Model Probability
2	0.8576	0.1909
3	0.7113	Inconsistent
4	0.6413	0.3080
5	0.5848	0.4181
6	0.4822	0.4822

Table 10.2: Steps in Matching Model to Real Scene Data

10.6 Heuristic Search Techniques

We developed a heuristic search strategy using the match between the *dntheta* parameters of *pairs* of features in the scene and model data described in section 10.4. In this case, matches between pairs of features in the scene and pairs of features in the model are sought initially. The set of matches is then extended by *one* feature at a time. The feature match selected for expansion at each level is determined by the highest model probability calculated on the partial match. In applying the heuristic search using the suggested set of heuristics it is found that most of the tables formed would be inconsistent as the antecedent probabilities, π_j, must be at least as large as the probability that the rule is false, $1 - \pi_R$. This condition arises from the observation that the world where the rule is false, but all antecedents true, cannot have negative probability. This condition implies that when the rule strength is weak for consistency all antecedents must have high probability. The requirement is too rigorous when seeking partial matches for future extension. By replacing the proposition that the model is present with the proposition that part of the model is present, the strength of the rule can reasonably be increased to 1.0 allowing weaker antecedent probabilities.

Using the heuristic search technique for the first example, the correct match is found more rapidly as anticipated. Of more interest is table 10.2. which shows the effect of applying the technique to the segmented scene data of figure 10.4(c).

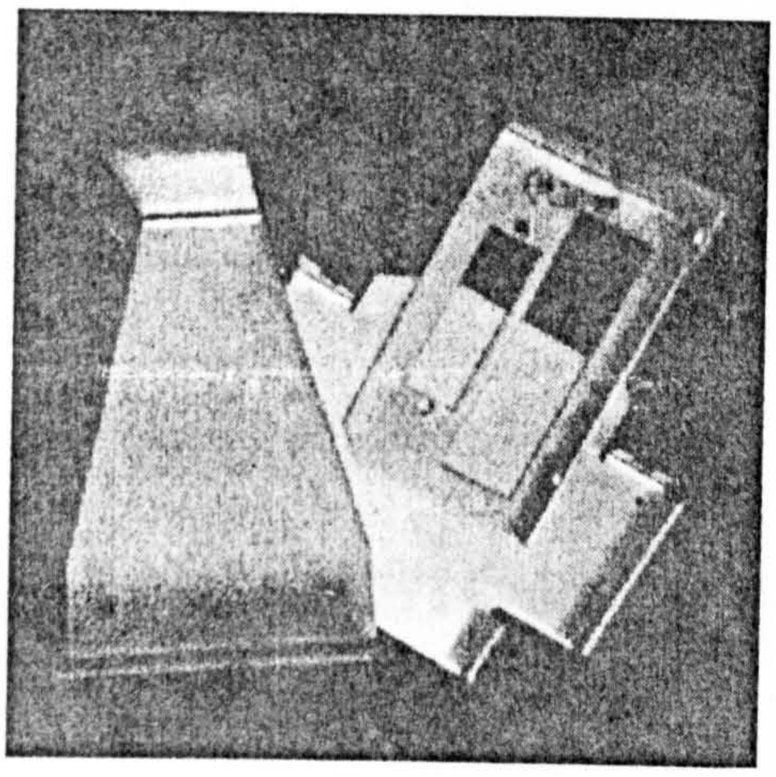
Convergence to the correct match and better probability estimates were found using equation 10.1. For the case shown the model was located in the correct position and its probability estimated as 0.4822.

10.7 Conclusion

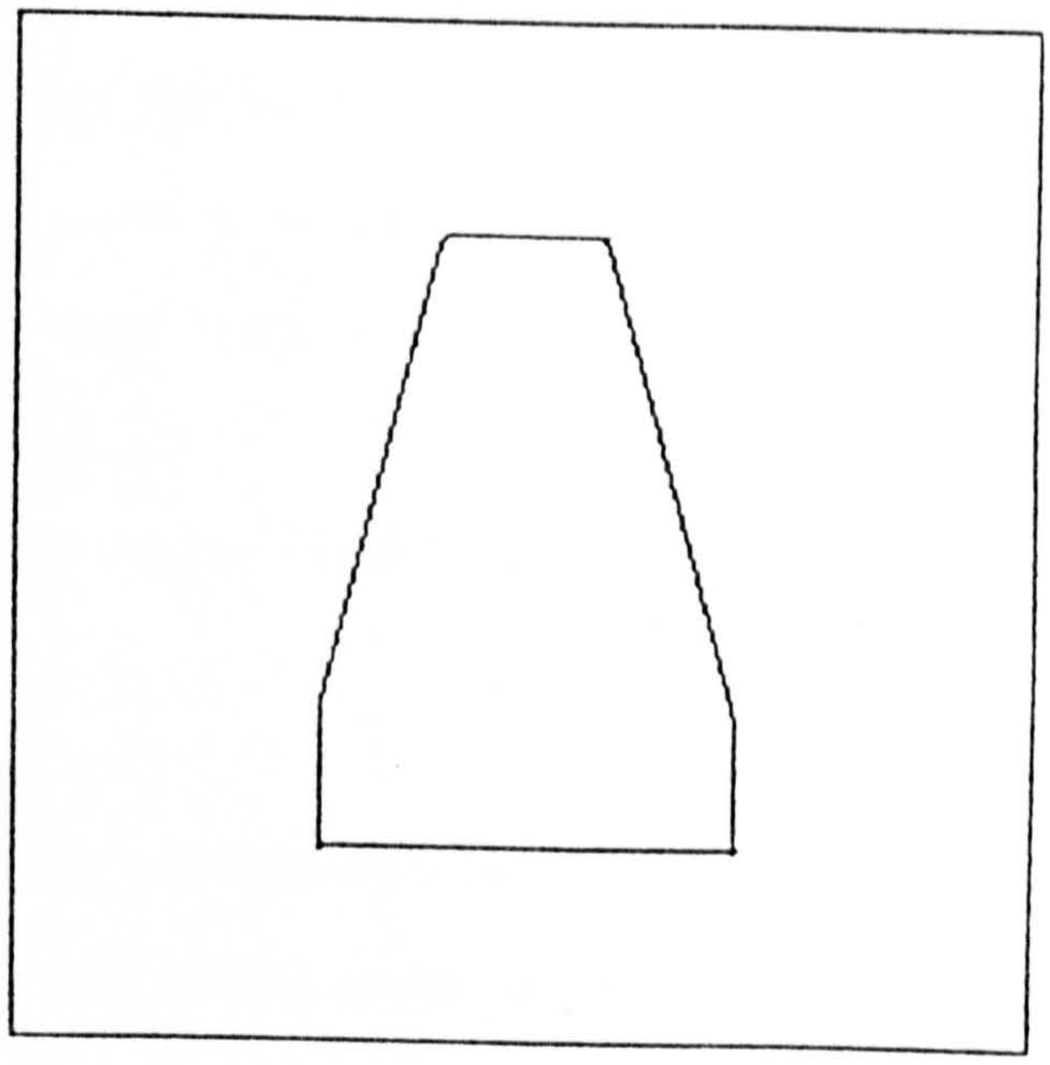
The techniques of Probabilistic Logic have been applied to interpret segmented scenes in comparison with stored models, in particular to ascertain the presence and position of objects in the original image. The basic pieces of evidence employed are the primitive linear scene features and the pairwise relations formed between these features. The pairwise constraints are useful in assessing scenes containing rigid bodies because of their invariance subject to rotation and translation of objects within the scene, and because of the capability of dealing with partial occlusion. Using the particular combination of Probabilistic Logic and the maximum entropy formalism described here, we derived a probabilistic value of the likelihood of occurrence of a known object in a scene based on heuristically determined probabilistic values for the match between scene and model pairwise relations.

The application of this technique provides a practical method to derive a measure of belief in the existence of an object in the scene, which is justifiable provided the source probabilities and heuristics employed are also meaningful. If the resulting probabilities are to give other than relative values, it is necessary to estimate the probabilities that linear features will be derived by the lower level image processing procedures, and the probability that consequent pairwise relations between them will exist, a process which is still subjective in the absence of complete models for image data in relatively unconstrained environments.

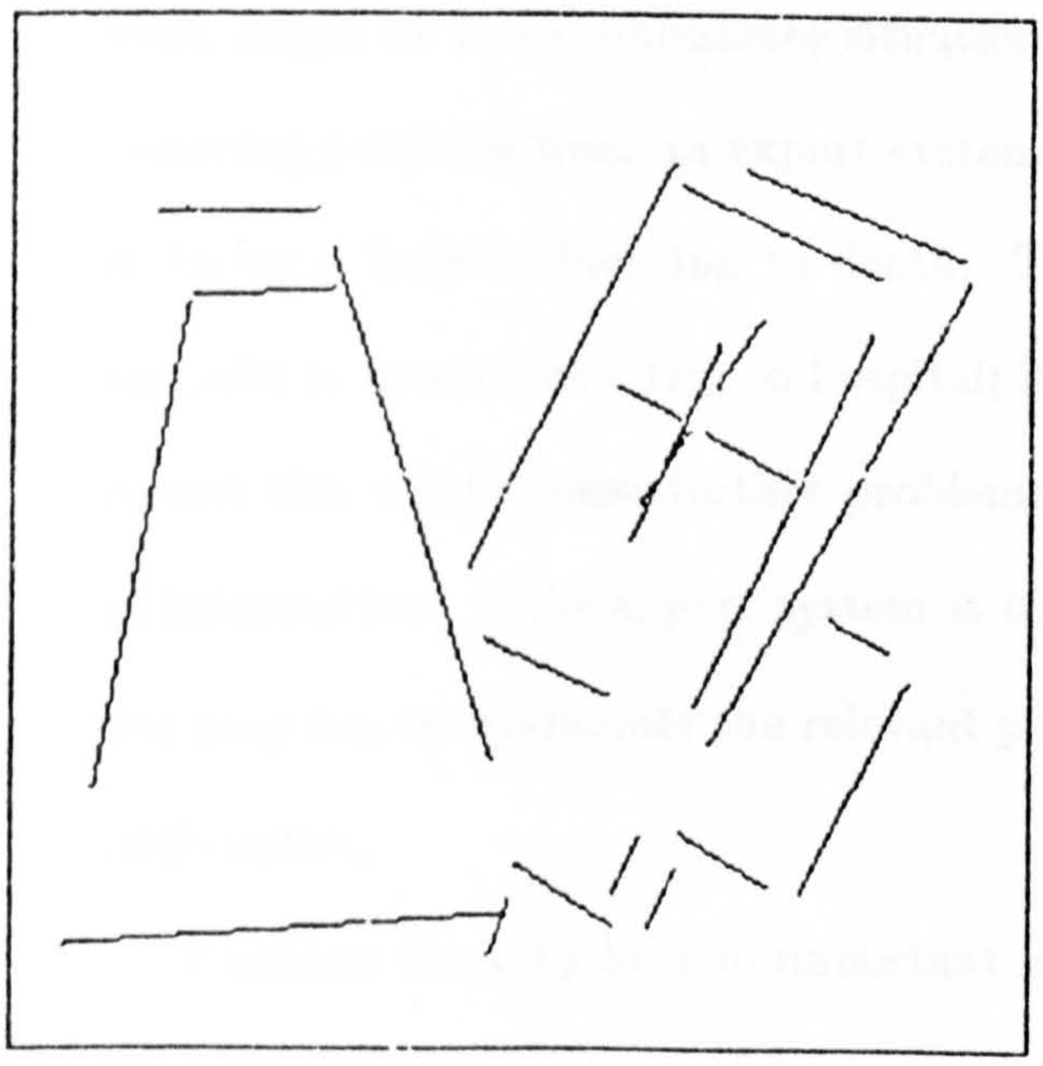
In the absence of complete information heuristics are used to judge the quality of individual matches between these relations and to evaluate the strength of the rule that a set of evidence implies the presence of a model. These heuristics can be constructed fairly easily on the basis of a known set of models. The derived probabilities can then be used to determine the best fit between the models and a scene, but not the absolute probability for the presence of a model. This is demonstrated by the use of different antecedent heuristics in section 10.6, where the models are correctly located but the calculated values for the probabilities differ.



a) Scene



b) Model



c) Segmentation

Figure 10.4: Real Scene, One Model and Segmentation

Chapter 11

Conclusion

11.1 Introduction

Expert systems have been developed to give intelligent and informative expert judgements when applied in difficult, perhaps critical, situations. An example of an expert system at work might be in an ambulance situation at the scene of an accident. A paramedic may be receiving guidance from an expert system as to what to do with a casualty who is in danger of losing a limb or bleeding to death. The paramedic might need to know how likely the casualty is to survive a trip to hospital; if it would be wise to cauterise the wound, whether or not this would cause further problems at the hospital; or a host of other critical pieces of information. If the expert system is to genuinely help the paramedic in this situation, it not only has to enumerate the relevant possibilities, but also to order them in terms of their importance.

I believe there to be two important aspects of a good expert system. The first is that internally to the system resides a semantically reliable framework for representing uncertainty and performing inference with uncertain evidence. The second is that the system is capable of Meta-Level reasoning. That is, it has knowledge of what is important in the evidence given, and is able to structure reasoning goals. Such a system would be able to understand the notion of priority.

I also feel that, although we are working towards such systems, they are still some way

off. This thesis has argued that it is practical to reason with uncertainty using Probabilistic Logic, which is a generalisation derived from predicate calculus and statistics. In the next sections I will summarise what has been achieved in this thesis, and how it could be built upon in such a way as to bring us closer to the type of reasoning with uncertainty that the paramedic could reasonably expect from the ambulance expert system.

11.2 Summary of Results Achieved in this Thesis

Uncertainty in Mathematics

Present day methods for reasoning with uncertainty have grown from two major domains. The first is mathematical inference techniques. The second is automatic deduction techniques for the digital computer. Any attempt to integrate the two faces major complexity problems (semi-decidability and intractability). An understanding of these problems places us in a position to see why simplification methods, such as MYCIN, PROSPECTOR and the simplification approximations of Bayesian Inference, have had to be used in the early expert systems, and why ultimately, any system of reasoning both with mathematical logic and probability theory must also employ simplification strategies of one sort or another.

On Nilsson's Probabilistic Logic

I have investigated the nature of probabilistic entailment with a view to answering criticisms levelled against the structure proposed by Nilsson: i.e. the inability to use a complete probability model, and the conservative estimate of probability using the half split. I have presented an interpretation of Probabilistic Logic slightly altered from Nilsson's proposed model. This new interpretation allows for the inclusion of conditional probabilities, an extended role for the maximum entropy formalism, and a new proof of the absolute bounds of an entailment problem.

The extension introduced allows Probabilistic Logic to use conditional probabilities in such a way that it is now possible to specify a complete probabilistic model in Probabilistic Logic, as for Bayesian Theory, and so to get point probability results. Also, a proof is given

as to how to deduce the bounds of an entailment without resorting to tracing the path of a convex hull in multi-dimensions. A presentation of Probabilistic Logic is made which can incorporate heuristic information and rule integration into the reasoning process.

On The Maximum Entropy Formalism In Nilsson's Probabilistic Logic

The maximum entropy formalism is the vehicle with which uncertainty is represented and judged in this thesis. The formalism is introduced and discussed in chapter 5. The conception of merging the theory of the maximum entropy principle, and the representation schema of Nilsson's probabilistic logic looks from the outset to be an impossible task when one wants to model all possible worlds; and the other wants to assign each of these worlds a least commitment probability value commensurate with the probability constraints given in the problem formulation. Maximum entropy problems usually require iterative solution. It is shown that when Nilsson's probabilistic logic is extended to allow the inclusion of conditional probabilities, there is a polynomial time algorithm for evaluating the terms of the non-linear equations.

The Relationship Between Probabilistic Logic and Bayesian Inference

With the extensions of chapter 4 Nilsson's probabilistic logic is able to use a complete probability model, if one is available, and to give the same results as Bayesian Inference. The two formalisms use different conditional information and are compared in chapter 6. The findings are that a valid statistical meaning for probabilistic entailment is more in line with the definition of conditional probability than with a generalisation of the rule of modus ponens and Bernoulli's rule of indifference. The difference between the two is that Bayesian inference is derived from a knowledge of the hypothesis whereas Nilssonian inference is derived from a knowledge of the evidence. In this regard the Nilssonian inference model is more amenable to expert systems situations where the evidence may be uncertain or incomplete than the Bayesian model.

On Incidence Calculus As A Probabilistic Logic

Bundy's incidence calculus is shown to be an approximation of Nilsson's Probabilistic Logic. Chapter 7 examines the effect of using a semantic tree theorem prover to produce all of the possible worlds for an uncertainty problem before making the probability assignment. This is a departure from Bundy's proposed model in which an arbitrary number of points are chosen, and these points become possible worlds when all the semantic relations can be represented within them (the Nilsson method in reverse). Four algorithms are presented, in this chapter, which allow a complete implementation of Incidence Calculus within the framework of Probabilistic Logic. It is shown how to involve the expert in order to deal with theorem proving problems which will lead to non-terminating proofs. The new system can be useful in situations where a complete implementation of Nilsson's probabilistic logic is impossible or impractical.

Meta-Level Reasoning

The entropy of a probability distribution is a measure as to how probability is spread over the possible worlds of an uncertain data set. The lower the entropy of a distribution, the more the probability is concentrated in small areas of the uncertainty set (i.e. the more certain it is). When the maximum entropy formalism is applied to an uncertain set and the entropy of the resultant distribution is low, then it is clear that the uncertainty is small in the result. Consequently, entropy can be used for judging the "certainty" of a maximum entropy probability distribution. This allows for the possibility of Meta-Level Inferencing: choosing the most informative rule to expand. A certainty measure is introduced to facilitate Meta-Level Inferencing in Nilsson's probabilistic logic. The proposed model also shows how "possibility" can be related to "probability".

Heuristics In Probabilistic Logic

I have considered the measured use of heuristics in Probabilistic Logic. As with all heuristics, these are applied when there is a serious deficit of reliable knowledge, or a major complexity problem which must be simplified. And, as with all heuristics, these mechanisms are designed

upon semantic considerations of the problem domain area. It is also shown how the aggregate factors of the maximum entropy solution can be used to effectively reduce the probability bounds in an entailment solution.

An Application For Probabilistic Logic In Two Dimensional Vision

The techniques of Probabilistic Logic have been applied to interpret segmented scenes in comparison with stored models, in particular to ascertain the presence and position of objects in the original image. The results are shown in chapter 10. Using the particular combination of Probabilistic Logic and the maximum entropy formalism described in this thesis, a heuristic value of the likelihood of occurrence of a known object in a scene has been derived from heuristically determined probabilistic values for the match between scene and model pairwise relations. The application of this technique provides a practical method to derive a measure of belief in the existence of an object in the scene, which is justifiable provided the source probabilities and heuristics employed are also meaningful.

The heuristics were constructed easily on the basis of a known set of models. The heuristically derived "probabilities" were used to determine the best fit between the models and a scene, but not the absolute probability for the presence of a model. This was demonstrated by the use of different antecedent heuristics and different segmentations, where the models are correctly located but the calculated "probabilities" differ.

11.3 Future Work

The work presented in this thesis is at at that level of reasoning with uncertainty which tries to represent beliefs and act on these intelligently. I believe it is important to move on from this level of reasoning in an attempt to characterise the higher level issues relating to the process of reasoning, and in particular, to the process of expert reasoning. I can see the following three ways of extending this work.

11.3.1 Applications and Investigations

Fuzzy Logic. From the work in chapter 8 it becomes semantically possible to fit Fuzzy logic inside the framework of Probabilistic Logic. It would be interesting to apply these results in a working system and make a comparison with a corresponding Fuzzy implementation.

Bayesian Networks. Since there is such a simple structure to the algorithm which finds the solution to the entropy equations in Probabilistic Logic, I feel it would be worthwhile to search for analytical solutions to entropy equations derived from problems of Bayesian Networks. It would also be interesting to know how informative a deduction from a Bayesian Network actually is.

Nonmonotonic Logics. The analysis in chapter 6 shows equivalence between Bayesian Inference and Probabilistic Logic. It also shows under what conditions the conditional probabilities of the entailment rules may be interchanged. Such a system, in its versatility at being able to cope with changing antecedent information is non-monotonic in nature. Thus it would be interesting to demonstrate how Probabilistic Logic could be used to reason nonmonotonically with uncertainty.

Heuristic Reasoning. The search for applications which would genuinely benefit from a heuristic reasoning system is also a possibility for future research.

11.3.2 The Importance of Efficient Parallel Processing

With the completion of the probability model for Probabilistic Logic we are left again facing problems of computational complexity; which are manifest in both SPACE and TIME. Spacial problems become overwhelming when the logical uncertainty space becomes too wide. As for example when too many of the proposition sentences are uncertain. Luckily, there are only two possibilities for a logical sentence— 'true' and 'false'. However, when the uncertainty ranges over n sentences, the system must necessarily deal with 2^n possible worlds. A situation which can easily become out of hand. Temporal problems enter the system either through building the semantic tree using a theorem prover, or solving complex non-linear entropy equations,

or simply through precisely estimating the probability of every possible world when there are many proposition sentences.

For example, it would not be practical to implement a possibility space suitable to completely map out an uncertainty space for five hundred rules and then to prune this space until a conclusion is reached. In such a situation the system would have somewhere in the region of 2^{500} (10^{150}) possible worlds to deal with— each containing five hundred proposition sentences— and one of these would be the true state of the world. (Although this number may seem arbitrary, MYCIN has more than five hundred rules in its database at present, and although there will be some redundancies of overlap, these would be more than made up for by including uncertain proposition antecedents in the tree.)

The fact that experts are capable of coming quickly to a conclusion from evidence of a highly complex nature testifies to the fact that time computations should be simplifiable. In the brain this is achieved through coordination of parallel processing. The computational analogy in Probabilistic Logicis that each entailment could be performed on a single processor. Each entailment producing results similar to those shown in appendix D. The *coordination* of the partial results from each processor is another area for research. Foreseeable problems will be those of: what to do when there are more rules than processors; how to make the best use of processors; and how to collate and integrate the results from the entailments back into the reasoning process.

11.3.3 A Theory of Knowledge Structuring

Another obvious area for the continued development of reasoning with uncertainty will be the development of theories of knowledge structuring. Such theories should be able to incorporate the dynamic aspects of expert reasoning: efficient parallel processing, goal formation, independent lines of analysis, the ability to summarise the current state of reasoning.

The development of such theories of knowledge structuring will probably have roots in psychology as well as mathematics. They will need to encompass both default reasoning (drawing conclusions from tentative evidence) as well as autoepistemic reasoning (reasoning about one's beliefs at the moment). What Probabilistic Logic offers in this endeavour is a

semantically justifiable structure which can easily be experimented upon.

11.4 General Conclusions

Present day expert reasoning systems perform a trade-off between semantic clarity and algorithmic performance in time. In this thesis I have examined Nilsson's probabilistic logic, a paradigm for reasoning with uncertainty which is built up from the simplest principles of dealing with uncertainty— namely to enumerate all of the possibilities of an uncertain situation, and then to apply some qualitative judgements among them using an appropriate utility measure. This system has been extended to be capable of operating over a complete probability model, which in turn has given Probabilistic Logic an extended role in the area of reasoning with uncertainty.

In particular when Probabilistic Logic is combined with the maximum entropy formalism, we have a system capable of producing point probability values, similar to those produced by MYCIN for example. These results can be thought of as "guesses" as to the true value of the probability based on the amount of information available. This "guess" is based on the principle of maximum entropy which imposes the least amount of assumptions on the available information and always gives a consistent estimate, and as such it has a sound mathematical calibre. I have also examined the utility of this guess in various conditions of uncertainty and related it to the bounds of uncertainty in which it is placed.

Appendix A

Thesis Nomenclature

Each of the operators and operations summarised below will be introduced more formally in the thesis proper. This section collects together the typographical conventions used throughout the text.

Logical Notations

LOGICAL AND ..... &

LOGICAL OR ..... ∨

LOGICAL NOT ..... ~

LOGICAL IMPLICATION ..... ⇒

CONDITIONAL PROBABILITY .....|

Statistical Notations

The probability of X ..... $p(X)$

The prior probability of X ..... $p'(X)$

The probability of X conditioned on Y ..... $p(X|Y)$

where p(X | Y) is defined as: $p(X|Y) = p(X\&Y)/p(Y)$

Interpretation Tables

Sentence	a	b	c	d	Probability
A	1	1	1	1	1
B	1	0	1	1	π_2
C	1	0	0	1	π_3

Table A.1: An Example Interpretation Table

Shown in table A.1 are three logical sentences A, B and C; which have been assigned the probabilities π_1, π_2 and π_3 respectively. The uncertainty leads to four possible worlds, here labelled a, b, c and d. In general, possible worlds are denoted with lower case letters, and logical sentences with upper case letters. The logical conditions which hold in a possible world are read in the columns of the matrix. Ones in the columns represent the logical value 'TRUE', and zeros represent 'FALSE'. So that, for example, in world c A is true, B is true and C is true.

The probabilistic equations from the above table are:

$$\begin{aligned} a+b+c+d &= 1 \\ a+c+d &= \pi_2 \\ a+d &= \pi_3 \end{aligned}$$

The probability of a sentence is the sum of the probabilities of all the possible worlds in which the sentence is true. Sentence A is true in each world, and is consequently a tautology (with probability 1). The probabilities attached to sentences B and C provide equational constraints on an acceptable probability distribution.

Appendix B

Prerequisites

Knowledge Modelling in Mathematical Logic

Aristotle introduced two of the most important of the tools necessary for systematically investigating this area: a logical calculus for reasoning (which led ultimately to predicate calculus); and the systemised notion of possible worlds, (which led to the development of Modal Logics). Consider the following example which demonstrates the utility of these two formalisms.

Janet is constantly forgetting what day of the week it is, but she does know that:

1. if the church bells are ringing it is Sunday;
2. if it is Sunday there will be brocolli bake for lunch;
3. any other day there is porridge for lunch.

This knowledge of Janet's is simplified by making some abbreviations. For example, as below where the token in bold case may be used to denote the logical proposition on its right.

CBR The church bells are ringing.
ISU It is Sunday.
BBL There will be brocolli bake for lunch.
PFL There will be porridge for lunch.

These tokens are basic *sentences* of the predicate calculus. To represent a deductive step the notion of implication, or alternatively, entailment, is used. In predicate calculus a proposition A implies a proposition B if whenever A is true B is true (but not necessarily conversely).

Proposition A is called the *antecedent proposition*, and proposition B is called the *consequent proposition*. In this regard, the three sentences above may be rewritten:

1. The ringing of the church bells implies that it is Sunday;
2. Today being Sunday implies that there will be brocolli bake for lunch;
3. Today not being Sunday implies that there will be porridge for lunch.

A further stage in the abstraction process is to represent the function *implies that* with a symbolic operator. I will use the symbol ⇒ for this purpose. The sentences which represent Janet's knowledge may be considered as rules and can be expressed as follows:

1. **CBR ⇒ ISU**
2. **ISU ⇒ BBL**
3. **~ ISU ⇒ PFL**

Let us assume that Janet hears the church bells ringing, that is, **CBR** is true. From this we can conclude that **ISU** is true, and from this that **BBL** is true. So that on hearing church bells, Janet can conclude that there will be brocolli bake for lunch.

Modal logic [9, 58] allows propositions to be possibly true as well as certainly true. In the example above, the church bells might also ring at a marriage. Now, the integrity of the reasoning process is lost when we conclude that every time Janet hears church bells there will be brocolli bake for lunch. If instead of saying that it is a rule that when the church bells are ringing we can imply that it is Sunday, we say that when the church bells are ringing we can imply that it is possible that it is Sunday, the integrity of the reasoning process is regained. Sentences are now no longer just of the values "true" and "false" but can also be "possibly true" and "necessarily true". When Janet hears the church bells it is only possible that there will be brocolli bake for lunch, and therefore still possible that lunch will be porridge.

The *real world* is only one of a number of possibilities, each of which can be shown consistent with the uncertain sentences. These possibilities for the real world are known as *possible worlds*. A drawback with modal logic is that it is not able to use any other information so as to say that one possible world is more likely than another. That is, there is no way to judge the quality of the uncertainty spread amongst the possible worlds.

Uncertainty Management

In the eighteenth century Jakob Bernoulli proposed a concrete manner of placing judgement of uncertainty on a scale from 0 to 1; hence formalising a concept of probability. Two definitions of probability came from this work.

Range Theory If there are n equi-possible states for an event, and A is true in m of these, then the probability of A is m/n.

Frequency Theory If you take a large number of measurements, say n, of an event, and A is true in m of these, then as n tends to infinity, so also does the ratio of m/n tend to the probability of A.

The theory of probability developed various other concepts of probability, but all follow simple axioms, and can be used in a more descriminating for choosing between a host of possibilities, and furthermore, to quantify the strength of judgement of a proposition.

One further aspect of expert reasoning is the expert's deployment of intuitional or inspirational procedures for reasoning, based upon personal *rules of thumb*. This final aspect of expert system reasoning moves further away from a logical basis for action towards the cultivation of fruits of experience, and psychological preference. In modelling this aspect of reasoning with uncertainty, the expert system must be capable of structuring the knowledge, and acting in what might appear to a novice to be an unpredictable manner. In artificial intelligence, this requirement is addressed by the emerging theory of *heuristic reasoning*. The word heuristic has good and bad connotations in Artificial Intelligence. It means inspired guesswork.

Appendix C

Entropy Aggregates for Rules of Probability 1

In this appendix is presented a proof that the aggregate factors in an entailment problem whose probability is 1 are given by the equations:

$$a_j \ (j = 1 \ to \ n) = \frac{\pi_j}{1 - \pi_j}, \tag{C.1}$$

$$a_R = \frac{1}{\prod_{j=1}^{n} (1 + a_j)}. \tag{C.2}$$

C.0.1 Inductive Proof of Entropy Algorithm

The proof proceeds in two stages. First, to show that for each of the terms $a_j \ (j = 1 \ to \ n)$ there is a direct match of terms on the numerator and denominator of the expression: $(\pi_j)/(1 - \pi_j)$, i.e. all the unknown worlds where sentence A_j is true divided by all the worlds where A_j is false. The second stage is for the final factor a_R and is based on the worlds in which the rule is true, and a recursive expression for describing the contribution of each of the possible worlds to this probability: $a_R \prod_{j=1}^{n} (1 + a_j) = 1$.

Base Case (n=1)

From the equations of table C.1:

Sentence	Possible Worlds		Equations
A_1	0	1	$a_1 a_R = \pi_1$
$A_1 \Rightarrow B$	1	1	$a_R + a_1 a_R = 1$

Table C.1: Sentences, Worlds, and Equations

$$a_1 = \frac{a_1 a_R}{a_R} = \frac{\pi_1}{(1-\pi_1)} \tag{C.3}$$

$$a_R(1+a_1) = 1; \;\Rightarrow\; a_R = \frac{1}{1+a_1} \tag{C.4}$$

And the above equations satisfy the algorithm with n = 1.

Step

The algorithm is true for n antecedents, now to prove it true for n+1 antecedents.

The new premise A_{n+1} is added to the antecedent arm of the rule, and placed after premise A_n in the premise list. We now have aggregate factors $a_1, ..., a_{n+1}, a_R$.

1. In each row there are now 2^{n+1} possible worlds, where there used to be 2^n. The difference between the tree for n+1 propositions and n propositions, being that in row n+1 there are now 2^n 1's and 2^n 0's, and the rule is pushed down to position n+2.

   For the half of the tree with 0's in row n+1 we proved that there is a direct match to give each of the previous a_i's. For the other half, we use the same enumeration, and find that the factor for proposition n+1 cancels out on top and bottom. Furthermore the numerator still only holds the worlds where sentence A_j is true, and the denominator the worlds where A_j is false. Therefore the equation still holds.

   Is the formula true for new row n+1?

   The new tree was made up of two identical copies of the old tree, one of which has a 1 in row n+1, the other of which has a zero in row n+1. Consequently, again it is possible to cancel the terms of the true worlds divided by the false worlds so that there is only a factor of a_{n+1} left.

So, for each of the antecedents, A_1 to A_{n+1} the expression for the associated aggregate factor is:

$$a_j \ (j = 1 \ to \ n+1) = \frac{\pi_j}{1-\pi_j}.$$

2. The expression for all the worlds where the rule is true is: $a_R \prod_{j=1}^{n} (1 + a_j)$

When we include the new row, we have a new multiplicative factor: We have two copies: one with an a_{n+1} in row n+1, and one with a 1. So the new expression for all the worlds is:

$$a_{n+1} a_R \prod_{j=1}^{n} (1 + a_j) + a_R \prod_{j=1}^{n} (1 + a_j) \ = \ (1 + a_{n+1}) a_R \prod_{j=1}^{n} (1 + a_j) \qquad \text{(C.5)}$$

$$= a_R \prod_{j=1}^{n+1} (1 + a_j) \qquad \text{(C.6)}$$

The probability for these worlds all summed together is 1, and so consequently, the factor associated with the rule of inference is:

$$a_R \ = \ \frac{1}{\prod_{j=1}^{n+1} (1 + a_j)}$$

and again the expression has been succesfully extended.

Appendix D

An Example of The Use of Heuristics

The data in this appendix refers to the rules and associated antecedent probabilities shown in chapter 8 tables 8.2 and 8.3. The first section of each of the figures is the rule number and its probability. The next shows the probabilities assigned to the antecedents. In the third section is a list of the possible worlds with the rule true, the associated maximum entropy probability for the possible world, and then two columns showing the half split heuristic result and the n-split result. Note that in all these figures there is one world missing. That is the world with all of the antecedents true and the rule false. It's probability is always 1 less the probability of the rule, and the conclusion is never true in this world so it is not included. Another point of note is that in the possible world with all of the antecedents true and the rule true, the probability of this world is not reduced from that assigned by the maximum entropy formalism. This is because the conclusion can only be true in this world.

The fourth section then gives the heuristically estimated probability of the conclusion from the n-split and the half split; plus the entropy information: first the entropy of the resultant maximum entropy distribution, then the minimum possible entropy and then maximum possible entropy for the rule with this number of antecedents. Finally, the specificity of the rule and it's antecedent probabilities is shown.

D.1 Heuristics for Rules 1-5

```
-------------------------------------------------------
Rule 1: 0.8
-------------------------------------------------------

Antecedents: [0.83,0.57,0.91]
-------------------------------------------------------

Worlds:
   1:  [0,0,0]    0.01028     0.0051399     0
   2:  [0,0,1]    0.081095    0.040548      0.027032
   3:  [0,1,0]    0.0088453   0.0044227     0.0029484
   4:  [0,1,1]    0.06978     0.03489       0.04652
   5:  [1,0,0]    0.038095    0.019048      0.012698
   6:  [1,0,1]    0.30053     0.15026       0.20035
   7:  [1,1,0]    0.03278     0.01639       0.021853
   8:  [1,1,1]    0.25859     0.25859       0.25859
-------------------------------------------------------

(Nsplit = 0.57;    Hsplit = 0.5293).
Entropy: 1.7478;   Minent: 0.5004;   Maxent: 2.1639

-------------------------------------------------------

Specificity: 0.74986
-------------------------------------------------------
```

```
-------------------------------------------------------
Rule 2: 0.6
-------------------------------------------------------

Antecedents: [0.78,0.45,0.95]
-------------------------------------------------------

Worlds:
   1:  [0,0,0]    0.016806   0.0084028    0
   2:  [0,0,1]    0.18486    0.092431     0.06162
   3:  [0,1,0]    0.0015278  0.00076389   0.00050926
   4:  [0,1,1]    0.016805   0.0084027    0.011204
   5:  [1,0,0]    0.029028   0.014514     0.0096759
   6:  [1,0,1]    0.3193     0.15965      0.21287
   7:  [1,1,0]    0.0026389  0.0013194    0.0017593
   8:  [1,1,1]    0.029028   0.029028     0.029028
-------------------------------------------------------

(Nsplit = 0.32667;    Hsplit = 0.31451).
Entropy: 1.4115;   Minent: 0.67301;   Maxent: 1.9207

-------------------------------------------------------

Specificity: 0.59191
-------------------------------------------------------
```

```
-----------------------------------------------------
Rule 3: 0.7
-----------------------------------------------------

Antecedents: [0.9,0.8,0.6,0.9]
-----------------------------------------------------

Worlds:
   1:  [0,0,0,0]    0.0023324   0.0011662    0
   2:  [0,0,0,1]    0.013994    0.0069971    0.0034986
   3:  [0,0,1,0]    0.0017493   0.00087464   0.00043732
   4:  [0,0,1,1]    0.010496    0.0052478    0.0052478
   5:  [0,1,0,0]    0.0058309   0.0029155    0.0014577
   6:  [0,1,0,1]    0.034985    0.017493     0.017493
   7:  [0,1,1,0]    0.0043732   0.0021866    0.0021866
   8:  [0,1,1,1]    0.026239    0.01312      0.019679
   9:  [1,0,0,0]    0.013994    0.0069971    0.0034986
   10:  [1,0,0,1]    0.083965    0.041982     0.041982
   11:  [1,0,1,0]    0.010496    0.0052478    0.0052478
   12:  [1,0,1,1]    0.062973    0.031487     0.04723
   13:  [1,1,0,0]    0.034985    0.017493     0.017493
   14:  [1,1,0,1]    0.20991     0.10496      0.15743
   15:  [1,1,1,0]    0.026239    0.01312      0.019679
   16:  [1,1,1,1]    0.15743     0.15743      0.15743
-----------------------------------------------------
(Nsplit = 0.5;    Hsplit = 0.42871).
Entropy: 2.0818;   Minent: 0.61086;   Maxent: 2.5517

-----------------------------------------------------

Specificity: 0.75792
-----------------------------------------------------
```

```
-------------------------------------------------------
Rule 4: 0.9
-------------------------------------------------------

Antecedents: [0.83,0.9]
-------------------------------------------------------

Worlds:
   1:  [0,0]    0.018889    0.0094445    0
   2:  [0,1]    0.15111     0.075555     0.075555
   3:  [1,0]    0.081111    0.040556     0.040556
   4:  [1,1]    0.64889     0.64889      0.64889
-------------------------------------------------------

(Nsplit = 0.765;     Hsplit = 0.77444).
Entropy: 1.0752;    Minent: 0.32508;    Maxent: 1.5727

-------------------------------------------------------

Specificity: 0.6012
-------------------------------------------------------



-------------------------------------------------------
Rule 5: 0.9
-------------------------------------------------------

Antecedents: [0.45,0.8]
-------------------------------------------------------

Worlds:
   1:  [0,0]    0.12222     0.061111     0
   2:  [0,1]    0.42778     0.21389      0.21389
   3:  [1,0]    0.077778    0.038889     0.038889
   4:  [1,1]    0.27222     0.27222      0.27222
-------------------------------------------------------

(Nsplit = 0.525;     Hsplit = 0.58611).
Entropy: 1.4032;    Minent: 0.32508;    Maxent: 1.5727

-------------------------------------------------------

Specificity: 0.86414
-------------------------------------------------------
```

D.2 Rules 6,7,8 with half split

```
-------------------------------------------------
Rule 6: 0.8
-------------------------------------------------

Antecedents: [0.5293,0.9]
-------------------------------------------------

Worlds:
   1:  [0,0]    0.058838   0.029419    0
   2:  [0,1]    0.41186    0.20593     0.20593
   3:  [1,0]    0.041163   0.020581    0.020581
   4:  [1,1]    0.28814    0.28814     0.28814
-------------------------------------------------

(Nsplit = 0.51465;    Hsplit = 0.54407).
Entropy: 1.3438;   Minent: 0.5004;   Maxent: 1.6094

-------------------------------------------------

Specificity: 0.76046
-------------------------------------------------
-------------------------------------------------
Rule 7: 0.7
-------------------------------------------------

Antecedents: [0.77444,0.78]
-------------------------------------------------

Worlds:
   1:  [0,0]    0.07089    0.035445    0
   2:  [0,1]    0.15467    0.077335    0.077335
   3:  [1,0]    0.14911    0.074555    0.074555
   4:  [1,1]    0.32533    0.32533     0.32533
-------------------------------------------------

(Nsplit = 0.47722;    Hsplit = 0.51266).
Entropy: 1.4866;   Minent: 0.61086;   Maxent: 1.5813

-------------------------------------------------

Specificity: 0.90242
-------------------------------------------------
```

```
-------------------------------------------------
Rule 8: 0.8
-------------------------------------------------

Antecedents: [0.42871,0.77444]
-------------------------------------------------

Worlds:
   1:  [0,0]    0.16108    0.080538    0
   2:  [0,1]    0.41021    0.20511     0.20511
   3:  [1,0]    0.064485   0.032242    0.032242
   4:  [1,1]    0.16422    0.16422     0.16422
-------------------------------------------------

(Nsplit = 0.40157;    Hsplit = 0.48211).
Entropy: 1.455;   Minent: 0.5004;   Maxent: 1.6094

-------------------------------------------------

Specificity: 0.86072
-------------------------------------------------
```

D.3 Rules 6,7,8 with n split

```
-------------------------------------------------
Rule 6: 0.8
-------------------------------------------------

Antecedents: [0.57,0.9]
-------------------------------------------------

Worlds:
   1:  [0,0]    0.05375    0.026875    0
   2:  [0,1]    0.37625    0.18812     0.18812
   3:  [1,0]    0.04625    0.023125    0.023125
   4:  [1,1]    0.32375    0.32375     0.32375
-------------------------------------------------

(Nsplit = 0.535;    Hsplit = 0.56187).
Entropy: 1.3541;   Minent: 0.5004;   Maxent: 1.6094

-------------------------------------------------

Specificity: 0.76975
-------------------------------------------------
```

Rule 7: 0.7

Antecedents: [0.765,0.78]

Worlds:

1:	[0,0]	0.073857	0.036929	0
2:	[0,1]	0.16114	0.080571	0.080571
3:	[1,0]	0.14614	0.073071	0.073071
4:	[1,1]	0.31886	0.31886	0.31886

(Nsplit = 0.4725; Hsplit = 0.50943).
Entropy: 1.4933; Minent: 0.61086; Maxent: 1.5813

Specificity: 0.90936

Rule 8: 0.8

Antecedents: [0.5,0.765]

Worlds:

1:	[0,0]	0.14688	0.073438	0
2:	[0,1]	0.35312	0.17656	0.17656
3:	[1,0]	0.088125	0.044062	0.044062
4:	[1,1]	0.21187	0.21187	0.21187

(Nsplit = 0.4325; Hsplit = 0.50594).
Entropy: 1.514; Minent: 0.5004; Maxent: 1.6094

Specificity: 0.91397

Bibliography

[1] J.B. Adams. A probability model of medical reasoning and the MYCIN model. *Mathematical Biosciences*, 32:177–186, 1976.

[2] Y. Bard. Estimation of state probabilities using the maximum entropy principle. *IBM Journal of Research and Development*, 24:563–569, 1980.

[3] Y. Bard. Maximum entropy principle, classical approach. In *Encyclopedia of Statistical Sciences*, pages 336–338, John Wiley & Sons, U.S.A., 1982. (Kotz and Johnson eds.).

[4] Y. Bard. A model of shared DASD and multipathing. *Communications of the ACM*, 23:564–572, 1980.

[5] T. Bayes. An essay towards solving a problem in the doctrine of chances. In E.S. Pearson and M.G. Kendall, editors, *Studies in the History of Statistics and Probability*, pages 134 — 153, Griffin, London, 1970.

[6] R. E. Bellman and L. A. Zadeh. Local and fuzzy logics. 103 – 165, 1977.

[7] J. Bernoulli. In *Ars Conjectandi*, Montmort, 1713.

[8] I. Biedermann. Human image understanding: recent research and a theory. *Computer Vision, Graphics and Image Processing*, 32:29–73, 1985.

[9] R. Bradley and N. Swartz. *Possible Worlds.* Basil Blackwell Publisher, Oxford, 1979.

[10] B. Buchanan and E. Shortliffe. *Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, Reading, MA., 1984.

[11] B.G. Buchanan, G. Sutherland, and E. Geirgenbaum. Heuristic DENDRAL: a program for generating explanatory hypotheses in organic chemistry. In *Machine Intelligence*, Elsevier, New York, 1969.

[12] A. Bundy. Correctness criterea of some algorithms for uncertain reasoning using incidence calculus. In *Journal of Automated Reasoning*, pages 109–126, D. Reidel, Dordrecht, 1986.

[13] A. Bundy. Incidence calculus: a mechanism for probabilistic reasoning. *Journal of Automated Reasoning*, 1:263–283, 1985.

[14] C.L. Chang and R.C.T. Lee. *Symbolic Logic and Mechanical Theorem Proving.* Academic Press, New York, 1973.

[15] P. Cheeseman. In defense of probability. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, William Kaufman, Los Angeles, January 1985.

[16] P. Cheeseman. A method for computing generalised Bayesian probability values for expert systems. In *Proceedings of the 6th International Joint Conference on Artificial Intelligence*, William Kaufman, Los Angeles, 1983.

[17] C.K. Chow and T.J. Wagner. Approximating discrete probability distributions with dependance trees. *IEEE Transactions on Information Theory*, 462–467, 1968.

[18] W.J. Clancey. The epistomology of a rule-based expert system: a framework for explanation. *Artificial Intelligence*, 20:215 — 252, 1983.

[19] W.F. Clocksin and C.S. Mellish. *Programming in Prolog.* Springer-Verlag, 1984.

[20] G.F. Cooper. Nestor: a computer based medical diagnostic aid that integrates causal and probabilistic knowledge. *Technical Report: Medical Computer Science Group*, 1984.

[21] G.F. Cooper. Probabilistic inference using belief networks is np-hard. In *Report KSL-87-27*, Medical Computer Science Group, Stanford University, 1987.

[22] R.A. Corlett and S.J. Todd. A monte-carlo approach to uncertain inference. In P. Ross, editor, *Proceedings of AISB-85*, pages 28–34, 1985.

[23] A.A. Cournot. In *Exposition de la theorie des chances et des probabilities*, Hachette, Paris, 1843.

[24] D.R. Cox and D.V. Hinkley. *Theoretical Statistics.* Chapman and Hall, London, 1974.

[25] R. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14:1 — 13, 1946.

[26] F.N. David. Dicing and gaming (a note on the history of probability. In E.S. Pearson and M.G. Kendall, editors, *Studies in the History of Statistics and Probability*, pages 1 — 17, Griffin, London, 1970.

[27] P.J. Davis and R. Hersh. In *The Mathematical Experience*, pages 330–339, Birkhauser, Boston, 1981.

[28] B. de Finetti. Probabilities of probabilities: a real problem or a misunderstanding. In Aykac and Brumet, editors, *New Developments in the Applications of Bayesian Methods*, pages 1–10, North Holland, Amsterdam, 1977.

[29] A. DeMoivre. In *Doctrine of Chances*, London, 1718.

[30] A.P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:157–175, 1967.

[31] Dromey. *Software Engineering.* Prentice Hall, New Jersey, 1981.

[32] D. Dubois and H. Prade. New results about properties and semantics of fuzzy-set-theoretic operators. In P.P. Wang and S.K. Chang, editors, *Fuzzy Sets. Theory and Applications to Policy Analysis and Information Systems*, pages 59–75, Plenum, New York, 1980.

[33] R. Duda, P. Hart, and N. Nilsson. Subjective Bayesian methods for rule-based inference systems. *Proceedings of the 1976 National Computer Conference*, 45:1075–1082, 1976.

[34] R.O. Duda, P.E. Hart, N.J. Nilsson, R. Reboh, and J. Sutherland. Development of a computer-based consultant for mineral exploration. In *Annual Report (projects 5821 and 6415)*, SRI International, Menlo Park, California, 1977.

[35] R. Fagin and J.Y. Halpern. Reasoning about knowledge and probability. In M. Vardi, editor, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 277–293, Morgan Kaufmann, Los Altos, California, 1988.

[36] E. Feigenbaum and J. Feldman. In *Computers and Thought*, McGraw-Hill, New York, 1963.

[37] E. Feigenbaum, H.P. Nii, and P. McCorduck. *The Rise of the Expert Company.* Vintage Books, New York, 1989.

[38] W. Feller. In *An Introduction to probability theory and applications*, Wiley, New York, 1968.

[39] T. Fine. *Theories of Probability.* Academic Press, New York, 1973.

[40] B. De Finetti. In *Theory of Probability*, Wiley, New York, 1974.

[41] B. De Finetti. Foresight: its logical laws, its subjective sources. *Annals of H. Poincare Institute*, 7:1 — 68, 1937.

[42] M.A. Fischler and R.C. Bolles. Random sample concensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[43] R.A. Fisher. The underworld of probability. *Sankhya*, 18:201 — 210, 1957.

[44] G. Frege. Begriffsschrift, a formula language modelled upon that of arithmetic, for pure thought (1879). In J. Van Heijenoort, editor, *From Frege to Godel: a source book in mathematical logic, 1879 — 1931*, pages 1–82, Harvard University Press, Cambridge, Mass., 1967.

[45] R.A. Frost. *Introduction to Knowledge Base Systems.* Collins, London, 1986.

[46] M.R. Genesereth and N. Nilsson. In *Logical Foundations of Artificial Intelligence*, Morgan Kaufmann, Palo Alto, CA, 1987.

[47] S.A. Goldman and R.L. Rivest. Making maximum entropy computations easier by adding extra constraints. In Lemmer and Kanal, editors, *Uncertainty in Artificial Intelligence*, pages 133–148, North Holland, Amsterdam, 1986.

[48] I.J. Good. In *Probability and the weighing of evidence*, Griffin, London, 1950.

[49] G.A. Gorry and G.O. Barnett. Experience with a model of sequential diagnosis. *Computers and Biomedical Research*, 1:490–507, 1968.

[50] G.A. Gorry and G.O. Barnett. Sequential diagnosis by computer. *Journal of the American Medical Association*, 205:849–854, 1968.

[51] B. Grosof. An inequality paradigm for probabilistic logic. In *Uncertainty in Artificial Intelligence*, pages 259–275, Elsevier Science Publishers, 1986.

[52] H. Guggenheimer and R.S. Freedman. Foundations of probabilistic logic. In *Proceedings of National Conference on Artificial Intelligence*, pages 939–941, 1987.

[53] J.Y. Halpern and M. Rabin. A logic to reason about likelihood. *Artificial Intelligence*, 32(3):379–406, 1987.

[54] B. Harris. Entropy. In *Encyclopedia of statistical Sciences*, pages 512–516, John Wiley & Sons, U.S.A., 1982. (Kotz & Johnsons eds.).

[55] D.R. Hofstadter. In *Godel, Escher, Bach: an eternal golden braid*, pages 246–272, The Harvester Press, 1979.

[56] J.E. Hopcroft and D. Ullman. *Introduction to automata theory, languages, and computation.* Addison-Wesley, Philippines, 1979.

[57] E.J. Horvitz and D. Heckerman. The inconsistent use of measures of certainty in artificial intelligence research. In Lemmer and Kanal, editors, *Uncertainty in Artificial Intelligence*, pages 137–151, North Holland, Amsterdam, 1986.

[58] G.E. Hughes and M.J. Cresswell. In *Introduction to Modal Logic*, Methuen, London, 1968.

[59] E.T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 1968.

[60] E.T. Jaynes. Where do we stand on the maximum entropy formalism. In *Proc. Maxent Workshop*, Cambridge, 1979.

[61] H. Jeffreys. In *Theory of Probability*, Oxford University Press, 1939.

[62] L. Johnson and E.T. Keravnou, editors. *Expert Systems Technology: A Guide.* ABACUS Press, Turnbridge Wells, 1985.

[63] A. Jones, A. Kaufmann, and H.J. Zimmermann. *Fuzzy Sets Theory and Applications.* D. Reidel, Dordrecht, Holland, 1986.

[64] L.N. Kanal and J.F. Lemmer. In *Uncertainty in Artificial Intelligence*, North-Holland, Amsterdam, 1986.

[65] T.B. Kane. Enhancing the inference mechanism of nilsson's probabilistic logic. *International Journal of Intelligent Systems*, 5(5):487–504, 1990.

[66] T.B. Kane. Maximum entropy in nilsson's probabilistic logic. In *IJCAI 1989*, pages 442–447, Morgan Kaufmann, California, 1989.

[67] T.B. Kane. Reasoning with maximum entropy in expert systems. In W.T. Grandy and L.H. Schick, editors, *Maximum Entropy and Bayesian Methods*, pages 201–214, Kluwer Academic Publishers, Boston, 1991.

[68] T.B. Kane, P. McAndrew, and A.M. Wallace. Model-based object recognition using probabilistic logic and maximum entropy. *International Journal of A.I. and Pattern Recognition*, 5(3):425 — 437, 1991.

[69] J.M. Keynes. In *A Treatise on Probability*, McMillan, London, 1921.

[70] J.H. Kim and J. Pearl. A computational model for combined causal and diagnostic reasoning in inference systems. In *Proceedings of the 8th International Joint Conference on AI*, pages 190–193, Morgan Kaufmann, Los Angeles, 1983.

[71] A.N. Kolmogorov. *Foundations of the Theory of Probability.* Chelsea Publishing, New York, 1950.

[72] K. Konolige. Circumscription ignorance. In *Proceedings of AAAI*, University of Pittsburg, Pensylvania, 1982.

[73] Kotz and Johnsons, editors. *Encyclopedia of statistical Sciences.* Volume 1-5, John Wiley & Sons, U.S.A., 1982.

[74] H.E. Kyburg. Bayesian and non-bayesian evidential updating. *Artificial Intelligence*, 31:271–294, 1987.

[75] P.S. Laplace. In *Ouvres Completes*, Paris, 1774.

[76] P.S. Laplace. In *Theoretical Analysis of Probability*, Paris, 1812.

[77] P.S. Laplace. *A Philosophical Essay on Probabilities.* Dover Publications, New York, 1951.

[78] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. In G. Shafer and J. Pearl, editors, *Readings in Uncertain Reasoning*, pages 415–448, Morgan Kaufmann, California, 1990.

[79] J. Lemmer and S. Barth. Efficient minimum information updating for Bayesian in expert systems. In *Proceedings of National Conference on Artificial Intelligence*, pages 424–427, 1982.

[80] R.D. Levine and M. Tribus, editors. *The Maximum Entropy Formalism.* MIT Press, Cambridge, Mass, 1979.

[81] P. McAndrew and A.M. Wallace. Interpretation of 2d scenes using a general relational model. In *Proc. 3rd Alvey Vision Conference*, pages 107–115, Cambridge, 1987.

[82] P. McAndrew and A.M. Wallace. Rapid invocation and matching of 2d images to 3d models using curvilinear data. In *Synopsis submitted to Proc. of 3rd Int. Conf. on Image Processing and its Applications*, Warwick, July, 1989.

[83] M. McLeish. A note on probabilistic logic. *AAAI*, 215–219, 1988.

[84] K. Mendelssohn. *The Quest for Absolute Zero (the meaning of low temperature physics).* Taylor & Francis, London, 1977.

[85] R. Von Mises. On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics*, 18:309–348, 1947.

[86] R. Von Mises. *Probability, Statistics and Truth.* Allen and Unwin, London, 1928.

[87] A. Newell and H.A. Simon. In *Human Problem Solving*, Prentice-Hall, Englewood Cliffs, NJ, 1972.

[88] N. Nilsson. In *Principles of Artificial Intelligence*, Morgan Kaufmann, Palo Alto, CA, 1980.

[89] N. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28 Nr. 1, 1986.

[90] J.B. Paris and A. Vencovska. A note on the inevitability of maximum entropy. In *International Journal of Approximate Reasoning*, 1989.

[91] J.B. Paris and A. Vencovska. On the applicability of maximum entropy to inexact reasoning. *International Journal of Approximate Reasoning*, 3:1–34, 1988.

[92] R.S. Patil. Artificial intelligence techniques for diagnostic reasoning. In *Exploring Artificial Intelligence*, pages 347–380, Morgan Kaufmann, San Mateo, California, 1988.

[93] J. Pearl. Bayesian decision methods. In *Encyclopedia of AI*, Wiley Interscience, New York, 1987.

[94] J. Pearl. Evidential reasoning under uncertainty. In *Exploring Artificial Intelligence*, pages 381–418, Morgan Kaufmann, San Mateo, California, 1988.

[95] J. Pearl. Evidential reasoning using stochastic simulation of causal models. In *Artificial Intelligence*, pages 245–257, Elsevier Science Publishers, North Holland, 1987.

[96] J. Pearl. On logic and probability. *Computational Intelligence*, 4(1):99–103, 1988.

[97] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, 1988.

[98] E.S. Pearson and M.G. Kendall, editors. *Studies in the History of Statistics and Probability.* Volume 1, Griffin, London, 1970.

[99] A. Perez and R. Jirousek. Constructing an intentional expert system (INES). In *Medical Decision Making*, pages 307–315, Elsevier Scientific Publishers, 1985.

[100] J. Pollock. *Knowledge and Justifications.* Princeton University Press, Princeton, 1974.

[101] H. Prade. A synthetic view of approximate reasoning techniques. In *Proc. 8th IJCAI*, pages 130–136, Karlsruhe, West Germany, 1983.

[102] H. Prade and C.V. Negoita (Eds). *Fuzzy Logic in Knowledge Engineering.* Verlag TUV Rheinland, Koln, 1986.

[103] J.R. Quinlan. Inferno: a cautious approach to uncertain inference. *The Computer Journal*, 26, 1983.

[104] F.P. Ramsey. Truth and probability. In *The Foundatiopns of Mathematics and other essays*, Kegan Paul, London, 1931.

[105] J.A. Robinson. A machine oriented logic based on the resolution principle. *Journal of the ACM*, 12:25 — 41, 1965.

[106] W.D. Ross. *The Works of Aristotle Translated into English.* Volume 1, 1928.

[107] B. Russell and A.N. Whitehead. In *Principia Mathematica*, Cambridge University Press, Cambridge, 1910.

[108] E. Sanchez and L. A. Zadeh (Eds). *Approximate Reasoning in Intelligent Systems, Decision and Control.* Pergamon Press, Oxford, 1987.

[109] L.J. Savage, editor. *The foundations of statistical inference.* Wiley, New York, 1962.

[110] L.J. Savage. *The foundations of statistics.* Wiley, New York, 1954.

[111] M.S. Schwartz, J. Baron, and J.R. Clarke. A causal Bayesian model for the diagnosis of appendicitis. In *Uncertainty in Artificial Intelligence*, pages 423–434, Elsevier Science B.V. (North-Holland), 1988. (Eds. Lemmer, Kanal).

[112] P.S. Sell. In *Expert Systems- A practical introduction*, Macmillan, London, 1985.

[113] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, 1976.

[114] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–623, 1948.

[115] J.E. Shore and R.W. Johnson. An axiomatic derivation of the maximum entropy principle. *IEEE Transactions on Information Theory*, January, 1980.

[116] B. Silver. *Meta-Level Inference.* Elsevier Science Publishers B.V., Amsterdam, 1986.